Univerzita Karlova

Filozofická fakulta
Ústav jazyků a komunikace neslyšících

Bakalářská práce

Zdenka Ondráčková

Korpusy znakových jazyků

Corpora of Sign Languages

2017 Mgr. Lenka Okrouhlíková, Ph.D.

Poděkování

Ráda bych poděkovala své vedoucí bakalářské práce Mgr. Lence Okrouhlíkové, Ph.D. za odborný dohled, cenné rady a vstřícnost, kterou mi v průběhu zpracování celé této práce věnovala. Poděkování patří též mým nejbližším za podporu nejen při psaní této práce, ale i za oporu během celého studia.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, 17. července 2017

……

Zdenka Ondráčková

Abstrakt

Cílem této bakalářské práce bylo popsat specifika tvorby a struktury korpusu znakového jazyka a navrhnout podobu korpusu českého znakového jazyka. Daná specifika jsou ilustrována na dvou konkrétních zahraničních korpusech znakového jazyka. V úvodní části práce jsou uvedeny základní informace o korpusové lingvistice a stručně shrnuty základní charakteristiky korpusů znakových a mluvených jazyků. Práce se zabývá obecným popisem nezbytných kroků při tvorbě korpusu znakového jazyka. Stěžejní část práce je věnována dvěma konkrétním zahraničním korpusům znakového jazyka – korpusu australského znakového jazyka a korpusu německého znakového jazyka. U obou korpusů jsou zpracovány informace týkající se jejich tvorby a struktury, což zahrnuje sběr a zpracování dat, výběr a případnou tvorbu vhodného softwaru. V části věnované korpusu australského znakového jazyka je detailněji popsán postup při zpracování již získaných jazykových dat – anotaci, na rozdíl od korpusu německého znakového jazyka, jehož popis je více zaměřen na získávání dat – elicitaci. V závěru této bakalářské práce je nastíněna možnost vzniku korpusu českého znakového jazyka a návrh jeho možné podoby.

Klíčová slova: korpus, korpus AUSLAN, korpus DGS, korpus ČZJ, znakový jazyk

Abstract

The purpose of this thesis is to describe specifics of sign languages corpora creation and structure and to design a form of a corpus of Czech Sign Language. These specifics are illustrated on two specific foreign corpus of sign languages. The first part of the thesis tells basic information about corpus linguistics and sums up characteristics related to a sign language corpus and also a corpus of a spoken language. The thesis deals with a general description of the steps that are necessary in the process of creating a sign language corpus. Major part of the thesis is devoted to describing two foreign corpora of sign languages – the corpus of Australian Sign Language (Auslan) and the corpus of German Sign Language (DGS). Both of the corpora include information about how they were created and their structure, i.e. data collection and data processing, a selection and creation of an appropriate software. While the part dedicated to Auslan describes the process of annotation of language data in detail, the part about DGS is more focused on the elicitation process. The possibility of forming a Czech Sign Language corpus is outlined in the conclusion of the thesis, as well as a proposal of its possible form.

Keywords: corpus, AUSLAN corpus, DGS corpus, Czech Sign Language corpus, sign language

Obsah

Seznam zkratek

ASL	American Sign Language
CODA	Children of Deaf Adult
ATLAS	Automatic Translation into sign LAnguageS
LIS	italský znakový jazyk
ČNK	Český národní korpus
ECHO	European Cultural Heritage On-Line
IMDI	ISLE Meta Data Initiative
HamNoSys	Hamburský notační systém
FACS	Facial Action Coding System
ASR	Automatic Speech Recognition
Auslan	australský znakový jazyk
ELAR	Endangered Languages Archive
ELDP	Endangered Languages Documentation Programme
SOAS	School of Oriental and African Studies
RH	right hand
LH	left hand
AMBI	ambidextrous
ELAN	EUDICO Lingvistický Anotátor
EUDICO	EUropean DIstributed Corpus
eaf	ELAN Annotation File
DS	depicting sign
L	locative
M	movement
H	handling
S	size and shape
G	gesture
NMS	non-manual sign
M	mouthing
MG	mouth gestures
FS	fingerspelling
A	addresssee
T	target
O	other

Z	cannot be coded
CA	constructed action
NorV	noun or verb
V	verb
nonA	non-argument
DGS	Deutsche Gebärdensprache, německý znakový jazyk
IDGS	Institut für Deutsche Gebärdensprache
iLex	integrated lexicon
URL	Uniform Resource Locator
SQL	Structured Query Language
HTML	HyperText Markup Language

Úvod

Cílem této bakalářské práce je přinést prvotní a rámcové informace o tvorbě a struktuře korpusu znakového jazyka. V českém prostředí nebyla dosud vytvořena žádná obsáhlejší práce, která by byla věnována tomuto tématu, proto bylo nutné vycházet z dostupných zahraničních zdrojů v opoře o české materiály zabývajících se korpusy obecně. Nejdříve jsou popsány základní charakteristiky korpusu a korpusové lingvistiky. Následují specifika korpusů znakových jazyků v opozici ke korpusům mluvených jazyků. Nejrozsáhlejší část bakalářské práce je věnována popisu a analýze korpusu australského znakového jazyka a korpusu německého znakového jazyka. Korpus australského znakového jazyka byl vybrán z důvodu množství dostupných materiálů, které se věnovaly popisu anotačního procesu do poměrně velkých detailů. Korpus německého znakového jazyka byl zvolen z důvodu přístupných podkladů a literatury, které se zabývají procesem elicitace jazykových dat. Na těchto dvou korpusech bylo záměrem ukázat jednotlivé kroky tvorby korpusu, ovšem tak, aby se informace zbytečně neopakovaly a nepřekrývaly. Závěrečnou část tvoří můj návrh tvorby korpusu českého znakového jazyka, který zahrnuje možnosti, jak by šly práce na korpusu realizovat.

1. Jazykový korpus

Nejdříve je nutné ujasnit si, co se rozumí pod pojmem jazykový korpus. Podle Čermáka (1995, s. 119) lze korpus definovat jako „[…] *vnitřně strukturovaný, unifikovaný a obvykle i oindexovaný a ucelený rozsáhlý soubor elektronicky uložených a zpracovávaných jazykových dat většinou v textové podobě, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž pak je také považován za reprezentativní*."

Avšak korpus může být také vymezen coby „[…] *rozsáhlý soubor autentických textů (psaných nebo mluvených) převedený do elektronické podoby v jednotném formátu tak, aby v něm bylo možné jednoduše vyhledávat jazykové jevy, zejména slova a slovní spojení (kolokace)*" (Cvrček, Richterová, 2014).

První zmíněná definice se zaměřuje spíše na podobu zpracování jazykových dat v korpusu, což je jeden z prvních a nejdůležitějších kroků, který si musí tvůrci korpusu ujasnit. Druhý výklad již klade důraz na to, že je korpus souborem textů a vzhledem k paralele korpusu mluveného jazyka s korpusem znakového jazyka, by označení „texty" šlo vztáhnout i na projevy ve znakovém jazyce.

Je však nutné mít na mysli, že ve znakovém nelze vytvářet psané texty,[1] v pravém smyslu slova. Čermák (1995, s. 119) se taktéž zmiňuje o vyhledávací funkci platformy, která by měla být klíčovým bodem při tvorbě softwaru. S přihlédnutím na specifika znakového jazyka jsou data takového korpusu objemově velká z toho důvodu, že především obsahují nahrané videonahrávky znakujících respondentů, a proto je nutné zabezpečit kvalitní software.

Fenlon (2015, s. 157) definuje korpus znakového jazyka jako „*rozsáhlý soubor dat sestávající ze spontánních a elicitovaných znakových projevů, které jsou digitalizované a současně jsou doplněné o lingvistické anotace. Jazyková data, která jsou ve strojově čitelné formě, jsou také v nejvyšší možné míře reprezentativní, jak z jazykového, tak z uživatelského hlediska.*" Takto je stručně popsán základní charakter korpusu znakového jazyka – znakové projevy respondentů, které jsou ve formě videonahrávek obohacené o lingvistické informace.

[1] Tímto nejsou myšleny psané texty ve smyslu např. psané češtiny českých neslyšících, touto problematikou se mimo jiné hojně zabývá Macurová (1998, s. 180), (1995, s. 23–33) apod. V českém prostředí vzniká korpus DEAF, který obsahuje psané texty českých neslyšících. Za psané texty ve znakovém jazyce by se daly pokládat zápisy znakového projevu notačním systémem, např. SignWriting.

1.1 Korpusová lingvistika

Korpusová lingvistika je odvětví lingvistiky, které se zabývá zpracováním jazykových dat uložených v elektronické podobě. Vzniknout mohla díky rozvoji techniky, resp. počítačovým programům, do kterých bylo a je možné ukládat textové soubory a dále s nimi pracovat. (Pala, 1996, s. 8) V současné době jsou však zpracovávaná data tak rozsáhlá, že je nutné ukládat je na servery, které pojmou takto velké soubory.

Korpusová lingvistika je poměrně nový obor v oblasti jazykového výzkumu pojící se s možnostmi, které nabízí pokročilejší počítačové technologie. V minulosti byl každý soubor dat, na kterých byla provedena lingvistická analýza, nazýván korpus. Nicméně s příchodem výpočetní techniky a korpusové lingvistiky, se užití termínu korpus omezuje na soubor textů, které jsou strojově čitelné. Johnston (2009, cit. dle Pfau, 2012, s. 1033) tvrdí, že: *„Korpusová lingvistika je založena na předpokladu, že zpracování velkého množství anotovaných textů může odhalit vzorce v užívání jazyka a jeho struktury, které nejsou dostupné skrze uživatelovu intuici nebo prostřednictvím odborné detailní lingvistické analýzy konkrétních textů.“*

Korpusová lingvistika a její nástroje jsou velmi cenné, díky nim lze lépe uchopit a poznat strukturu jazyka. Již Čermák (2005, s. 15) předznamenává využitelnost elektronického zpracování jazykových dat v mnoha jeho rovinách. Zmiňuje se o uplatnění tzv. futurálních korpusů, které budou postaveny na audio-vizuálním zobrazení jazyka a díky nim bude možné zachytit a zkoumat jazyk z mnoha úhlů. Čermák (2005, s. 15) tímto projevuje zájem o zachycení mimojazykového kontextu, ve kterém se jazyk užívá. V tomto případě však lze hovořit o komplexním vizuálním zachycení znakového jazyka s jeho manuálními i nemanuálními artikulátory, které jsou integrální složkou jazyka a jistě neméně důležitou složkou pro výzkum.

Zajímavý postřeh přináší Hanke (2016, s. 89), který říká, že zatímco výzkumníci se stále snaží přijít s dalšími novými strukturami znakového jazyka, které by mohly být studovány, tak jejich dovednosti v oblasti práce s vizuální technikou v rámci korpusové lingvistiky nejsou zcela samozřejmé. Je tedy důležité, aby měl výzkumník velmi dobré lingvistické poznatky o znakových jazycích, ale zároveň se zajímal o problematiku jeho možného zpracování v rámci rozvíjející se techniky a spolupracoval v týmu lidí, kteří mají tyto potřebné znalosti.

V korpusové lingvistice *„kvantitativní analýza jde ruku v ruce s kvalitativní analýzou“* (Leech, 2000, s. 715), jelikož se empiričtí lingvisté zabývají skutečnými jevy jazyka, získaných z nahrávek mluvených nebo psaných textů, při kterých uplatňují postup zdola nahoru. Z analýzy jednotlivých citací odvozují obecnosti, které vedou k formulacím abstraktních hypotéz o jazyce. Jejich úkolem je shromažďovat a formulovat jazykové zákonitosti potřebné k porozumění jakéhokoli textu (Pfau, 2012, s. 1034).

Stejné cíle platí i pro korpus znakového jazyka. Protože však jde o jazyk, který je vizuálně motorický, objevující se zejména v komunikaci tváří v tvář, je tedy spíše srovnatelný s korpusem mluveného jazyka než s korpusem psaného jazyka a podle Leecha (2000, s. 684, cit. dle Pfau, 2012, s. 1033–1034) „*jsou zde dva způsoby navrhování korpusu mluveného jazyka, aby se dosáhlo reprezentativnosti. Jedním je zvolit nahrávky mluvené řeči, aby byly zastoupeny různé typy aktivit, kontextů a žánrů, do kterých může být mluvený projev klasifikován. Toto by šlo nazvat vzorkem založeným na žánru projevu. Druhá metoda je vzorkování skrze populaci mluveného jazyka, u které chceme, aby byla reprezentována z hlediska proměnných, jako je oblast, pohlaví, věk a socioekonomický ukazatel obyvatel. Takto lze představit vyrovnaný průřez populací a je možné pojmenovat ho jako demografický vzorek*.“ Při tvorbě korpusů znakových jazyků byla dosud používána spíše druhá metoda.

1.2 Terminologie v korpusové lingvistice

Než bude následovat podrobnější popis toho, co je korpus a jak vzniká, je důležité seznámit se s terminologií, která se v souvislosti s korpusovou lingvistikou používá.

Korpus, který je tvořen soubory textů, ať už mluvených, znakovaných nebo psaných by nebyl korpusem, bez přidané lingvistické hodnoty. Texty je nutné segmentovat na jednotlivá slova či v tomto případě znaky a přiřadit jim lingvistické informace. Bez těchto jazykových značek by nebylo možné texty zařadit do korpusu a následně využít k výzkumům. (Cvrček, Richterová, 2016)

Pro lepší pochopení následujícího textu, který se bude věnovat popisu tvorby korpusů, je nutné přiblížit a vysvětlit specifické termíny korpusové lingvistiky, mezi které mimo jiné patří i lemma, tag či token.

1.2.1 Segmentace

Segmentace neboli členění psaného textu na menší úseky je ve většině případů prováděno automatickými metodami. Na rozdíl od korpusů znakových znaků, kde segmentace, stejně jako anotace, probíhá manuálně.

Daný text je nejčastěji segmentován nejdříve na věty a posléze ho lze dělit i na menší jednotky, tedy morfémy. Český národní korpus členění na slova „ […] *obvykle nazývá tokenizace, neboť slovo je základní vyhledávací jednotkou (angl. token) v korpusu*.“ (Cvrček, Richterová, 2016)

Korpusy mluvených jazyků zpravidla segmentují nikoli věty, ale spíše smysluplné úseky promluvy, které jsou identifikovány v proudu řeči (Čermák, 2016). Podobný přístup je uplatňován při segmentaci v korpusech znakových jazyků, kdy se znakový projev nečlení na věty, ale na logické významové pasáže a posléze i jednotlivé znaky.

1.2.2 Anotace

Anotace je proces, během něhož se ručně nebo automaticky přiřazují „[…] *interpretační lingvistické údaje, strukturní údaje a/nebo metatextové údaje k textovým datům korpusu*“ (Cvrček, Richterová, 2014c).

Za lingvistické interpretační údaje se označují morfologické značky (tagy), které se přiřazují slovním tvarům. Dále také syntaktická či sémantická data související s danými slovními výskyty. Strukturní údaje označují např. původ textu či vymezují začátek a konec věty procesem segmentace. Tyto údaje následně usnadňují vyhledávání v korpusech. (Petkevič, 2016)

1.2.3 Lemma

Reprezentativní slovníková podoba lexému neboli lemma je charakterizované určitou gramatickou formou, což je např. u substantiv nominativ singuláru a u sloves infinitiv (Hladká, 2016).

Nicméně tuto definici nelze stoprocentně vztáhnout na lingvistické jednotky v korpusech znakových jazyků, protože u znaků znakového jazyka nelze jednoznačně identifikovat jejich reprezentativní tvar nejen z důvodu neexistující kodifikace jazyka. Problémem je i lemmatizace[2] specifických znaků ve znakovém projevu, pro něž dosud neexistují jednoznačné překlady v jazyce mluveném. Ovšem i v těchto korpusech se určují lemmata, protože na jejich základě se posléze sestavují slovníky.

V českém jazyce tedy lemma „[…] *vzniká abstrakcí morfologických vlastností slovního tvaru* […] *představuje tedy množinu forem lišících se pouze tvaroslovnými afixy, příp. pravopisnou variantou*“ (Cvrček, 2016).

[2] Článek zabývající se problematikou lemmatizace ve znakovém jazyce: JOHNSTON, Trevor. Corpus linguistics and signed languages: no lemmata, no corpus. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Německo: Institute for German Sign Language and Communication of the Deaf, 2008, s. 82-87.

1.2.4 Token

„*Token je nejmenší jednotka textu, většinou se jedná o grafické slovo* […], *resp. o jednu jeho konkrétní realizaci*“ (Cvrček, 2014). Jednotlivé tokeny v korpusu se také mohou označovat jako pozice a proces segmentace tokenů v textu se označuje jako tokenizace (Cvrček, Richterová, 2014).

V kontextu znakových jazyků se jedná o každý jednotlivý znak, který mluvčí produkuje, není zde tedy požadavek grafického záznamu. Ovšem o to obtížnější je segmentace projevu do jednotlivých tokenů, protože není určené, co je jeden znak, resp. kde znak začíná a končí, na rozdíl od grafického zápisu slova, které lze rozpoznat ohraničujícími mezerami.[3] (Hanke, Storz, 2008a, s. 64)

1.2.5 Tag

Tagem se nazývá morfologická značka, která v sobě zahrnuje gramatické informace o slovu či znaku v jeho konkrétní realizaci v textu.

Morfologická značka je ve většině případů souborem symbolů, které se skládají z kombinace písmen a čísel, které zastupují a nesou informaci o gramatických kategoriích. Tyto kategorie zahrnují například slovní druh, jmenný rod, číslo, čas nebo negaci daného slova. (Cvrček, Richterová, 2014)

V korpusech znakových jazyků se toto označení liší, protože u znaků není možné jednoznačně určit jejich gramatickou kategorii. Proto se zaměřují spíše na určení fonetické a fonologické formy znaku či značení orálních komponentů.

1.2.6 Glosa

V korpusu znakového jazyka se vyskytuje specifický atribut – glosa. Přestože znakový jazyk nemá písemnou formu, je nutné ho v nějaké písemné formě zanést do korpusu kvůli následnému anotačnímu zpracování. Za glosování lze označit proces, při kterém se přepisuje jeden jazyk do druhého pomocí psané formy a hledá se tak nejbližší a nejvýstižnější ekvivalent výchozího výrazu. Daná psaná informace je glosa.

Překlad je přibližný a poměrně zjednodušený. Glosy se většinou píší kapitálkami. (Johnston, 2009, s. 91)

[3] Článek věnující se problematice tokenizace: GREFENSTETTE, Gregory a Pasi TAPANAINEN. What is a word, What is a sentence? Problems of Tokenization. In: *The 3rd International Conference on Computational Lexicography and Text Research*[online]. Budapešť, 1994, s. 79-87 [cit. 2017-07-14]. Dostupné z: https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/GrefenstetteTapanainen1994.pdf

1.2.7 Metadata

Text v korpusu je doplněn metadaty, což jsou standardizované a strukturované informace, které se týkají obsahu, původu nebo lingvistické interpretace daného textu. Metadata se mohou vztahovat „*k celému korpusu, konkrétnímu textu, k jeho části nebo k jednotlivému slovnímu tvaru. Na úrovni celých textů jde např. o bibliografické údaje, označení žánru a původu textu, údaje o nahrávce, licenci nebo záznam toho, jakými verzemi kterých nástrojů byl text zpracován*.“ (Cvrček, Richterová, 2013b)

1.2.8 Strukturní jednotky a atributy

Je nezbytné, aby každý korpus byl hierarchicky strukturován. Uspořádání korpusu se dělí na různé strukturní jednotky a k jedné takové jednotce je vždy připojen jeden nebo více strukturních atributů.

Je rozdíl mezi strukturními jednotkami psaných a mluvených textů a samozřejmě také texty znakového jazyka v korpusu. V korpusech obsahujících psané texty jsou strukturními jednotkami text či soubor textů, dále odstavec a věta.

V korpusu mluveného jazyka se strukturní jednotky dělí na celkový dokument, promluvu mluvčího a členění promluv mluvčího na kratší úseky. Podobným způsobem se strukturuje korpus znakového jazyka.

Ke každé strukturní jednotce se váží strukturní atributy. Všechny jednotky v korpusu mají atribut, který slouží k identifikaci strukturní jednotky. Nicméně tak, jak se liší strukturní jednotky v korpusech psaných či mluvených, potažmo znakových, tak se různí i strukturní atributy u jednotlivých typů korpusů. Psané korpusy se zaměřují na uvedení následujících atributů: název dokumentu, autor dokumentu, vydání, místo a rok vydání, překladatel či zdrojový jazyk. Strukturní atributy korpusů mluvených jazyků, jistě použitelné i u korpusů znakových jazyků, obsahují: identifikaci uceleného rozhovoru, rok nahrání rozhovoru, počet mluvčích, typ promluvy, pohlaví, věková kategorie či vzdělání mluvčího. (Cvrček, Richterová, 2017)

1.3 Korpus znakového jazyka

V lingvistice znakového jazyka je korpusová lingvistika stále ještě považována za mladou disciplínu, která se ale rychle rozvíjí. Johnston (2008) vyjádřil potřebu korpusu znakového jazyka následovně: „*Korpus znakového jazyka výrazně zlepší popis znakových jazyků a umožní tvorbu analýz znakového jazyka založených právě na korpusu. Korpus je důležitý pro testování jazykových hypotéz ve všech jazykových výzkumech na všech úrovních, od fonologie po diskurz* […]. *Tohle je zvláště platné pro komunitu neslyšících, která je mladou jazykovou komunitou. Ačkoliv introspekce a pozorování může pomoci vyvinout hypotézy týkající se používání a struktury jazyka, protože znakový jazyk nemá psanou formu ani pevně stanovené standardy jazyka, tak intuice či výzkumníkova pozorování mohou selhat z důvodu absence jednoznačného konsenzu rodilých uživatelů znakového jazyka v otázkách fonologických nebo gramatických, v rámci běžného užívání. Dřívější spoléhání na intuici několika mála respondentů a na izolované kontextové příklady byly v tomto ohledu problematické. Výzkum v rámci znakových jazyků dramaticky vzrostl v posledních třech až čtyřech desetiletích, ale pokroku v této oblasti bylo bráněno překážkami v oblasti zpracování a sdílení dat*.“

Jeden z prvních, ne-li úplně první velký projekt korpusu znakového jazyka začali Lucas a Bayley a jejich tým na americkém znakovém jazyce (dále také jen ASL[4]). V průběhu roku 1995 byla shromážděna data ze sedmi měst USA, která byla určena jako reprezentativní oblasti. Ve všech těchto městech byly velké komunity neslyšících a také tam byly internátní školy pro neslyšící. Do projektu bylo celkově zahrnuto 207 mužů a žen. Byli rozděleni do tří věkových skupin 15−25, 26–54 a více než 55 let. Všichni se naučili znakový jazyk doma, na internátě či ve škole před dosažením věku 6 let. Z každé oblasti měla určená kontaktní osoba označit rodilé uživatele ASL, kteří žili v komunitě nejméně 10 let. Kontaktní osoba, sám neslyšící člověk žijící v dané komunitě, sestavila skupinu o dvou až sedmi znakujících osobách. Všechna setkání byla nahrávána a každá schůzka sestávala ze tří částí. První část byla věnována volné konverzaci, která trvala přibližně hodinu, přičemž výzkumník nebyl přítomen. Ve druhé části byli vybráni dva respondenti a byli tázáni na různé otázky neslyšícími výzkumníky. Rozhovor mimo jiné zahrnoval témata týkající se sociálních vztahů a vzorců jazykového chování. Nakonec jim bylo ukázáno 34 obrázků, které měly pomoci k elicitaci jednotlivých znaků pro objekty nebo činnosti, které byly reprezentovány na obrázcích.

[4] ASL = American Sign Language

Důležitým prvkem bylo to, aby během setkání nebyl přítomný žádný slyšící výzkumník, protože: „*Bylo prokázáno, že mluvčí ASL jsou velmi náchylní k audiologickému a etnickému statutu výzkumníka* […]. *Tato citlivost se odrážela přechodem z ASL do znakované angličtiny v přítomnosti slyšící osoby*“ (Lucas, Bayley, 2005).

Ve stejné době byl vyvinut katalogizační systém a počítačová databáze, kde se sbírala a uchovávala metadata, tedy informace o tom, kdy a kde byla každá skupina dotazována a osobní informace o respondentech (jméno, věk, edukační základ, povolání atd.).

Projekty korpusu znakového jazyka začaly vznikat i v jiných státech, např. v Austrálii, Číně, Francii, Irsku, Itálii, Německu, Velké Británii a na dalších místech jsou plánovány. Některé z korpusů se zaměřují na sociolingvistické variace znaků, ale většina z nich má více cílů a získaná data neplánují využít pouze pro lingvistický popis jazyka. Usilují i o zachycení starší podoby znakového jazyka, což lze využít pro budoucí výzkumy, např. k dokumentaci diachronní změny znaků či k tvorbě materiálů pro výuku znakového jazyka. (Pfau, 2012, s. 1034–1036)

1.3.1 Typy korpusů znakových jazyků

Je zřejmé, že jakýkoli korpus musí mít dané nějaké zaměření, aby bylo jasné, co uživatelům nabízí, co v něm můžou najít a očekávat. Proto se přistupuje k vytváření tematicky zaměřených korpusů, které reprezentují určitou oblast jazyka. Můžeme se tedy setkat s korpusy obecnými a specifickými, synchronními a diachronními, korpusy jazyka psaného a mluveného, korpusy jednojazyčnými a vícejazyčnými apod. Takovéto dělení je přínosné nejen laickým uživatelům, ale především odborníkům, kteří podle dostupných dat dokáží vypracovat studie týkající se poznatků o jazyce, ale také slouží k jeho detailnějšímu popisu.[5] Případně se z těchto zjištěných informací vytváří další sekundární výstupy jako například slovník znakového jazyka.[6]

Dělení není u korpusů znakových jazyků tolik používané, jako u korpusů mluvených jazyků, což je nejspíše způsobené jednak velikostí korpusů, poměrně nedávným vznikem zájmu o tuto problematiku, ale také tím, že jsou tyto korpusy většinou jednojazyčné, obecné a synchronní.

[5] Na korpusových datech je založena např. studie britského znakového jazyka od Cormier a Schembri z roku 2014 - Describing sociolinguistic variation in verb directionality in British Sign Language: A corpus-based study nebo také výzkum The use of space with indicating verbs in Auslan od autorů de Beuzeville, Johnston a Schembri z roku 2009.

[6] Slovník znakového jazyka založený na korpusu již vznikl např. v Německu, Austrálii či Polsku.

Například korpus britského znakového jazyka obsahuje 40 000 lexikálních jednotek a naproti tomu Britský národní korpus zahrnuje na 100 milionů slov, takže je názorně vidět, že je zde velký obsahový nepoměr. Navzdory této jasné disproporci ve velikosti korpusů znakových jazyků vznikají i v těchto případech specificky zaměřené korpusy, které jsou však poměrně ojedinělé. Na začátku tvorby libovolného jazykového korpusu by tedy mělo být jasné, jaký je jeho účel a podle toho odvíjet jeho zaměření.

Pro ukázku následuje výběr korpusů znakových jazyků, které se zaměřují na určité jazykové hledisko. Výzkumníci si vybrali jednu konkrétní oblast, kterou prozkoumali a ze získaných dat vytvořili korpus. Tyto korpusy jsou uvedeny pouze pro představu, že i v minoritní oblasti korpusů znakových jazyků vznikají specializované korpusy. V Asii byl vytvořen korpus hongkongského znakového jazyka, který se zaměřuje na produkci jazyka neslyšících dětí. Tyto děti byly nahrávány při komunikaci s dospělými. (Fung H-M, 2008, s. 22)

Dalším korpusem je korpus japonského znakového jazyka, který je zaměřený na rozhovory rodilých mluvčích japonského znakového jazyka. Hlavním cílem budování tohoto korpusu je vytvoření elektronických slovníků, které z něj následně mají vycházet. Později byly přidány rozhovory mezi rodilými mluvčími japonského znakového jazyka a CODA dětmi.[7] Tato databáze obsahuje 1 096 znaků. (Nagashima, 2008, s. 141)

V Evropě existuje paralelní korpus[8] italského znakového jazyka, který v sobě zahrnuje italský znakový jazyk a italský jazyk. Vznikl v rámci projektu ATLAS,[9] který skrze virtuálního tlumočníka automaticky umožňuje překlad italských textů do italského znakového jazyka. (Bertoldi, 2010, s. 21)

V USA vznikl projekt na podporu rozvíjení jazykových znalostí neslyšících dětí. Projekt byl pojmenován CopyCat a v jeho rámci byly pořízeny nahrávky třiceti neslyšících žáků. Žáci hráli počítačovou hru, při které bylo nutné komunikovat, takže se vyjadřovali americkým znakovým jazykem. Celkově bylo shromážděno 5 829 znakovaných frází vztahujících se nejen k videohrám, které poté byly zařazeny do korpusu znakového jazyka. (Brashear, 2010, s. 29)

[7] CODA – Children of Deaf Adult tedy slyšící dítě neslyšícího dospělého
[8] paralelní korpus – korpus, který obsahuje stejné texty ve dvou či více jazykových mutacích
[9] ATLAS – Automatic Translation into sign LAnguageS

Kromě dělení korpusu podle typologie zachycených osob, tedy dětí a dospělých, nebo paralelního korpusu, které jsou svým zaměřením ve velké míře podobné korpusům mluveného jazyka, je nutné přemýšlet o specifičnosti znakového jazyka, která by měla být reflektována při tvorbě takového korpusu.

Bylo by jistě prospěšné vytvořit v rámci korpusu další subkorpusy zaměřené například na specifické znaky, nepřímá pojmenování či klasifikátory. Je zřejmé, že takovéto subkorpusy by nebyly moc obsáhlé, ale zato by byly velmi potřebné vzhledem ke specifičnosti těchto jazykových prvků.

Pro představu uvedu údaje z Českého národního korpusu,[10] který obsahuje přes 3,6 miliardy českých slov a 1,5 miliardy cizích slov, takže je jasné, že se dělí na další subkorpusy. Z dostupných českých subkorpusů zmíním ty nejobsáhlejší, tedy SYN2015 (synchronní, psaný, 100 mil. slov), ORAL2013 (neformální mluvená čeština, synchronní, 2,8 mil. slov), DIAKORP (diachronní, psaný, 3,4 mil. slov), InterCorp (cizojazyčný, více než 30 jazyků, 1,46 mld. slov) a další, které se neustále rozšiřují a doplňují. Avšak existují i specificky zaměřené korpusy, které se zaměřují například na díla jednoho autora, v českém prostředí to je Korpus Karla Čapka či Bohumila Hrabala a samozřejmě i vznikající Korpus psané češtiny českých neslyšících. (Cvrček, Richterová, 2017) Zde lze pozorovat paralelu ke korpusům znakových jazyků, které se taktéž specificky vymezují a nejsou příliš obsáhlé.

1.3.2 Rozdílnost korpusů z hlediska jazykových dat

V rámci tvorby jakéhokoliv korpusu je nutné získat jazykový materiál. Ovšem způsoby získávání a následně i zpracování jazykových dat se v případě korpusu znakového a mluveného jazyka výrazně liší. Velmi podstatný je již výběr respondentů. Při tvorbě korpusu znakového jazyka jsou respondenty většinou konkrétně vybrané osoby, jejichž mateřským jazykem je znakový jazyk a které jsou ve studiu natáčeny kamerami, aby bylo možné zachytit a následně analyzovat jejich projev. Je jasné, že toto není pro respondenty zcela přirozené prostředí, když jsou ve speciálně zřízeném studiu a míří na ně kamery (podrobněji o nahrávání viz níže). Navíc jsou jejich projevy získány skrze speciálně připravené elicitační materiály, na které během setkání reagují. Taktéž je nezbytné, aby projevy ve znakovém jazyce byly přeloženy do psaného národního jazyka, aby šlo s daty pracovat. Následné zpracování těchto dat probíhá ručně, což znamená, že je nutná přítomnost několika anotátorů, kteří dané nahrávky znakovaných projevů segmentují a přiřazují jim lingvistické značky.

[10] Seznam korpusů ČNK: http://wiki.korpus.cz/doku.php/cnk:uvod

V protikladu je korpus mluveného jazyka, který je sestaven ze psaných či mluvených textů, které jsou převedené do elektronické podoby. Psané texty se již nemusí překládat a mluvené texty se pouze přepíší do elektronické podoby, takže zde neprobíhá převod projevu z jednoho jazyka do druhého.
Jazykové projevy lze získat v jejich přirozeném prostředí bez nutnosti elicitačních podnětů. Data jsou zpracovávána automaticky skrze softwary, ve kterých jsou nahrány programy, které dokáží rozpoznat použité lexémy a přiřadit jim lingvistické značky, takže tento typ zpracování šetří mnoho času.

1.3.3 Metodologie sběru jazykových dat

V lingvistickém výzkumu lze použít dva typy metodologií, kvalitativní a kvantitativní. Lingvisté se shodují v tom, že tyto metodologie jsou odlišné a často nekompatibilní. Tyto přístupy lze označit jako kontinuum výzkumných metod, kdy na jednom konci je kvalitativní introspekce, a na druhé straně se ocitají experimenty, jako typické kvantitativní metody. Do druhé skupiny se zahrnují metody, jako je pozorování a popis na základě systematické elicitace, což jsou metody vhodné pro výzkumy v rámci znakových jazyků. Avšak objevují se zde i výzkumy, které jsou založené na datech zpracovaných v korpusu, odkud jsou získávány informace pro příkladové jazykové struktury. Nicméně ve výzkumech znakových jazyků se používají i další techniky k získání kýžených informací, které jsou založené na intuici informanta. Z nich lze jmenovat:

- rozpoznání a oprava chyb − rozpoznání chyb působí jako poměrně jednoduchý úkol, avšak dosud nebyl hojně využit ve výzkumech znakových jazyků. V tomto úkolu je informantům představeno více podobných projevů ve znakovém jazyce a následně jsou požádáni o označení chyb, které se podle nich v projevu objevují, a pokud tam nějaké odhalí, mají je opravit. Nicméně tento úkol je velmi obtížný, protože většina znakových jazyků nemá kodifikovanou formu a navíc mnoho uživatelů znakového jazyka nedostalo vzdělání o jejich mateřském jazyce, takže pro některé informanty může být velmi problematické reflektovat svůj jazyk.
- gramatické úsudky − v tomto typu úkolu jsou informantům představeny promluvy ve znakovém jazyce a ti se mají vyjádřit, zda je daný projev po gramatické stránce správně či nikoliv. Pokud je jejich odpověď negativní, jsou požádáni o interpretaci správné struktury.

- sémantické úsudky – informantům je představeno několik lexémů. Následně informanti mají vyjádřit, v jakém kontextu se dané lexémy typicky vyskytují, či v jakých projevech je považují za vhodné. (Pfau, 2012, s. 1025–1026)

1.3.4 Elicitace dat

Jak již bylo zmíněno, v rámci tvorby korpusu znakového jazyka se jazyková data získávají přímou elicitací od respondentů. Existují různé techniky a materiály, které výzkumníci používají k elicitaci jazykových dat. V následujícím výčtu jsou uvedeny některé elicitační úkoly, které jsou různou měrou využívány.

Úkoly:

- nahrávání přirozeného projevu – tento úkol je také nazýván jako „úkol s malou nebo žádnou kontrolou“. Respondent tedy nemá žádné omezení a může hovořit na jakékoliv téma. V této souvislosti je velmi důležité, aby výzkumník myslel a dával pozor na tzv. paradox pozorovatele. Tato teorie tvrdí, že i když je pozorovatel velmi pozorný k tomu, aby neovlivňoval jazykovou aktivitu respondenta, tak pouhá jeho přítomnost může mít dopad na účastníky výzkumu, kteří s velkou pravděpodobností budou produkovat projev jiným způsobem, než kdyby tam pozorovatel nebyl.
- volné a vedené rozhovory – při volné promluvě je informantovi poskytnuto pouze téma a je požádán, aby hovořil na toto téma. Nicméně i v tomto případě je nutná jistá kontrola k zajištění toho, aby následně bylo možné elicitovat určité větné struktury. Příkladem by mohlo být to, že jsou informanti požádáni o vyprávění jejich dřívějších životních zážitků, a tak jsou získány z jejich projevů číselná data, bez nutnosti oznamovat jim požadavek zapojení číslic do projevu. Při vedeném rozhovoru mají informanti např. nakreslit vlastní rodokmen a mluvit o vztazích v rodině, čímž jsou elicitovány termíny zaměřené na příbuzenské vztahy.
- hraní rolí a simulační hry – toto jsou úkoly, ze kterých lze lehce elicitovat gramatické struktury týkající se tázacích vět ve znakovém jazyce. Jeden informant převezme roli tazatele a druhý se ujme pozice dotazovaného, takže získaná data obsahují mnoho otázek. Samozřejmě mohou být vytvořeny jiné typy úkolů na hraní rolí poskytujících produkci otázek. Simulační hry jsou většinou hrány ve větším měřítku (s více účastníky) a jsou méně přesně definovány. Účastníci dostanou pouze zadání a musí v průběhu času improvizovat, podle toho, kam se konverzace ubírá. V tomto případě nemá výzkumník velkou kontrolu nad projevy respondentů, ale i tak může získat kýžená data, např. informace o využití znakovacího prostoru při střídání rolí.

- komunikační hry – komunikační hry lze použít i ve výzkumech znakového jazyka. Příkladem může být hra hraná dvěma lidmi, kteří se mají podívat na obrázky, které obsahují několik rozdílů. Hráči nevidí obrázek toho druhého a snaží se uhodnout, jaké jsou ty rozdíly, ptaním se na otázky týkající se obrázků. Další možností je hra hraná skupinou lidí, ze kterých jeden účastník myslí např. na slavnou osobu a ostatní musí hádat identitu této osoby dotazováním se otázkami, na které lze odpovědět ano či ne.
- převyprávění příběhu – ve výzkumech znakových jazyků je převyprávění příběhu běžným úkolem pro elicitaci dat. Jsou známy čtyři formy představující převyprávění příběhu:
    - převyprávění obrázkového příběhu – informantům je představen obrázkový příběh, který sestává z kreseb, a informanti mají popsat zobrazené události. Tyto obrázkové příběhy neobsahují žádný typ lingvistické informace, což znamená, že se zde nevyskytuje psaná podoba jazyka. Následující příběhy již byly využity v různých výzkumech znakových jazyků. Například The Horse Story (Hickmann, 2003) byl původně určen pro výzkum mluvených jazyků, ale posléze byl použit i k elicitaci dat znakového jazyka. Také příběh Frog, where are you? (Mayer, 1969) neobsahuje slovní popis děje, a proto je používán k elicitaci dat ve znakovém či mluveném jazyce. Dané příběhy lze posléze využít k výzkumům, které budou porovnávat oba jazyky – mluvený i znakový.
    - převyprávění filmového příběhu – kromě obrázků se využívají i filmové příběhy, které jsou respondentům přehrány celé nebo pouze jejich část a oni je posléze převypráví. Většinou se využívají animované filmy nebo hrané filmy, které neobsahují prvky mluveného ani psaného jazyka, nebo jich zahrnují zanedbatelné množství. Takovými příklady jsou animované seriály Simpsonovi, Růžový panter či Tweety a Sylvester. Jsou to animované pohádky určené pro televizní produkci, avšak existují i filmy, které byly vytvořeny přímo pro použití v lingvistickém výzkumu. Kupříkladu The Pear Story, šestiminutový film vytvořený W. Chafem a jeho týmem v roce 1980, slouží k elicitačním účelům pro výzkumníky na celém světě. Film obsahuje zvukové stopy, nikoli slova.

- převyprávění psaného příběhu – v tomto úkolu již byly využity např. příběhy z Ezopových bajek. Práce s tímto typem přeloženého textu může mít dva nedostatky. Prvním je morfologicko−syntaktické hledisko cílového jazyka, které může být ovlivněno zdrojovým jazykem a za druhé je nutné ujistit se, že informant má dostatečné znalosti v obou jazycích. Přínosem těchto úkolů by mohlo být to, že s přeloženými texty je možné pracovat v rámci paralelních korpusů, např. pro překladové studie.
- převyprávění znakovaného příběhu – některé povídky v nizozemském znakovém jazyce již byly využity jako elicitační materiál pro sběr dat v nizozemském znakovém jazyce. Respondentům byl přehrán znakovaný příběh a oni ho měli převyprávět.

- popis obrázků – informant si prohlédne obrázek nebo sérii obrázků a následně odpovídá na otázky, které jsou vytvořeny tak, aby se daly elicitovat struktury kýžené pro danou studii. V lexikografickém výzkumu se tento typ úkolu vyskytuje poměrně běžně, ale používá se i pro získávání gramatických struktur. (Pfau, 2012, s. 1027–1031)

Výběr elicitačních úkolů má velký vliv na výsledky výzkumu. Povaha údajů, se kterými se pracuje, může ovlivnit názory a postoje respondenta, ale také stanovisko výzkumníků při rozhodování, jak přistupovat k analýze získaného materiálu daného znakového jazyka.

1.3.5 Metadata specifická pro znakový jazyk

Při tvorbě korpusu je nezbytně nutné, aby se sbírala i metadata související s danými lingvistickými daty. Ve stávajících projektech korpusů znakových jazyků byla použita IMDI[11] databáze metadat. Databáze byla vytvořena v rámci projektu ECHO[12] na Max Planck Institute for Psycholinguistics v Nizozemsku.

Příkladem získávání metadat může být případ Costella, Fernándeze a Landy (2008), kteří „*Nahrávali informanty v různých situacích a kontextech, např. při spontánní konverzaci, řízeném rozhovoru a elicitačních úkolech vycházejících z připravených materiálů. Každé nahrané setkání je uloženo v IMDI databázi, což zajišťuje, že všechna související metadata jsou zaznamenána. Metadata spjatá s informantem jsou například tato:*

- *věk, místo narození a pohlaví;*
- *stupeň sluchové ztráty, sluchová vada u rodičů, typ sluchové pomůcky (jestli ji má);*

[11] IMDI = ISLE Meta Data Initiative, databáze standardně používaná k popisu multimediálních a multimodálních jazykových prostředků
[12] ECHO = European Cultural Heritage On-Line

- *věk, od kdy se učí znakový jazyk;*
- *jazyk užívaný v rodině;*
- *vzdělání (věk, vzdělávací program, typ školy);*

taktéž se zadává specifičnost kontextu, v jehož rámci se nahrávání uskutečňovalo:

- *typ komunikačního aktu (dialog, storytelling, otázky a odpovědi);*
- *stupeň formality prostředí;*
- *místo a sociální kontext;*
- *téma.“* (Pfau, 2012, s. 1036)

Pokud někdo z respondentů nesouhlasí se zveřejněním vlastní nahrávky, je možné ji i s metadaty nezveřejňovat. Nicméně vyvstávají otázky, zda by nemělo být určeno nějaké datum vypršení doby anonymnosti nahrávky, kdy by posléze mohla být nahrávka zveřejněna i s příslušnými metadaty. Dále se také diskutuje o jednotlivých údajích v rámci metadat, která by se měla zjišťovat. Jisté informace by měly být buď upraveny, nebo vymazány z důvodu nejasnosti či nevhodnosti dat. Mezi tyto údaje je zařazován vzdělávací model, resp. jeho hodnota „orální vzdělávání“. Není totiž jasné, zda se jedná o vzdělávací politiku či jazykové prostředí. Taktéž není jistota u hodnoty sluchového postižení respondenta. Nedoslýchavý respondent se totiž může považovat za neslyšícího a tento fakt reflektovat v metadatech, což ovšem mění hodnoty zanesených dat. Zmiňuje se i fakt, že pro někoho může být tato položka příliš osobní, a proto by se měla úplně z metadat vyjmout. (SLCN Metadata, 2009)

1.3.6 Výběr informantů

Kompetence mluvčích v jednom jazyce se může výrazně lišit. Tato situace je s velkou pravděpodobností podobná ve všech jazycích a komunitách, avšak výrazněji se projevuje v komunitě neslyšících komunikující znakovým jazykem, což zřejmě souvisí s faktem, že se 90–95 % neslyšících dětí rodí slyšícím rodičům, kteří neznají a nepoužívají znakový jazyk. Znakový jazyk si tedy děti začnou osvojovat až při příchodu do školy pro sluchově postižené. Je nutné tedy hledat respondenty, kteří budou schopni poskytnout takový typ projevu, který bude možné zařadit do korpusu.

Přirozená data jsou požadována pro všechny jazyky, a jak uvádí Costello, Fernández a Landa (2008) „*neexistuje žádná dohodnutá definice rodilého mluvčího, ani žádné vysvětlení tohoto termínu.*“ Nejlepší možná volba vzoru rodilé znakující osoby je informant, který pochází z (nejméně) druhé generace neslyšící znakující rodiny. Nicméně v malé komunitě, jako je komunita neslyšících je těchto „ideálních“ informantů velmi málo.

V takových případech si tedy výzkumníci stanoví několik kritérií, které nerodilí znakující respondenti musí splňovat. Tato kritéria zahrnují například brzké vystavení znakovému jazyku, kdy se uvádí osvojení jazyka do věku 3 v některých případech až 7 let.
Další podmínkou může být vzdělání ve škole pro neslyšící žáky s požadavkem internátního ubytování, denní používání znakového jazyka či déletrvající členství v komunitě Neslyšících. (Pfau, 2012, s. 1036–1038)

1.3.7 Nahrávání jazykových dat

Při tvorbě korpusů znakových jazyků musí výzkumníci řešit metodické záležitosti, které jsou pro tyto korpusy specifické. Například to, že data nemohou být nahrávána jako audionahrávka, ale musí být zachyceny jako videonahrávky. Výzkumník se také musí postarat o vysokou kvalitu nahrávek, ale zároveň musí dbát na to, aby co nejméně ovlivnil jazykovou produkci respondenta. Je známo, že uživatelé jazyka jsou ovlivněni formalitou prostředí, což je srovnatelné u výzkumů týkajících se znakového i mluveného jazyka. Ovšem znakový jazyk má ještě další specifikum spojené s minoritním zastoupením ve společnosti, a tudíž jeho ovlivňováním majoritním mluveným jazykem, což přidává faktor, který musí být brán v úvahu. Problém související s touto otázku je tendence neslyšících osob přizpůsobovat svoji jazykovou produkci (slyšícímu) účastníkovi dialogu.

Projevuje se tedy snaha připravit takové prostředí, které se co nejvíce vyhýbá ovlivňování respondenta, což zahrnuje následující podmínky:

- sběr dat v klidném prostředí,
- neslyšící konverzační partner, který nemusí být známý danému respondentovi, avšak přítomnost neznámého člověka, hlavně pokud je vysoce vzdělaný, může mít dopad na jazykovou produkci, proto je lepší pracovat s lidmi, kteří se vzájemně znají, nicméně nemají příliš blízký vztah (manžel, manželka, sourozenec), protože tento kontakt může způsobit produkci pro ně specifického jazyka, který by ovšem nebyl pro korpus reprezentativním vzorkem,
- vyhnout se přítomnosti slyšících osob při natáčení,
- používat pouze nezbytně nutné množství kamer a vyhnout se dalšímu nahrávacímu zařízení a světlům,
- vyřazení prvních deseti minut nahrávky z důvodu možného počátečního dyskomfortu respondentů z neznámého prostředí. (Pfau, 2012, s. 1038–1040)

1.3.8 Technické vybavení

V rámci natáčení je nutné také zabezpečit takové podmínky prostředí, které přispějí ke kvalitní videonahrávce. Základem je zajištění následujících faktorů:

- oblečení – respondenti by měli mít oblečení kontrastující s jejich barvou pokožky, tedy světlého či tmavého odstínu, podle barvy kůže respondenta. Šperky jsou považovány za rozptylující, a tudíž by měly být co nejvíce eliminovány.
- pozadí – pozadí prostředí ovlivňuje viditelnost znakového projevu, takže je vhodné použít jednobarevné nevzorované pozadí, většinou zelené nebo modré barvy, které je důležité pro následné klíčování pozadí v počítačovém programu.
- usazení v prostoru – je nutné mít na paměti, že když znakující osoba sedí, tak se nepatrně mění znakovací prostor na rozdíl od situace, kdy při znakování člověk stojí. Toto může být důležitým faktorem pro pozdější fonetický a fonologický výzkum znakového jazyka.
- kamery – množství kamer a jejich umístění je určeno specifičností výzkumné otázky, např. analýza nemanuální aktivity při znakování vyžaduje kamery s možností velmi blízkého přiblížení na informantův obličej.
- pozice kamer ve vztahu ke znakujícím(u) – vždy je nutné zajistit zachycení celého prostoru. Někteří výzkumníci se vyhýbají plně frontálnímu nahrávání, kdy raději volí natáčení pod jistým bočním úhlem, aby se respondenti necítili příliš nepohodlně. Kamerovým pohledem shora je usnadněna případná analýza vztahu mezi rukama a tělem, což může být důležité při výzkumu využití prostoru ve znakovém jazyce.
- použití elicitačního materiálu – znakující by během produkce neměl držet žádné věci a ani by neměl začít znakovat, když hledí do elicitačního materiálu. (Pfau, 2012, s. 1040–1041)

1.3.9 Anonymita

Velkou nevýhodou nahrávání znakového projevu je to, že je spojené se ztrátou anonymity respondentů. Při prezentování nebo publikování práce si výzkumníci přejí ilustrovat výsledky s příklady z natočených nahrávek, nicméně ne všichni znakující schvalují to, aby byl jejich projev zveřejněn. Na nahrávkách není možné např. digitálně překrýt respondentům obličej, kvůli nezbytnosti viditelnosti nemanuálních komponentů ve znakovaném projevu. Možností by byla reprodukce projevu někým jiným, ale to by bylo vhodné jen v případě zopakování izolovaných znaků, avšak s delšími promluvami by byl nejspíše problém z hlediska ztráty či změny znaků v projevu.

Proto se řeší otázka anonymity tím, že výzkumníci nechávají respondenty podepsat prohlášení, ve kterém souhlasí se zveřejněním nahrávky či případně zahrnují možnost vystřihnutí jisté části nahrávky v případě, že ji nebudou chtít respondenti zveřejnit. (Pfau, 2012, s. 1041)

1.3.10 Reprezentativnost korpusu

Nyní je vhodné objasnit si, které prvky by měl obsahovat jazykový korpus. Existuje řada kritérií, která by korpus měl splňovat, aby se dal považovat za důvěryhodný zdroj informací. Jedním z nejdůležitějších znaků je reprezentativnost korpusu. Reprezentativnost souvisí s obsahem korpusu, tedy jazykovým vzorkem zastoupeným v něm a jazykovou realitou, tudíž jazykem reálně užívaným. Jestliže korpus zahrnuje různé soubory textů, které svým množstvím odpovídají všem varietám daného jazyka, poté lze říci, že je korpus reprezentativní. (Cvrček, Richterová, 2013a) Čermák (1995, s. 124) se mimo jiné také zmiňuje o tzv. interních a externích kritériích. Za interní považuje ta, která se dotýkají lingvistiky, což je např. formálnost či neformálnost textu. Externí čili nelingvistická kritéria se naopak zaobírají nejazykovou stránkou, takže si všímá původu textů či jejich připravenosti.

Je nasnadě se ptát, jak obsáhlý by měl korpus být, aby mohl být považován za reprezentativní a zároveň, aby měl nějaký přínos pro uživatele. Šulc (2001, s. 53) říká, že za standard se dá považovat korpus o velikosti 100 milionů slovních výskytů.

Reprezentativnost korpusů mluvených jazyků je kromě jiného určována počtem zpracovaných slov, ovšem u korpusů znakových jazyků toto kritérium nelze zcela plošně užít, proto se autoři raději zaměřují na množství respondentů, od kterých získávají kýžený jazykový materiál. Z dostupných materiálů lze zjistit, že se reprezentativnost korpusů vyjadřuje v jistém počtu respondentů zahrnutých v elicitaci jazykového projevu či v počtu lemmat, která jsou v databázi zpracována. Častěji se uvádí množství respondentů, u nichž se stanoví jistá kritéria, která mají být dodržena vzhledem k obsahu databáze, a tak byla částečně zabezpečena reprezentativnost vzorku.

Jednotlivé korpusy znakových jazyků se poměrně shodují v počtu respondentů zahrnutých do sběru jazykového materiálu. Jednotlivá čísla se pohybují v řádu kolem 100 účastníků, i když lze nalézt i výjimky jako je korpus švédského znakového jazyka, který zahrnul do sběru jazykových dat pouhých 42 účastníků (Borstell, 2016, s. 20). Naproti tomu při tvorbě korpusu britského znakového jazyka pracovali s 249 mluvčími znakového jazyka (Schembri, 2013, s. 141).

Rozdílnost těchto dat je způsobena množstvím lidí, kteří jednak mohou být zařazeni do daného projektu jako vhodní participanti, ale také tím, zda jsou ochotni zúčastnit se takové činnosti s ohledem na případnou ztrátu anonymity.

Dalším možným vodítkem, které nám může být nápomocné v určování, zda je či není daný korpus reprezentativní, je množství lemmat, které databáze obsahuje. Korpus řeckého znakového jazyka zahrnuje 6 000 lemmatizovaných hesel s plánem dosáhnout čísla 10 000 (Eftihimiou, 2016, s. 64). Korpus britského znakového jazyka se dostal až na 40 000 anotovaných lexikálních jednotek, což je pochopitelné k již výše zmíněnému počtu respondentů (Schembri, 2013, s. 147). Samozřejmě rozsah není zárukou kvality korpusu, proto se přistupuje k požadavku získání takového vzorku populace, který by věrně odrážel povahu dané skupiny.

Již byly zmíněny otázky týkající se kritérií, která se stanovují pro lepší zachycení variantnosti znakového jazyka se zřetelem k věku, pohlaví, vzdělání, místu bydliště či jazykovému zázemí. Jazykovým zázemím se myslí fakt, zda daný člověk vyrůstal v rodině, kde byli slyšící či neslyšící rodiče, čímž se dostáváme k dalšímu specifickému kritériu a to k tomu, zda je daný jedinec rodilým mluvčím znakového jazyka, případně, kdy se naučil znakovat. Snaha všech směřuje k tomu, aby v korpusu bylo zastoupeno přibližně stejné procento mužů i žen v jistých věkových skupinách, které jsou většinou 3 až 4, stupňují se po 10 až 15 letech a začínají zahrnovat respondenty až kolem 20. roku života.

Dále je dobré se ptát, co by daný korpus měl obsahovat, či lépe, jaké typy projevů, zda elicitované skrze speciálně připravené materiály či volně vedené a zachycené rozhovory. Ve většině případů se přistupuje k zachycení obou typů těchto projevů, aby se daly postihnout různé oblasti jazyka a předešlo se nechtěnému vynechání podstatného jazykového jevu.

1.3.11 Nemanuální prostředky znakového jazyka v korpusu

Ve znakových jazycích se vyskytují dva typy artikulátorů, manuální[13] a nemanuální.[14] Soudí se, že manuální artikulátory nesou v první řadě významy lexikální a nemanuální artikulátory jsou primárně zaměřené na gramatické významy. Vyvstává tedy otázka, zda a jakým způsobem by se měly tyto jednotlivé prvky znakového jazyka postihnout a zařadit do korpusu.

[13] manuální artikulátor – ruka, obě ruce
[14] nemanuální artikulátor – obličej, hlava, horní část trupu

V některých případech se lze setkat s problematikou toho, zda do korpusu zařazovat i gesta, protože gesta byla označována za nejazykové prostředky, tudíž nepatřící do přirozeného jazyka, kterým ale znakový jazyk je. Avšak toto tvrzení se začalo zpochybňovat a britský lingvista McNeill (1985, s. 350) tvrdí, že i gesta se dají analyzovat, proto je nutné s nimi ve znakových jazycích počítat a některé jazykové korpusy je tak mohou zahrnovat do své databáze, jako se to například stalo v korpusu švédského znakového jazyka.

Na příkladu dvou korpusů bude nastíněna problematika zpracování nemanuálních prostředků v dané jazykové databázi. Pracovníci korpusu francouzského znakového jazyka se zaměřili na možnost zpracování nemanuálního artikulátoru – obočí. Jedna z obtíží byla ta, že u popisu pohybů obočí nelze zachytit intenzitu pohybu ani dobu trvání pohybu obočí, jak by to bylo možné u pohybu rukou. Dále se potýkali s nedostatečně přesným popisem tohoto prvku uvedeným v notačních systémech jako je HamNoSys[15] nebo SignWriting,[16] proto pro svoji potřebu navrhli novou metodu, jak popsat pohyby obočí. Užili systém FACS,[17] který byl vyvinut pro deskripci emocí skrze mimiku obličeje a pracuje na principu rozeznání výrazů tváře na základě pohybů obličejových svalů. Avšak FACS rozlišuje pouze dvě polohy obočí - zvednuté a snížené, z toho důvodu si do rozpoznávacích kritérií přidali ještě jedno postavení a to neutrálně posazené obočí. Také se vyskytnul problém, jak přesně zachytit pohyby obočí u respondentů, s různě širokým obočím, aby systém úspěšně rozeznal daný pohyb. Proto bylo nutné upravit nastavení programu, kdy byla zvětšována tolerance krajních bodů zachycovaných na obočí, aby měl software jistou zálohu pro rozeznání jednotlivých pohybů. Celkově bylo nutné označit 18 bodů na obou obočích k zachycení kýžených informací, a to vše nezávisle na pohybech hlavy. (Chételat-Pelé, 2008, s. 29−30)

V Německu se zabývali možnostmi, jak zachytit a popsat mluvní komponenty[18] německého znakového jazyka, protože jsou nedílnou a významnou součástí znakového projevu. Zde se zaměřili na rozpoznávání mluvních komponentů u slyšících tlumočníků německého znakového jazyka na videonahrávkách získaných z televizního pořadu vysílaného v letech 2009 – 2011, ve kterém tlumočili počasí do německého znakového jazyka. Nejdříve bylo nutné zjistit, zda se mluvní komponenty slyšících tlumočníků neliší od těch, které užívají neslyšící rodilí mluvčí znakového jazyka.

[15] HamNoSys – Hamburský notační systém, který využívá ikonicitu užívaných symbolů pro zápis znaků znakového jazyka
[16] SignWriting – notační systém založený na kombinaci ikonických symbolů představující části těla a pohyby, které jsou zobrazeny kreslenými obrázky
[17] FACS – Facial Action Coding System
[18] mluvní komponent - pohyby úst artikulující slovo, slabiku či hlásku z příslušného národního jazyka, které jsou simultánně produkovány se znaky znakového jazyka

Posléze bylo rozhodnuto použít projevy slyšících tlumočníků vzhledem k vázanosti mluvních komponentů na mluvený jazyk. Následně ještě bylo nutné, aby v projevu byly dostatečně rozlišeny orální komponenty[19] od mluvních komponentů. Při rozpoznávání jednotlivých slov bylo důležité, aby byly přesně odlišeny pohyby rtů a tváří, ale také zubů a jazyka. Proto v oblasti úst bylo popsáno deset bodů, které byly od sebe vzdáleny v určitém intervalu a tři v oblasti uvnitř úst k zachycení polohy jazyka a zubů. Následně tedy slovo, které určil program ASR[20] sestavený pro automatické rozpoznávání slov, bylo přepsáno pomocí programu GIZA++ běžně používaném ve statistickém strojovém překladu ze zdrojového jazyka do cílového, a tak byl získán žádaný mluvní projev v psané podobě. Navíc databáze slov v ASR je sestavena tak, že zahrnuje dostatečně velkou množinu možných výslovností pro každé německé slovo, takže bylo jisté, že se postihne každé pronesené slovo. (Koller, 2014, s. 89–90)

Toto jsou jen zlomky záležitostí, které musí řešit každý, kdo chce vytvořit plnohodnotný korpus znakového jazyka a musí tedy počítat i s takovými prvky, jako je zpracování nemanuálních prostředků a jak je co nejlépe zachytit a zpracovat v rámci korpusu.

[19] orální komponent - pohyby úst, které nepocházejí z národního mluveného jazyka
[20] ASR - Automatic Speech Recognition

2. Korpus australského znakového jazyka

V následujícím textu jsou popsány a vysvětleny jednotlivé kroky a myšlenky, které byly použity při vytváření korpusu australského znakového jazyka (dále jen Auslan). Vytvořit korpus znakového jazyka je nejen pro lingvisty výzvou. Tvůrci takového korpusu se potýkají s mnoha překážkami, které musí řešit, ty se týkají nejen jazykového hlediska, ale i softwarového zpracování dat. Existence korpusu nejen australského znakového jazyka je potřebná k jasnému ukotvení popisu znakového jazyka. Touto cestou je možné podložit lingvistické teorie o znakovém jazyce hmatatelnými důkazy, které lze doložit a nespoléhat se tak pouze na nahodilá pozorování, která nejsou nikde zaznamenána.

Korpus je neméně důležitý pro testování hypotéz v lingvistických výzkumech na všech úrovních od fonologie přes morfologii, syntax až k pragmatické rovině. Lze uvést několik důvodů, proč je analýza jazyka založená na korpusových datech obzvláště důležitá v oblasti znakového jazyka. Za prvé jsou znakové jazyky ještě stále považovány za mladé jazyky menšinových komunit, ovšem s vlastními gramatickými jevy. Dále je nutné podotknout, že reprezentativnost projevů ve znakovém jazyce s využitím psaných glos znamená, že získaná primární data dříve zůstávala v podstatě nepřístupná dalším výzkumníkům a nebyla tedy následně k dispozici pro smysluplné výzkumy, které by plně využily jejich potenciálu, protože zápis glosami nevystihuje pravou podstatu znakového jazyka a navíc ne všichni výzkumníci uměli pracovat s těmito zápisy. Takže ačkoliv pozorování a introspekce mohou být stále cennou pomůckou pro lingvisty, kteří rozvíjí hypotézy týkající se užívání a struktury znakového jazyka, je také nutné uznat, že podobná intuice může kdykoliv selhat. (Johnston, 2009, s. 87−95)

Ke zdokumentování znakového jazyka se přistupuje také s perspektivou zachycení jazyka komunity pro budoucí využití, které dosud nebylo možné. Tvorba korpusu tedy pramení z potřeby zachytit jazyk v té formě, která se nyní užívá, vzhledem k tomu, že není k dispozici mnoho zachycených historických projevů ve znakovém jazyce.

Taktéž není možnost zachytit znakový jazyk psanou formou, což znemožňuje jeho retrospektivní zkoumání a v neposlední řadě dochází k přerušení mezigeneračního transferu jazyka, z toho důvodu, že se neslyšící děti rodí do slyšících rodin, kde se znakový jazyk neužívá či se vzhledem k rozvoji kompenzačních pomůcek potřeba znakového jazyka jako komunikačního prostředku dostává do pozadí zájmu.

Další využití skýtají sesbíraná data v tom, že na jejich základě lze tvořit plnohodnotné a smysluplné výukové materiály či pomůcky nejen pro lektory znakového jazyka, ale samozřejmě také pro studenty, kterým by dostupné podklady mohly sloužit pro samostudium. (Johnston, 2016)

2.1 Auslan archiv

Nyní bude popsáno pozadí tvorby korpusu Auslan. Korpus je založen na digitálním video archivu,[21] sestávajícího z reprezentativního vzorku nahrávek projevů rodilých neslyšících spojených s anotacemi a metadaty (Johnston, Schembri, 2006). Samotný archiv vznikal mezi lety 2004−2006, od roku 2008 je součástí Archivu ohrožených jazyků ELAR[22] a od roku 2012 se archiv stal přístupný veřejnosti.

Auslan korpus byl financován skrze projekt The Hans Rausing Endangered Languages Project a zásluhu na tom měl Trevor Johnston, tak jako za v podstatě celou tvorbu korpusu australského znakového jazyka. Johnston měl dvě hlavní představy, a to vytvořit a zajistit referenční archiv Auslanu a také vytvořit lingvistický korpus, který bude využit k rozličným lingvistickým účelům. (The Auslan Corpus, 2012)

Tento archiv obsahuje dva datové soubory. Jeden obsahuje data sesbíraná v rámci projektu, který zkoumal sociolingvistickou variantu jazyka v Auslanu prováděným Trevorem Johnstonem a Adamem Schembrim v letech 2003 až 2005. Druhá část obsahuje data získaná v rámci Endangered Language Documentation Project financovaným z The Hans Rausing Endangered Languages Documentation Programme (ELDP) na School of Oriental and African Studies (SOAS) na Londýnské univerzitě, který probíhal v letech 2004−2007. Oba soubory dohromady obsahují 200 hodin videonahrávek, na kterých je zachycena produkce australského znakového jazyka rodilými nebo skororodilými[23] neslyšícími uživateli tohoto jazyka. (Johnston, 2016)

[21] Archiv videí: https://elar.soas.ac.uk/Collection/MPI55247
[22] ELAR = Endangered Languages ARchive
[23] skororodilí mluvčí znakového jazyka – ty osoby, které si neosvojily znakový jazyk hned od narození, ovšem osvojily si ho před 7. rokem života

2.2 Auslan korpus

Jak již bylo zmíněno výše, Auslan korpus je založen na Auslan archivu, ve kterém jsou uloženy videonahrávky projevů ve znakovém jazyce. Než začala práce na samotném oslovování respondentů a následném sběru dat, výzkumníci si museli nejdříve ujasnit, jaké populační zastoupení musí být v korpusu obsaženo, tedy určit požadavky na místo pobytu, věkové rozhraní či pohlaví účastníků, aby se korpus dal považovat za reprezentativní.

Nakonec se celkový počet všech respondentů, kteří se zúčastnili elicitace dat v letech 2004−2007, vyšplhal na číslo 256. Byli to lidé z různých míst Austrálie, v různých věkových kategoriích od 18 do 85 let, ale s vyrovnaným počtem zúčastněných mužů a žen. Do tohoto vzorku ale nebyly zahrnuty osoby CODA a ti, kteří se naučili znakový jazyk po 7. roce života.

Respondenti se zúčastnili natáčení, které trvalo kolem tří hodin a probíhalo tak, že vždy, když to bylo v rámci elicitace nutné, probíhala interakce mezi dvěma respondenty. Setkání vedl rodilý mluvčí znakového jazyka, který byl dostatečně proškolen v dané problematice, a konkrétní respondenti ho obvykle dobře znali. (The Auslan Corpus, 2012) Toto kritérium je důležité z jazykového hlediska, aby bylo zaručené, že respondenti budou rozumět zadání jednotlivých částí elicitace, ale také, že budou mít pocit komfortu a nebudou cítit stud nebo antipatie k osobě, která povede celé sezení.

Jednotlivá zadání na produkci znakového jazyka byla rozdělena do jedenácti kategorií, podle různě zaměřených úkolů, které souvisely s odlišnými funkcemi jazyka. Zadání se týkala následujících oblastí (Johnston, 2009, s. 1−17):

- hláskování svého jména prstovou abecedou a ukázáni jmenného znaku a jeho vysvětlení, pokud nějaké existuje;
- převyprávění Ezopovi bajky (bajka Zajíc a želva nebo Chlapec, který křičel „vlk“) založené na psaném textu, který respondenti dostali týden před natáčením;
- převyprávění nezapomenutelné životní události;
- vyjádření se na témata hluchoty a znevýhodnění, rodičovské preference, nové reprodukční možnosti, změny v komunitě Neslyšících za posledních 30 let;
- volná konverzace;
- zhlédnutí a převyprávění kresleného filmu Tweety a Sylvester (díl Canary Row);
- prohlédnutí si a vyprávění obrázkového příběhu z knihy (Žábo, kde jsi);
- zhlédnutí příběhu v australském znakovém jazyce a jeho převyprávění;
- zhlédnutí a popis videí, kde se hýbaly hračky (užití klasifikátorů) a kde lidé dělali nějaké činnosti (dvojice podstatné jméno a sloveso);

- zhlédnutí a popis 18 obrázků, na kterých byly zobrazeny různé činnosti (činnosti mohly být popsány tak, že šlo změnit pořadí znaků ve větě a význam věty se nezměnil a ty, které bylo nutné popsat pomocí vět s pevným znakosledem (zadání pro elicitaci pořadí znaků ve větě));
- hraní hry „označ rozdíl“ při užití dvou mírně odlišných obrázků (zaměřené na tvorbu otázek).

Celkově bylo natočeno 200 hodin materiálu, který byl rozčleněn do 1 100 videonahrávek. Samozřejmě bylo nutné tyto videonahrávky opatřit lingvistickými informacemi, protože vytvořený archiv těchto znakovaných projevů netvoří korpus v jeho pravém smyslu slova. Archiv jsou pouze natočená a uložená data, která nepřináší žádnou další nosnou hodnotu. (Johnston, 2016).

V únoru 2016, prošlo již 459 videonahrávek, z celkových 1 100, primárním zpracováním, tj. základní anotací, přiřazením ID glos a přiřazením volného překladu příslušného znakového projevu do anglického jazyka. To představuje pouze kolem 14 hodin zpracovaného materiálu z již zmíněných 200 hodin a více než 105 000 glosovaných tokenů znaků. Určité množství dat prošlo už i sekundárním a terciálním zpracováním, konkrétně 50 videonahrávek bylo zpracováno jako součást výzkumného projektu, který se zabýval gramatickým využitím prostoru v Auslanu, dalších 50 videonahrávek bylo součástí analýzy zjišťující gramatikalizaci znaků souvisejících s dokončením děje v Auslanu, tzn. všechny mluvní a orální komponenty související s těmito znaky byly anotovány. A v dalších 100 videonahrávkách, byly na větné úrovni přesně vymezeny jednotky, u kterých byly identifikovány argumenty pro větné členy. Tento poslední projekt byl zaměřen zejména na gramatiku australského znakového jazyka. (Johnston, 2016, s. 8)

2.2.1 Pojmenování souborů v korpusu

Všechna korpusová data musí být systematickým způsobem pojmenována, aby se s nimi dalo následně rychle a jednoduše pracovat. Proto lze původní digitální nahrávku, ze které pochází následně sestříhaný klip, jednoduše identifikovat a dohledat v případě, že by data musela být znovu nějak zpracovávána či opětovně digitalizována.

Pro názvy jednotlivých prvků v korpusu se vytvořily značky, které se následně připojily k příslušným datům.

U všech znakujících osob se předpokládá, že mají dominantní[24] pravou ruku a proto se informace o dominantnosti připojuje k datům, pouze pokud tomu tak není. Takže se připojuje značka LH (left hand) pro levou ruku a AMBI (ambidextrous) pro oboustrannou dominanci rukou, ale tento případ se vyskytnul v tomto korpusu pouze jednou.

Nahrávaná setkání v rámci sběru dat do korpusu byla prostorově rozdělena na dvě části. Osoba nalevo byla označena značkou A, a člověk napravo značkou B. Zachycené aktivity (rozhovor, konverzace, převyprávění apod.) byly očíslovány c1 (clip) až do c11.

Názvy přípon souborů se shodují se souvisejícími videonahrávkami, podle jejich typu formátování, tedy např. .mov, .wmv, .dv, .mp4 atd., anotované soubory nesou označení .eaf, soubory metadat jsou pojmenovány příponou .imdi. Funkce .imdi metadat nebyla v korpusu použita, protože jsou metadata uložena separovaně v jiném tabulkovém procesoru.

V pracovní verzi korpusu, která však není veřejně dostupná, mají k sobě data přiřazena značky pro metadata pohlaví (_M nebo _F), pro věk (_#) a pro to, zda jsou rodilí mluvčí nebo skoro rodilý. _NN (near native) pro skororodilé a _N (native) pro rodilé mluvčí. (Johnston, 2016, s. 10)

2.2.2 Metadata

Jak již bylo uvedeno, k videonahrávkám v archivu jsou připojena metadata, proto by bylo dobré uvést podrobnější informace k této záležitosti. Metadata jsou uváděna na všech anotačních úrovních a jsou ukládána v databázi IMDI.[25] Databáze IMDI byla vyvinuta v Psycholingvistickém institutu Max Planck v Nizozemsku. Zde byl také vyvinut software ELAN, důležitý pro fungování korpusu, ale o tomto softwaru bude zmínka později. Metadata v tomto případě obsahují informace o médiu a datech v následujících kategoriích:

- účastník (bydliště, pohlaví, věk, vzdělání atd.);
- obsah (různé jazykové úkoly, použité elicitační materiály);
- médium (formát a typ);
- projekt (název, jazyk, metodologie);
- setkání (název úkolu, účastníci atd.);
- související psané zdroje (pokud existují elicitační materiály pro multimediální soubory).

[24] dominantní ruka - ruka, která vykonává větší pohybovou aktivitu či preferovaná ruka pro motorické úkony (pro praváka většinou pravá a naopak)
[25] IMDI - ISLE Meta Data Initiative (ISLE - International Standard for Language Engineering)

V korpusu je možné uplatňovat vyhledávací postupy, které kombinují vyhledávací možnosti ELANu a kritéria zanesená v metadatech IMDI. Tak je například možné vyhledat všechny výsledky pro znak BAD-LUCK (neštěstí), který produkovali muži nad 50 let z jižní části Austrálie. (The Auslan Corpus, 2012)

2.2.3 Základní proces zpracování jazykových dat

Pro lepší představu následuje stručný popis základního procesu tvorby korpusu, který bude dále podrobněji vysvětlen. Nejdříve probíhá sběr jazykových dat od respondentů. Tato data jsou následně zpracována ve vybraném softwarovém programu. Anotace je náročný a zdlouhavý proces, který zahrnuje zpracování jazykových dat, která mají podobu videonahrávek znakového projevu. Znakový projev se přeloží do psané podoby příslušného národního jazyka, dále se segmentuje (tzn. tokenizace) do jednotlivých znakových tokenů a následně jsou k tokenům připojeny příslušné glosy. Nakonec jsou přiřazeny jednotlivé tagy.

2.2.4 Notace, přepis, anotace a tagování

Jak je známo, znakové jazyky jsou vizuálně motorické jazyky, které nemají akceptovanou písemnou formu nebo standardizovaný notační systém, který by umožňoval záznam toho, co bylo proneseno. Při tvorbě korpusu Auslan se ukázalo, že je nutné vyřešit otázku zpracování dat, tedy jak získaná jazyková data zakódovat ve strojově čitelné texty. (Johnston, 2009, s. 87−95)

Dříve přepisy znakových projevů přinesly pouze malé procento zpracovaných souborů, nereprezentativních a počítačově nečitelných. Detailní fonetický nebo fonologický přepis, kterým se zabývalo mnoho výzkumníku, zkonzumoval spoustu jejich času a úsilí a výsledkem byla pouze malá část zpracovaných textů bez určení jednotek na jakékoliv lingvistické úrovni kromě určení samotného znaku. (The Auslan Corpus, 2012)

2.2.4.1 Notace a přepis

Notace dle Johnstona (2016) je definována buď jako zápis jednotlivých slov či znaků nebo jako systém symbolů použitých pro tento účel. Příkladem notace je systém HamNoSys.

Přepis je grafické zachycení promluvy, která byla produkována znakovým či mluveným jazykem, tj. projev, který byl znakován nebo pronášen hlasem. Obecně lze říci, že přepis je obvykle používán jako referenční bod pro lingvistickou analýzu, např. při tvorbě textu pro psané jazyky, pro fonologické analýzy nebo pro gramatické analýzy.

Také slouží jako zachycení psané formy pro původní projevy, které jsou převedeny do strojově−čitelného souboru a díky tomu mohou být následně zpracovány v počítači. Pokud jsou tedy slova či znaky znakového jazyka přepsány do psané podoby, tak mohou být následně anotovány na různých lingvistických úrovních.

Jeden z hlavních účelů přepisu a notace je umožnit čtenáři, aby mohl původní projev s menší či větší přesností zpětně reprodukovat pomocí zapsaných grafických symbolů, což také závisí na stupni zaznamenaných detailů v notaci a přepisu. V rámci znakových jazyků se ale také lze setkat s avatarem,[26] který by měl být schopný opakovat to, co dotyčný člověk předtím znakoval. (Johnston, 2016)

Poněkud překvapivě je ještě stále výzkumníky znakových jazyků předpokládáno, že přepis znakovaného projevu je prvním a nezbytným krokem k tvorbě korpusu znakového jazyka. Avšak vzhledem k tomu, že jsou přepisem a notací obvykle zabrány hodiny práce při zpracovávání pouhé minutové videonahrávky, výsledkem je v podstatě nevyužitelný výstup strojově−čitelného textu, což představuje nejen významné plýtvání zdrojů skrze promrhání možnosti využití výhodného potenciálu nové technologie, ale také to reprezentuje zásadní nepochopení povahy moderního lingvistického korpusu. (Johnston, 2010, s. 110)

2.2.4.2 Anotace a tagování

Zpočátku byla anotace brána pouze jako druh přidaného komentáře, který byl přiřazen k již existujícímu psanému textu, ať už šlo o přepis mluvené promluvy nebo o nějaký jiný projev, tj. v tomto případě vyjádření znakovým jazykem. Anotace byla většinou dvojjazyčný komentář v obtížných zahraničních textech a byla vytvořena proto, aby pomáhala k pochopení daného textu.

Z lingvistického hlediska jsou anotace vytvořeny jako takové mini lingvistické komentáře, které jsou důležité proto, aby identifikovaly jednotky jazyka. Anotace připojují fonologické, morfologické, syntaktické a sémantické informace o znaku, podle účelu zamýšleného výzkumu. Jako takové jsou anotace pro lingvisty neocenitelnou pomůckou při rozeznávaní schémat na různých jazykových úrovních.

Anotace a tagy poskytují lingvisticky relevantní informace o jednotce jazyka. Nicméně to, co se nyní běžně nazývá tagování, tak se vlastně vztahuje k typu automatické anotace, kdy se připojí značka k psanému textu poté, co slova či znaky byly digitalizovány a následně zpracovány pomocí počítače.

[26] avatar – vizuální prezentace osoby v počítačové animaci

Např. přiřazení tagu k psané anglické větě: *Joanna stubbed out her cigarette with unnecessary fierceness*. (Joanna zbytečně prudce típla cigaretu.) může být z větší části provedeno automaticky s odvoláním na využití elektronického slovníku angličtiny ve spojení s jednoduchými pravidly slovních kolokací a rozdělením slovních jednotek. Jako tagy se využívají značky, které velkými písmeny ve zkratce označují gramatickou informaci jednotlivých znaků.

Příklad tagů

Tag	Význam	
_NP	singular proper noun	(vlastní jméno, jednotné číslo)
_VBD	past tense form of lexical verb	(sloveso, minulý čas)
_RP	adverbial particle	(příslovečné určení)
_PP$	possessive pronoun	(přivlastňovací zájmeno)
_NN	singular common noun	(podstatné jméno, jednotné číslo)
_IN	preposition	(předložka)
_JJ	adjective	(přídavné jméno)
_.	full stop	(tečka)

Příklad tagování slov ve větě

Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN unnecessary_JJ fierceness_NN ._.

(Johnston, 2010, s. 19−20)

Na anotování by se dalo nahlížet jako na nikdy nekončící proces ve dvou významech. Zaprvé anotace není nikdy kompletní ve smyslu, že není imunní vůči korekci a zadruhé není ucelená v tom smyslu, že se liší perspektivy (teoretické i praktické), které mohou být použity při náhledu na stejný projev, což umožňuje, aby byl anotován různými způsoby.

Očekává se, že se anotace v korpusu budou postupem času neustále opravovat a rozšiřovat. Anotace jsou revidovány druhým anotátorem a následné úpravy by nikdy neměl provádět ten stejný anotátor. Dále také mohou být již existující anotace rozšiřovány nebo obohacovány dalšími výzkumníky skrze následující jiné anotace, kterými videonahrávka prochází. Následně přidané anotace se mohou týkat buď identifikování jednotlivých znaků či složených znakových konstrukcí jako jsou věty nebo fráze. Dále mohou být přidány nové lingvistické informace, které byly nově rozpoznány, a v předchozí anotaci je nezaznamenali.

Opakovaná anotace činí jednotlivé položky a tím i celý korpus velmi detailním a bohatým na data vhodná k výzkumu.

Když se při kontrole anotací najde chyba, je označena slovem error (chyba) a je k ní připojen komentář, což umožňuje rychle rozpoznat možné chyby ještě předtím, než je rozhodnuto, jestli by byla korekce oprávněná, např. porovnáním s lexikální databází nebo v diskuzi s původním anotátorem, výzkumným týmem či skupinou neslyšících užívajících znakový jazyk apod. Tímto se vyhne tomu, že by něco bylo změněno v anotacích na dalších úrovních, což by vedlo k nepředvídatelným nebo neviditelným nesrovnalostem, které by narušily integritu dat. Taktéž šetří čas, když jeden anotátor nebo výzkumník může zafixovat něco, co by další anotátor, který by jistou položku považoval za chybu, mohl později vymazat a takhle by to mohlo pokračovat v naprosto neproduktivním cyklu. Anotace nejsou nikdy formálně brány jako finální a ověřené jednou osobou nebo např. komisí jazykových „expertů“ (rodilý uživatelé jazyka, učitelé, lingvisté), vždy do nich lze zasáhnout a něco upravit. (Johnston, 2016, s. 6−7)

2.2.5 Strojově čitelný korpus a anotační značky

Je nutné, aby byla nahrávka projevu jazyka, ať už mluveného nebo znakového – tedy toho, který se převážně odehrává tváří v tvář – strojově čitelná a správně časově přiřazená k anotacím tak, aby byl korpus náležitě funkční. Digitální multimediální anotační software umožňuje v nahrávce přesně takové segmentace a značení časových jednotek znakovaných projevů.

V současných digitálních multimediálních souborech jsou přepisy a anotace přístrojově čteny, ale také jsou propojeny se zdrojovým projevem, který je buď auditivní či vizuální. Vzhledem k tomu, že znakové jazyky nemají psanou podobu, standardizovaný notační systém a ani mezinárodní fonetickou abecedu jako mluvené jazyky, je tedy produktivnější vytvářet základní stupeň anotací, které identifikují znakové jednotky v projevu pomocí glos. (Johnston, 2009, s. 90−91)

2.2.5.1 Glosa, ID glosa a lemma

Jak bylo zmíněno výše, je důležitou podmínkou, aby každý projev v korpusu byl strojově čitelný, aby se daly jednoznačně a důsledně identifikovat jednotlivé znaky. Identifikací jsou míněny na glosách založené anotace použité v korpusu jako ID glosy, tj. identifikační glosy (Johnston, 2016, s. 6)

Glosa je tedy typ anotace. Je to stručný jedno až dvouslovný „překlad“ v jednom jazyce pro slovo nebo morfém v dalším jazyce. Překlad musí být relativně přibližný a zjednodušující. V korpusu Auslanu je glosovacím jazykem angličtina a je konvencí psát glosy velkými písmeny.

Je nutné, aby v korpusu byla anotační úroveň, kde by znaky byly identifikovány jednoznačně a důsledně, protože nelze produktivně využít glosy, které se mohou lišit kontext od kontextu. V tomto korpusu je tento typ identifikačních glos označován jako ID glosa. ID glosa je (anglické) slovo, které je běžně používané k označení znaku v rámci korpusu, bez ohledu na význam znaku v určitém kontextu nebo zda byl význam nějakým způsobem pozměněn.

Proces přiřazování ID glos k lexikální formě znaku v korpusu je v podstatě podobný procesu lemmatizace – stejně jako lemmatizace redukuje skloňované a časované formy slov do jejich základních forem, tj. lemmat, tak ID glosování v podstatě pomíjí odlišné varianty nebo systematické modifikace ve formě znaku, pokud tyto formy nejsou ve prospěch základní citátové formy znaku. Nicméně je možné, že se vyskytnou různé glosy pro stejný znak a mohou být tak použity v různých kontextech k odrazu významu tohoto znaku v textu, což je odlišuje od ID glosy, která označuje znak bez ohledu na jeho význam. (Johnston, s. 91, 2009) Např. pokud osoba ukazuje znak DŮM (znak ikonicky odkazuje ke tvaru střechy a stěn), ale ve skutečnosti má znak význam „domov“ nebo znak představuje zvětšenou formu znaku DŮM, avšak osoba má na mysli znak „zámek“, aniž by byla pozměněna forma k rozeznání a rozlišení lexému jazyka, ID glosa DŮM je užita v obou případech k identifikování znaku v anotační glose. Konzistentně aplikovat značky tohoto typu znamená, že je možné prohledávat více anotačních souborů a najít všechny výskyty konkrétního znaku za účelem zjištění, ve kterých prostředích a způsobech je znak používán. Toto lze praktikovat pouze v případě, že všechny relevantní znaky mají stejnou ID glosu v korpusu. S respektem k rozlišování mezi glosováním a překladem, význam znaku je přiřazen k textu skrz glosování pouze nepřímo kvůli tomu, že ID glosa, která je primárně určena k identifikování znaku, ve skutečnosti používá anglické slovo, které nese vztah k významu znaku. Jinými slovy, ID glosa není vybrána náhodně či nahodile, protože volba anglického slova je vysoce motivována.

Nicméně, ID glosa není stále zamýšlena jako překlad. Překlady jsou vytvářeny na jim vlastních věnovaných úrovních v ELAN anotačních souborech. Takže když znakující produkuje znak ÚSPĚCH, ale myslí tím „dosáhnout něčeho“, tak je to stále anotováno s ID glosou ÚSPĚCH, a když člověk znakuje DŮLEŽITÝ, ale myslí tím „hlavní“ nebo „důležitost“ je to stále značeno jako znak DŮLEŽITÝ.

Selhání či nepoužití ID glos a standardizovaného glosovacího postupu v multimediální korpusové anotaci může způsobit dva problémy. Zaprvé, konzistence a souměřitelnost dat, která jsou anotována, tj. glosována různými výzkumníky nebo dokonce jedním výzkumníkem v různých fázích nemůže být zajištěna žádným jiným způsobem. Zadruhé by se soubor dat stal fakticky neomezeným, pokud by tady nebyl žádný nátlak na „významově založené glosování", protože každý artikulovaný znak, který může mít odlišnou formu, by mohl potenciálně mít svou vlastní odlišnou glosu odrážející jeho význam v daném kontextu. Jednoznačná identifikace formy znaku, což je jedna z hlavních motivací při tvorbě lingvistického korpusu v moderním smyslu, např. pro účely vyhledávání a kvantifikace typů znaků a jejich tokenů, by tedy nemohlo být dosaženo.

Bez konzistentního používání ID glos by nebylo možné používat korpus produktivně a mnoho času stráveného anotacemi by bylo fakticky plýtváním, protože korpus přestane být nebo nikdy nebude strojově čitelný v žádném významném smyslu. Což by tedy nebyl korpus, který by lingvisté dnešní doby chtěli a usilovali o jeho realizaci. Spíše by to byla pouze sbírka referenčních textů. (Johnston, 2010, s. 24–25)

2.3 ELAN

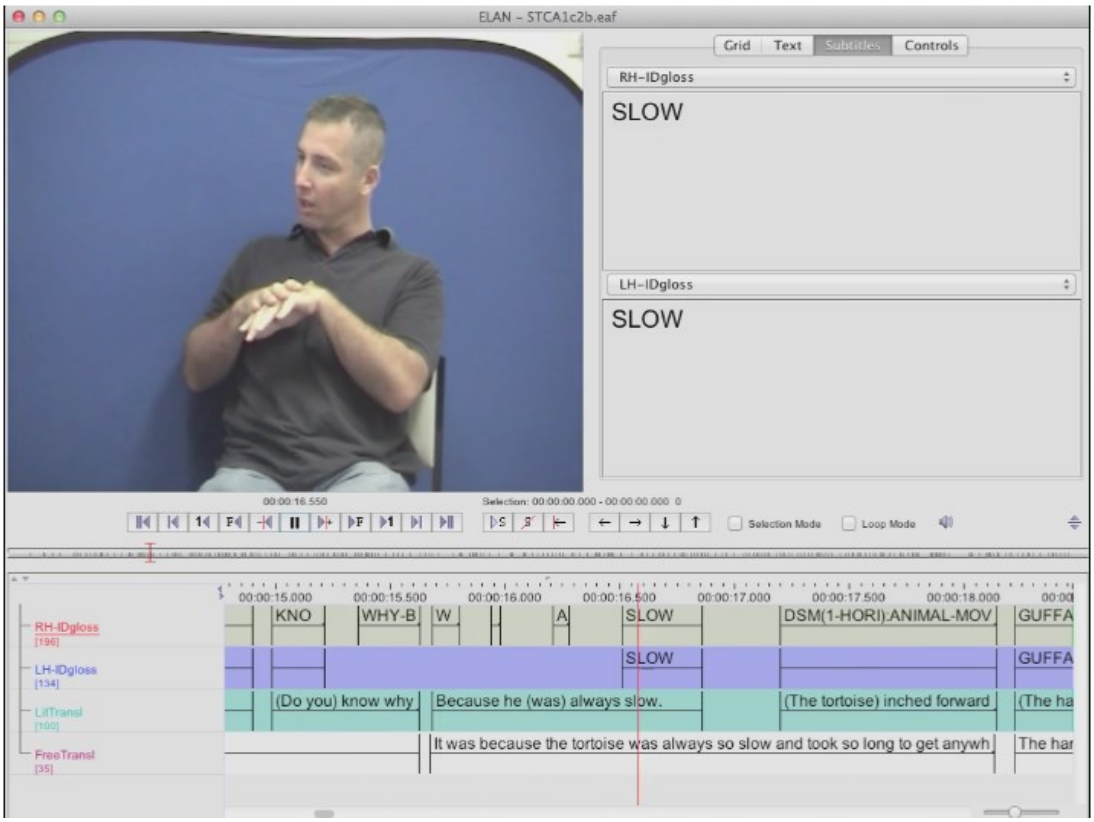
Digitální videonahrávky v korpusu jsou anotovány skrze software nazvaný ELAN (EUDICO[27] Lingvistický ANotátor), který byl vytvořen v Max Plank Institutu pro psycholingvistiku. Původně byl vyvinut pro výzkum jazyků a gest jako nástroj pro kompletní editaci tagů.

Tento anotační software umožňuje přesné časové sladění anotací s příslušnými videonahrávkami na vícero uživatelsky upřesněných úrovních. Jeho výhodou je, že podporuje zobrazení videa s jeho anotacemi – propojení anotací s médiem, na kterém je nahrán, neomezené množství anotačních úrovní, které lze vytvořit, tvorbu různých značek, export anotací jako tabulkově vymezených textových souborů a také má doplňující možnost importovat anotované textové soubory. (Johnston, 2010, s. 21–23)

Taktéž umožňuje propojení anotací s dalšími anotacemi, protože je v něm možné vyhledávání skrze vícero anotačních souborů. Podporuje různé vlastnosti dat, jako je například přepis a glosování v různých zápisových systémech. Rovněž dovoluje práci s databázovými programy a propojení s dalšími široce používanými lingvistickými softwary. (Johnston, 2006)

[27] EUDICO - EUropean DIstributed COrpus

Obr. 1: Anotovaná videonahrávka v ELANu pro znak SLOW



2.3.1 Úrovně ELANu

V současné době existuje poměrně velké množství úrovní pro anotaci dat, které jsou využívány v rámci korpusu softwarem ELAN. Minimální počet a typ úrovní, které by bylo nutné ustanovit, aby byla korpusová data kompletní pro jazykový výzkum, je ještě nutné upřesnit. To je částečně způsobeno nutností provedení určitých pokusů a omylů, aby se zjistilo, co bude nejužitečnějším druhem anotací. Avšak je jisté, že kdykoliv mohou být do softwaru přidány další úrovně pro anotaci. Proto je vhodné mít šablonu, aby bylo možné vyhovět požadavkům mnoha výzkumníků, takže daný stejný anotační soubor může být jednoduše a opakovaně užit pro různé účely. (Johnston, 2016, s. 9)

Každý anotační soubor ELAN, jehož přípona v rámci souborů vypadá následovně .eaf,[28] je zamýšlen tak, aby mohl být rozšiřován a obohacován různými výzkumníky skrze opakované anotace. Anotace obvykle začínají informací na úrovni, kde se identifikuje a pojmenuje znak, což je již výše zmíněná úroveň ID glosy. (Johnston, 2010, s. 21–23)

[28] eaf - ELAN Annotation File

V ELANu se užívají tři základní vrstvy pro anotaci. Úroveň pro pravou ruku, která se pro lepší přehlednost značí červenou barvou a úroveň pro levou ruku, která má žlutou barvu a na těchto úrovních se uvádí identifikační glosy. Třetí – zelená − úroveň je určena pro volný překlad daného znakového projevu. Tyto tři základní úrovně jsou v korpusu vždy uvedeny, ale je samozřejmě možné i žádoucí doplnit je o další vrstvy a to např. o úroveň nemanuálních komponentů, úroveň věnovanou využití úst, tedy mluvních a orálních komponentů, rovinu větných celků či rovinu pro střídání rolí ve znakovém jazyce. Všechny tyto vrstvy jsou samozřejmě také odlišeny různými barvami.

Při anotaci se postupuje tak, že se nejdříve uvádí základní informace a až posléze více specifické. Začíná se tedy od pravé a levé ruky, pro které se vytvoří primární anotace, posléze danou videonahrávku může převzít další anotátor, který anotuje další, např. mluvní či orální komponenty, které se v projevu objevují. Za specifický údaj se v ELANu označuje např. chování při znakování, tedy pohyb těla, hlavy, očí a obočí apod. Skrze anotační software ELAN je tedy možné přidávat sesbíranému materiálu informační hodnotu. (Johnston, 2016)

2.3.2 Současné rozšíření ELANu

Neustále se pracuje na zkvalitňování tohoto softwaru, a proto se zlepšuje rozhraní, tedy možnost prohlížení velkého množství úrovní a zvýšení počtu klávesových zkratek.
Dále usnadňuje tvorbu nových anotací a manipulaci s již existujícími anotacemi, např. jednoduché kopírování vybraných anotací na další úrovně. Zlepšily se i možnosti vyhledávání, které nyní umožňují komplexní vyhledávání skrze více anotačních souborů. V neposlední řadě se zájem zaměřil na spolupráci mezi výzkumníky a nyní je možný přenos dat z vybraných úrovní do dalších úrovní, které vytvořili jiní výzkumníci. A jako další nesporné plus softwaru je možné přidat a zviditelnit anotátorův příspěvek na každé anotační úrovni. (Johnston, 2006)

2.4 Cílená anotace – 3 fáze zpracování dat[29]

Přetvoření archivovaných médií do lingvisticky přínosných dat v korpusu se děje ve třech fázích – primární, sekundární a terciární.

[29] Informace v této části jsou čerpány z dokumentu Auslan Corpus Annotation Guidelines, s. 13–85, z r. 2016 od T. Johnstona.

2.4.1 Primární zpracování dat

Primární zpracování dat probíhá ve dvou fázích nebo by se také dalo říct na dvou úrovních a zahrnuje základní anotaci a detailní anotaci. Základní úroveň korpusové anotace se týká rozčlenění znakových projevů v Auslanu do smysluplných jednotek, které jdou volně přeložit do psané angličtiny, což je vlastně proces tokenizace těchto projevů. Poté se k těmto jednotkám přiřazují glosy. Detailní úroveň korpusové anotace zahrnuje anotování daných jednotek na dalších úrovních lingvistických a komunikačních aktivit, včetně nemanuálních aktivit.

2.4.1.1 Základní anotace

V korpusu se preferují minimálně tři úrovně při anotaci souborů v korpusu. Jedna úroveň pro volný překlad a dvě pro úrovně pro ID glosy, každá pro anotaci jedné ruky. Všechny nově anotované soubory jsou tvořeny tímto způsobem. Nicméně v počátcích anotování, tedy v letech 2004–2008, byly přidávány k získaným datům pouze ID glosy ve snaze vytvořit co nejvíce glosovaného textu v nejkratším možném čase. Tato základní data jsou následně postupně obohacována překlady, jakmile to čas a zdroje dovolí. Nicméně ze zkušeností při tvorbě korpusu vychází to, že je výhodnější dělat volný překlad během počáteční primární anotace a analýze dat a nikoli později.

2.4.1.1.1 Úroveň volného překladu

Psaný volný překlad projevu ve znakovém jazyce je prováděn jako úplně první krok při tvorbě základního anotačního souboru. Volný překlad je umístěn v anotačních souborech tak, že je časově sladěn s příslušnými částmi znakového projevu. Dané jednotky mohou být jednotkami promluvy, které jsou jednoduché nebo složité komplexní věty, které se projevují ve formě souvislého celku, založeného na významu nebo na daném pronesení jako jsou např. pauzy, kývání hlavou nebo vizuálně motorická intonace a rytmus projevu v jistém okamžiku. Vytvoření překladu také znamená vytvořit typ paralelního textu, který se posléze může hodit k výzkumu znakového jazyka, když finance, odborníci nebo čas budou někdy v budoucnu k dispozici.

2.4.1.1.2 Tokenizace videa pro základní glosování

Mluvení či znakování je souvislý proud slov nebo znaků a jako mezi jednotlivými slovy nejsou tichá místa, když mluvíme, kromě přirozených nebo úmyslných pauz, tak nejsou ani mezery mezi jednotlivými znaky při znakování. Znakující neprodukují a ani nemohou rázně znakovat jeden znak za druhým s plynulým vrácením se do neutrální pozice mezi jednotlivými znaky, dokonce ani nemohou artikulovat znaky bez přechodných pohybů mezi jednotlivými znaky.

Ignorování nebo upravování přechodných pohybů by mylně naznačovalo, že by tato doba byla bez znakovací aktivity. Nicméně jistá mezera, alespoň rámcová, by měla být ponechána mezi znaky pro anotace, abychom si mohli být jisti, že časový přesah nebo zarovnání je správně identifikováno. Důvodem tohoto je, že se objevují hraniční anotace, které mohou způsobovat falešné výsledky vyhledání.

Obecná pravidla pro začátek znaků:

a) když ruka nebo ruce změní směr pohybu, v případě že dokončí celý pohyb relevantní k artikulaci právě artikulovaného znaku a /nebo;

b) když ruka nebo ruce začnou měnit tvar ruky, za předpokladu, že ten tvar není součástí tohoto znaku.

Znak končí:

a) bezprostředně před tím, než ruka nebo ruce začnou měnit směr pohybu, mající ukončeny všechny pohyby relevantní k artikulaci daného znaku;

b) těsně před tím, než ruka nebo ruce začnou měnit tvar ruky, za předpokladu, že ten tvar není součástí znaku;

c) když se ruka nebo ruce začnou vracet do klidové pozice, tj. založené ruce, paže u boků nebo na nějakém podpůrném povrchu či objektu nebo jsou paže volně spuštěny podél těla.

Pauza, ve které je ruka nebo ruce drženy v určité stálé pozici a se stejným tvarem rukou, se považuje za pokračování artikulace znaku, pokud se to jeví jako smysluplné.
Anotační prostor pokračuje, dokud je držení uvolněné a ruce se nevrátí do klidové pozice nebo se ruce nezačnou pohybovat k provedení dalšího znaku.

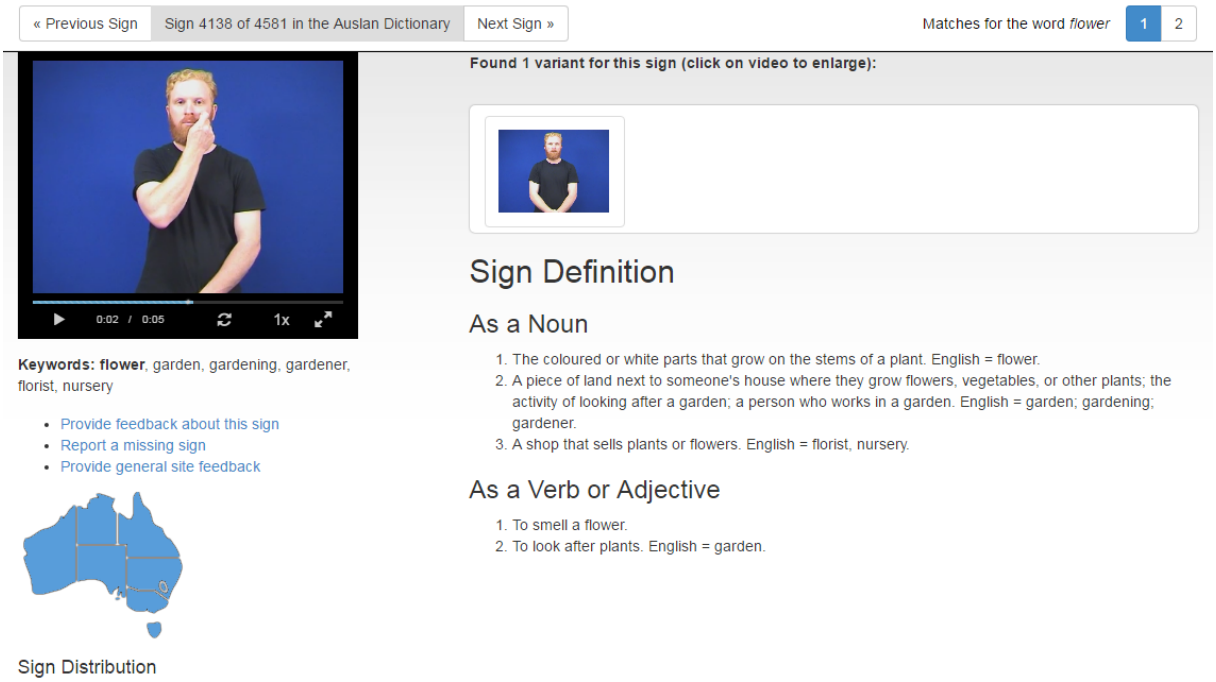
2.4.1.1.3 Úrovně glosování

Když je videonahrávka přeložena, tak je následně segmentována, rozdělena do jednotlivých tokenů a glosována. V ELANu jsou pro glosy určeny dvě úrovně, každá pro jednu ruku.

Jednoruční znak je anotován pouze na té ruce, která artikuluje znak. Ve dvouručním znaku nesou obě ruce stejnou ID glosu, ovšem na jim určených úrovních a jednotlivé úrovně na sobě nejsou závislé. Je nezbytné, aby jazykové jednotky stejného typu byly důsledně a jednoznačně identifikovány a každý token tak měl stejnou identifikační glosu, která bude jedinečná pro ten daný typ znaku. Aby bylo možné takhle účinně a efektivně provádět glosování, je třeba sestavit referenční lexikální databázi, která bude dokumentovat lexikální položky jazyka.

Anotátoři korpusu používají lexikální databázi Auslanu, která je volně dostupná jako Auslan Signbank.[30] Samozřejmě, že žádný slovník nemůže být kompletní, takže když se objeví nový znakový token v korpusu a zdá se, že nebyl ještě zaevidován jako konvenční lexikální jednotka jazyka, tak je přidán do Auslan Signbank. Tento proces je nezbytně kruhovitý.

Obr. 2: Auslan Signbank – výsledek hledaného výrazu FLOWER (květina)



[30] Auslan Signbank: http://www.auslan.org.au/dictionary/

Při stavbě korpusu se v ideálním případě neočekává, že by se začal znakový projev glosovat bez toho, aby se nejdříve provedl lexikografický a lexikologický výzkum jazyka a následně nebyla popsána slovní zásoba, která by byla zanesena v databázi nebo slovníku. Avšak ne všechny znaky, se kterými se setkáme ve znakovém projevu, jsou konvenčními znaky, které by měly být uvedeny ve slovníku. Znaky se liší ve stupni konvenčnosti a rozsahu od plně lexikalizovaných, přes částečně lexikalizované až k nelexikalizovaným znakům.

Stručně řečeno, plně lexikalizované znaky jsou vysoce konvencionalizované znaky jak ve své formě, tak i ve svém významu. Také jsou relativně konzistentní napříč různými kontexty. Plně lexikalizované znaky mohou být jednoduše zahrnuty do slovníku.

Částečně lexikalizované znaky jsou kombinací konvencionalizovaných a produktivních znaků. V literatuře, která se věnuje lingvistice znakových jazyků, zahrnují do kategorie produktivních znaků znaky známé jako klasifikátory a ukazovací znaky. Tyto znaky nemohou být uvedeny ve slovníku v nějaké zjednodušené formě a ani jim nemůžou být lehce přiřazeny ID glosy, protože vše záleží na významových souvislostech v projevu. Znaky, které jsou částečně lexikalizované, mají jednu nebo obě z těchto dvou důležitých charakteristik – mají malou nebo žádnou konvencionalizovanou nebo jazykově specifickou významovou hodnotu a navíc nesou formační komponenty, např. tvar ruky, místo, orientace atd. a za druhé mají význam, který je ovšem v jistém směru nekompletní, tzn. potřebují k sobě kontext promluvy, aby bylo možné doplnit význam znaku.

Nelexikalizované znaky jsou v podstatě gesta, která vypadají, že nemají žádný, jazykově specifický, konvencionalizovaný význam, který by se s nimi pojil. V tomto kontextu je gestem jakýkoliv záměrný komunikační tělesný úkon, manuální i nemanuální, s minimálním konvencionalizovaným významem či formou.

Konvence glosování jsou odlišné pro každý z těchto typů znaků, tak aby se snadněji identifikovaly a snadno se zařazovaly nebo vyřazovaly v rámci každého korpusového vyhledávání.

2.4.1.1.4 Plně lexikalizované znaky

Lexikalizované znaky jsou jednoduše identifikovatelné skrze užití ID glosy, která je napsána velkými písmeny nebo kapitálkami. ID glosa je načtena ze Signbank nebo je vytvořena nová glosa, pokud by se v této databázi nevyskytovala. K načtení ID glosy musí anotátor prohledat databázi užitím anglického klíčového slova, tj. překladu - ekvivalentu, který je spojen se znakem.

ID glosa znaku je obvykle jedno z klíčových slov, které je spojené se znakem. Pokud znak vyžaduje více než jedno přesné anglické slovo ke glosování, pak je mezi tato slova vložena pomlčka a tudíž mezi nimi není ponechána žádná mezera.

Obr. 3: Příklad užití ID glosy GROW-UP

Je snaha o to, aby každá ID glosa byla specifickým slovem nebo skupinou slov, která by určovala znak znakového jazyka. Nicméně se stává, že některá běžná, vysoce frekventovaná anglická slova musí být užita více než jednou, při glosování stejně běžných nebo vysoce frekventovaných znaků Auslanu. Ovšem neslyšící by mohli vnímat přiřazení jednoho slova k více znakům za znehodnocení znakového jazyka. V těchto případech je tedy znak připojen ke glose až po nějaké době a je k němu připojena poznámka, která upozorňuje na jeho možný další výskyt. Tento připojený komentář pomáhá anotátorům virtuálně si zapamatovat ID glosu.

Ve starších pokynech pro anotaci ID glos je uvedeno, že primárně glosovaným znakům byla znovu připojována již užívána slova a jednoduše jim byla přiřazena pořadová čísla v pořadí, jak ta slova přicházela, např. BEFORE1, BEFORE2, BEFORE3 (před). Nastavený systém se ukázal jako příliš nepřehledný. Anotátoři si těžko pamatovali tato čísla.

V současnosti je tento typ ID glos postupně v korpusu nahrazován příznakovými slovy, která nejlépe vystihují určitý znak. Kupříkladu v Auslanu jsou nejméně dva znaky, které jsou glosovány slovem FINISH (konec).

Jeden je tvořen tvarem ruky pro znak „good“ (dobrý) a druhý je produkován tvarem ruky jako znak „spread“ (rozšíření) nebo také „five“ (pět) a ty jsou glosovány následovně FINISH.GOOD (konec.dobrý) a FINISH.FIVE (konec.pět).

Vzhledem k existenci korpusu anotovaného v softwaru ELAN a možnostem použití snímků obrazovky nebo hypertextových odkazů v moderních digitálních médiích, se očekává, že uplatnění jednoduchých psaných glos pro znakový jazyk bude stále méně časté, pokud se jim v budoucnu nebude vyhýbat úplně. Užívání glos v podobně omezujících souvislostech vlastně znamená, že glosy zkreslují data znakového jazyka. Jejich používání je tedy v určitém smyslu kontraproduktivní.

2.4.1.1.4.1. Úroveň významu

V korpusu se zaznamenávají čtyři možné významy znaku na jedné úrovni. Za prvé, tato úroveň zaznamenává význam znaku, když z jakéhokoliv důvodu není k dispozici ID glosa. Anotátor vybere nejjednodušší anglické slovo ke glosování tohoto znaku, které se zdá být nejvhodnější k danému kontextu, připojí své iniciály k této dočasné glose a přidá několik slov vysvětlujících význam do „významové“ úrovně, např. ID glosa CONTRITION (lítost) by byla označena iniciály anotátora - TJ (Trevor Johnston) a znamená to něco jako „contrition“ (lítost), „remorse“ (výčitka svědomí), „regret“ (žal) nebo „sorrow“ (zármutek).

ID glosa by tedy na této úrovni vypadala následovně:

ID-gloss CONTRITION-TJ
ID glosa lítost-TJ

Meaning contrition/remorse/regret/sorrow
Význam lítost/výčitka svědomí/žal/zármutek

Pokud je nově identifikovaný znak následně rozpoznán jako nový nebo nezaznamenaný znak, tak je pro něj vytvořeno místo v lexikální databázi a znaku je přiřazena příslušná ID glosa. Existující glosy v korpusu pro jiný znak jsou následně korigovány či opraveny skrze univerzální vyhledávání.

Další významová úroveň uvádí význam pro znaky, ke kterým ještě musí být přiřazeno klíčové slovo z lexikální databáze.

Třetí úroveň zaznamenává kontextově specifický význam ID glos, které se vyskytují výjimečně nebo nejsou dosud zaznamenané, ale můžou se vzácně vyskytnout. V tomto smyslu je tedy anotátorův překlad zaznamenán jako ID glosa, kdyby ho bylo potřeba v budoucnu použít.

Za čtvrté, významová úroveň může být použita jako „držitel místa“ pro znaky s neznámou ID glosou, protože anotátor ještě nebyl schopný umístit ji do lexikální databáze, což může být později opraveno a doplněno.

2.4.1.1.4.2 Variantní formy znaků

Vzhledem k tomu, že žádné slovo nebo znak není vždy pronášen nebo produkován úplně stejně, tak by mělo být jasné, že drobné individuální odchylky při produkci znaků jsou při glosování ignorovány. Nicméně tato individuální odchylka musí být rozlišena od mnoha změn či modifikací ve slově či znaku, které vědomě a významově rozlišují význam, který může být gramatický nebo lexikální.

2.4.1.1.4.3 Jednoruční nebo dvouruční podoba znaku

V korpusu se neoznačuje pravá a levá ruka jako dominantní nebo nedominantní ruka. Jsou doslovně označeny jako pravá ruka (RH – right hand) a levá ruka (LH – left hand). Dominance ruky znakujícího je zaznamenána v metadatech pro konkrétního jednotlivce a v pojmenování daného anotačního souboru.

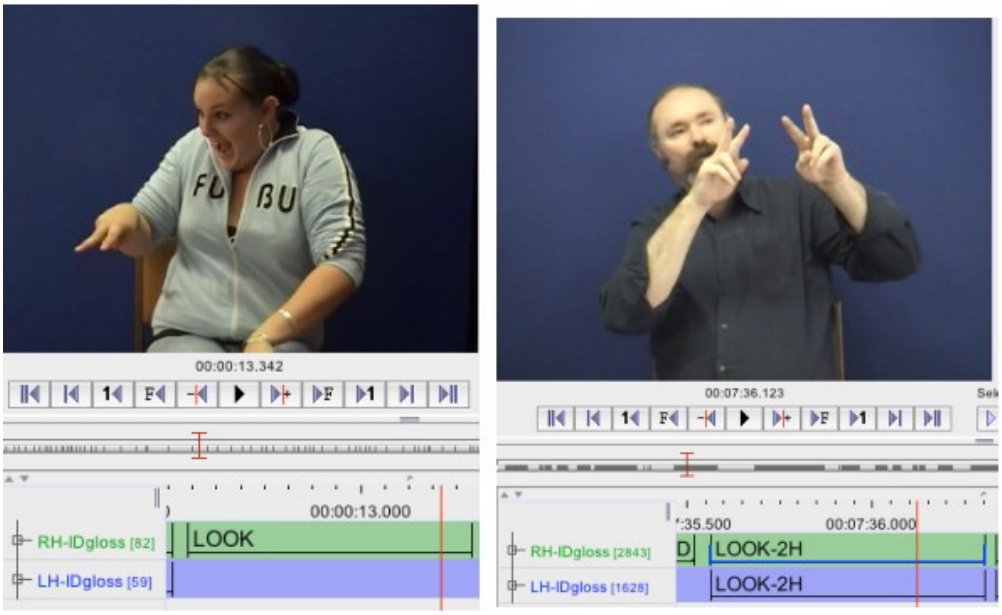
Pokud je znak dvouruční jako např. znak OWL (sova), ID glosa je napsána do dvou řádků, kdy každý řádek odpovídá jedné ruce. Při produkci jednoručního znaku je anotace vepsána do toho řádku, který náleží té ruce, která znak vykonává, i když by to byla nedominantní ruka znakujícího. Jestliže každá ruka provádí jiný znak, poté se různé anotační glosy vepisují do řádků pro každou ruku, přesně tak, jak jsou prováděny a jak k sobě patří.

Obr. 4: Anotovaná podoba dvouručního znaku OWL

Obr. 5: Anotace v případě, že každá ruka nese jiný znak

V současné době, pokud se znak zapisuje do slovníku a do databáze jako jednoruční znak, protože tak byl produkován, ale ve skutečnosti je to znak dvouruční, pak je anotace doplněna značkou -2H za danou glosou. A naopak, pokud je znak zapisován do slovníku a databáze jako dvouruční znak, ale ve skutečnosti by měl být produkován jen jednou rukou, tak je za glosu přidána značka -1H.

Obr. 6: Anotovaná podoba stejného znaku, který je produkován jednou nebo oběma rukama



Obohacování a rozšiřování korpusu umožňuje potvrdit nebo vyvrátit informace zachycené v databázi. Například mnoho znaků má jednoruční i dvouruční variantu provedení znaku a je tedy občas obtížné říct, která forma je běžněji či nepříznakově používaná nebo alespoň to, kterou variantu můžeme označit za citátovou formu. Takto se na základě korpusu ukázalo, že znak GLASSES (brýle) je vlastně více používaný jako jednoruční znak než dvouruční, což následně vedlo k opravě v databázi slovníku a posléze i k přizpůsobení anotací v korpusu.

Pokud znak zahrnuje změny tvaru u obou rukou a zároveň se mění i počet rukou během artikulace, tak se nejprve zapisuje tvar ruky a posléze následuje informace o počtu rukou. Tento typ doplňujících informací se obvykle připojuje k ukazovacím znakům či klasifikátorům.

2.4.1.1.4.4 Čísla a číselné inkorporace

Pokud znakující vyjadřuje nějaký číselný údaj, např. rok 1987, tak to je glosováno pomocí slov a ne číslic, tedy:

NINETEEN-EIGHTY-SEVEN ne 1987

ONE-NINE-EIGHT-SEVEN ne 1987

Když je číslo inkorporováno do znaku, tj. znaky pro hodiny, roky, týdny, dny, věk atd., tak je opět glosováno použitím slov a ne pomocí číslic. Obvykle znaky, které inkorporují čísla, mají standardní formu znaku, což znamená, že vyjadřují jednu jednotku něčeho, nějakého času či období. Například znak WEEK (týden) také znamená „one-week“ (jeden týden) a je jednoduše glosován jako „WEEK“. Je-li inkorporováno jiné číslo, tak se zapisuje do kulatých závorek za znak. K tomuto kroku se přistupuje jednoduše proto, že by mohlo docházet ke zmatení programu v procesu nahrávání, pokud by se tam uváděla pouze čísla místo textu.

WEEK(TWO) týden (dva)	*ne*	TWO-WEEKS dva týdny	*nebo*	2-WEEKS 2-týdny
WEEK-AGO(TWO) týden před (dva)	*ne*	TWO-WEEKS-AGO dva týdny před	*nebo*	2-WEEKS-AGO 2-týdny-před
AGE-YEARS(FOURTEEN) stáří let (čtrnáct)	*ne*	FOURTEEN-YEARS-OLD čtrnáct let starý	*nebo*	14-YEARS-OLD 14-let-starý
O'CLOCK(TWO) hodiny (dvě)	*ne*	TWO-O'CLOCK dvě hodiny	*nebo*	2-O'CLOCK 2-hodiny
YEAR(THREE) rok (tři)	*ne*	THREE-YEARS-AGO tři roky před	*nebo*	3-YEARS-AGO 3-roky-před
YESTERDAY(FOUR) včera (čtyři)	*ne*	FOUR-DAYS-AGO čtyři dny před	*nebo*	4-DAYS-AGO 4-dny-před

2.4.1.1.4.5 Negativní inkorporace

Mnoho sloves, v australském znakovém jazyce, která mají negativní význam, má tento příznak kvůli inkorporaci záporu do znaku. ID glosa těchto znaků je ve slovníku vedena jako obecná nepříznaková glosa, která je následována glosou pro negaci. Tento systém tak činí jednodušší nejen vyhledávání, ale taktéž i klasifikaci znaků podle jejich významu a pojmenování. Je to tedy snadnější, když jsou znaky glosovány následovně: KNOW-NOT (vědět ne) spíše než DON´T KNOW (nevědět), takže glosy KNOW (vědět) a KNOW-NOT budou řazeny hned za sebou, přesně podle abecedního pořádku.

KNOW-NOT	*ne*	DON'T-KNOW
vědět ne		nevědět
WANT-NOT	*ne*	DON'T-WANT
chtít ne		nechtít
WILL-NOT	*ne*	WON'T
bude ne		nebude

2.4.1.1.4.6 Jmenné znaky

Jmenným znakům předchází značka NS:[31] následována řádným vlastním jménem. Takže jmenný znak člověka, který se jmenuje Peter, bude ID glosou zapsán následujícím způsobem NS:PETER.

Dále se mohou přidat ještě doplňující informace, ale to není vyžadováno. Pokud by byl jmenný znak založený na určitém písmenu prstové abecedy nebo na zajímavé podobě znakové formy, poté by se tato informace dala přidat za glosu, např. NS:PETER(P-shake)
(P-třást).

Pokud je jmenný znak identický s lexikalizovaným znakem, poté je výchozí identifikovaný znak zapsán za glosu jmenného znaku do závorek, např. NS:MISSKENTWORTH(HAIR-BUN) (drdol).

[31] NS – name sign (jmenný znak)

2.4.1.1.4.7 Znaky znakované angličtiny a znaky převzaté ze zahraničí

Lexikální znaky, které se jeví jako převzaté ze znakovaného systému,[32] např. australské znakované angličtiny, nebo jiného zahraničního znakového jazyka a které nejsou obecně považovány za součást Auslanu, mají k sobě připojenou ID glosu, která zahrnuje tuto informaci. Př. GAVE.SE[33] − takto vypadá ID glosa znaku GAVE (dát), který pochází ze znakované angličtiny.

Pokud by se zdálo, že znak je nedávnou výpůjčkou z jiného znakového jazyka, poté by nebyl zařazen v lexikální databázi Auslanu, a proto by ani neměl přiřazenou ID glosu. Když by šlo o výpůjčku, tak je snaha přiřadit znaku co nejpřesnější glosu vztahující se k danému kontextu, která by byla následována jménem znakového jazyka, ze kterého pochází, např. vypůjčený znak COOL (chladný) z amerického znakového jazyka, by byl zapsán: COOL.ASL.

2.4.1.1.5 Částečně lexikalizované znaky

Přiřazování ID glos částečně lexikalizovaným znakům není jednoduché, protože se nemůžeme obrátit na lexikální databázi a odtud čerpat příslušnou ID glosu. V této situaci tedy neexistuje citátová forma znaku. Místo použití standardních identifikačních glos k identifikování tokenu jako tokenu lexikálního znaku, jsou tak tyto tokeny glosovány pomocí kombinace obecného a idiosynkratického[34] prvku kvůli jejich jedinečnosti. Vyhledávání frekvence nebo kolokací je řízeno systém, který je založený na vyhledávání té části glosy, která je identifikovatelná.

2.4.1.1.5.1 Ukazovací znaky

Všechny znaky vyjadřující ukazování začínají zkratkou PT.[35] Následně se specifikují podle typu ukazovacího znaku. V korpusu nalezneme deset typů těchto znaků, ale uvedu pouze pár vybraných pro ukázku. Typ znaku zaměřený na ukazování na osobu funguje jako zájmeno, např. on, oni, je specifikováno jako 1., 2. a 3. osoba, dále znak ukazující na místo funguje jako příslovce místa, např. tady, tam, anebo znak, který ukazuje na věc, osobu, místo a fungují jako zájmeno a zároveň určuje místo. Ani jedna informace není upřednostňována nebo separována od druhé, př. on tam, oni tam, ono tady atd.

[32] znakovaný systém - komunikační systém založený na znacích daného znakového jazyka, ale využívající gramatiku příslušného mluveného jazyka
[33] SE – signed english
[34] idiosynkratický – nesystematický
[35] PT - point

Pokud se tvar ruky ukazovacího znaku odlišuje od těch ukazovacích znaků, které jsou klasicky popsané v určitých kontextech, a pokud si anotátor přeje tuto informaci zahrnout do glosy, tak se přidává na konec glosace.
PT:PRO1SG(B) = ‘me’ produkován rukou, která má plochý tvar s prsty u sebe
PT:POSS1SG(5) = ‘my’ produkován rukou, která má plochý tvar s roztaženými prsty

2.4.1.1.5.2 Klasifikátory

Lze říci, že klasifikátory nemají jeden určitý význam, který by byl uveden ve slovníku, protože jejich smysl je příliš obecný a tudíž nedostatečně informativní nebo se naopak může ukázat, že je příliš významově omezený a na kontextu závislý a tudíž nedostatečně lexikalizovaný.

Anotace je poté tedy rozdělena do dvou částí, kdy je nejprve určeno, že daný znak je klasifikátor a následně se přiřazuje informace, o jaký typ klasifikátoru jde. Klasifikátory začínají předponou DS[36] a přidává se k nim písmeno, které určuje jejich podtyp L[37] – místo, M[38] – pohyb, H[39] – držení věci, S[40] – velikost a tvar. Poslední dvě kategorie klasifikátorů pro velikost a tvar a především ty pro držení věcí je občas velmi těžké rozeznat od gest.

Klasifikátory mají po značce DS určeno, jaký nesou tvar ruky, což se zapisuje do závorek, protože tvar ruky je jeden z nejcharakterističtějších rysů těchto znaků. Navíc určení tvaru ruky pomáhá v určení a analýze těchto znaků. Také se může připojit kód pro orientaci ruky, zvláště při popisu běžných a opakujících se klasifikátorů. Následně je doplněna informace o významu znaku. Obecně zápis klasifikátoru vypadá následovně: DSL/S/M/H/G(HANDSHAPE): BRIEF-DESCRIPTION-OF-MEANING-OF-SIGN.
(DSL/S/M/H/G(tvar ruky):stručný-popis-významu-znaku)

Popis klasifikátoru může být poněkud obecnější, např. *UPRIGHT-HUMAN-MOVES* (člověk ve vzpřímené pozici se pohybuje), ale někdy mohou být až moc specifické jako např. *THE-PERSON-ON-THE-RIGHT-WITH-LONG-HAIR-MOVES-SLOWLY-DIAGONALLYTO-THE-LEFT-OUT-THE-DOOR-IN-ANGER* (osoba s dlouhými vlasy nacházející se napravo se vztekle pomalým tempem pohybuje diagonálním směrem ven ze dveří).

Je tedy jasné, že by se měla udržovat vyváženost mezi obecností a přílišnou konkrétností:

[36] DS – depicting sign
[37] L − locative
[38] M − movement
[39] H − handling
[40] S – size and shape

DSM(1):HUMAN-MOVES DSM(1): člověk-se hýbe	spíše než	DSM(1):SHEPHERD-RUNS-LEFT DSM(1):pastýř-utíká-vlevo
DSM(B):ANIMAL-CRAWLS DSM(B):zvíře-se plazí	spíše než	DSM(B):TURTLE-MOVES-SLOWLY DSM(B):želva-se hýbe-pomalu

Jednotlivé typy klasifikátorů se nevylučují, takže se v jistých situacích může vyskytovat jedna forma znaku, která ovšem bude představovat různé typy klasifikátorů užívaných v rozdílných kontextech. Klasifikátory se samozřejmě lépe určují při pohledu na celkový kontext než na pouhý jeden znak. Pokud se ve znakovém projevu vyskytne klasifikátor, při kterém je nutné použít obě ruce, tak se v těchto případech přiřazují identické anotační glosy pro dominantní i nedominantní ruku.

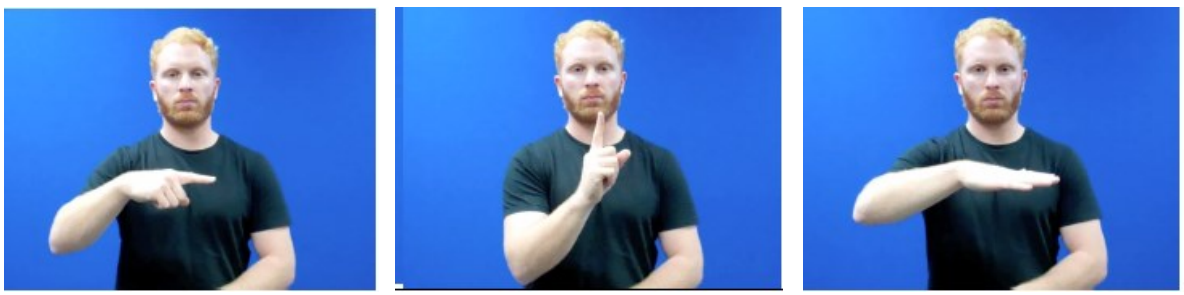
Nicméně, někdy se stane, že klasifikátory jsou komplexní simultánní konstrukcí a každá ruka tedy nese vlastní sémantickou informaci, takže anotátor musí popsat význam klasifikátoru na každé ruce zvlášť. Bez ohledu na specifikaci pro každou ruku, celková glosa by měla zachytit kompletní význam klasifikátoru.

Glosy pro klasifikátory jsou pravidelně kontrolovány, a když se zjistí, že forma a význam klasifikátoru byly glosovány odlišně a jsou v podstatě stejné, tak se glosy regulují tím způsobem, že se více zobecní, jsou tedy snadněji identifikovatelné, počitatelné, uspořadatelné apod.

V korpusu se rozlišuje 24 typů klasifikátorů. Rozlišují se tvary ruky a její orientace, ale také polohy a směřování prstů, tudíž např. vertikální a horizontální poloha či směřování dlaně nahoru nebo dolů. Kromě těchto pozic se také rozlišují různé typy pohybů. Poloha a počet prstů se značí následujícím způsobem:

- vertikálně vztyčený jeden prst (1-VERT)
- horizontálně vztyčené dva prsty (2-HORI)
- otevřená ruka, jejíž dlaň směřuje dolů (B-HORI)
- otevřená ruka, jejíž dlaň směřuje na stranu (B-LATERAL)

Obr. 7: Příklady tvarů a orientací ruky či prstů užívaných v klasifikátorech



2.4.1.1.5.3 Tvar nedominantní ruky

V korpusu se také rozlišují tvary nedominantní ruky, které se užívají jako referenční body, ke kterým se během znakování odkazuje. Daný tvar ruky může být držen po celou dobu artikulace každé položky, kterou provádí dominantní ruka, nebo když je potřeba obou rukou při dvouručním znaku, tak je tento tvar ruky opuštěn a posléze se zase objevuje v držení nedominantní ruky.

Na začátku anotační glosy se uvádí značka, která identifikuje příslušný tvar ruky a následuje krátký popis, čeho se daný tvar ruky týká. Například se při produkci drží určitý tvar ruky, na které je zdvižený konkrétní počet prstů, jako při počítání, což odkazuje k určitým subjektům nebo nápadům, které jdou po sobě. Tudíž natažený ukazováček odkazuje k první věci z nějakého výčtu, následně se přidávají další prsty, pro každou následující jednotku. Když počet objektů koresponduje s počtem zdvižených prstů, pak anotace, která se týká řekněme 4 objektů a je zobrazena 4 zdviženými prsty, vypadá takto: LBUOY(I):FOURTH.
Pokud je nutné počítat na obou rukou, tak anotace vypadá následovně:
RH ID-gloss PT:LBUOY
LH ID-gloss LBUOY(5)
PT značí, že se používá i druhá ruka.
Je důležité, že se zde vyskytuje nějaký ukazovací znak a ten je artikulován dominantní rukou a nedominantní ruka ho doplňuje.

2.4.1.1.6. Nelexikalizované znaky

2.4.1.1.6.1. Manuální gesta

Při komunikaci znakovým jazykem znakující nepoužívají vždy pouze jeden konvencionalizovaný znak za druhým, aniž by občas nepoužili gesto, takže jejich projev není vždy úzce lingvisticky definovatelný. Gesta mohou být kulturně sdílená či idiosynkratická, která se běžně vyskytují jak ve znakovaném, tak i v mluveném projevu.

Některá gesta, která jsou běžně užívaná v kultuře mluvených jazyků, jsou vysoce konvencionalizovaná a jsou sdílená i v komunitě neslyšících. V důsledku toho nejsou klasifikována jako gesta, ale jako znaky a jsou tedy zahrnuta do slovníku znakového jazyka a může jim být přiřazena ID glosa.

Nicméně existují i gesta, která jsou také kulturně sdílena, ale nestala se lexikální součástí znakové zásoby Auslan. Nejsou zahrnuta ve slovníku a ani jim nebyla přidělena ID glosa. Ta jsou klasifikována jako manuální gesta.

Nelexikalizovaná gesta, která mohou být kulturně sdílena nebo mohou být idiosynkratická, musí být identifikována při primární anotaci.

Není zde důvod, aby se anotátoři zdráhali kategorizovat jako gesta manuální a nemanuální chování, které se nesnadno zařazuje do kategorie konvencionalizovaných znaků nebo do skupiny klasifikátorů. Analýza a identifikace gest hraje důležitou roli v tom, jakou úlohu vlastně mají tato gesta v australském znakovém jazyce.

U gest lze rozeznat jak význam, tak formu daného gesta, podle toho, jak obvyklé se gesto zdá být používané a zapisuje se následovně: TYPE:MEANING (TYPE:význam). A protože jsou gesta z velké části nekonvenčními znaky, tak ve většině případů, když někdo identifikuje znak jako gesto, je nutné popsat jeho význam, což samozřejmě vždy závisí na daném kontextu. Anotace tak začíná značkou G[41] a vypadá takto G:DESCRIPTION-OF-MEANING (G:popis-významu), např. G:HOW-STUPID-OF-ME (G:já-hlupák) nikoli G:HIT-PALM-ON-FOREHEAD (G:udeřit-se-dlaní-do-čela).

Lze to uvést na příkladu gesta „WELL“ (dobře). Toto gesto má anotaci G(5-UP):WELL – 5-UP. Toto je velmi běžné gesto jak v kulturním, tak v lingvistickém prostředí. Užívají ho slyšící i neslyšící, v různých částech země. Má mnoho různých významů a funkcí dokonce i ve znakových jazycích. V Auslanu se tím většinou myslí „well“ (dobře). V jiném prostředí to znamená něco jako „don´t know“ (nevím) a v jiných zase „shocked“ (šokovaný).

Při zpracovávání anotací je k dispozici velké množství příkladů, a když některé z těchto gest má trochu odlišnou formu nebo funkci, tak dochází k rekategorizaci a následnému přepisování glos.

Toto je jedna z výhod používání korpusu jako prostředku popisu znakového jazyka, ale taky to klade požadavek na anotátory, aby neustále přiřazovali ID glosy všem typům znakových jednotek, jak plně lexikalizovaným, částečně lexikalizovaným, tak i nelexikalizovaným znakům a gestům.

2.4.1.1.6.2 Nemanuální znaky

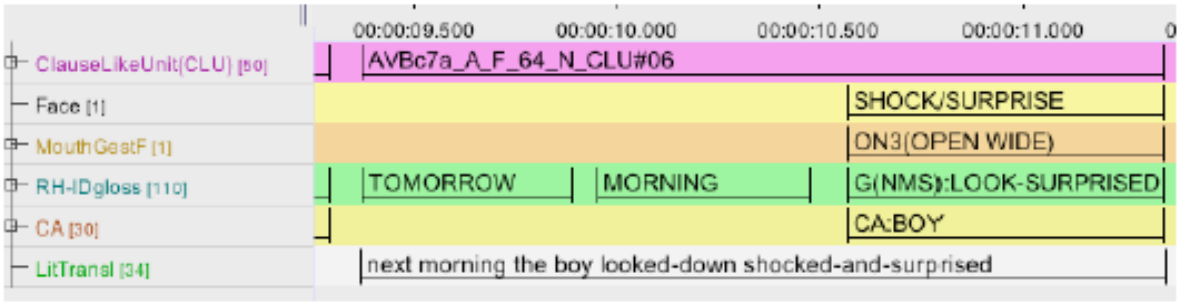
Některé znaky nemají jako hlavní nosič ruce, ale jsou soustředěny na tělo, hlavu nebo obličej a nezahrnují tedy žádnou manuální aktivitu. Od začátku se ID glosy primárně pojily se znaky, které se produkovaly manuálně, nicméně nemanuální znaky se nepropojovaly s úrovněmi, na kterých se tyto ID glosy vyskytovaly.

[41] G - gesture

Dříve se výzkumy prováděly přes anotace uvedené v ELANu, kde se realizovaly úrovně glos, a tak se mohlo stát, že se přehlédl nějaký počet nemanuálních gestických jednotek. Nemanuální znaky byly považovány za „výplň“ mezi znaky a to budilo dojem, že se v té době mezi znaky z lingvistického hlediska nedělo nic zajímavého.

Nemanuální znaky se objevují jako anotace na úrovni hlavy, obličeje, úst a těla, resp. tam, kde je třeba. Je nutné zmínit, že značka G může být užita v souvislosti s NMS[42] v kulatých závorkách k připomenutí běžnému pozorovateli, že existuje určitá neznaková nemanuální gestická aktivita, která může být reflektována na příslušných úrovních. Pokud samostatné nemanuální gesto zahrnuje pohyb úst, tak se používá označení M[43] nebo MG[44] místo značky G.

Obr. 8: Anotace gesta LOOK-SURPRISED (vypadat překvapeně)



2.4.1.1.6.3 Glosování slov hláskovaných prstovou abecedou

Kdykoliv je užita prstová abeceda k vyhláskování slova, tak je to anotováno pomocí značky FS[45] za čímž následuje dvojtečka a to dané spelované slovo, tedy FS:WORD (FS:slovo). Pokud by bylo při hláskování nějaké písmeno vypuštěno, ať už kvůli rychlosti nebo omylem znakujícího, avšak cílové slovo by bylo známé, tak je zachyceno jako to celkové slovo, nikoli jak bylo vyhláskováno. Nicméně pokud se tyto „chyby“ vyskytují ve významném měřítku, jako např. opakované ortografické chyby nebo zkratky několika písmen, tak ta reálně hláskovaná písmena jsou umístěna do závorek za dané hláskované slovo:

FS:WORD(WOR) *ne* FS:WOR
FS:slovo

FS:WORD(WRD) *ne* FS:WRD
FS:slovo

[42] NMS – non-manual sign
[43] M – mouthing
[44] MG – mouth gestures
[45] FS – fingerspelling

FS:SO(SI) *ne* FS:SI
FS:takže

Zdá se, že neustálá redukce při hláskování slov skrze prstovou abecedu mnoha znakujícími je náznakem zvnitřňování a lexikalizace ve znakovém jazyce. Glosy tak bude třeba v budoucnu přizpůsobovat tomuto trendu. Pokud je hláskováno více slov za sebou, tak je každé slovo glosováno zvlášť:

FS:MISS FS:KENTWORTH *ne* FS:MISSKENTWORTH
FS:slečna

Jestliže je produkováno pouze jedno písmeno v rámci hláskování, někdy i dublovaně, které může znamenat více věcí podle různých kontextů, tak se do anotace zapíše celé slovo. Avšak je nutné myslet na to, že některá dublovaná písmena jsou normálními lexikalizovanými znaky a mají tak svoji ID glosu v databázi, např. dublované D znamená znak DAUGHTER (dcera). Fakt, že jsou tyto znaky odvozeny z hláskování pomocí prstové abecedy, se odráží v lexikální databázi.

FS:M-MONTH (měsíc), FS:M-MINUTE (minuta), FS:M-MILE (míle), FS:Y-YEAR (rok), FS:Y-YARD (yard)

FS:GG-GOVERNMENT (vláda), FS:GG-GOVENOR-GENERAL (guvernér), FS:GG-GARAGE (garáž)

2.4.1.1.6.4 Znaky, kterým nelze rozumět nebo nejsou čitelné

Pokud se v projevu účastníka objeví nějaký znak, jehož provedení není jasné a je nemožné identifikovat, co za znak to má být, natož určit, jestli jde o plně, částečně či nelexikalizovaný znak, tak se vytvoří zvláštní anotační pole pro tento znak a přiřadí se mu glosa INDECIPHERABLE (nerozluštitelný). Což naznačuje, že danou formu a význam znaku nelze jistě určit.

2.4.1.1.7 Tokenizace videa

2.4.1.1.7.1 Stínování, anticipace, perseverace

Když nedominantní ruka napodobuje jedním nebo více rysy komponenty dominantní ruky, tj. částečně kopíruje tvar ruky nebo pohyb, které ve standardní situaci nese jednoruční znak při užití dominantní ruky a je to prováděno neúmyslně nebo alespoň bez záměrného komunikačního záměru nebo rozeznatelného přidaného významu, poté není tato aktivita nedominantní ruky brána v potaz.

Stejně tak, pokud se nedominantní ruka na malý moment stane součástí již probíhajícího znaku dominantní ruky, protože se „připravuje“ na produkci dalšího znaku, tak tato malá aktivita nebude anotována jako součást znaku do kterého zasahuje.

Pokud se zdá, že se na nedominantní ruce objevuje perseverace, tedy pokračování části právě artikulovaného znaku, který se pomalu vrací do neutrálního tvaru nebo pozice, tak se v této situaci neprodlužuje doba pro anotaci, aby se zamezilo zahrnutí těchto slábnoucích prvků do znaku, hlavně když další znak už jasně začal nebo je artikulován na druhé ruce a tato ruka artikuluje znak bez zjevného napojení na perseverující ruku. Anotují se tedy pouze informace pro dominantní ruku, protože pohyby nedominantní ruky nemají v tomto kontextu význam.

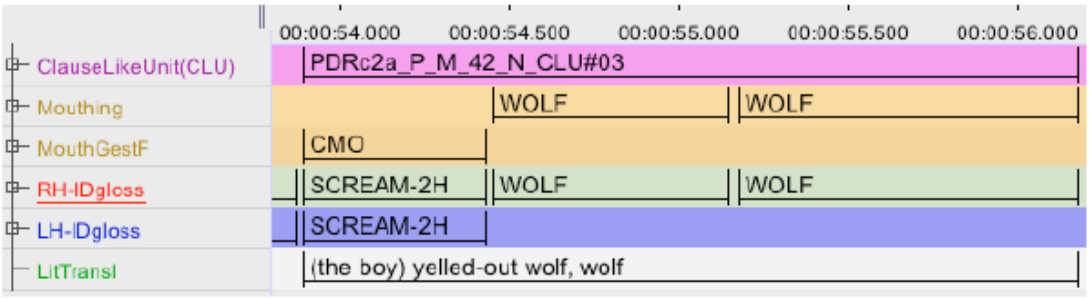
Stručně řečeno, anotace je vždy vytvořena pro obě ruce ve dvouručních znacích nebo pro každou ruku zvlášť, když se zdá, že každá ruka dělá něco záměrného či smysluplného, i když znak není dvouruční.

2.4.1.1.7.2 Opakování

Někdy je celý znak opakován a občas je pouze pohyb daného znaku modifikován právě tímto opakováním. Je často těžké rozlišit mezi těmito dvěma situacemi. Každá má jiný dopad na význam znaku. Pokud znak vypadá, že by se dal přeložit pomocí jednoho anglického slova, které by evokovalo gramatickou změnu, např. opakování znaku „WAIT“ (čekat) může být přeloženo jako „waiting“ (čekání) místo „wait“ nebo také jako fráze, tedy opakované „WAIT“ přeloženo jako „čekat po opravdu dlouhou dobu“, poté je užita pouze jedna anotace a glosa. A v tomto případě bude glosa WAIT. Modifikace čili opakování znaku je bráno jako přirozená gramatická změna. Gramatická informace je kódována na dalších, tomuto věnovaných, úrovních anotace.

Nicméně pokud znak vypadá opravdu tak, že se opakuje, tj. je produkován více než jednou a může být stejně tak přeložen opakujícím se anglickým slovem, poté bude každá jednotka anotována odděleně. Pokud si anotátor není jistý, je doporučeno, aby se utvořila poznámka na příslušné úrovni.

Obr. 9: Anotace opakovaného znaku „WOLF“ (vlk)



2.4.1.1.7.3 Složeniny a slovní spojení

Dva znaky, které jsou pravidelně znakovány dohromady, tak mohou být brány buď jako složenina, ale mohou být také slovním spojením.

Slovní spojení je běžné spojení dvou znaků nebo slov. Když se vyskytne jedno slovo, tak se očekává ve spojení s druhým v ustáleném pořadí, např. „black and white“ (černá a bílá) nebo „I think“ (já myslím) v angličtině nebo znak „KNOW PRO2SG“ (ty víš) v Auslanu. Slovní spojení jsou psána jako dvě oddělené anotace a nezáleží na tom, jak často se objevují společně nebo jak rychle jsou dva znaky za sebou znakovány.

Pokud se však vyskytuje nějaká fonologická redukce mezi dvěma členy, tak se to obvykle bere jako složenina. Složeninu lze obvykle zapsat jako anotaci pro jeden znak. Většina složenin se již zařazuje s rozličnými ID glosami do lexikální databáze, např. MOTHER^FATHER (otec matka) je standardně v Auslanu složenina pro PARENTS (rodiče) a WRONG^MIND (špatná mysl) je složeninou pro GUILTY (vinný). Samozřejmě se uvádí ID glosy PARENTS a GUILTY.

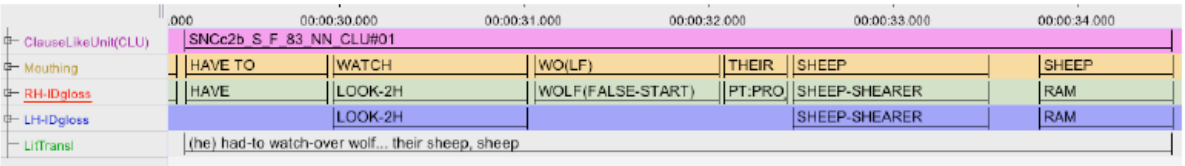
Pokud tyto znaky nemohou být ve slovníku uvedeny jako složenina, tak znak bude zapsán jako jeden znak se dvěma položkami, které jsou odděleny symbolem (^).

2.4.1.1.7.4 Falešné začátky a opravy

Během projevu v mluvených a znakových jazycích, hlavně v nepřipravené komunikaci tváří v tvář se objevuje mnoho případů falešných začátků. Mluvčí nebo znakující začne artikulovat slovo nebo znak, ale nedokončí to z různých důvodů. To je obvykle okamžitě následováno pár slovy či znaky k opravě toho, co bylo zamýšleno sdělit v prvním případě. Pokud by to byl jasný případ, tak se jako přípona za ID glosu připojí slova „false-start“ (chybný začátek) v závorkách.

Určit falešné začátky pomáhá v tom, proč některé prvky nejsou nebo by neměly být zahrnuty do struktury tagování. Také to později umožňuje extrahovat tyto typy chyb z korpusu pro různé analýzy podle jejich charakteristik, načasování a následných oprav.

Obr. 10: Anotace znaku, který je po chybné produkci bezprostředně opraven



2.4.1.2 Dodatečná detailní anotace

Znakové jazyky nejsou produkovány pouze rukama. Uživatelé znakového jazyka využívají také okolního prostoru kolem sebe prostřednictvím již dobře známých nemanuálních komponentů, jako je držení těla, pohyby hlavou, pohled očí, výrazy tváře, mluvní a orální komponenty. Jak již bylo zmíněno, tak pro všechny tyto aspekty existují určité, jim věnované, úrovně v korpusu. Tyto nemanuální komponenty je nutné anotovat, aby posléze bylo možné určit jejich roli v lexikálně−gramatické rovině každého znakového jazyka.

Nemanuální komponenty se mohou pojit jak s individuálními znaky, tak i s celými frázemi, větami nebo delšími významovými jednotkami. Z tohoto důvodu jsou všechny anotační úrovně v ELANu nezávislé a jednotlivé složky nejsou vzájemně vázány nějakou podmínkou.

2.4.1.2.1 Anotace nemanuálních komponentů nebo prozódie

Hlavní úrovně užívané při anotaci nemanuálních komponentů jsou roviny pro tělo, obličej, tvář, pohled, oči a obočí, mluvní komponenty a orální komponenty.

2.4.1.2.1.1 Úroveň těla

Zdá se, že pohyby těla při znakování mají více funkcí a korpusová anotace nám pomáhá popsat a kategorizovat tyto funkce. „Úroveň těla“ se užívá ke kódování pohybů, které jsou něčím charakteristické a zdají se být lingvisticky významné.

Změny jsou charakterizovány s ohledem na neutrální pozici, která se určuje v oblasti hrudníku, na vertikální vzpřímené ose směrem k adresátovi. Popisují se následující pohyby: naklánění trupu v určitém směru a/nebo otáčení trupu, často ve velmi malých odchylkách.

Stručně řečeno, tyto pohyby těla jsou obvykle použity k určení té části projevu, jednoho znaku či sekvence znaků, která se věnuje nějakému objektu, účastníkovi nebo místu promluvy a naznačuje směr pohybu nebo orientaci trupu, např. vpravo, vlevo, vzadu nebo vepředu ve znakovacím prostoru. Objekty mohou být reálné nebo vymyšlené, konkrétní nebo abstraktní, živé nebo neživé.

Posun těla sám o sobě může ustanovit objekt v nějakém místě, ale obvykle se využívá asociace s již vytvořeným odkazem v projevu, což může být buď umístěním objektu do prostoru s ukázáním do tohoto místa, kde je objekt v centru dění, tj. zrovna se o tom znakovalo, dále artikulací znaku, který není prováděn na těle a vztahuje se k danému místu a již předchozím posunem těla. V následujícím příkladu byl doktor umístěn vlevo od znakujícího a kněz vpravo a pohyby těla ukazují následující:

ID-gloss ID-glosa	UNDERSTAND rozumět	SCIENCE věda	UNDERSTAND rozumět	SCIENCE věda
Head hlava	nod přikývnout		shake kroutit	
Body tělo	left:doctor vlevo:doktor		right:priest vpravo:kněz	
FreeTransl	*The doctor understood science, whereas the priest didn't understand science*.			

volný překlad Doktor vědě rozumí, zatímco kněz vědě nerozumí.

2.4.1.2.1.2 Úroveň obličeje

Tato úroveň popisuje výrazy tváře na obecné rovině. Tato úroveň vymezuje časové rozpětí, jak dlouho daný výraz trvá. Výrazy tváře mohou být detailněji popsány na dalších úrovních nemanuálních komponentů, např. hlava, pohled, oči, obočí a ústa.

2.4.1.2.1.3 Úroveň hlavy

Tato úroveň je užívána pro kódování pohybů hlavy, které jsou něčím charakteristické a/nebo lingvisticky významné. Stejně jako ostatní úrovně nemanuálních komponentů, tak i úroveň hlavy je kódována s respektem k neutrální pozici – hlava je zpříma a čelem k adresátovi. Anotační úroveň určuje čas, kdy probíhá toto nemanuální chování.

2.4.1.2.1.4 Úroveň pohledu

Tato úroveň je užívána pro kódování pohybu pohledu očí, který je něčím charakteristický a/nebo lingvisticky významný. Stejně jako ostatní úrovně nemanuálních komponentů, tak i úroveň hlavy je kódována s respektem k neutrální pozici – znakující se dívá na adresáta. Anotační úroveň určuje čas, kdy probíhá toto nemanuální chování.
Kódy, které se používají: A[46] − adresát, T[47]– cíl, O[48]– další, nebo Z[49] – nemůže být kódováno.

2.4.1.2.1.5 Úroveň očí a obočí

Tato úroveň je užívána pro kódování pohybů očí a obočí, které jsou něčím charakteristické a/nebo lingvisticky významné. Stejně jako ostatní úrovně nemanuálních komponentů, tak i tato úroveň je kódována s respektem k neutrální pozici – uvolněné a otevřené oči a uvolněné obočí. Jsou kódovány společně na jedné úrovni k jejich zřejmému propojení, např. zvednuté obočí s otevřenýma očima, snížené obočí se zúženýma očima. Anotační úroveň určuje čas, kdy probíhá toto nemanuální chování.

2.4.1.2.1.6 Mluvní komponenty

Mluvní komponenty jsou pohyby rtů, které vyslovují slovo nebo část slova a jsou anotovány na této úrovni. I přesto, že mají mluvní komponenty tuto svoji vlastní úroveň, tak všechny mluvní komponenty jsou nejprve anotovány výběrem ID glosy, jakmile je přidána anotace. Kde si jsou anotátoři jisti, že bylo slovo vysloveno a jsou zde jisté artikulační pohyby, ale nejsou schopni určit, o jaké slovo se jedná, tak je toto slovo anotováno jako nečitelné.

2.4.1.2.1.7 Orální komponenty

Orální komponenty jsou všechny pohyby úst, které nejsou mluvními komponenty. V korpusu se rozlišují dva základní typy komponentů, které jsou anotovány, ale ty se ještě nadále dělí. V projevu identifikují a přidělují kódy např. pro nafouknuté tváře, nebo jen jednu tvář, když je jeden ret vysunutý, oba rty jsou vysunuté, dále také, když je jazyk před rtem nebo se jazyk dotýká tváře uvnitř úst atd. Nejzákladnější je ale rozdělení, zda jsou ústa otevřená či zavřená a následně se dělí dle výše uvedených kritérií.

[46] A − addressee
[47] T − target
[48] O − other
[49] Z – cannot be coded

2.4.1.2.2 Anotace jednotek delších než jsou jednotlivé znaky

Volný překlad a segmentace projevu do jednotlivých znakových tokenů je tím nejzákladnějším, co se požaduje, aby se získala surová data a dalo se s nimi dále pracovat. Samozřejmě lingvistická analýza vyžaduje práci s celým projevem, ne pouze s jednotlivými znaky. Promluvy obvykle obsahují více než jeden znak a jsou produkovány tak, že jsou spolu kumulovány kvůli artikulaci, významu a lingvistické struktuře.

Užívají se různě dlouhé jednotky promluvy ke sdělení určité informace. Někdo „říká" něco někomu skrze kódování lexikálně−gramatickým konstrukčním schématem v jednom jazyce. Mnoho jednotek má stejný status jako významové chunky[50] a tyto jednotky jsou mnohem lépe identifikovatelné jako věty.

2.4.1.2.3 Anotace napodobované činnosti a napodobovaného rozhovoru

Pohyby těla také nazývané jako střídání rolí, které jsou anotované na úrovni těla, jednoduše ustanovuje asociaci mezi tím, co bylo znakováno a umístěním vztahujícím se k tělu, kde se pohybovalo nebo hýbalo.

2.4.1.2.3.1 Napodobované činnosti

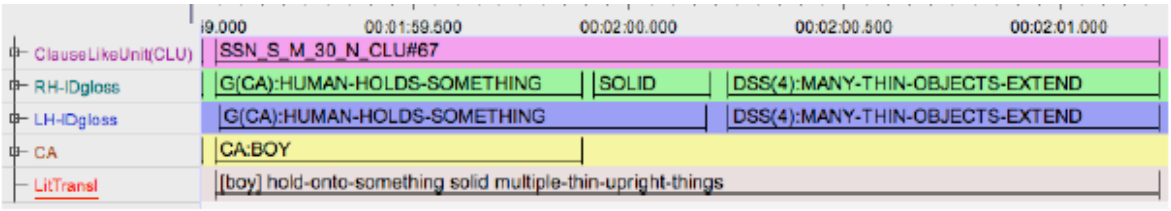
V literatuře odkazují napodobované činnosti k užití expresivních gest, které napodobují projev nějakého jiného znakujícího. Termín napodobované činnosti byl do lingvistiky znakových jazyků vnesen Winstonem v roce 1991, protože odkazují k činnostem, které nejsou pouhou nápodobou něčí činnosti, ale jsou skutečným vybraným hraním, tj. jsou rekonstrukcí něčího znakového projevu.

Během napodobované činnosti znakující kopíruje nebo se dá také říct, že cituje činnosti či výrazy. To se projevuje v obličejových výrazech, pohybech těla a hlavy a/nebo v pohybech rukou a paží, které nejsou součástí lexikálních znaků v Auslan slovníku nebo se neobjevují v klasifikátorových konstrukcích. Mnoho gest je vlastně příkladem napodobované činnosti, protože během nějaké doby znakující předvádí činnosti charakteristické pro nějakou roli. Např. při produkci manuálního znaku, jako je „SEARCH" (hledat), znakující může pohybovat hlavou ze strany na stranu k ukázání činnosti, že člověk něco hledá, nebo místo produkce konvencionalizovaného znaku „WINK" (mrknout), znakující může zvolit opravdové mrknutí k ukázání toho, že někdo mrknul.

[50] chunk – různě dlouhý úsek informace (písmena, slova, krátké věty)

Když se určí oblast, kde se napodobuje činnost, tak se na anotační úroveň napíše značka CA,[51] následuje jméno osoby nebo bytosti, která je reálná nebo vymyšlená a jejichž chování bylo napodobováno.

Obr. 11: Anotace znaku, který napodobuje činnost, kdy člověk něco drží



2.4.1.2.3.2 Napodobovaný rozhovor

Lze taktéž reprodukovat něčí projev, ať již mluvený nebo znakový. Fakticky někdo kopíruje či cituje někoho, kdo něco pronášel. Jedná se o typ přesné citace a velmi se podobá opakování slov, která někdo pronesl, což také zahrnuje napodobení hlasu, intonace, hlasitosti a přízvuku originálu. Např.: *He said "Soooo… WHO do you think YOU are?!"* spíše než *He asked me who did I think I was.* (On řekl „Taaaakže… KDO si myslíš, že JSI?! spíše než Zeptal se mě kdo si myslím, že jsem.) což je nepřímá řeč. To, co mluvčí a znakující dělají během napodobování rozhovoru je to, že reprodukují projev, ale nikdy ne přesně. Je to tedy napodobování.

2.4.2 Sekundární zpracování dat

Sekundární zpracování dat zahrnuje přidávání dalších informací čili tagů k anotacím, které již byly vytvořeny v primárním zpracování. Zahrnuje subkategorie různých velikostí, od jednotlivých znaků, frází až po věty a určení jejich jednotlivých složek. Tento proces pojímá přiřazování fonologické, morfologické, sémantické, syntaktické a pragmatické informace daným lingvistickým formám a určení těchto údajů záleží na účelu dané analýzy.

2.4.2.1 Tagování znakových tokenů

Tagování znakových tokenů pokrývá lingvisticky relevantní informace, jako je specifikace fonetické a fonologické formy, stupeň shodnosti tokenu a citátové formy znaku, specifikuje význam znakového tokenu vzhledem k danému kontextu, přiřazování gramatické třídy apod.

[51] CA – constructed action

S respektem k formě znaku je ID glosa rozšířena o fonetickou a fonologickou anotaci na úrovni přepisu. Kódování fonetických a fonologických prvků může být uvedeno buď jako sada na jedné úrovni anotace nebo na více úrovních, kde může být každý významný aspekt, jako je např. tvar ruky, orientace, pohyb atd. popsán nezávisle na sobě.

2.4.2.2 Citátová modifikace nebo úroveň variace

ID glosy identifikují typy znaků a tím upravují lexikální znaky jako by byly v citátové formě. Samozřejmě, tohle je poměrně vzácné, protože znaky nejsou produkovány izolovaně, ani nejsou odloučeny od individuálních zvláštností v projevu jednotlivců, kteří znakují. Avšak zároveň se znaky mohou lišit od svých citátových forem, protože byly záměrně a systematicky upraveny kvůli zprostředkování různých významů. Úroveň citátové modifikace je užívána pro tagování znaku v jeho neupraveném (flektivním) či upraveném (citátovém) tvaru. Do současné doby byla v korpusu tato úroveň užita pouze pro kódování modifikace znaku vzhledem k prostoru.

2.4.2.3 Sémantické tagování

2.4.2.3.1 Úroveň významová

Tato úroveň stručně uvádí význam znaku, když není z nějakého důvodu uvedena ID glosa, např. když ji anotátor nemůže umístit do slovníku, nebo se znak ukáže jako nový, ještě nezachycený znak. Dále také přidává význam znaku, ke kterému ještě musí být uvedeno klíčové slovo, tj. ještě není ověřen a nyní se zdá, že bude přidán do lexikální databáze.
Následně zachycuje kontextově−specifický význam ID glosy, který se vyskytuje vzácně a také vyjadřuje význam znaku, jehož ID glosa neukazuje jeho kontextovou gramatickou funkci. Jinými slovy, ID glosa nemusí nutně určovat gramatickou třídu znaku, např. podstatné jméno, sloveso, přídavné jméno atd.

Gramatickou třídu u většiny znaků není jednoduché určit z jejich morfologie nebo syntaxe, tj. znaky se nijak vnitřně nemění, ani nemají rozlišovací morfémy připojené ke znakům, které by jednoznačně určily gramatickou třídu. Znak australského znakového jazyka může být užit mnoha způsoby, např. jako podstatné jméno i sloveso, takže je nutné, aby se anotátor řídil podle kontextového spojení tokenu s gramatickou třídou, spíše než aby usuzoval podle ID glos.

Je třeba mít na paměti, že všechna užití na významové úrovni propojují token znaku s významem. Na úrovni volného a doslovného překladu jsou také vyjádřeny významy znaků, ale to se děje na větné úrovni.

2.4.2.3.2 Úroveň gramatické třídy

V korpusu lze identifikovat 29 gramatických tříd pro tagy znaků. Jsou v nich třídy pro podstatná jména, která se ještě dále dělí na prostá a popisná, dále zájmena, slovesa, která se také samozřejmě ještě dále dělí na směrová, prostá, popisná atd., kategorie pro čísla, negace, pro příslovce, znaky pro pozdrav atd. 14 z nich má orientační charakter.

Pokud je třeba, lze v ELANu přidávat nové kategorické značky. Přiřazení gramatických kategorií k jednotlivým znakům nemůže být učiněno bez kontextu a tedy významu znaku. Věty pomáhají určit roli jakéhokoliv znaku v projevu a tedy jakou gramatickou třídu aplikovat. Nicméně mnoho znaků nelze jednoznačně určit, navzdory jejich určení jako jistého prvku ve větě. V těchto případech se nepoužívá značka „indeterminate" (neurčitý), ale použije se nadřazená kategorie, takže např. NorV,[52] což znamená, že znak může vystupovat jako podstatné jméno nebo sloveso.

[52] NorV – noun or verb

Obr. 12: Gramatické třídy tagů

CV tag	Expanded	Description
Signs that name, identify or show entities		
NorV	Noun or Verb	A sign which could be analysed as either a noun or a verb but there is not enough evidence to decide either way.
NP	Noun: Plain	A noun sign which cannot be re-located in space. These nouns are usually also body anchored.
NLoc	Noun: Locatable	A noun sign that can be re-located in space, but probably cannot be moved through space.
ND	Noun: Depicting	A partly lexical sign that denotes or describes an entity or participant.
Pro	Pronoun	Points to referent or to establish a referent.
Loc	Locative	Points to a location or to establish a location.
Signs that name or show processes		
NorV	Noun or Verb	A sign which could be analysed as either a noun or a verb but there is not enough evidence to decide either way.
VP	Verb: Plain	A verb sign which cannot be physically moved about in space. These verbs are usually body anchored.
VD	Verb: Depicting	A partly lexical sign that denotes or describes a process, activity or relationship.
VIDir	Verb: Indicating Directional	A verb sign that can change its start and end positions in the signing space. It can be moved meaningfully through space (this usually means can also be located). This also implies location modification.
VILoc	Verb: Indicating Locatable	A verb sign that can change its location in the signing space. Tends to be used for signs that cannot also change direction.
Signs that modify entities or processes		
Adj	Adjective	Modifies a noun.
Adv	Adverb	Modifies a verb or an entire clause or sentence.
Aux	Auxiliary	Co-occurs with a main verb, and expands its meaning in some way.
Num	Number	A sign for a number, used to describe quantities (esp. times and dates)
Det	Determiner	A sign that usually co-occurs with its referent signed explicitly before, after or simultaneously with the point. The signer is marking that the referent is known or specific in some sense (e.g., like 'the' in English).
Loc	Locative	Points to a location or to establish a location.
Signs that link signs, phrases or clauses		
Conj	Conjunction	Joins other signs or sign phrases or clauses.
Prep	Preposition	Grammatical words that fulfil a wide range of functions (esp. linked to meanings associated with direction and location). Essentially they are equated with English prepositions.
Buoy	Buoy	A handshape held up to represent/mark a referent that is being mentioned.
WH-Rel	Relative pronoun	A question sign used in a non-interrogative function, such as a relative pronoun to introduce a complement phrase.
Signs that have other functions		
Neg	Negator	Negates another sign (usually a verb). Normally considered a type of auxiliary but since there is no copula in Auslan it could be used to negate an adjective.
WH-ProQ	Wh- Pronoun Question sign	A pronoun question sign such as WHO, WHAT, WHERE, WHEN, HOW-MUCH, WHAT-AGE, etc.
Interact	Interactive	An expression of emotion or attitude and usually appears on its own, appears not to enter into any structural/syntactic relationship with any other surrounding elements (i.e., not part of a grammatical sequence of other signs).
DM	Discourse marker	Marks stages or transitions in a text.
Fragment	Fragment	A unit that appears not to enter into any structural/syntactic relationship with any other surrounding elements (i.e., not part of a grammatical sequence of other signs).
Salutation	Salutation	Conventional sign or signs used in greeting or leave taking.
Title	Title	Precedes the name of a person, showing their social role or status.
Unsure	Unsure	Used to show an attempt has been made at categorization but no decision was arrived at.

2.4.2.4 Anotování a tagování související s větami

Věta se skládá z predikátu, tj. slovesa, které určuje procesy nebo vztahy a argumentu tedy podstatného jména, které určuje účastníky děje. Ostatní jednotky věty, jako jsou ustálené obraty, příslovečná určení (času, místa atd.), gesta, přídavná jména, čísla apod. určují okolnosti dané situace. Jsou tagovány jako ne-argumenty.

Hlavním účelem této úrovně je identifikovat hlavní predikáty, tj. sloveso nebo slovesa a hlavní samostatné manuální či nemanuální znakové jednotky, které vystupují jako argumenty ve větě a určit jejich typ, počet a pořadí jejich výskytu ve větě.

2.4.2.4.1 Manuální znaky – argumenty

Identifikovatelné znaky jsou anotovány na pravé ruce RH-Arg a na levé ruce LH-Arg úrovních. Argument je značen jako A, pokud je víc než jeden, je k němu přidáno číslo, sloveso V[53] a taktéž je k němu přidáno číslo, když jich je více a ne-argumenty jako nonA.[54]

Př.	ID-gloss	PT:PRO3SG	BUY koupit	CAR auto	YESTERDAY včera
	CLU	TJ1aCLU#01			
	Arg	A1	V	A2	nonA

	ID-gloss	PT:PRO3SG	BUY koupit	BIG velký	RED červený	CAR auto	YESTERDAY včera
	CLU	TJ1aCLU#01					
	Arg	A1	V	nonA	nonA	A2	nonA

Pokud by se ve větě opakoval stejný argument, tak by byl anotován pomocí stejné anotační značky, např. A, nepřiřazovalo by se mu číslo, tedy A1, ale jednoduše by se mu přiřadilo opět A. U sloves je to stejné.

2.4.2.4.2 Úroveň sémantické role argumentů

Sémantické role znaků lze dělit do kategorií podle různých kritérií, které se mohou překrývat. Neexistuje zde žádná konečná a univerzální kategorizace. Tyto role se pohybují od základních jako je „Původ“, „Umístění“ či „Cíl“ k potenciálně dost velkému seznamu se specifickými sémantickými rolemi, které se pojí například pro každé sloveso zvlášť.

[53] V − verb
[54] nonA − non-arguments

Obr. 13: Příklad úrovní sémantických rolí

Semantic-role tier tag	Explanation
VERBS (PROCESSES, COMPLEMENTS, or RELATIONS)	
ACTION	verb that names an activity (Aktionsart: Activity, Achievement, Accomplishment)
STATE	verb that predicates an attribute or condition of something which is in principle non-inherent in the nature of that thing, often it describes a state or asserts the existence of something (Aktionsart: State)
EQUIVALENCE	verb that equates two things as the same, often it describes a state (Aktionsart: State)

Tag úrovně sémantické role Vysvětlení

SLOVESA (PROCESY, ROZVÍJEJÍCÍ VĚTNÉ ČLENY nebo POMĚRY)

ČINNOST	sloveso, které označuje činnost (aktivita, výkon, provedení)
STAV	sloveso, které přisuzuje vlastnost nebo stav něčeho, co je principiálně nepodstatné v základu té věci, často popisuje stav nebo něco tvrdí o existenci něčeho
ROVNOCENNOST	sloveso, které srovnává dvě věci jako stejné, často popisuje stav

Příklad věty s jedním argumentem

ID-gloss	PT:PRO3SG	STUDY studovat	ALL-DAY-LONG celý den
CLU	TJ1aCLU#01		
Arg	A	V	nonA
MacroRole	ACTOR účastník	PROCESS proces	
SemRole	AGENT činitel	ACTION činnost	
FreeTransl	*He studied all day long.*		
Volný překlad	Celý den studoval.		

V korpusu se rozeznává 5 základních sémantických úrovní, které se dále dělí. Jsou to úrovně pro slovesa (týkající se průběhu nebo vztahů, např. slovesa pojmenovávající nějakou aktivitu), slovesa (týkající se nějakého hlediska, např. slovesa označující souvislou činnost, která byla dokončena), herec (např. osoba, která promlouvá, nebo např. osoba, jejímž směrem se něco hýbe), vlastnosti (např. téma – argument, o kterém se něco tvrdí, komentář – argument, který něco tvrdí o tématu), pomocné prvky (např. umístění – místo, kde je něco umístěno, způsob – způsob, jakým je něco uděláno).

2.4.2.4.3 Úroveň doslovného překladu

Doslovný překlad se týká anotace celé věty, spíše než jednoho znaku. Doslovný překlad často není gramaticky správná angličtina. Tento překlad se snaží ukázat, jak je určitý význam reprezentován ve větě a hlavně prezentovat to, co je více či méně explicitně vyjádřeno ve zdrojovém a cílovém jazyce. Identifikování významu každé věty tak, jak se jeví, může doslovnému překladu pomoci učinit strukturu konstrukce více zřejmou a přístupnou k analýze.

Doslovný překlad také může poskytnout informaci o využití prostoru v Auslanu nebo také může informovat o přítomnosti nebo nepřítomnosti vlastností, které se mohou objevit ve volném anglickém překladu a mohou být nesprávně považovány za přítomné nebo chybějící v originálním projevu v australském znakovém jazyce. Tato úroveň je pravděpodobně užitečná pro lingvisty, kteří nahlíží do těchto projevů, ale nerozumí Auslanu.

Glosses	LEAVE	BEFORE	EIGHT-O'CLOCK	WILL	ARRIVE	LUNCH
glosy	odejít	před	osm hodin	bude	dorazit	oběd
CLU	TJ1aCLU#01			TJ1aCLU#02		
Literal	(we) leave before eight o'clock?			(we) will arrive (before) lunch		
doslovný	(my) odejdeme před osmou hodinou?			(my) dorazíme (před) obědem		
Free	*If we leave before eight o'clock, we will get there before lunch.*					
volný	Pokud odejdeme před osmou hodinou, tak tam dorazíme před obědem.					

2.4.3 Terciální zpracování dat

Možnosti, které se nabízejí anotováním digitálních videí v korpusu znakového jazyka, tak jak to bylo popsáno výše, umožňují pracovat se získanými daty skrze vyhledávání a třídění primárních a sekundárních anotací, jako je např. frekvence určitých vlastností. Tyto informace pak mohou být přidány do korpusu, např. připojení tagů k již existujícím ID glosám, aby se korpus obohatil a umožňoval tvorbu více sofistikovaných analýz, které budou brát tyto hodnoty v potaz.

Budoucí rozvoj funkcí ELANu nejspíše usnadní toto všechno. Kupříkladu bude možné vytvořit anotace založené na „překrývajících se hodnotách" na již existujících anotačních úrovních. Výzkumníci budou schopni specifikovat, kdy se anotace překrývají na úrovních X, Y a Z a zda by nová anotace měla být vytvořena na úrovni W, a poté i určit anotaci nebo tag, který by mohl být automaticky vložen do nově vytvořeného pole. Využití této techniky může korpus obohacovat tak, jak by to bylo nemožné pro člověka, protože by daná jazyková data nemohl kódovat v nějakém přiměřeném čase.

2.4.3.1 Úroveň frekvence

V ELANu je možné vyhledávat skrze vícenásobná anotační data a díky tomu lze vytvořit statistické frekvence pro anotace a tudíž i pro ID glosy. Při exportu znaků do databáze nebo do konkordančních[55] programů, mohou být znaky zařazeny do frekvenčních skupin, např. velmi vysoká, vysoká frekvence, střední, nízká frekvence apod., které jsou založené na těchto statistikách. Nicméně tato samotná informace může být vložena do ELANu jako tag na frekvenční úrovni.

2.5 Oprava či změna dat v korpusu

Část základních anotačních dat z korpusu je veřejně dostupná po přihlášení do archivu Auslan. Registrovaní výzkumníci mají umožněn přístup k editovaným anotacím v korpusu pro jejich vlastní výzkum výměnou za souhlas, že vrátí obohacená data, tj. data s přidanými anotacemi.

V případě přístupu do editovaného korpusu k anotačním datům, výzkumník nebo anotátor, který si myslí, že našel chybu, může jednoduše identifikovat tuto chybu vložením komentáře o této možné chybě na úroveň komentářů. To umožňuje korpusovému manageru určit možné chyby ještě předtím, než se rozhodne, zda je nutná náprava. Tím se zamezí riziku změn a následnému nepředvídatelnému řetězovému působení transformací v anotacích na dalších úrovních, které by vedly k nevysvětlitelným nebo dokonce neviditelným nesrovnalostem, které by porušily integritu dat.

Také šetří čas, když jeden anotátor nebo výzkumník může „zafixovat“ něco, co další anotátor považuje za chybu a posléze to může vrátit zpět. Tento stejný proces se používá během počátečního tvoření primárních anotačních dat, tzn. anotace nejsou pevně zafixované či nezměnitelné.

2.6 Shrnutí procesu anotace

Pro přehlednost jsou zde jen velmi krátce shrnuty předešlé informace týkající se cílené anotace dat. Anotace probíhá ve třech na sebe navazujících fázích.

Při primární anotaci se vytváří volný překlad znakového projevu, připojují se ke znakům identifikační glosy, které musí být velmi dobře vytvořeny, podle konkrétních pravidel.

[55] konkordance − představuje všechny výskyty hledaného jevu v korpusu spolu s okolním kontextem

Jako nejlepší se jeví videonahrávku nejdříve volně přeložit a posléze přiřazovat ID glosy, protože glosa je přehlednější a pro slyšící uživatele korpusu i jednodušší ji vyhledat podle toho, že se zjednodušeně řečeno spojuje slovo a znak jako rovnocenná jednotka. Taktéž se v této fázi rozlišuje mezi znaky plně lexikalizovanými, což jsou konvencionalizované znaky znakového jazyka, pro něž se relativně snadno najde ekvivalentní slovo mluveného, resp. psaného jazyka. Další kategorií jsou nelexikalizované znaky, které by šly označit jako gesta. Třetí skupinou znaků, které se rozeznávají, jsou částečně lexikalizované znaky, což jsou znaky, které spadají mezi dvě předešlé skupiny a mezi ně patří např. klasifikátory či ukazovací znaky.

V sekundární fázi se přidávají další anotace k již existujícím anotacím, také se tagují tokeny, přičemž je důležité pamatovat na kontext projevu. Dále dochází k segmentaci promluv a subkategorizaci stavby znaku, jednotek promluvy, části promluv a vět. V této fázi je také možné přidat přepis např. v HamNoSys, ovšem v korpusu Auslan se toto nerealizuje.

A v poslední terciární fázi se inkorporují informace odvozené z vyskytujících se rozličných hodnot z primárního a sekundárního zpracování dat v korpusu, což je například frekvence užívání určitého znaku. (Johnston, 2016)

3. Korpus německého znakového jazyka

Tato kapitola je zaměřena na popis druhého, poměrně rozsáhlého korpusu a to korpusu německého znakového jazyka (dále také jen DGS[56]). V moderní lexikografii se zvyšuje povědomí o tom, že by slovníky měly být založeny na korpusových datech. V lexikografii znakových jazyků bylo dosud vytvořeno pouze několik korpusů a ještě méně z nich bylo využito pro tvorbu slovníku.[57] Proto se v Německu zaměřili nejen na tvorbu korpusu, ale i na následné vytvoření překladového slovníku z německého znakového jazyka do německého jazyka a naopak, který by byl založen na analýze dat z korpusu.

Projekt korpusu má za cíl zdokumentovat užívání DGS v běžné komunikaci a také zachytit informace o kultuře Neslyšících. Korpus by měl obsahovat velké množství různých projevů a gramatických struktur, kvůli budoucímu výzkumnému využití. Ti, kdo pracují na daném projektu, předpokládají, že materiál nebude přínosný jen pro výzkumníky, ale také pro samotné členy komunity Neslyšících. Předpokládá se zveřejnění 50 hodin natočeného materiálu a jeho příslušný přepis. (Hanke, 2010, s. 178−185) Korpus současně nemá sloužit pouze jako podklad pro různé výzkumy či databáze pro zpracování slovníku, ale má být využit ke vzdělávacím účelům a tvorbě materiálů k výuce znakového jazyka (Rathmann, 2016).

Práce na korpusu DGS začala v roce 2009 na Academy of Science v Hamburku a za vůdčí osobnost tohoto projektu lze označit Christiana Rathmanna, který je nepostradatelným členem týmu už jen z toho důvodu, že on sám je neslyšící lingvista (Hanke, 2010, s. 178−185).

Korpus DGS na rozdíl od korpusu Auslan nevychází z již dříve vytvořeného archivu znakových projevů, ze kterého jsou posléze čerpána data do korpusu. Němečtí výzkumníci se přímo zaměřili na sběr jazykového materiálu, který byl určen pro korpusovou analýzu a po následném zpracování i jako podklad pro vznik slovníku znakového jazyka. Tak jako v korpusu australského znakového jazyka, i nyní bylo nutné rozhodnout, jakým způsobem, kde a od koho se bude kýžený materiál elicitovat.

Inspirace pro tvorbu tohoto korpusu vycházela z již vzniklých korpusů znakových jazyků a to konkrétně korpusu australského znakového jazyka a korpusu nizozemského znakového jazyka. Práce na těchto korpusech začala již dříve, proto mohli němečtí výzkumníci vycházet z jistého know-how tvorby.

[56] DGS − Deutsche Gebärdensprache
[57] Slovník založený na korpusu znakového jazyka byl vytvořen např. v Polsku, Švédsku či Dánsku.

Další korpusy začaly postupně vznikat v Polsku či ve Francii a s těmito státy výzkumníci z Německa komunikovali a spolupracovali, a proto bylo v těchto případech možné předkládat různé návrhy, které se již osvědčily. Kooperace se například vyplatila v tom, že výzkumníci mohli využít elicitační materiál, u kterého již bylo jisté, že v praxi funguje, což se stalo při spolupráci mezi kolegy z Německa a Polska, kdy polská strana získala povolení k použití těchto materiálů.

Do elicitace jazykového materiálu bylo celkově zapojeno 330 informantů z celého Německa. V souhrnu bylo nahráno 540 hodin materiálu ve znakovém jazyce, což odpovídá přibližně 2,5 milionu tokenů. V rámci získaných metadat byl zaznamenáván věk zúčastněných, který se pohyboval v následujících rozmezích 18–30, 31–45, 46–60 a 61 a výše. Současně byl kladen důraz na rovnoměrné zastoupení mužů a žen zapojených do elicitace. Dále byly vytvořeny skupiny, které indikovaly věk osvojení znakového jazyka respondenty v rozpětích 0–3, 4–6 a období nad 6 let věku, která byla nejméně početná. Důležitým aspektem bylo také místo, kde informanti vyrůstali, kam chodili do školy či četnost jejich stěhování. Tyto všechny informace byly sledovány kvůli reprezentativnosti získaných dat. (Rathmann, 2016)

Sběr dat probíhal na 12 předem určených sběrných místech rovnoměrně rozmístěných po celém Německu, kdy daná místa byla charakteristická relativně vysokou populační hustotou neslyšících osob a taktéž byla tato místa dostupná okolním městům a vesnicím, takže byla vcelku dobře zajištěna reprezentativnost z hlediska geografie. Mobilní studio tvořila technika, se kterou se tři roky cestovalo do těchto vybraných míst a natáčel se s ní jazykový materiál s vybranými respondenty. (Hanke, 2010, s. 178−185)

Důležitou otázkou je také způsob získávání kontaktů na jednotlivé respondenty. Nejprve bylo vybráno 22 kontaktních osob, tedy 1–2 osoby z každého sběrného místa, které spolupracovaly s korpusovým týmem a pomohly vybrat vhodné respondenty pro projekt korpusu. Tyto kontaktní osoby měly informace o daném regionu, chodily do místních škol, znaly tedy neslyšící ze své komunity a mohly určit osoby hodící se do výzkumu. (Rathmann, 2016)

Při jednom natáčení se sešly vždy tři osoby – neslyšící moderátor, který vedl celé setkání a dva respondenti, od kterých probíhala elicitace jazykových dat. Celé jedno setkání trvalo přibližně 7 hodin, včetně tří přestávek (Prillwitz, 2008, s. 159−164).

Důležitou roli při sběru jazykového materiálu představuje anonymita respondentů. Na natočených materiálech nelze u respondentů zakrýt obličej z důvodu užívání nemanuálních komponentů, proto informanti dávají svolení k použití tohoto materiálu.

Daným svolením je informovaný souhlas, který je současně přeložen do znakového jazyka. Informovaný souhlas seznamuje respondenty s účelem výzkumu a využitím získaných jazykových dat. Taktéž jsou zde informace o tom, že celé nahrávky nebudou zveřejněny pro širokou veřejnost a pokud se respondentům nějaká část jejich nahrávky nezdá vhodná k použití, mohou ji nechat vyjmout a nebude tak využita buď jen pro publikování, nebo i v rámci celého výzkumu. Stěžejní je, aby respondenti porozuměli celému projektu a později nevznikala žádná nedorozumění. (Rathmann, 2016)

3.1 Elicitace jazykových dat

Při elicitaci bylo nutné zajistit každému informantovi stejné informace, proto byl natočen figurant, který pro každý úkol znakoval instrukce v německém znakovém jazyce. Tyto instrukce byly posléze informantům prezentovány na obrazovce spolu s elicitačními materiály. Každý informant měl před sebou umístěnou svoji obrazovku, a proto bylo možné přehrát každému informantovi jinou instrukci, pokud měli v rámci úkolu rozdílné role.

3.1.1 Role moderátora a prezentace podnětů

Stejně jako při tvorbě korpusu Auslan, i tady je nutné eliminovat paradox pozorovatele, proto celé setkání vede neslyšící moderátor a ve studiu není během nahrávání přítomna žádná další osoba. Moderátor je zodpovědný za představení jednotlivých úkolů, stejně tak i za bezproblémový průběh natáčení.

Každý úkol je krátce představen samotným moderátorem, ale následně je ještě vše detailně vysvětleno skrze instrukce prezentované ve videonahrávkách v německém znakovém jazyce, které jsou pouštěny na obrazovkách před každým jednotlivým úkolem. Tato přednatočená vysvětlení mají zajistit, aby všichni informanti dostali přesně ty stejné informace a nic nebylo vypuštěno či nechtěně změno. Před samotným plněním elicitačního úkolu je možné vyjasnit si případné nesrovnalosti s moderátorem, pokud něčemu respondenti nerozumí.

Materiály používané jako podněty během prezentace úkolů zahrnují různé mediální formáty jako obrázky, kresby či videonahrávky. Ty jsou promítány na obrazovkách jako slidy v poloautomatické prezentaci, kdy je jeden slide následován dalším v přesně stanovené rychlosti, avšak podle potřeby může moderátor do této rychlosti zasáhnout a částečně ji kontrolovat. A jak již bylo řečeno, záleží na jednotlivých úkolech, zda je prezentace pro oba informanty stejná nebo odlišná.

Cílem úkolu je, aby skrze použitý podnět byla vyvolána konverzace mezi oběma informanty, zatímco moderátor je pouze pozoruje a zasahuje výlučně v případě, pokud to je nezbytně nutné. Moderátorovou povinností je kontrolovat čas, který je určen na jednotlivé úkoly, kvůli dostatečnému času na celé sezení. Jsou také vytvořeny speciální úkoly navíc, které můžou, ale nemusí být zahrnuty do elicitačních materiálů, podle průběhu setkání a zbývajícího času. Vedení elicitace během celého setkání vyžaduje od moderátora jeho plnou pozornost a zodpovědnost.

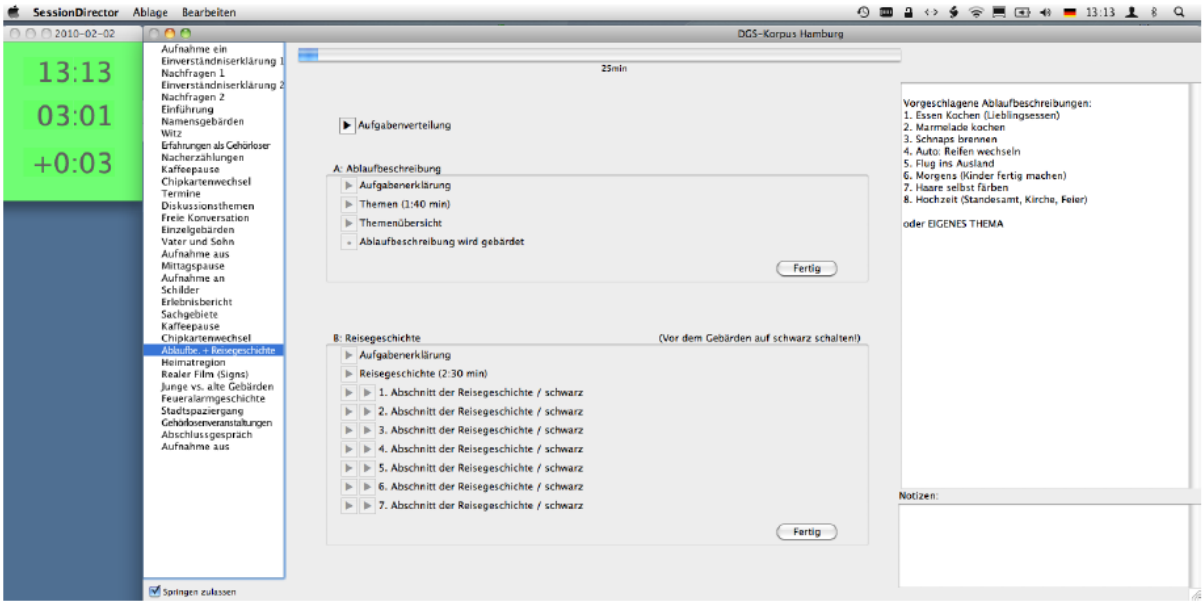
3.1.2 Program „session director“

Program session director umožňuje moderátorovi prezentovat slidy respondentům a řídit tak jejich přehrávání. Jak již bylo zmíněno, ve studiu jsou obrazovky umožňující přehrávání pokynů k elicitaci a jsou tam celkem tři – jedna pro moderátora a po jedné pro každého respondenta. Obrazovky určené pro informanty jsou umístěny čelně k těmto neslyšícím osobám a prezentují se na nich materiály, určené přímo jim. Moderátorova obrazovka je určena pro přehrávání pokynů v rámci session director, ale také se mu tam zobrazují další dvě okna, na kterých pozoruje prezentace, které běží na obrazovkách informantů. Kromě zobrazení seznamu úkolů umí session director v hlavním okně zobrazit detail aktuální otázky, na které se pracuje. To zahrnuje i ukazatele průběhu úkolu, který udává čas strávený na daném úkolu ve vztahu k naplánovanému času, stejně tak prezentuje pořadí podúkolů. Pořadí úkolů je již předdefinované, ale pouze moderátor může aktivovat další úkol spouštěcím tlačítkem, které má na své obrazovce. Pokud moderátor zjistí, že informantům není jasná některá část vysvětlení vztahující se k úkolům, může opakovaně spustit danou instrukci. Takto moderátor může průběžně kontrolovat, zda respondenti rozumí promítaným pokynům.

Není proveditelné a ani žádoucí, aby byl každý úkol přesně časově naplánován. Avšak je nutné, aby se dodržela naplánovaná doba celkového setkání, proto je pro tento účel v programu zobrazováno další okno, které udává aktuální čas, uplynulý čas a čas, který ukazuje, zda se sezení zpožďuje či předchází časový plán. Okno programu umožňuje měnit barvy, které slouží jako upozornění, pokud je překročen čas určený na daný úkol. Jestliže se setkání z nějakého důvodu zpožďuje v čase, tak není možné, aby se redukoval čas pro každý následující úkol na minimum, ale rovnou se přistoupí k přeskočení podúkolů nebo dokonce celého úkolu. Nastavení programu session director vede moderátora v rozhodnutích, který úkol je onen speciálně vytvořený a tudíž je zařazen navíc, tím že ho program označí, a proto může být následně vypuštěn.

V zásadě je pořadí úkolů, zahrnující přestávky, předem naplánováno. Nicméně moderátor má volnost v tom, že může přeskupit posloupnost otázek nebo změnit předpokládaný čas trvání setkání. Tato možnost se vyplatí například v momentě, kdy jeden z informantů přijde pozdě na setkání. Výzkumníci ovšem mají zkušenosti i s opačnou situací, kdy se informanti při plnění úkolů tolik baví, že jim nevadí zůstat déle, aby nepřišli o žádný přichystaný úkol.

Obr. 14: Okno zobrazující se moderátorovi v rámci programu session director



3.1.3 Průběh setkání – soubory

Když setkání začne, program session director načte XML soubor obsahující popis průběhu setkání. Pro každý úkol a podúkol tento soubor definuje očekávanou a maximálně přijatelnou délku trvání a také zahrnuje identifikátory slidů, které budou zobrazeny informantům, buď jako celý soubor slidů pro oba, nebo jako oddělené soubory pro každého informanta zvlášť.

Mimo to definuje relativní důležitost každého úkolu, kterou session director případně využije k označení úkolů, které by mohly být přeskočeny. Pokud je nutné, aby měl moderátor psané podklady, týkají se např. otázek, které má respondentům klást, je možné tento psaný text zobrazovat na moderátorově obrazovce nezávisle na probíhající prezentaci zadání úkolu.

3.1.4 Instruktáž moderátora

Moderátor musí umět pracovat s programem session director naprosto bezpečně a bez jakéhokoli zaváhání. S programem je seznamován v rámci instruktážních kurzů. Přitom musí program užívat opakovaně, v rámci cvičných setkání, s různě kooperujícími či ne příliš spolupracujícími zkušebními informanty.

Řízení času je velmi složitý aspekt setkání, a proto moderátor musí umět regulovat dobu elicitace tak, aby se vykompenzovaly případné časové nesrovnalosti. Zpětná vazba z prvního setkání pomohla zavést novou funkci „pauza“ k zastavení úkolu, pokud je nutná spontánní přestávka. (Hanke, 2010, s. 106−109)

3.2 Tvorba elicitačních materiálů

Jak již bylo zmíněno, jedno sezení trvá přibližně sedm hodin, během kterých je pomocí různých úkolů elicitován jazykový materiál od neslyšících respondentů. Tvorbě korpusu DGS předcházelo poměrně dlouhé období příprav elicitačních úkolů a určení jejich pořadí.

Nejprve byla pozornost věnována již existujícím elicitačním materiálům, které byly dříve použity v rámci jiných projektů. Přestože existuje mnoho různých způsobů, jak sesbírat jazyková data od respondentů skrze podněty jako jsou obrázky, fotografie či filmy, je nutné, aby výzkumníci znakového jazyka prošli různé elicitační materiály, které by ve výsledku poskytly ucelený přehled těchto podkladů. Mnoho elicitačních materiálů je sdíleno mezi lingvisty, avšak žádný není veřejně dostupný, což je pochopitelné z důvodu snížení možnosti přípravy respondentů na dané setkání.

V rámci komunity neslyšících byla tedy vedena anketa zaměřující se na sesbírání potřebných informací směřujících k tvorbě žádoucího elicitačního materiálu. Byl vytvořen dotazník, který se zaměřoval na zjištění následujících detailů:

- nejvhodnějšího forma elicitačního materiálu (obrázek, animovaný film atd.);
- obsah či téma elicitačního materiálu (téma rozhovoru, obsah obrázkového příběhu apod.);
- srozumitelnost výzkumných otázek;
- jasnost specifických otázek, které budou kladeny informantům.

Na základě těchto dotazníků byli výzkumníci schopni kategorizovat různé typy elicitačních materiálů, které chtějí použít, do následujících skupin:

- jazykový vstup (seznam izolovaných slov, napsané jednoduché věty, napsané texty, znakovaná videa);
- obrázky (z animovaného filmu nebo pohádky, jednotlivé obrázky, obrázkové příběhy, fotografie);
- filmy, animace;
- témata pro volné rozhovory nebo diskuze (pohádky, bajky);
- hry;

- kombinace obrázků a slov.

Daná analýza průzkumu dovolovala popsat výhody a nevýhody různých podnětů. Mimo to vyšlo najevo, že existují materiály, které jsou zvláště vhodné pro porovnávající lingvistické studie, protože již byly využity ve studiích jak pro znakové jazyky, tak pro mluvené jazyky, např. obrázková kniha Frog, where are you? (Mayer, 1969), animovaná pohádka Tweety a Sylvester (Warner Brothers, 1950) apod. Průzkum také ukázal, že je při přebírání elicitačních materiálů ze zahraničí nutné respektovat kulturní rozdíly. Někteří lingvisté jako elicitační podnět používají Ezopovy bajky, jako tomu bylo v případě korpusu Auslan. Ovšem v Německu jsou mnohem více známy pohádky bratří Grimmů, proto jsou zde Ezopovi bajky nevhodným elicitačním materiálem, kvůli neznalosti daných textů. Lingvisté by však měli mít neustále na paměti, že i když se jim nějaký příběh může jevit jako známý, tak nemusí být v povědomí komunity Neslyšících. Průzkum také poukazuje na to, že pokud je většina podnětů podobného typu, může se postupně elicitace pro informanty stát nudnou záležitostí.

Nakonec byly z jiných projektů přejaty úkoly týkající se příběhů Frog, where are you?, Pear Story a animovaná pohádka Tweety a Sylvester. První dva byly původně použity jako podněty ve studiích mluvených jazyků (Berman, Slobin, 1996) a následně byly přejaty výzkumníky znakových jazyků. Animovaná pohádka Tweety a Sylvester byla použita v lingvistické studii pro porovnání klasifikátorových konstrukcí v 9 znakových jazycích.[58]

Jiné existující materiály nešlo použít nebo se nehodily z hlediska zamýšleného účelu projektu. Například dostupné podněty pro shodová slovesa a negace ve znakovém jazyce (materiály z Centre for Sign Linguistics and Deaf Studies in Hong Kong) byly vytvořeny k elicitaci izolovaných vět. A v tomto případě by pouhé věty bez kontextu nebyly tím, co je smyslem korpusu, tedy umožnění výzkumníkům analyzovat znaky a lingvistické struktury v širších souvislostech. Z tohoto důvodu bylo nutné vytvořit nové podklady zaměřující se na požadované jevy.

Důležitou otázkou je také vyřešení autorských práv s přejatými i s nově vytvořenými materiály. Například dva obrázkové příběhy musely být ze souboru vyňaty, protože nakladatel nedal povolení k použití těchto materiálů.

V průběhu vytváření úkolů probíhalo testování, které ověřovalo relevantnost jednotlivých úkolů. Tyto testy byly provedeny slyšícími výzkumníky a studenty − asistenty s neslyšícími kolegy na IDGS[59] jako informanty.

[58] studie A Cross-linguistic Study of Sign Language Classifiers od D. Brentari z roku 2001
[59] IDGS - Institut für Deutsche Gebärdensprache (Institut německého znakového jazyka)

Po každém testu byli informanti tázáni, zda se během práce na úkolech cítili pohodlně, zda pochopili instrukce a pokud ne, co by doporučili vylepšit. Také byli dotazováni, zda považují tyto úkoly za vhodné a použitelné pro potenciální neslyšící respondenty. Všechna setkání v rámci testování byla nahrávána a analyzována podle následujících hledisek:

- Cítí se informanti v rámci úkolů komfortně?
- Rozumí informanti nahraným instrukcím? Jsou sděleny všechny nezbytné informace?
- Rozumí informanti materiálům, které jsou jim předloženy? Vidí to, co chceme, aby viděli?
- Kolik času je třeba na to, aby informanti dokončili úkol?
- Kolik znakového projevu informanti produkují v každém úkolu?
- Produkují informanti očekávaný druh jazykového výstupu?

Testy odhalily, že v určitých případech nebyly první verze instrukcí plně srozumitelné. Toto následně vedlo k několika etapám změn a opakovanému testování před tím, než byla hotová finální verze elicitačních materiálů. V některých úkolech pretesty ukázaly, že natočené instrukce k jednotlivým úkolům neobsahovaly dostatečné množství informací.

Jednou z problematických věcí bylo odkazování ve znakovém projevu. Znakující osoba v natočené instrukci oslovovala informanty ukazováním přímo dopředu na ně a odkazovala k druhému informantovi ukazováním za záda kvůli zasedacímu pořádku ve studiu. Ačkoli se odkazování k informantům stanovené ve znakovaných instrukcích hodilo do reálného prostoru, kde měla elicitace probíhat, tak informanti nerozuměli tomuto využití prostoru ve videonahrávce. Z tohoto důvodu moderátor představil použitý odkazovací systém na úplném začátku setkání, aby nemohlo docházet k nedorozumění. Pretesty také informovaly o tom, že někteří informanti měli tendenci znakovat směrem spíše k moderátorovi místo ke svému konverzačnímu partnerovi, se kterým měli vést dialog. Moderátor se tedy musí speciálně trénovat na to, aby uměl předcházet těmto situacím.

Je také nutné upravovat použitý materiál po stránce grafické, takže se změnila velikost písma a i některé obrázky byly nahrazeny vhodnějšími, které jednoznačněji vyjadřovaly svůj záměr.

Pretesty ukázaly, že obrázkové příběhy je nutné v průběhu jejich převyprávění informanty skrýt. V opačném případě totiž znakující sleduje obrázkový příběh místo toho, aby se díval na konverzačního partnera.

Po prvním testovacím období byly vybrány jednotlivé úkoly a byly sestaveny do pořadí, které bylo pokládáno za vhodné, vzhledem k délce setkání, které mělo mít pět a půl hodiny elicitační části s hodinou a půl na přestávky.

Po skončení pretestů byla ještě provedena dvě kompletní testovací setkání, každé trvalo jeden celý den. Na prvním setkání byly informanty neslyšící studenti – asistenti a na dalším sezení dvě neslyšící osoby, které nebyly žádným způsobem spojené s IDGS. Kontaktní osoba zodpovědná za oblast Hamburku, kde se později také natáčelo, moderovala obě setkání.

Hlavním cílem těchto kompletních testovacích setkání bylo simulovat elicitační setkání v situacích, které by byly blízké reálnému sběru materiálu, jak jen bylo možné. Sledovány byly zejména tyto faktory:

- Jak dlouho trvají jednotlivé úkoly?
- Jak dlouho trvá celé elicitační setkání?
- Jsou přestávky ve správných chvílích? Jak moc je setkání pro informanty stresující?
- Funguje pořadí úkolů? Ovlivňují se úkoly navzájem?
- Funguje interakce mezi moderátorem a informanty?
- Funguje program session director tak, jak bylo zamýšleno – prezentování úkolů a podnětů? Ví, informanti, co a kdy mají dělat?
- Zvládnou neslyšící lidé vypracovat úkoly, když mají různé edukační zázemí?

Jedním z podstatných výsledků z testovacích setkání bylo zjištění, že úkoly zabírají méně času, než v pretestech a informanti celkově poskytli méně materiálu, než se očekávalo. To zřejmě bylo způsobeno tím, že neslyšící kolegové, kteří se zúčastnili pretestů věděli o záměru projektu a snažili se tedy produkovat mnohem více jazykového projevu a ochotně spolupracovali. Dalším důvodem mohlo být, že se testovací informanti více zaměřovali na každý úkol, zatímco „opravdoví“ účastníci věděli, že setkání obsahuje mnoho úkolů a bude trvat sedm hodin, takže se víc zaměřovali na dokončení a neprodlévali na jednotlivých úkolech.

Výsledkem tedy bylo, že se moderátor musí soustředit na to, aby se zadaný úkol nedokončil tak rychle, jak je jen možné, ale aby byl řádně využit čas a bylo možné získat očekávané množství jazykového materiálu. Důsledkem analýz těchto dvou setkání také bylo, že byla přizpůsobena očekávaná doba trvání stanovená pro jednotlivé úkoly, modifikovaly se úkoly přidáním podúkolů a podnětů a taktéž se změnilo pořadí úkolů k vyvážení doby mezi přestávkami.

3.3 Zadání úkolů

Poté, co moderátor objasní otázky týkající se informovaného souhlasu, který se vztahuje ke zveřejnění videí a zkontroluje s každým informantem vyplněný dotazník pro metadata, zaujmou moderátor a informanti místo ve studiu a začíná oficiální nahrávání. V následujícím textu představím koncepci jednotlivých úkolů.[60]

3.3.1 Jmenné znaky

Prvním úkolem je ukázání vlastního jmenného znaku informanta a vysvětlení jak vzniknul. Cílem tohoto úkolu je shromáždit jmenné znaky jako specifickou součást kultury neslyšících. Také to má informanty připravit na následující úkoly a představit je sobě navzájem. Pokud respondenti chtějí, mohou své jméno vyhláskovat prstovou abecedou. Celý proces trvá přibližně dvě a půl minuty.

3.3.2 Vtipy

Ještě před oficiálním setkáním je každý informant předem požádán, aby si připravil jeden vtip, který bude vyprávět svému komunikačnímu partnerovi. Výzkumníci přijali fakt, že jeden úkol budou respondenti znát ještě dříve, než natáčení začne a situovali ho hned na začátek natáčení. Toto zadání má také respondentům pomoci vpravit se do následujících hodin nahrávání a pomoci jim cítit se lépe tím, že znakují něco připraveného. Očekává se, že někteří z respondentů bude prezentovat vtipy týkající se znakového jazyka nebo hluchoty, což je velmi zajímavou a nepostradatelnou součástí kultury Neslyšících. V této části záleží na délce produkce vtipů obou informantů, ale trvání úkolu se pohybuje mezi 2 a 7 minutami.

3.3.3 Zkušenosti neslyšících

Moderátor požádá oba informanty, aby pohovořili na téma úzce spjaté se životem neslyšícího: školy pro neslyšící, internátní ubytování, domovy pro neslyšící seniory, sportovní kluby, spolky neslyšících atd. Zkrátka mají vyprávět příběhy ze svého života, které se specificky týkají neslyšících.

[60] Zdroj, ze kterého byly čerpány informace o elicitačních úkolech: HANKE, Thomas a Christian RATHMAN. Elicitation Methods in the DGS (German Sign Language) Corpus Project. In: *4th Workshop on the Representation and Processing of Sign Languages:: Corpora and Sign Language Technologies*. Hamburk: University of Hamburg, 2010, s. 178-185. ISBN 2-9517408-6-7. Dostupné z: https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/Nishio_et_al._LREC_2010_elicitation_methods.pdf

K této otázce není připraveno instruktážní video, ale místo toho si moderátor připraví otázky, které se hodí k profilu informanta, přičemž využije metadata z dotazníku. Cílem je zachytit typické zkušenosti ze života neslyšících. Pro tento úkol je vyhrazen čas 20 minut.

3.3.4 Převyprávění filmu a obrázkového příběhu

Informanti mají zhlédnout obrázkový příběh nebo krátký film bez jazykového vstupu a jsou požádáni o jeho převyprávění. Jsou vytvořeny dva různé soubory obsahující každý po dvou podnětech. Používají se obrázkové příběhy Frog, where are you? (Mayer, 1969), animovaná pohádka Sylvester a Tweety (Warner Brothers, 1950) a film Pear story (Chafe, 1980). Všechny tři příběhy byly jako podněty přebrány z jiných výzkumů, aby bylo v budoucnu případně možné provádět analýzy na porovnávání jazyků. Čtvrtým podnětem je komický příběh Haushaltshilfe (Hospodyně), který byl vysílán v německé televizi a byl určený pro neslyšící diváky. Toto je jediný podnět v německém znakovém jazyce, který byl použit v rámci setkání.

Materiály Frog, where are you? a Sylvester a Tweety jsou přehrány dvakrát. Poprvé jsou tyto dva podněty prezentovány celé, ale v rámci druhého přehrávání jsou rozděleny do krátkých sekvencí obrázků či několika klipů a informanti převypráví vždy pouze danou část příběhu.

Pro účely korpusu jsou některé podněty oproti originálu pozměněny. Například film Pear story obsahuje zvuk v pozadí, ale je přehráván beze zvuku. Příběh Haushaltshilfe je puštěn bez titulků, které byly původně jeho součástí. Celý úkol, jenž zahrnuje soubor s příběhy Frog, where are you? a Sylvester a Tweety, zabírá celkem 27 minut a soubor obsahující film Pear Story a Haushaltshife trvá průměrně 17 minut.

Obr. 15: Ukázka obrázků z elicitačního příběhu Frog, where are you?

3.3.5 Kalendářní jednotky

Informantům je ukázán jednotýdenní kalendář se zapsanými fiktivními schůzkami a jsou instruováni, aby si naplánovali dvě schůzky ve dvou různých časových intervalech. Také jsou tázáni na jejich aktivity během týdne. Cílovou slovní zásobou jsou dny v týdnu, časové údaje a různé aktivity s nimi spojené, jako je např. kontrola u lékaře, dovolená, setrvání v práci, sportovní aktivity, objednání opraváře domů apod. Toto je jeden z úkolů, u kterého se požaduje jistá interakce mezi respondenty. Záměrem je zjistit strategii plánování a vyjednávání. Úkol trvá přibližně 9 minut.

3.3.6 Rozhovor

V položce „rozhovor“ jsou představena čtyři konverzační témata a informanti si jedno vyberou. Témata zahrnují dva okruhy týkající se neslyšících, je to názor na kochleární implantát a pohled na komunitu neslyšících. Další dva obsahují obecná témata, např. náhled na zákaz kouření apod.

Záměrem je zapojit informanty do živého a emocionálního rozhovoru, ve kterém nebudou tolik přemýšlet o své produkci jazyka. Tato část trvá kolem 20 minut, v některých případech je moderátor nucen ukončit rozhovor a přesunout se na jiné téma.

3.3.7 Elicitace jednotlivých znaků

Zatímco celé elicitační setkání je primárně zaměřeno na produkci monologů nebo dialogů, tento úkol je orientován na získání oblastních variant izolovaných znaků. Informantům jsou ukázána slova napsaná v německém jazyce, která mohou být doplněna obrázky. Následně jsou respondenti vyzváni, aby znakovali ekvivalenty daných lexémů i v kontextu jedné krátké příkladové věty. Celkově bylo vybráno 34 lexémů, např. chléb, voda, žena, narozeniny či chyba.

Dále jsou elicitovány znaky pro měsíce, roční období a 11 barev. Záměrem je vysledovat regionální varianty u předem vyhlédnutých znaků, u kterých byla variantnost již dříve vypozorována. Tato elicitace trvá cca 12 minut.

3.3.8 Převyprávění obrázkového příběhu Vater und Sohn

Na závěr první části setkání je každý informant požádán, aby převyprávěl jednoduchý obrázkový příběh obsahující 5 nebo 6 obrázků, které byly převzaty z knihy Vater und Sohn (Otec a syn) od německého karikaturisty E. Ohsera.

Toto je jeden z volitelných úkolů a může být přeskočen, pokud se předchozí úkoly časově protáhnou. Na tento úkol jsou vymezeny zhruba 4 minuty.

V této části je naplánovaná přestávka, takže moderátor informuje respondenty o možnosti opustit místnost a vrátit se po přestávce. Z etických a praktických důvodů, moderátor explicitně řekne, že jejich rozhovor není v této době natáčen. Takto probíhá každá další přestávka.

3.3.9 Varující a zákazové značky

V prvním odpoledním úkolu informanti zhlédnou varující a zákazové značky sesbírané z různých míst na světě a jsou vyzváni, aby diskutovali o tom, co značky mohou znamenat. Informanti většinou tyto značky neznají a jsou nuceni hádat. Jedním z cílů je opět připravit informanty na pokračování natáčení a na následující náročné úkoly. Výzkumným cílem je získat věty, ve kterých bude užita negace v koherentním kontextu.

Původně úkol obsahoval 12 různých značek, avšak později byly ještě 4 značky přidány, protože se ukázalo, že je diskuze kratší než se očekávalo. Prokázalo se, že informanti potřebují 16 minut, aby si dostatečně prohlédli instruktážní klip a mohli diskutovat o všech 16 značkách.

Obr. 16: Elicitační obrázky – varující a zákazové značky

3.3.10 Co jsi dělal, když se to stalo?

V tomto úkolu mají informanti říct, co dělali a/nebo co cítili, když se dozvěděli o šokující nebo dojemné události představené v rámci daných témat. Jsou zahrnuty významné historické momenty, např. přistání na Měsíci či pád Berlínské zdi, dále významná fotbalová utkání ve světovém poháru, katastrofy jako zemětřesení či tsunami, nukleární nehoda v Černobylu nebo smrt slavných lidí. Jedno téma je specifické pro neslyšící a to tím, že se týká nečekané smrti Guntera Trubea, známého neslyšícího umělce, tedy události, která šokovala německou komunitu Neslyšících. Pro zlepšení elicitace jsou ke znakovaným instrukcím se zadáním promítány charakteristické obrázky, které mohou evokovat vzpomínky na danou událost.

Cílem je podpořit informanty, aby v monologu vyprávěli osobní zkušenosti a/nebo vedli dialog. Rovněž je zde výmluvně zachyceno to, jak se neslyšící, kteří mají omezený přístup k informacím, dozvídají o zprávách nebo událostech ve světě a jak si pro sebe tyto informace přebírají. Často bylo pozorováno, že neslyšící vzpomínali na televizní zprávy, ze kterých odhadovali, co se asi děje. Výzkumníci nechtějí informanty zahltit, proto si mají vybrat z šesti témat jedno, o kterém hovoří samostatně nebo si mají vybrat dvě témata, o kterých vzájemně debatují.

V pilotní fázi dávali mladší účastníci výzkumu zpětnou vazbu, že se neobjevují pouze současné události, ale také situace, které jsou starší než oni sami, což nehodnotili moc pozitivně. Nakonec bylo rozhodnuto netvořit specifické úkoly pro mladší účastníky, ale používat shodná zadání pro všechny, aby se nepoškodila flexibilita dat. Doba tohoto úkolu je přibližně 20 minut.

3.3.11 Tematické oblasti

Následující elicitační úkol má zahájit mezi respondenty konverzaci o různých tématech. Cílem je získat větší množství výpovědí pro výběr základní slovní zásoby DGS. Témata každodenních konverzací byla rozdělena do 25 okruhů jako je práce a zaměstnání, zdraví, rodina a vztahy, oslava, emoce a pocity, oblečení a móda, partnerství, láska, sexualita, škola a vzdělávání, sport a hry nebo cestování. Každý okruh je prezentován slovním spojením v psané němčině se čtyřmi až osmi fotografiemi nebo kresbami, které slouží k podněcování myšlenkových asociací. Je připraveno 8 různých souborů a každý soubor obsahuje 4 tematické oblasti. Páru informantů je z 8 souborů prezentován pouze jeden, který obsahuje 5 slidů.
Každý slide se věnuje jedné tematické oblasti spolu s ilustracemi k danému tématu a jeden závěrečný slide, který shrnuje ty všechny předchozí. Informanti si následně ze 4 nabízených témat vyberou jen dvě, kterým se dále věnují.

Pokud ovšem nemají o čem mluvit, moderátor začne klást předem připravené otázky, aby zahájil konverzaci. Otázky mohou vypadat následovně – „Co shledáváš dobrého na svém zaměstnání?“ „Je tady nějaké právo, které je zaměřené přímo na neslyšící?“ apod. Pokud zbývá čas, je představeno jedno další téma, které není součástí daného souboru k rozvinutí doplňující diskuze. Tento rozhovor zabere kolem 32 minut.

Obr. 17: Elicitační obrázky k tematické oblasti „zdraví“

3.3.12 Kombinovaný úkol

Toto je kombinovaný úkol, který je vymyšlen tak, že jeden informant má popsat nějakou činnost z paměti a druhý má převyprávět obrázkový příběh. V popisné části je informant požádán, aby si vybral jednu jemu známou aktivitu ze souboru 8 činností, a tu posléze popíše. Každá aktivita sestává ze sledu činností, např. výroba džemu, výměna pneumatiky na autě či zdobení vánočního stromku. Cílem je získat postup popsaný krok po kroku a jeho vysvětlení. Dalším záměrem je elicitace určitých slovních spojení.

Celkově jsou připraveny dva soubory po osmi aktivitách a každý z nich je prezentován na každém setkání, takže je pokryto 16 aktivit. Pokud informanti neznají postup ani jedné z nabízených aktivit, mohou si vybrat jakoukoli jinou, jim známou aktivitu.

Převyprávění obrázkového příběhu vypadá tak, že se informant podívá na obrázkový příběh o průvodci a účastnících zájezdu, kteří musí během cesty překonat několik problémů. V další části informant vidí několik obrázků a má je převyprávět druhému informantovi. Cílem je elicitovat různé způsoby, jak lze využít prostor pro vyjádření směru a mnohosti ve znakovém jazyce. Je tedy vytvořen originální speciální příběh sestávající ze 17 obrázků, protože žádný existující příběh nebyl pro elicitaci vhodný. Natáčení těchto úkolů trvá průměrně 17 minut.

Obr. 18: Ukázka obrázku z elicitačního příběhu o zájezdu

3.3.13 Regionální zvláštnosti

Informanti mají vzájemně konverzovat o zajímavých událostech či produktech v regionu, kde žijí. Proto je třeba, aby oba pocházeli ze stejného regionu a žili tam alespoň 10 let. Témata mohou zahrnovat festivaly v regionu, oblíbené turistické destinace, typické aktivity, slavné osobnosti, tradice a zvyky nebo typické produkty a speciality. Záměrem je sesbírat znaky pro názvy míst, slavných festivalů, regionálních specialit atd. Na tento úkol je vymezeno cca 20 minut.

3.3.14 Převyprávění filmu Sign

Oba informanti mají zhlédnout pětiminutový film a hovořit o něm. Tato instrukce je záměrně nejasná, aby se vyhnulo případnému podsouvání tématu konverzace. Film Sign je speciální tím, že se v něm nemluví. Dva herci komunikují skrze papíry, na kterých jsou napsaná anglická slova. Závěr filmu nechává na divákovi, aby se rozhodl, zda je hlavní herečka neslyšící nebo ne. V tomto úkolu se očekávají znaky vyjadřující lásku. Kvůli ujištění, že všichni obsahu filmu porozumí, jsou k anglickým slovům přidány německé titulky. Tohle je jeden z volitelných úkolů, který nemusí být plněn, pokud není čas a měl by trvat 8 minut.

3.3.15 Nové vs. staré znaky

Informanti jsou tázáni na popis toho, jaké jsou rozdíly mezi znaky mladší a starší generace neslyšících. Jedním z cílů je zachytit sociolingvistické varianty znaků, které nejsou zahrnuty v jiných úkolech. Navzdory užitečnosti tohoto materiálu se výzkumníci rozhodli označit tento úkol jako volitelný, kvůli výsledkům pretestu, ve kterém pozorovali rozpaky mezi respondenty. Stanovená délka této elicitace by měla být cca 7 minut.

V rámci elicitace jazykových dat do korpusu jsou ke konci setkání vloženy dva volitelné úkoly, převyprávění filmu Sign a nové versus staré znaky, avšak pouze v případě, že je zaručen dostatek času.

3.3.16 Události v komunitě Neslyšících

Elicitační setkání končí úkolem, který je speciálně zaměřen na neslyšící. V něm je každý informátor požádán, aby vyprávěl jednu událost spojenou s komunitou Neslyšících, které se zúčastnil. Témata se pohybují od národních událostí, jako jsou kulturní festivaly neslyšících k mezinárodním událostem jako je Deaflympiáda či Deaf Way. Jako pomůcka, aby si neslyšící případně vzpomněli, jsou použity německé názvy spolu s obrazovým materiálem, který jim je prezentován.

Pokud se informátor neúčastnil ani jedné této akce, může si vybrat jinou událost. Cílem je zdokumentovat kulturu Neslyšících a zahrnout tak do korpusu autentické příběhy s následným propojením konverzace. Celkově tento úkol trvá přibližně 21 minut.

Po tomto posledním úkolu setkání končí závěrečnou diskuzí, kdy je od informantů zjišťována zpětná vazba týkající se celkové elicitace.

Obr. 19: Elicitační obrázky, data, místa – události v komunitě Neslyšících (Deaflympiáda)

3.3.17 Dodatečné úkoly

Moderátor může ještě nakonec zařadit dva dodatečné úkoly, pokud není vyčerpán časový plán setkání. Jedním úkolem je převyprávění znakovaného příběhu o požárním alarmu v hotelu a další úkol je založen na popisu cesty, který vychází z obrázku mapy města. Pokud se moderátor rozhodne zařadit jeden nebo oba úkoly, tak jsou vloženy ještě před úkol „události neslyšících“, protože elicitace by vždy měla končit tématem, které je zaměřené na situaci neslyšících. (Hanke, 2010, s. 178−185)

Obr. 20: Elicitační obrázek – mapa města

Celá elicitace dat je tedy v prvé řadě zaměřena na získání jazykového materiálů, na kterém bude možno provádět další lingvistické výzkumy. Zároveň je digitálně zachyceno kulturní dědictví Neslyšících.

3.4 Natáčecí studio

Nahrávání projevu ve znakovém jazyce vždy balancuje mezi dvěma problémy. Na jedné straně se zaměřením na vysokou kvalitu nahrávací techniky a zachyceného záznamu, který je důležitý pro lepší a komplexnější přepisovací proces, umožněný automatickým nahráváním. Na druhé straně je to otázka vytvoření atmosféry, kde by se informanti cítili dost příjemně na to, aby se prostředí, které může být stresující, nepromítlo do jejich znakování.

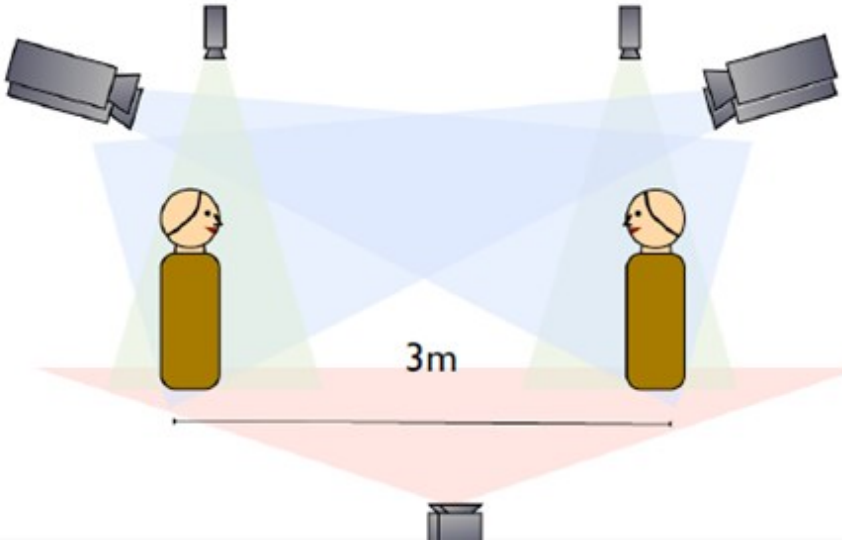
Ve studiu jsou respondenti většinou natáčeni před modrým pozadím[61] bez dalších rušivých objektů. V případě projektu korpusu DGS se objevilo další omezení, totiž to, že studio bylo nutné během 3 let dvanáctkrát přemisťovat, kvůli nahrávání v jednotlivých regionech. Není totiž žádoucí, aby se respondenti sjížděli na jedno místo, kvůli případnému mísení dialektů. Tudíž bylo vytvořeno studio, které bylo praktické a mobilní. (Hanke, 2010, s. 106–109)

[61] Zveřejněný subkorpus DGS: http://meine-dgs.de

3.4.1 Celkové uspořádání studia

Během nahrávání sedí pár informantů čelem k sobě, v přibližné vzdálenosti tři metry. Ve studiu je umístěno 8 kamer. Materiály související se zadáním úkolů jsou promítány pro každého informanta zvlášť na dvou obrazovkách, které jsou umístěny mezi nimi, situovány blízko zemi, aby se zamezilo tomu, že by jim obrazovky překážely ve výhledu na toho druhého. Moderátor sedí mezi informanty, ale více v pozadí a zadává úkoly a pozoruje konverzaci. Zásah do konverzace se děje velmi výjimečně, pouze pokud je to nezbytně nutné. Moderátor má před sebou také jeden monitor, na který ale ani jeden respondent nevidí, kvůli tomu, že se mu tam zobrazují informace v rámci programu session director. (Hanke, 2010, s. 106–109)

Obr. 21: Pozice kamer a respondentů ve studiu



3.4.2 Pozice kamer

Celkový počet kamer ve studiu je 8, tři mířící na jednoho informanta a dvě snímající celou scénu včetně moderátora. Dvě HD kamery poskytují čelní pohled na každého informanta, dvě kamery umístěné nad informanty umožňují pohled shora, což pomáhá určit např. některé tvary ruky, které nejsou zepředu vidět.

Další dvě kamery jsou umístěny z boku a zachycují tak další úhel znakovaného projevu a poskytují tedy záběry, které umožnují obrazovou analýzu k rekonstrukci 3D informací a pomáhají automatickému zpracování projevu. Poslední dvě kamery poskytují také HD kvalitu obrazu a zachycují celou scénu, tzn. oba informanty a také moderátora, který interaguje s oběma informanty.

Přepisovatel tak má zběžný přehled o celkové komunikaci a to mu pomáhá přesně identifikovat interakci mezi všemi třemi účastníky. Původním záměrem bylo pracovat pouze s pěti kamerami, což se ukázalo jako nedostačující, kvůli nezbytnosti zachytit všechny účastníky natáčení v potřebných úhlech. (Rathmann, 2016)

Obr. 22: Záběry na respondenty umožněné různými pozicemi kamer

3.4.3 Technické detaily

Pro mobilní studio je nutné zajistit místnost velkou nejméně 5 x 5 metrů a optimálně s výškou stropu tři metry, a ještě jednu malou místnost vedle této, aby bylo možné nastavit všechny kamery a další potřebné technické vybavení. Taktéž se tam musí vejít bez větších problémů všechny osoby participující na natáčení. (Hanke, 2010, s. 106–109)

3.5 Překlad, segmentace, přepis

Po získání jazykových dat je nutné, aby korpusová data prošla mnoha různými anotacemi a přepisovacími procesy, jejichž cílem je identifikovat znaky a zdokumentovat jejich vlastnosti. Tento proces je časově velmi náročný. Po překladu projevů z německého znakového jazyka do psané němčiny vznikne základní přepis, který slouží k segmentaci těchto promluv a identifikaci lexikálních jednotek. Toto všechno poskytuje prvotní přístup k datům. Předpokládá se, že přibližně 50 % přepisů bude ještě znovu přepsáno do mnohem větších detailů.

3.5.1 Překlad a segmentace

Úplně prvním krokem je elicitovaný materiál v německém znakovém jazyce přeložit do německého jazyka zkušenými tlumočníky znakového jazyka a synchronizovat to s projevy v DGS. Obsah rozhovorů je tedy zachycen a stává se vyhledatelným skrze psanou němčinu. Časové zarovnání také poskytuje první přibližnou segmentaci znakového projevu do smysluplných úseků. Mimo to pasáže, které jsou nějakým způsobem speciální s ohledem k lingvistickým nebo obsahovým záležitostem, jsou označeny a zdokumentovány ve zprávě od tlumočníků. Tyto poznámky, mimo další jiná kritéria, jsou zohledněny při přidělování priorit pro další analýzy na určitých úsecích. Samozřejmě jsou vzaty v potaz při výběru částí korpusu, které budou detailněji přepsány a případně publikovány. (Prillwitz, 2008, s. 159−164)

Segmentace je vlastně rozhodování, kde znak ve znakovém projevu začíná a kde končí. Výzkumní pracovníci měli jasně vymezené, jak znaky segmentovat. Bylo dohodnuto, že se přechodové znaky nepočítají jako součást znaku. Obvykle jsou totiž mezi dvěma znaky pauzy, během nichž se artikulátory pohybují z konce jednoho znaku na začátek dalšího znaku. Navíc se běžné znakování pohybuje ve frekvenci 500 snímků za sekundu a při přechodových znacích je frekvence kolem 50 snímků za sekundu, takže je možné, podle zmíněných rychlostí, poměrně přesně určit, kde znak začíná a končí. (Rathmann, 2016)

3.5.2 Základní přepis

Přepis prováděli studenti, kteří byli zároveň výzkumnými asistenty. Na jejich práci dohlíželi a kontrolovali je rodilí mluvčí německého znakového jazyka, aby byla jistota ve správnosti a přesnosti přepisu. (Rathmann, 2016)

Základní přepis slouží k segmentování a identifikaci lexikálních jednotek. Poskytuje tak první a snadný přístup k jednotlivým znakům. Předešlá segmentace vyplývající z překladu je revidována a vylepšena z hlediska znakového jazyka. Přepis se provádí v souladu s pokyny a kritérii, která jsou uvedena v manuálu, který mají přepisovatelé k dispozici.

Tokeny jsou manuálně přiřazeny ke znakům, které jsou shromážděny v lexikální databázi, která je součástí přepisovacího prostředí iLexu (o tomto softwaru detailněji viz níže). Každý znak je označen jedinečnou glosou a jeho forma je popsána pomocí zápisu v HamNoSys. V této fázi přepisu jsou tokeny produktivních znaků, stejně jako jinak specifických znaků, např. ukazovací znaky, prstová abeceda apod., připojeny k větší skupině a kódovány pouze na velmi obecné rovině. (Prillwitz, 2008, s. 159−164)

3.5.2.1 Výběr lemmatu a správa detailního přepisu

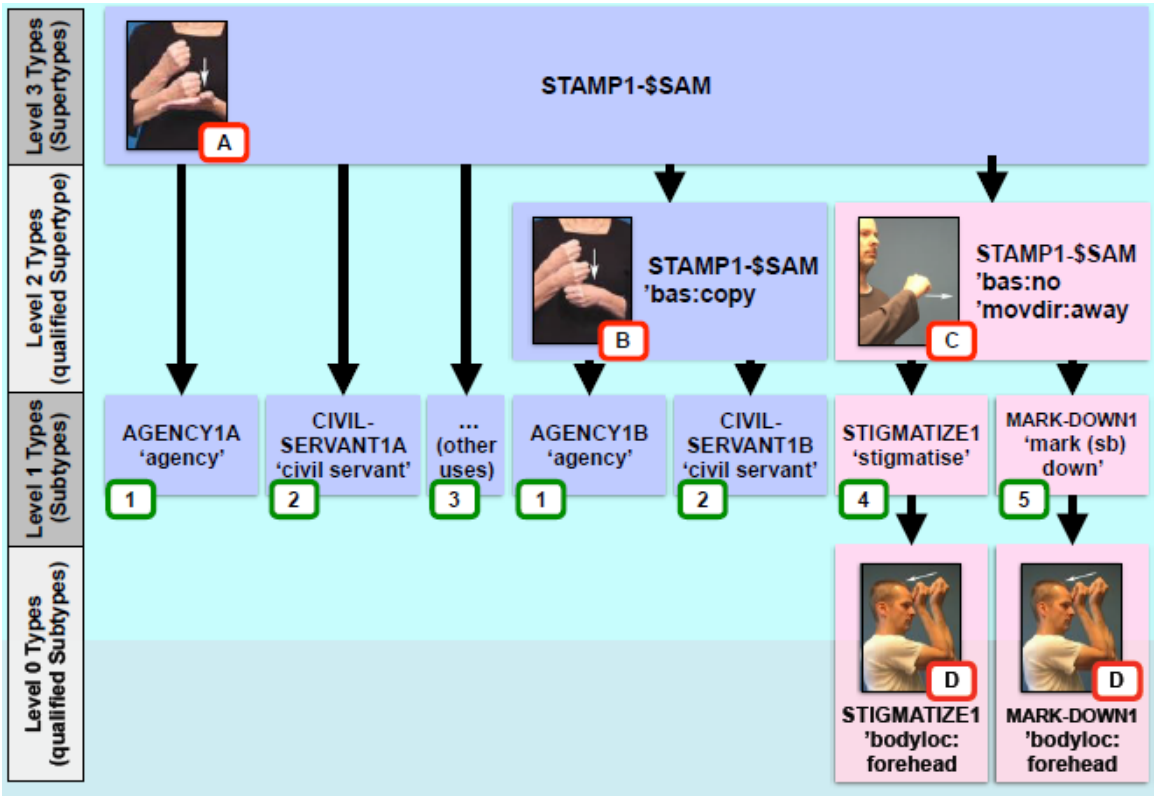
Po základním přepisu jsou data přístupná už i z pohledu tokenů. To umožňuje výběr a přiřazení lemmat. Následně jsou jednotlivé tokeny vybrány pro detailnější přepis. Jedním z hlavních kritérií pro zařazení znaku do slovníku je frekvence jeho výskytu a rozšíření mezi informanty. Pro každý vybraný znak tedy lingvista přezkoumá všechny výskyty a rozhodne, zda všechny nebo pouze některé vybrané tokeny musí být přepsány s většími detaily. Další vlivy ovlivňující volbu detailního přepisu je také vhodnost použití znaků v příkladech v překladovém slovníku a případy gramatických jevů, které mohou sloužit jako základ pro sestavení gramatické příručky ve slovníku. Detailní přepis umožňuje bližší pohled na znak a opodstatněnou analýzu gramatických jevů, která je více než nutná.

3.5.2.1.1 Detailní přepis

Ve druhé fázi přepisu jsou vybrané výskyty lemmat znovu detailněji přepsány i z pohledu kontextu. Očekává se, že přibližně polovina základních přepisů bude zpřesněna tímto detailním přepisem. Anotace a přepis v korpusu budou úzce spjaty s požadavky lexikální databáze potřebné pro produkci slovníku. (Prillwitz, 2008, s. 159−164)

Anotace se zaměřuje především na formu znaku, tedy na formu, která může zahrnovat fonologické varianty znaku. Příkladem může být znak STAMP (razítko), jež pojímá dvě fonologické varianty a dvě modifikované formy znaku. (Langer, Troelsgard, 2016, s. 144)

Obr. 23: Příklad znaku STAMP, který zahrnuje fonologickou variantu znaku A, B a modifikovanou formu znaku C, D



Nejen daný token, ale také okolní znaky, jsou přepsány tak, aby bylo možné zachytit kontextuální výskyt znaku, kolokace a další relevantní informace. Jsou tedy kódovány následující aspekty:

- mluvní a orální komponenty;
- mimika;
- klasifikace formy tokenu;
- kontextový význam tokenu;
- syntaktické kategorie znaku;
- hledisko využití prostoru pro ustanovení prostoru, umístění objektů ve znakovacím prostoru a vracení se k již určeným místům;
- produktivní jevy pro vizualizaci objektů a procesů.

Jak již bylo zmíněno, během detailního přepisu jsou promluvy segmentovány do menších jednotek či frází. Zajímavé pasáže, které se týkají obsahu nebo jazyka samotného jsou znovu zapsány v poznámkách. Kratší pasážím lze porozumět i bez kontextu a jsou značeny jako potenciální příklady vět pro budoucí použití ve slovníku.

Znaky jsou dále rozlišeny, tudíž jsou jejich tokeny podrobněji popsány a klasifikovány, například do:

- fonologických variant;
- forem, které jsou výsledkem gramatických procesů, jako je směrová orientace znaku, inkorporace množného čísla apod.;
- forem, jejichž výsledkem je modifikace znaku, např. v rámci metaforického používání znaku.

3.5.3 Přepisovací tým

V této části by bylo vhodné zařadit informace o přepisovacím týmu a jeho fungování. Otevření přístupu k profesionálnímu přepisu studentům na IDGS v tak velkém měřítku dříve nebylo realizováno. Základní i detailní přepis většinou provádí studenti. Studenti jsou školeni a neustále kontrolováni zkušenými neslyšícími rodilými mluvčími znakového jazyka, proto má tento proces několik výhod. Při takovém postupu je přepis neustále kontrolován a ověřován týmem zkušených rodilých znakujících, tzn. přepisy vidí nejméně dvě osoby, které pracují nezávisle na sobě. Tím je zaručena vysoká kvalita přepisu bez zdvojnásobení nákladů na něj. Studentům se tak nabízí prvotní přístup do korpusové lingvistiky a vhled do praktické lexikografie znakového jazyka. Kromě toho umožňuje kombinace zkušených neslyšících přepisovatelů a studentů přepsat větší množství dat v rámci jisté časové jednotky.

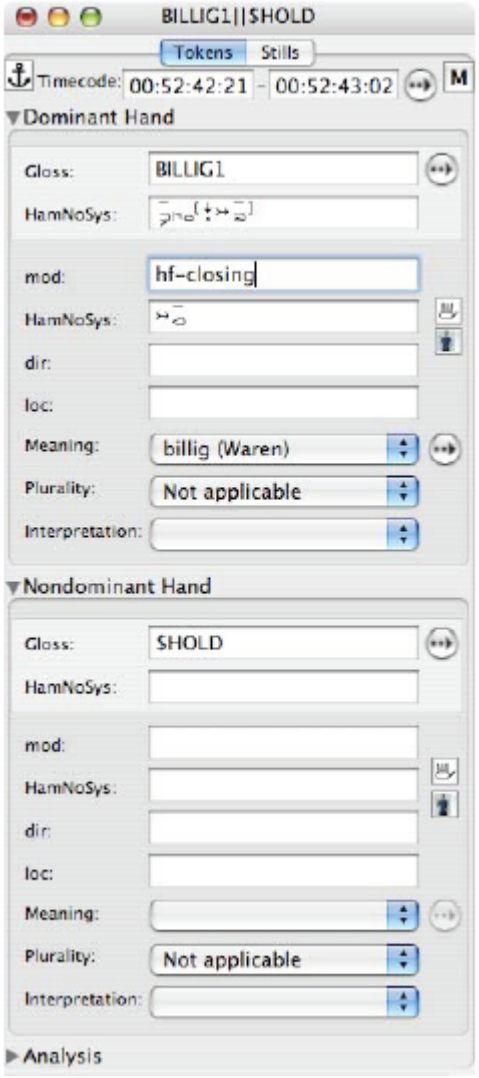
Použitím softwaru iLex je možné modelovat celý transkripční proces a především podporovat konzistenci párování token – lemma. Přepisovatelé také spolu mohou komunikovat v rámci přepisovacího prostředí skrze videochat a webové technologie 2.0, což přispívá ke stabilnímu toku informací mezi širokou skupinou přepisovatelů. (Prillwitz, 2008, s. 159−164)

3.6 iLex

iLex[62] je transkripční a anotační nástroj, který umožňuje práci s několika synchronizovanými videostreamy a dává uživateli možnost přepínat mezi různými pohledy (Rathmann, 2016). iLex dokáže převést zadaná tabulková data do různých typů grafů. Uživatel si tedy může zvolit body, ze kterých si chce vytvořit graf a přitom není třeba, aby určitá data někam kopíroval (Hanke, 2016, s. 89).

[62] iLex − integrated lexicon

Obr. 24: Okno v iLexu zaznamenávající token a příslušné informace

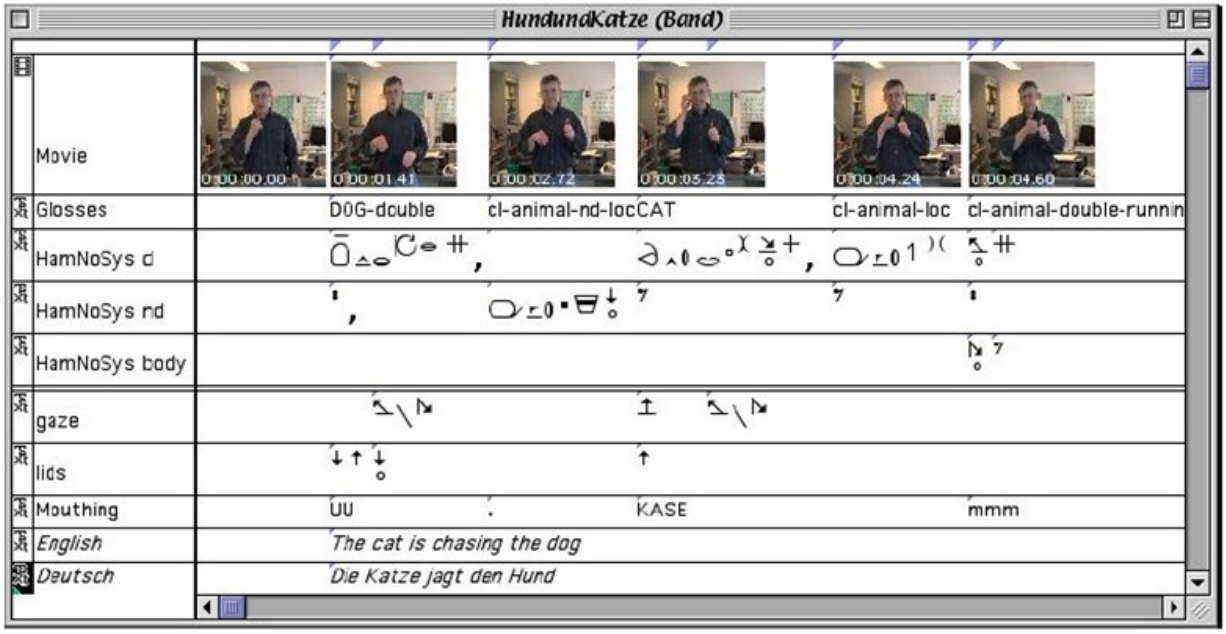


3.6.1 Vznik iLexu

V 90. letech byl prvním pokusem přepisovací nástroj syncWRITER,[63] který dovoloval uživateli propojit digitální videosekvenci se specifickými částmi přepisu. Nevýhodou tohoto nástroje bylo, že byl zaměřen především na prezentaci přepisu v graficky atraktivním provedení, ale nebyl stejně dobře vybavený pro jakoukoli analýzu projevu nebo lexikografický účel.

[63] syncWRITER – interlineární textový editor

Obr. 25: Okno textového editoru syncWRITER

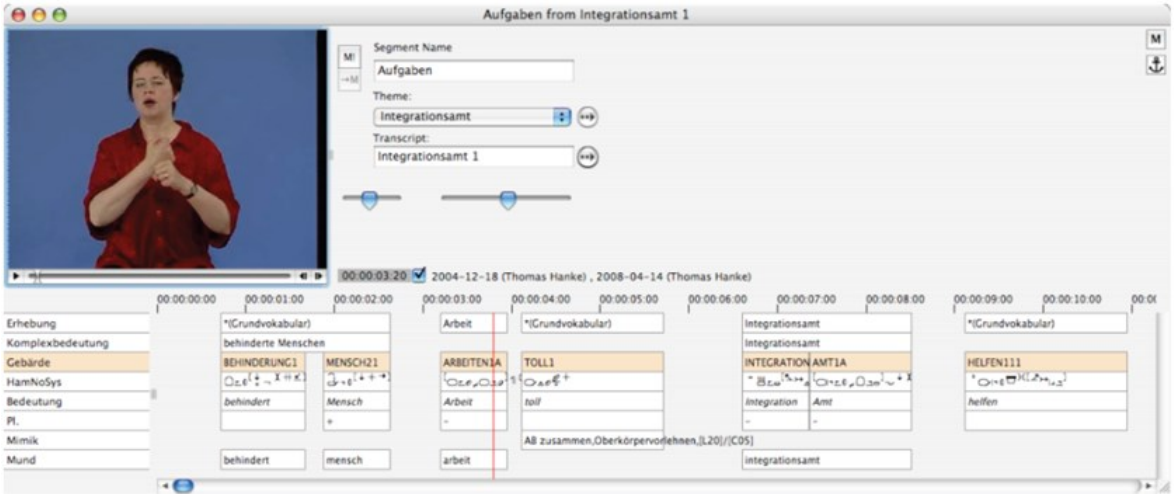


iLex tak kombinuje dva přístupy, je to přepisovací databáze pro znakový jazyk kombinovaná s lexikální databází. V iLexu se přepisy neskládají ze sekvence glos zapsaných a časově synchronizovaných s videonahrávkou, ale z tokenů. Kvůli neexistenci psané formy znakových jazyků není přiřazování přepsaných tokenů k příslušným znakům úplně jednoduchý proces, jako např. u mluvených jazyků, které mají mluvenou podobu s příslušnou psanou formou, takže je vyžadována přepisovatelova plná pozornost při propojování tokenu s daným znakem. (Hanke, 2008b, s. 1–3)

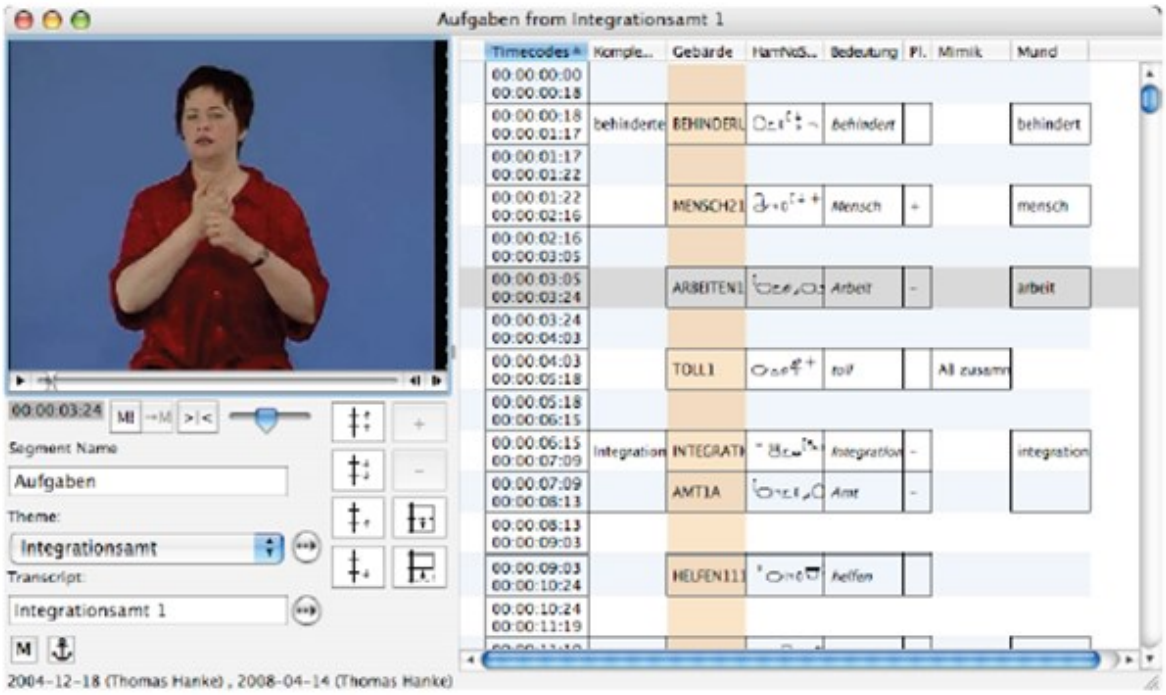
3.6.2 Zobrazení času v iLexu

iLex nabízí horizontální zobrazení přepsaných dat známá těm, kdo používá jiná přepisovací prostředí. Čas zarovnává zleva doprava a délka zobrazení tagů je úměrná jeho trvání. Tento pohled je doplněn vertikálním zobrazením, kde je čas zarovnán shora dolů. (Hanke, 2008, s. 64–67)

Obr. 26: Horizontální zarovnání času v iLexu



Obr. 27: Vertikální zarovnání času v iLexu



3.6.3 Importování dat z jiných přepisovacích systémů

Importování přepisů z jiných zdrojů, jako je ELAN, syncWRITER nebo SignStream je realizováno skrze ovládací menu v iLexu. Výsledky tohoto importovacího procesu jsou pouhé úrovně přepsaných textů a v dalším kroku je nezbytné převést textové úrovně popisující tokeny, ve většině případů to znamená glosy, k reálným úrovním tokenů.

V importovaných datech může dojít k nesrovnalostem, proto je vyžadována plná pozornost přepisovatele. iLex toto umožňuje skrze mapovací tabulku, kde jsou umístěny externí glosy a užitím jisté modifikace v programu iLexu. (Hanke, 2008b, s. 1–3)

3.6.4 Spolupráce přepisovatelů na projektu

Používání centrální databáze, kde je seznam lidi pracujících na projektu nebo alespoň na několika projektech v rámci jedné instituce, umožňuje účinnou kooperativní práci. Všichni spolupracují na tvorbě přepisů, tokenů apod. a je tak jednodušší sdílet informace o datech, když o sobě navzájem ví. Výsledek introspekce může být rychleji přístupný ostatním uživatelům skrze webkameru. Integrace kamery do programu tedy umožňuje sdílet znakované příklady bez nutnosti starat se o technické aspekty jako je videokomprese a příslušná metadata pro nový videomateriál jsou automaticky přidána do databáze.

Na všechna data je možné odkazovat pomocí URL.[64] Jednoduchým přetažením lze data z iLexu přesunout do Wiki nebo blogu, kde je vloženo URL a kdokoliv s přístupem do iLexu databáze může zobrazit data, která jsou diskutována jednoduchým kliknutím na tento link. (Hanke, 2008, s. 64–67)

Nevýhodou této spolupráce je samozřejmě to, že je nutné, aby se dodržovaly určité přepisovací konvence. Zatímco některé aspekty přepisovacího procesu mohou být individualizovány, jiná data potřebují být přístupná a tudíž srozumitelná všem uživatelům, rozšíření takové spolupráce proto musí proběhnout shodným způsobem. (Hanke, 2008b, s. 1–3)

Zkušenosti ukazují, že je zapotřebí několika schůzek se všemi přepisovateli, pokud se začíná pracovat na novém projektu a hlavně v případě dosud neřešeného úkolu, který se významně liší od ostatních projektů, aby všichni pochopili požadovaný průběh práce (Hanke, 2008, s. 64–67).

3.6.5 Technické zázemí

Název iLex se používá pro přepisovací databázi, ale také pro užívanou aplikaci, díky které lze vstoupit do databáze. Tato databáze je umístěna na virtuálním databázovém serveru. Vzhledem k SQL[65] databázi byl vybrán PostreSQL otevřený databázový serverový systém, který může být nainstalován na širokou paletu platforem.

[64] URL = Uniform Resource Locator (jednotná adresa zdroje)
[65] SQL = Structured Query Language, standardizovaný strukturovaný dotazovací jazyk, který je používán pro práci s daty v relačních databázích

Tento systém je spolehlivý a má zabudované dobré bezpečnostní mechanismy, také je podporován aktivní komunitou uživatelů a nabízí několik realizačních aspektů, které jsou v daném kontextu výhodné. Jednou z těchto výhod je například serverové zahrnutí skriptovacího jazyka, jakým je třeba Perl, který je využit pro produkci budoucího překladového slovníku.

Filmy, obrázky a ilustrace nejsou uchovávány v databázi a je na ně pouze odkazováno. Mohou být uloženy na uživatelově počítači nebo na centrálním souborovém serveru. Tento hybridní koncept uložení umožňuje uživateli pracovat z domova. Přístup do databáze mu je umožněn skrze zabezpečení virtuální soukromé sítě, což je pro uživatele výhodné, protože má přístup k videonahrávkám bez nutnosti přihlášení.

Software je volně přístupný pro MacOS X a stejně tak i pro Windows XP s německým a anglickým uživatelským rozhraním. Zdrojový kód je volně přístupný, kromě několika mála funkcí, které jsou použity z komerčně dostupných zásuvných modulů namísto zařazení vlastní služby. Software je možné nainstalovat i na notebook, avšak není-li na tomto počítači dostatek RAM paměti, sníží se rychlost zpracování ve srovnání se standardním zpracováním. (Hanke, 2008, s. 64–67)

3.6.6 Produkce slovníku

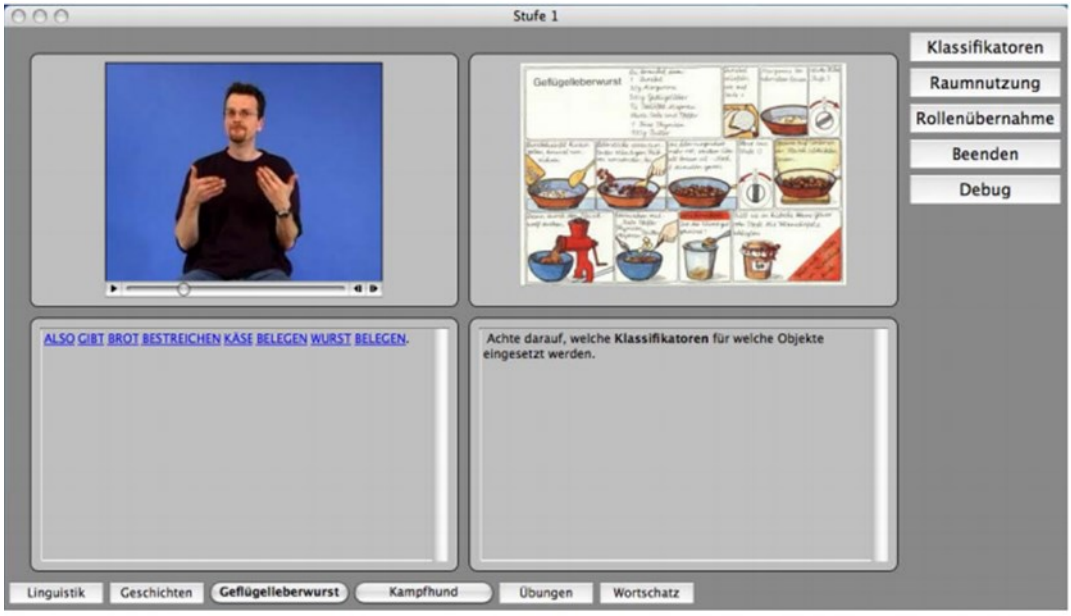
Všechna data potřebná k vytvoření slovníku jsou uchovávána v databázi jako výsledek přepisovacího procesu nebo pozdějších analyzujících kroků. To umožňuje automatickou produkci slovníku v rozumném čase. Pro tento účel je využíván Perl, což je internetový programovací jazyk, propojený s databází Adobe Indesign pro úpravu uveřejněného produktu a HTML šablonou k produkci webových aplikací. Pouhou změnou šablony nebo přidáním dalšího souboru lze kompletně změnit vzhled slovníku a reprodukovat tištěnou a online verzi během několika hodin. V současné době se vytváří další sada šablon pro optimalizování HTML výstupu pro iPhony a iPody, které slibují, že se stanou vhodnou platformou pro slovníky.

3.6.7 Produkce výukového materiálu

V minulosti byly vyprodukovány vysoce kvalitní výukové CD-ROMy znakového jazyka, které byly individuálně programovány, avšak vznikla zde potřeba jednoduše a snadno produkovatelného materiálu pro každodenní výuku. Ideálně něco, co by lektoři mohli používat a částečně přizpůsobovat dle svých potřeb sami.

Nejčastěji vytvářené jsou e-learningové materiály znakového jazyka, což jsou videonahrávky s přiřazenými vysvětlivkami a odkazy, např. do slovníku. Odkazy do slovníku a slovník sám o sobě může být produkován skoro bez další manuální intervence. Modul přehrávače samozřejmě pracuje samostatně a nepožaduje přístup do databáze iLexu. (Hanke, 2008b, s. 1–3)

Obr. 28: E-learningový materiál znakového jazyka – videonahrávka s odkazy



3.7. Překladový slovník DGS – německého jazyka

3.7.1. Analýza a kompozice slovníkových hesel

Slovník bude kompletně založen na korpusu německého znakového jazyka s ohledem na znaky, které budou zahrnuty jako lemmata. Většina příkladů použití znaků bude převzata přímo z korpusu. Nicméně informace uvedené u slovníkových hesel budou přesahovat příklady v korpusu. Data se budou systematicky excerpovat z výskytů získaných obecným popisem lexikálních jednotek. Tento popis bude zahrnovat následující aspekty:

- forma znaku (citátová forma);
- fonologické varianty;
- využití prostoru, morfologie znaku (např. plurální formy);
- syntaktická funkce;
- význam, možné překlady do němčiny;
- informace o dialektu;

- odkazy vztahující se ke stejným znakům, synonymům, antonymům;
- příklady ilustrující gramatiku, užití a různost významu znaku − příklady vzaté z korpusu nebo vymyšlené příklady, potvrzené rodilými mluvčími znakového jazyka.

Po první předběžné analýze se určí, které dodatečné informace jsou potřebné pro konečnou analýzu a složení záznamu. Jakmile bude detailní přepis požadovaných úseků dokončen, mohou být velké objemy dat porovnány a zpracovány jako slovníková hesla. Popsána a analyzována nebude pouze struktura slovní zásoby, ale také gramatické jevy, které budou sledovány více podrobně za účelem sestavení gramatiky použitelné ve slovníku, která bude založená na korpusu.

3.7.2. Elektronický slovník založený na korpusových datech

Slovník bude prvním obsáhlým a komplexním na korpusu založeným slovníkem německého znakového jazyka. Bude publikován v elektronické podobě a primárně bude sloužit těmto cílovým skupinám:

- studenti DGS, kteří jsou rodilými mluvčími němčiny, také slyšící lidé setkávající se s neslyšícími lidmi v rámci své práce, rodiče vychovávající neslyšící děti, studenti Deaf studies/tlumočení znakového jazyka a osoby se sluchovým postižením nebo lidé ohluchlí, pro které je DGS druhým jazykem;
- profesionální tlumočníci německého znakového jazyka do německého jazyka a naopak;
- rodilí mluvčí DGS − dospělí neslyšící, CODA;
- neslyšící děti nebo žáci, osvojující si DGS jako mateřský jazyk;
- učitelé znakového jazyka, lingvisté a každý, kdo se zabývá strukturou znakového jazyka na praktické či teoretické úrovni.

Aby slovník mohl sloužit různorodým skupinám, musí kombinovat různé typy funkcí, které umožňují rozmanité druhy použití. Ve slovníku vedle lemmatu nalezneme příkladové věty, aby bylo jasné užití a význam daného znaku. Německá část slovníku poskytne přístup k DGS pro slyšící uživatele skrze jejich rodný jazyk. Avšak i neslyšící uživatelé dostanou základní informace o německých slovech a příkladové věty k představě jejich užití. Je předpokládáno, že slovník bude obsahovat kolem 6 000 znakových hesel.

3.7.2.1 Gramatika použitá ve slovníku

Důležitou součástí komplexního slovníku je gramatická příručka. Odkaz do gramatické příručky činí hesla kratší, více kompaktní a z tohoto důvodu i jasnější. Budoucí uživatelé mohou těžit z toho, že gramatika je psána v cílovém jazyce, tzn. německém jazyce, ve kterém snadno porozumí termínům. Gramatická příručka bude taktéž založena na korpusu. Skutečnosti, které nemohou být rozhodnuty na základě pohledu na korpusová data, budou předána rodilým mluvčím DGS nebo by byla provedena dodatečná elicitace, aby se rozhodlo o jejich platnosti. Příručka by tedy měla obsahovat obecný přehled nejdůležitějších gramatických jevů DGS, doplněných příklady vzatých z korpusu.

3.7.2.2 Anotovaný korpus

V souladu s korpusově − lingvistickými cíli, bude veřejnosti online zpřístupněno kolem 50 hodin reprezentativního výběru korpusového materiálu. Aby byla data přístupná výzkumníkům, kteří nerozumí němčině, bude právě tato část korpusu přeložena do angličtiny. Všechny anotované znaky budou navíc obohaceny anglickými glosami, protože iLex poskytuje formát vhodný pro vzájemnou výměnu přepisu a metadat bez jakékoliv ztráty informací. V rámci kooperace bude kompletní korpus přístupný pro lingvisty a doktorandy, výměnou za dodatečné anotace, které mohou být využity v projektu. (Prillwitz, 2008, s. 159–164) Vydání slovníku je plánováno na rok 2020. V roce 2011 byl proveden výzkum mezi potenciálními uživateli, aby zjistili jejich potřeby a přání týkající se tohoto slovníku (Dictionary DGS, 2016).

4. Návrh korpusu českého znakového jazyka

V této kapitole bych ráda přednesla svůj návrh na vytvoření korpusu českého znakového jazyka, protože tak, jako v ostatních případech, i čeští odborníci na lingvistiku znakového jazyka, a především česká komunita Neslyšících by měla mít možnost přístupu k informacím, týkajících se českého znakového jazyka. O přínosu případně vytvořeného korpusu a jeho využití není pochyb, jak již bylo zmíněno ve dvou předešlých případech.

Přípravné práce na korpusu jsou jedny z nejdůležitějších, proto je třeba promyslet, jakou by měl mít strukturu, kolik respondentů bude možné zapojit, z kterých oblastí či jak je do projektu získat. Dále také, kde se bude natáčet, jaké použít elicitační materiály a který program na zpracování získaných jazykových dat zvolit. Zda navázat spolupráci s jinými státy, kde už podobné projekty probíhají apod.

Neméně důležitou otázkou je financování celého projektu. Korpus českého znakového jazyka by mohl být financován skrze projekty vyhlašované Ministerstvem školství, mládeže a tělovýchovy České republiky. Jedním z takových projektů by mohl být např. operační program Výzkum, vývoj a vzdělávání, který se mimo jiné zaměřuje na produkci kvalitních výsledků výzkumu, podporu kvalitního vzdělávacího systému a rozvoj jazykového vybavení, což jsou všechno aspekty, které by korpus potenciálně rozvíjel. (Dotační.info) Zřejmě by šly využít i dotační fondy Evropské unie, které by byly orientovány na tvorbu teoretických studií, výuku cizích jazyků, počítačové aplikace či překladatelství.

Je nezbytné stanovit, která metadata by se o respondentech zjišťovala a shromažďovala. Metadata by se mohla soustředit na následující informace: pohlaví, věk, vzdělání, velikost ztráty sluchu, příp. kompenzační pomůcky, mateřský jazyk a preferovaný jazyk respondenta v komunikaci. Všechna data jsou přínosná svým zaměřením na zjišťování podoby znakového jazyka.

Nejdříve je tedy nutné určit, kolik respondentů by mělo být zapojeno do elicitace jazykových dat a je také nezbytné myslet na možnosti, zda je reálné daného počtu, vzhledem k populaci neslyšících, dosáhnout. Předešlé korpusy měly počty respondentů pohybujících se v řádech dvou set až tří set neslyšících osob. Obávám se, že v českém prostředí je tento počet nedosažitelný, ať už vzhledem k počtu neslyšících, které by bylo nutné získat do daného výzkumu, což by asi nebylo úplně jednoduché vzhledem k tomu, že žádný podobný projekt v komunitě českých neslyšících ještě neprobíhal a tudíž by nejspíše nebudil moc velký zájem, ale také z důvodu splnění podmínek, nutných pro zachování diverzity a reprezentativnosti korpusu.

Reálné číslo, kterého by podle mého názoru bylo možné dosáhnout, se pohybuje kolem sta respondentů. Z věkového hlediska bych ponechala podobné členění, které používali v Německu a to kvůli rozumnému rozdělení do skupin, které definuje skupinu v jistém životním období. Samozřejmostí by byla vyrovnanost zastoupení žen a mužů.

Realizovala bych nápad, který použili v Německu, a to, že si určili kontaktní osoby, které poté navrhly možné potenciální respondenty v jednotlivých regionech. Tyto kontaktní osoby by mohly pocházet z jednotlivých organizací neslyšících, např. Česká unie neslyšících, která má své pobočky v Praze, Brně i Ostravě či organizací věnujících se výuce znakového jazyka, jako je např. pražská Pevnost, či brněnský Trojrozměr. Touto cestou by bylo pokryto velké území republiky. Je známo, že se většina lidí z komunity neslyšících zná, a proto by bylo přes tyto osoby snadnější získat kontakt na další neslyšící osoby, které by fungovaly jako respondenti do projektu. Tyto potenciální kontaktní osoby mají jisté edukační zázemí, aby mohly být lektory v daných institucích, takže by bylo tyto osoby snadnější vyškolit jako případné moderátory pro elicitační část projektu. Jak jsem již zmiňovala, rozdělení republiky na tři části – Čechy, Moravu a Slezsko a s nimi související města, tedy Praha, Brno a Ostrava je podle mého názoru dobrou volbou, z toho důvodu, že je zde kumulováno nejvíce neslyšících osob a také vzhledem k jejich geografické poloze jsou pro případné respondenty velmi dobře dostupné. Možná by bylo dobré Čechy rozdělit na menší územní celky, např. na jihočeskou část spojenou s Českými Budějovicemi, protože by bylo přeci jen pro respondenty z této části republiky obtížnější přijíždět až do Prahy. S tím se pojí i problematika natáčecího studia, které bych také pojala jako mobilní a potřebná technika by se tedy převážela z jednoho místa na druhé, vzhledem k nízkému počtu natáčecích míst.

V rámci elicitačních materiálů bych částečně převzala některé úkoly z již vzniklých a použitých materiálů, kvůli srovnatelnosti výzkumů v budoucnu, nicméně bych také zařadila nové úkoly, které by se více hodily do českého prostředí. Nejvíce bych čerpala z německých materiálů, protože se mi líbí jejich podklady i záměry z nich vyplývající. Zároveň tyto materiály byly použity i při tvorbě korpusu polského znakového jazyka, takže by se vycházelo z podkladů, které úspěšně využili oba naši zahraniční sousedé. Navrhla jsem zapojení 13 úkolů, které jsou zaměřené na získání jazykového materiálu, ale také na zachycení kultury Neslyšících. Setkání je rozděleno na tři bloky – ranní blok, ve kterém se řeší čtyři úkoly, které jsou spíše jednoduššího charakteru a zaměřené na uvolnění atmosféry mezi respondenty a vžití se do situace ve studiu, kde budou následující hodiny společně kooperovat.

Následně je naplánována přestávka a po ní následuje polední blok čítající pět úkolů, které jsou již více orientovány na získání lingvisticky významných informací. Poslední čtyři elicitační

úkoly jsou složeny tak, aby se střídal úkol složitější se snadnějším, z důvodu udržení pozornosti respondentů ke konci setkání.

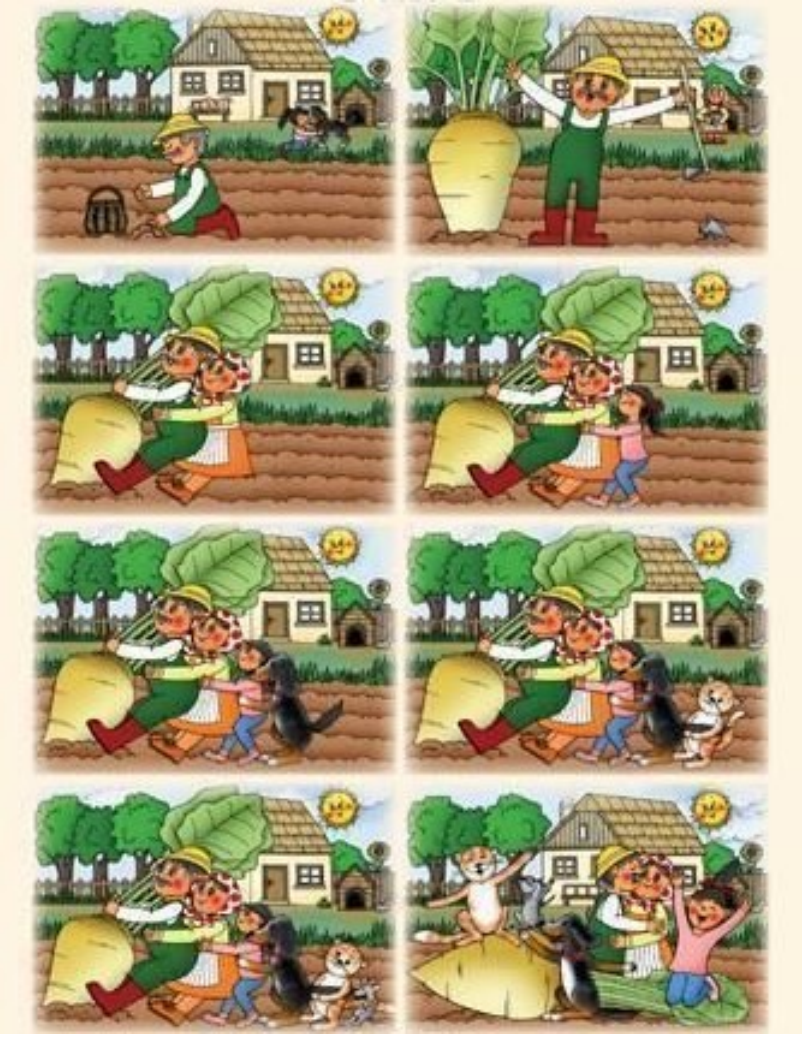
Jistě by bylo nutné provést pilotáž elicitačních úkolů na menším vzorku zkoumaných osob, zda jsou úkoly vhodně a srozumitelně formulovány. Taktéž by se vyzkoušela kvalita nahrávací techniky a připravenost moderátora na vedení setkání. Samozřejmě nesmí chybět vytvoření informovaného souhlasu a jeho představení respondentům.

Ve studiu jsou přítomni dva respondenti a jeden moderátor. Na úplném začátku moderátor seznámí respondenty s alespoň hrubým průběhem a časovým harmonogramem setkání. Následně se již zahájí natáčení. V rámci úkolů buď jednotlivci sami produkují jazykový projev, nebo vedou společnou konverzaci. Elicitační materiály pro respondenty jsou promítány na jednotlivých počítačových obrazovkách umístěnými před nimi, stejně jako v případě natáčecího studia v Německu.

Dopolední blok začíná představením respondentů, kteří užijí svůj jmenný znak a následně vysvětlí, jeho původ a vysvětlí, co znamená. Dále přejdou k vyprávění vtipu, který již mají předpřipravený, stejně jako v německém projektu. Následuje vyprávění zkušeností neslyšících osob, týkajících se jejich života, např. vyprávění o školách, kam chodili, zda bydleli na internátu nebo také to, jakou mají nyní práci. To vše jsou informace velmi cenné, vzhledem k historii vzdělávání neslyšících v České republice. Posledním úkolem v tomto bloku je první společná interakce, kde respondenti dostanou již předem vyplněnou stránku kalendáře s různými schůzkami a na nich je, aby se domluvili, kdy a kde se setkají. Po této části následuje přestávka.

Druhá část obsahuje pět úkolů. Začíná se s obrázkovou pohádkou O veliké řepě, která se respondentům promítá na obrazovkách, a respondenti ji mají převyprávět.

Obr. 29: Elicitační obrázky pohádky O veliké řepě



Dále se přejde k elicitaci jednotlivých znaků, které jsou zaměřeny na získání znaků, u kterých je předpoklad, že se jejich forma liší z pohledu genderových i regionálních rozdílů znakujících, např. znak VÁNOCE. Byl by předem připravený seznam glos, na které by se přímo dotazovalo respondentů, avšak byl by ponechán prostor, kdy by sami respondenti mohli uvádět znaky, které by považovali za odlišné. Následuje promítání příkazových a zákazových značek, aby se stejně jako v německém případu získaly znaky s negativní konotací. Další úkol je zaměřen na tematické oblasti, v rámci nichž se respondenti baví na vybrané téma. Posledním úkolem v tomto poledním celku je přehrání krátkometrážního snímku The Elevator (Výtah), po jehož zhlédnutí ho respondenti převypráví. Tento film nepoužili v žádném projektu, ale já bych ho zařadila, protože se v tomto snímku nehovoří a je vystavěn pouze na pohybu hlavního herce.

Navíc by se, podle mého názoru, při převyprávění tohoto filmu měly v projevu v dostatečné míře vyskytovat klasifikátorové konstrukce a využití znakovacího prostoru, což je pro analýzu znakového jazyka poměrně významný prvek. Tímto polední celek skončí a následuje druhá přestávka.

Obr. 30: Ukázka ze snímku The Elevator

Poslední odpolední část se věnuje nejprve elicitaci znaků, které respondenti považují odlišné z hlediska jejich používání u mladší a starší generace neslyšících. Probíhá tak, že znaky, u kterých se předpokládá např. posun místa artikulace, jsou předloženy jako psané glosy neslyšícím respondentům a ti následně nad těmito znaky konverzují, zda nějaké změny registrují či nikoliv. Tento úkol samozřejmě nepojme všechny znaky, které prošly změnou, proto je možné, aby sami respondenti uvedli další příklady znaků, které považují za odlišné z hlediska generačního používání. Pro odlehčení práce se zařadí úkol, ve kterém jeden respondent popisuje tomu druhému předem vyznačenou cestu na mapě. Druhý respondent na jeho pokyny reaguje zjišťovacími otázkami, aby se ujistil, zda na stejné, ovšem slepé mapě postupuje správně. Nyní se dostanou k části, která je zaměřena na elicitaci příkladových vět založených na použití specifických znaků českého znakového jazyka. Podklady k tomuto úkolu by byly využity z publikace P. Vysučka – Specifické znaky v českém znakovém jazyce. V této publikaci jsou vybrány a popsány specifické znaky, takže je vhodné vybrat přibližně 10 specifických znaků, ty respondentům předložit ve formě fotky znaku a přibližného českého překladu, které publikace obsahuje.

Respondenti následně produkují příkladové věty obsahující příslušné specifické znaky. Takto by se skrze produkované příkladové věty ověřilo, zda se na území České republiky a také v rámci generační škály respondentů užívají a chápou specifické znaky stejným způsobem. Tento úkol nebyl zahrnutý ani v jednom výše popisovaném korpusovém projektu, avšak já si myslím, že tento úkol je více než vhodný, z důvodu jeho specifičnosti z lingvistického hlediska. Následně by takto sesbírané znaky byly velmi vhodným materiálem k dalším lingvistickým výzkumům. Celé setkání je ukončeno volnou konverzací na libovolné téma, do kterého se dá případně zahrnout i reflexe průběhu celodenní elicitace.

Obr. 31: Příklad specifického znaku užitého v rámci elicitačních materiálů (pozn. nejčastější formulace překladu: ...zaskočilo mě to…, ...nevěděl jsem, co mám dělat…)

Do elicitačních úkolů jsem nezahrnula úkoly, které se např. týkaly převyprávění znakového příběhu, protože bych se obávala ovlivnění respondentů v používání některých znaků produkovaných v elicitační videonahrávce. Také bych nezařadila úkol týkající se vyprávění nezapomenutelných životních událostí, protože si myslím, že by toto zadání mohlo vyústit k vyprávění příliš osobních informací. S tímto by souvisel i již zmíněný, předem podepsaný, souhlas s natáčením a zveřejněním daných nahrávek, se kterým by respondenti byli seznámeni ještě před natáčením.

Dále je nutné rozhodnout, jaký program pro zpracování dat zvolit, zda ELAN, iLex či nějaký jiný. Když budu vycházet z programů, které jsou v této práci popsány, tak oba zmíněné softwary mají svá pro a proti. ELAN bych volila vzhledem k jeho anotačním funkcím a barevně rozlišeným úrovním, které slouží k lepší orientaci. Dalším jeho pozitivem je snadná manipulace s již vloženými daty, které lze kopírovat do dalších úrovní. Nedostatek ELANu vidím v tom, že nebyl původně vytvořen jako software pro anotaci znakových jazyků, není tedy plně přizpůsobený pro popis všech prvků.

V kontrastu, iLex byl přímo vyvinut pro anotaci znakového jazyka, proto má potenciál obsáhnout všechna potřebná kritéria při procesu anotace jazyka. Jeho výhody spatřuji v možnosti zasahovat do již vzniklých dat prostřednictvím poznámek v jednotlivých úrovních, ale např. i v metadatech. Další předností je možnost okamžitého sdílení dat v jedné databázi mezi všemi pracovníky a možnost vstupovat do souborů i mimo pracoviště. V neposlední řadě je pozitivem možnost nastavit software na dvě jazyková rozhraní, pro umožnění přístupu více výzkumníkům s různou jazykovou výbavou a v případě nutnosti je možné data přesunout i do ELANu. Z těchto důvodů bych volila pro korpus českého znakového jazyka software iLex.

Je také nutné vyřešit otázku toho, kdo by zpracovával získaná jazyková data. Dovoluji si tvrdit, že v České republice není mnoho lidí, kteří by měli zkušenosti s prací na podobném projektu. Ovšem dali by se zapojit ti výzkumníci, kteří již pracují na realizaci korpusu Deaf a ti by tak mohli předat své vědomosti získané touto prací dalším osobám, které by vytvářeli korpus českého znakového jazyka. Jistě by to byla dobrá zkušenost pro studenty a pracovníky z lingvistických oborů, pro studenty oboru Čeština v komunikaci neslyšících a pracovníky z Ústavu Českého národního korpusu. Spolupráce studentů, kteří mají lingvistické znalosti se studenty, kteří mají znalosti o znakovém jazyce, by jistě byla prospěšná. Nicméně by byla nutná kontrola edukovaných neslyšící supervizorů, kteří by dohlíželi na tyto průběžné práce.

Ráda bych zde ještě podotkla, že je tento návrh pouze rámcový a zcela předběžný. Nicméně by mohl sloužit jako prvotní námět k tvorbě korpusu českého znakového jazyka.

5. Závěr

Ve své bakalářské práci se zabývám popisem tvorby a struktury korpusů znakových jazyků. Na základě zejména zahraniční literatury a elektronických zdrojů jsem představila přístupy k vytváření korpusů znakových jazyků. Ještě jednou bych zde ráda uvedla, že tato bakalářská práce má být pouze prvním, a tudíž obecným vhledem do problematiky tvorby korpusů znakových jazyků. Každá dílčí fáze tvorby korpusu znakového jazyka by svým rozsahem jistě vydala na samostatnou práci, což by v tomto případě nebylo na místě.

Nejprve jsem se věnovala korpusové lingvistice a s ní spojenými odbornými pojmy, z důvodu sjednocení terminologie používané v celé práci. Posléze jsem se zaměřila na korpus znakového jazyka, jeho specifika, která zahrnovala odlišnou metodologii sběru jazykového materiálu, práci s respondenty či zpracování získaných jazykových dat, a to vše vztažené do opozice ke korpusu jazyka mluveného.

Po obecných informacích týkajících se korpusu znakového jazyka jsem se zaměřila na popis korpusu australského znakového jazyka. Nejprve jsem uvedla prvotní myšlenky, které vedly k tvorbě korpusu, a následně jsem popsala jednotlivé kroky, které bylo nutné provést, aby mohl daný korpus vzniknout. Prostor je věnován i anotačnímu softwaru ELAN, který byl použit, což je důležité zmínit, z důvodu možného použití jiných softwarů. Převážná část charakteristiky korpusu australského znakového jazyka je věnována popisu anotace získaných jazykových dat, která je poměrně detailní.

Druhým korpusem, na který je práce zaměřena, je korpus německého znakového jazyka. I tento korpus jsem popsala postupně od jeho vzniku a záměrů tvorby po specifické funkce, jakým je např. session director. V rámci charakteristiky tohoto korpusu jsem se více zaměřila na popis elicitačních úkolů, protože v prvním případě se moje pozornost soustředila spíše na proces anotace jazykových dat. Taktéž jsem uvedla informace o anotačním softwaru iLex, který používají v Německu. Nakonec je zmíněn překladový slovník DGS a německého jazyka, který vychází z korpusu německého znakového jazyka. Odlišnost struktury popisu obou korpusů je způsobená tím, že jsem nechtěla, aby se informace o korpusech zbytečně opakovaly, ale spíše doplňovaly a obohacovaly jinými údaji.

Poslední částí bakalářské práce je můj návrh vzniku a podoby korpusu českého znakového jazyka, který vychází z informací, které jsem získala při psaní celé této práce.
Mé návrhy se týkají specifičnosti získávání respondentů, ale i elicitačních úkolů, z nichž některé jsem navrhla speciálně pro potenciální korpus českého znakového jazyka. Následně se zabývám otázkou, který software na zpracování videí použít a proč.

Taktéž jsem zmínila možnou spolupráci na tomto projektu skrze více odborníků i z řad studentů. Domnívám se, že poznatky v této práci by mohly být prakticky využity při počáteční tvorbě korpusu českého znakového jazyka, kvůli jejich celistvému uspořádání a shrnutí podstatné části informací, týkajících se tvorby korpusu znakového jazyka.

Seznam literatury a zdrojů:

BERMAN, Ruth a Dan SLOBIN. Relating events in narrative: a crosslinguistic developmental study. *Journal of Child Language*. 1996, **23**(3), 715−723.

BERTOLDI, N. a kol. On the Creation and the Annotation of a Large-scale Italian-LIS Parallel Corpus. In: *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Německo: Institute for German Sign Language and Communication of the Deaf, 2010, s. 19−22.

BORSTELL, Carl, Johanna MESCH a Moa GARDENFORS. Towards an Annotation of Syntactic Structure in the Swedish Sign Language Corpus. In: *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*. Hamburg: University of Hamburg, 2016, s. 19−24.

BRASHEAR, H. a kol. CopyCat: A Corpus for Verifying American Sign Language During Game Play by Deaf Children. In: *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Německo: Institute for German Sign Language and Communication of the Deaf, 2010, s. 27−32.

BRENTARI, Diane. *A Crosslinguistic Study of Sign Language Classifiers*. Cambridge: Cambridge University Press, 2001.

CORMIER, Kearsy a Adam SCHEMBRI. Describing sociolinguistic variation in verb directionality in British Sign Language: A corpus-based study. In: *Economic and Social Research Council* [online]. 2014 [cit. 2017-07-15]. Dostupné z: http://www.researchcatalogue.esrc.ac.uk/grants/ES.K003364.1/read

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: reprezentativnost. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2013a [cit. 2017-05-03]. Dostupné z: http://wiki.korpus.cz/doku.php/pojmy:reprezentativnost?rev=1379079160&vecdo=cite

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: metadata. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2013b [cit. 2017-06-23]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:metadata&rev=1379076564

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: korpus. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2014a [cit. 2017-05-03]. Dostupné z: http://wiki.korpus.cz/doku.php/pojmy:korpus?rev=1416829573&vecdo=cite

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: segmentace. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2014b [cit. 2017-04-30]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:segmentace&rev=1416830148

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: anotace. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2014c [cit. 2017-04-30]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:anotace&rev=1416826135

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: token. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2014d [cit. 2017-04-30]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:token&rev=1416830704

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: tag. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2014e [cit. 2017-04-30]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:tag&rev=1416830549

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: struktura_korpusu. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2016 [cit. 2017-04-29]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:struktura_korpusu&rev=1472978350

CVRČEK, Václav. Lemma. In: *CzechEncy – Nový encyklopedický slovník češtiny online* [online]. Praha: NLN, 2016 [cit. 2017-04-30]. Dostupné z: https://www.czechency.org/slovnik/LEMMA

CVRČEK, Václav a Olga RICHTEROVÁ. Pojmy: atributy_strukturni. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2017 [cit. 2017-06-23]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:atributy_strukturni&rev=1496321982

CVRČEK, Václav a Olga RICHTEROVÁ. cnk:uvod. In: *Český národní korpus* [online]. Praha: Příručka ČNK, 2017 [cit. 2017-04-30]. Dostupné z: http://wiki.korpus.cz/doku.php/cnk:uvod?rev=1493217800&vecdo=cite

ČERMÁK, F. Jazykový korpus: Prostředek a zdroj poznání. *Slovo a slovesnost*. Praha: 1995, (56): 119−140.

ČERMÁK, František. Korpus, informace a lingvistika. In: *Přednášky z 48. běhu Letní školy slovanských studií.* Praha : Filozofická fakulta Univerzity Karlovy, 2005 s. 15−24.

ČERMÁK, František. Segmantace. In: *CzechEncy – Nový encyklopedický slovník češtiny online* [online]. Praha: NLN, 2016 [cit. 2017-04-29]. Dostupné z: https://www.czechency.org/slovnik/SEGMENTACE

DEBEUZEVILLE, Louise, Trevor JOHNSTON a Adam C. SCHEMBRI. The use of space with indicating verbs in Auslan: A corpus-based investigation. *Sign Language*. 2009, **12**(1), 53−82. DOI: 10.1075/sll.12.1.03deb. ISSN 1387-9316. Dostupné také z: http://www.jbe-platform.com/content/journals/10.1075/sll.12.1.03deb

Dictionary. In: *Auslan Signbank* [online]. Sydney: Macquarie University, 2015 [cit. 2017-05-10]. Dostupné z: http://www.auslan.org.au/dictionary/words/flower-1.html

Dictionary. In: *DGS - KORPUS* [online]. Hamburk: University of Hamburg, 2016 [cit. 2017-04-02]. Dostupné z: https://www.sign-lang.uni-hamburg.de/dgskorpus/index.php/dictionary.html

EFTHIMIOU, Eleni, Stavroula - Evita FOTINEA a A-L. DIMOU. From a Sign Lexical Database to an SLGolden Corpus – the POLYTROPON SL Resource. In: *7th Workshop on the Representation and Processing of Sign Languages:: Corpus Mining*. Hamburg: University of Hamburg, 2016, s. 63−68.

FENLON, Jordan, Adam SCHEMBRI a Trevor JOHNSTON. Documentary and Corpus Approaches to Sign Language Research. *Research methods in sign language studies: a practical guide*. Chichester: Wiley Blackwell, 2015, 156-172. Guides to research methods in language and linguistics. ISBN 978-1-118-27142-1.

FUNG H-M, C. a kol. Simultaneity vs. Sequentiality: Developing a transcription system of Hong Kong Sign Language acquisition data. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Německo: Institute for German Sign Language and Communication of the Deaf, 2008, s. 22−27.

GLIENNA, Greg. In: *Youtube* [online]. 26.01.2010 [cit. 2011-06-28]. Dostupné z: https://www.youtube.com/watch?v=Q-TQQE1y68c&index=7&list=PLEv1DPaINTKsHCwDWH-mnAD7T9ddwwEUT

HANKE, Thomas. ILex – A tool for Sign Language Lexicography and Corpus Analysis. In: *Third International Conference on Language Resources and Evaluation* [online]. Paříž: ELRA, 2002, s. 923−926 [cit. 2017-06-27]. Dostupné z: http://www.lrec-conf.org/proceedings/lrec2002/

HANKE, Thomas a Jakob STORZ. iLex – A database tool integrating sign language corpus linguistics and sign language lexicography. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Hamburk: University of Hamburg, 2008a, s. 64−67.

HANKE, Thomas a Jakob STORZ. iLex - A database tool for integrating sign language corpus linguistics and sign language lexicography. In: *Construction and Exploitation of Sign Language Corpora*. Hamburk: University of Hamburg, 2008b, s. 1−3.

HANKE, Thomas, Lutz KONIG a Sven WAGNER. DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Hamburk: University of Hamburg, 2010, s. 106−109. ISBN 2-9517408-6-7.

HANKE, Thomas a Christian RATHMAN. Elicitation Methods in the DGS (German Sign Language) Corpus Project. In: *4th Workshop on the Representation and Processing of Sign Languages:: Corpora and Sign Language Technologies*. Hamburk: University of Hamburg, 2010, s. 178−185. ISBN 2-9517408-6-7.

HANKE, Thomas. Towards a Visual Sign Language Corpus Linguistics. In: *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* [online]. Hamburk: University of Hamburg, 2016, s. 89−93 [cit. 2017-06-17]. Dostupné z: https://www.sign-lang.uni-hamburg.de/lrec2016/programme.html

HICKMANN, Maya. *Children's discourse: person, space and time across languages*. Digitally printed version. Cambridge: Cambridge University Press, 2003. ISBN 978-052-1065-108.

HLADKÁ, Zdeňka. Lemma. In: *CzechEncy – Nový encyklopedický slovník češtiny online* [online]. Praha: NLN, 2016 [cit. 2017-04-30]. Dostupné z: https://www.czechency.org/slovnik/LEMMA

CHÉTELAT-PELÉ, Emilie, Annelies BRAFFORT a Jean VÉRONIS. Annotation of Non Manual Gestures: Eyebrow movement description. In: *3rd Workshop on the Representation and Processing of Sign Languages:: Construction and Exploitation of Sign Language Corpora*. Hamburg: University of Hamburg, 2008, s. 28−32.

JOHNSTON, Trevor a Onno CRASBORN. The use of ELAN annotation software in the creation of signed language corpora. In: *Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art* [online]. Lansing: MI, 2006 [cit. 2017-02-22]. Dostupné z: http://emeld.org/workshop/2006/papers/johnston-crasborn.pdf

JOHNSTON, Trevor. Corpus linguistics and signed languages: no lemmata, no corpus. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Německo: Institute for German Sign Language and Communication of the Deaf, 2008, s. 82−87.

JOHNSTON, Trevor. Creating a Corpus of Auslan within an Australian National Corpus. In: *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus,* [online]. Somerville: MA: Cascadilla Proceedings Project, 2009, s. 87−95 [cit. 2017-02-13]. ISBN 978-1-57473-435-5. Dostupné z: http://www.lingref.com/cpp/ausnc/2008/

JOHNSTON, Trevor. The Auslan corpus with the benefit of hindsight. In: *Sign Linguistics Corpus Network Workshop* [online]. London: DCAL, 2009, s. 1−17 [cit. 2017-02-14]. Dostupné z: www.ru.nl/publish/pages/570578/slcn_london_johnston.pdf

JOHNSTON, Trevor. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* [online]. 2010, **15**(1), 106−131 [cit. 2017-02-15]. Dostupné z: http://www.aclweb.org/anthology/Y08-1002

JOHNSTON, Trevor. Showreel for the deposit "Auslan Corpus". In: *Endangered Languages Archive at SOAS University of London* [online]. London: SOAS University of London, 2015 [cit. 2017-05-10]. Dostupné z: https://elar.soas.ac.uk/Record/MPI1035333

JOHNSTON, Trevor. Adding Value to, and Extracting Value from, a Signed Language Corpus through Strategic Annotations. In: *Mezinárodní letní škola znakových jazyků 2014: Aktuální otázky v lingvistice znakových jazyků* [online]. Praha: Univerzita Karlova v Praze, 2016 [cit. 2017-02-13]. ISBN 978-80-7308-671-8. Dostupné z: http://cisl.ff.cuni.cz/

JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

KOLLER, Oscar, Hermann NEY a Richard BOWDEN. Weakly supervised automatic transcription of mouthings for gloss-based sign language corpora. In: *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*. Hamburg: University of Hamburg, 2014, s. 89−94.

LANGER, Gabriele, Thomas TROELSGÅRD a Jette KRISTOFFERSEN. Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In: *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* [online]. Hamburk: University of Hamburg, 2016, s. 143−152 [cit. 2017-06-17]. Dostupné z: https://www.sign-lang.uni-hamburg.de/lrec2016/programme.html

LEECH, Geoffrey. Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. *Language Learning* [online]. 2000, 50(4), 675 - 724 [cit. 2017-05-28]. Dostupné z: http://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xpapers/Leech_spoken_grammar.pdf

MAYER, Mercer. *Frog, where are you?*. New York: Dial Books for Young Readers, c1969. ISBN 978-080-3728-813.

MCNEILL, David. So you think gestures are nonverbal? *Psychological review* [online]. 1985, **92**(3), 350−371 [cit. 2017-01-17]. ISSN 1939-1471. Dostupné z: http://www.communicationcache.com/uploads/1/0/8/8/10887248/so_you_think_gestures_are_nonverbal.pd

Metadata for sign language resources. In: *SLCN workshop 2. Metadata* [online]. Nijmegen: Radboud University, 2009, s. 1−4 [cit. 2017-06-21]. Dostupné z: http://www.ru.nl/slcn/workshops/2_metadata/

NAGASHIMA, Y. a kol. Construction of Japanese Sign Language Dialogue Corpus: KOSIGN. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Německo: Institute for German Sign Language and Communication of the Deaf, 2008, s. 141−144.

Operační program Výzkum, vývoj a vzdělání 2014-2020. *Dotační.info* [online]. [cit. 2017-06-29]. Dostupné z: http://www.dotacni.info/operacni-program-vyzkum-vyvoj-a-vzdelavani-2014-2020/

PALA, Karel. Informační technologie a korpusová lingvistika (1). *Zpravodaj ÚVT MU* [online]. 1996, 6(3), 8−11 [cit. 2017-01-19]. ISSN 1212-0901. Dostupné z: http://webserver.ics.muni.cz/bulletin/articles/58.html

PETKEVIČ, Vladimír. Anotace. In: *CzechEncy – Nový encyklopedický slovník češtiny online* [online]. Praha: NLN, 2016 [cit. 2017-04-30]. Dostupné z: https://www.czechency.org/slovnik/ANOTACE

PFAU, Roland, Markus STEINBACH a Bencie WOLL. Sign Language: An International Handbook. In: *Handling sign language data: Data collection.* Boston: De Gruyter Mouton, 2012, 1023−1043. ISBN 978-3-11-020421-6.

PRILLWITZ, Siegmund a Thomas HANKE. DGS corpus project – Development of a corpus based electronic dictionary German Sign Language / German. In: *3rd Workshop on the Representation and Processing of Sign Languages:: Construction and Exploitation of Sign Language Corpora*. Hamburk: University of Hamburg, 2008, s. 159−164.

RATHMANN, Christian. Projekt korpusu německého znakového jazyka: design korpusu. In: *Mezinárodní letní škola znakových jazyků 2014: Aktuální otázky v lingvistice znakových jazyků*. Praha: Univerzita Karlova v Praze, 2016. ISBN 978-80-7308-671-8.

RUTKOWSKI, Paweł, Joanna ŁACHETA a Piotr MOSTOWSKI. Korpus polského znakového jazyka. Section for Sign Linguistics University of Warsaw: Pracownia Lingwistyki Migowej. Prezentace. [cit. 2017-06-16].

SCHEMBRI, Adam, Jordan FENLON, Ramas RENTELIS a Sally REYNOLDS. Building the British Sign Language Corpus. *Language Documentation & Conversation* [online]. 2013, 7(7), 136−154 [cit. 2017-01-16]. DOI: http://hdl.handle.net/10125/4592. ISSN 1934-5275. Dostupné z: http://scholarspace.manoa.hawaii.edu/handle/10125/4592

ŠULC, Michal. Tematická reprezentativnost korpusů. *Slovo a slovesnost* [online]. 2001, 62(1), 53−61 [cit. 2017-01-19]. Dostupné z: http://sas.ujc.cas.cz/archiv.php?art=4000 reprezentativnost korpusu

The Auslan Corpus [online]. Australia: Creative Commons BY-NC-ND 4., 2012 [cit. 2017-02-13]. Dostupné z: http://www.auslan.org.au/about/corpus/

VALCHAŘOVÁ, Veronika. *Pohádka O veliké řepě.* Pinterest. [cit. 2011-06-28]. Dostupné z: https://cz.pinterest.com/pin/260575528417203663/

VYSUČEK, Petr. *Specifické znaky v českém znakovém jazyce*. 2. opr. vyd. Praha: Česká komora tlumočníků znakového jazyka, c2008. ISBN 978-80-87153-53-6

Příloha: Seznam obrázků a jejich zdrojů

Obr. 1: Anotovaná videonahrávka v ELANu pro znak SLOW
JOHNSTON, Trevor. Showreel for the deposit "Auslan Corpus". In: *Endangered Languages Archive at SOAS University of London* [online]. London: SOAS University of London, 2015 [cit. 2017-05-10]. Dostupné z: https://elar.soas.ac.uk/Record/MPI1035333

Obr. 2: Auslan Signbank – výsledek hledaného výrazu FLOWER (květina)
Dictionary. In: *Auslan Signbank* [online]. Sydney: Macquarie University, 2015 [cit. 2017-05-10]. Dostupné z: http://www.auslan.org.au/dictionary/words/flower-1.html

Obr. 3: Příklad užití ID glosy GROW-UP
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 17 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 4: Anotovaná podoba dvouručního znaku OWL
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 20 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 5: Anotace v případě, že každá ruka nese jiný znak
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 21 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 6: Anotovaná podoba stejného znaku, který je produkován jednou nebo oběma rukama
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 21 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 7: Příklady tvarů a orientací ruky či prstů užívaných v klasifikátorech.
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 34−35 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 8: Anotace gesta LOOK-SURPRISED (vypadat překvapeně)
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 43 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 9: Anotace opakovaného znaku
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 47 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 10: Anotace znaku, který je po chybné produkci bezprostředně opraven
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 48 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 11: Anotace znaku, který napodobuje činnost, kdy člověk něco drží
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 58 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 12: Gramatické třídy tagů
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 65 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 13: Příklad úrovní sémantických rolí
JOHNSTON, Trevor. *Auslan Corpus Annotation Guidelines* [online]. Australia: Macquarie University, 2016, s. 74 [cit. 2017-02-13]. Dostupné z: http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Obr. 14: Okno zobrazující se moderátorovi v rámci programu session director
HANKE, Thomas, Lutz KONIG a Sven WAGNER. DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Hamburk: University of Hamburg, 2010, s. 109. ISBN 2-9517408-6-7.

Obr. 15: Ukázka obrázků z elicitačního příběhu Frog, where are you?
RUTKOWSKI, Paweł, Joanna ŁACHETA a Piotr MOSTOWSKI. *Korpus polského znakového jazyka*. Section for Sign Linguistics. University of Warsaw: Pracownia Lingwistyki Migowej. Prezentace.

Obr. 16: Elicitační obrázky – varující a zákazové značky
RUTKOWSKI, Paweł, Joanna ŁACHETA a Piotr MOSTOWSKI. *Korpus polského znakového jazyka.* Section for Sign Linguistics. University of Warsaw: Pracownia Lingwistyki Migowej. Prezentace.

Obr. 17: Elicitační obrázky k tematické oblasti „zdraví"
RUTKOWSKI, Paweł, Joanna ŁACHETA a Piotr MOSTOWSKI. *Korpus polského znakového jazyka*. Section for Sign Linguistics. University of Warsaw: Pracownia Lingwistyki Migowej. Prezentace.

Obr. 18: Ukázka obrázku z elicitačního příběhu o zájezdu
HANKE, Thomas a Christian RATHMAN. Elicitation Methods in the DGS (German Sign Language) Corpus Project. In: *4th Workshop on the Representation and Processing of Sign Languages:: Corpora and Sign Language Technologies*. Hamburk: University of Hamburg, 2010, s. 183. ISBN 2-9517408-6-7.

Obr. 19: Elicitační obrázky, data, místa – události v komunitě neslyšících (Deaflympiáda)
HANKE, Thomas a Christian RATHMAN. Elicitation Methods in the DGS (German Sign Language) Corpus Project. In: *4th Workshop on the Representation and Processing of Sign Languages:: Corpora and Sign Language Technologies*. Hamburk: University of Hamburg, 2010, s. 184. ISBN 2-9517408-6-7.

Obr. 20: Elicitační obrázek – mapa města
RUTKOWSKI, Paweł, Joanna ŁACHETA a Piotr MOSTOWSKI. *Korpus polského znakového jazyka.* Section for Sign Linguistics. University of Warsaw: Pracownia Lingwistyki Migowej. Prezentace.

Obr. 21: Pozice kamer a respondentů ve studiu
RATHMANN, Christian. Projekt korpusu německého znakového jazyka: design korpusu. In: *Mezinárodní letní škola znakových jazyků 2014: Aktuální otázky v lingvistice znakových jazyků*. Praha: Univerzita Karlova v Praze, 2016. ISBN 978-80-7308-671-8.

Obr. 22: Záběry na respondenty umožněné různými pozicemi kamer
RATHMANN, Christian. Projekt korpusu německého znakového jazyka: design korpusu. In: *Mezinárodní letní škola znakových jazyků 2014: Aktuální otázky v lingvistice znakových jazyků*. Praha: Univerzita Karlova v Praze, 2016. ISBN 978-80-7308-671-8.

Obr. 23: Příklad znaku STAMP, který zahrnuje fonologickou variantu znaku A, B a modifikovanou formu znaku C, D
LANGER, Gabriele, Thomas TROELSGÅRD a Jette KRISTOFFERSEN. Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In: *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* [online]. Hamburk: University of Hamburg, 2016, s. 143-152 [cit. 2017-06-17]. Dostupné z: https://www.sign-lang.uni-hamburg.de/lrec2016/programme.html

Obr. 24: Okno v iLexu zaznamenávající tokeny a příslušné informace
HANKE, Thomas a Jakob STORZ. iLex - A database tool for integrating sign language corpus linguistics and sign language lexicography. In: Construction and Exploitation of Sign Language Corpora. Hamburk: University of Hamburg, 2008b, s. 1−3.

Obr. 25: Okno textového editoru syncWRITER
HANKE, Thomas. ILex – A tool for Sign Language Lexicography and Corpus Analysis. In: *Third International Conference on Language Resources and Evaluation* [online]. Paříž: ELRA, 2002, s. 924 [cit. 2017-06-27]. Dostupné z: http://www.lrec-conf.org/proceedings/lrec2002/

Obr. 26: Horizontální zarovnání času v iLexu
HANKE, Thomas a Jakob STORZ. iLex – A database tool integrating sign language corpus linguistics and sign language lexicography. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Hamburk: University of Hamburg, 2008a, s. 65.

Obr. 27: Vertikální zarovnání času v iLexu
HANKE, Thomas a Jakob STORZ. iLex – A database tool integrating sign language corpus linguistics and sign language lexicography. In: *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Hamburk: University of Hamburg, 2008a, s. 65.

Obr. 28: E-learningový materiál znakového jazyka – videonahrávka s odkazy
HANKE, Thomas a Jakob STORZ. iLex - A database tool for integrating sign language corpus linguistics and sign language lexicography. In: Construction and Exploitation of Sign Language Corpora. Hamburk: University of Hamburg, 2008b, s. 3.

Obr. 29: Elicitační obrázky pohádky O veliké řepě
VALCHAŘOVÁ, Veronika. *Pohádka O veliké řepě.* Pinterest. [cit. 2011-06-28]. Dostupné z: https://cz.pinterest.com/pin/260575528417203663/

Obr. 30: Ukázka ze snímku The Elevator
Short film "The Elevator". In: *YouTube* [online]. 2010 [cit. 2017-06-29]. Dostupné z: https://www.youtube.com/watch?v=Q-TQQE1y68c

Obr. 31: Příklad specifického znaku užitého v rámci elicitačních materiálů (pozn. nejčastější formulace překladu: …zaskočilo mě to…, …nevěděl jsem, co mám dělat…)
VYSUČEK, Petr. *Specifické znaky v českém znakovém jazyce*. 2. opr. vyd. Praha: Česká komora tlumočníků znakového jazyka, c2008. ISBN 978-80-87153-53-6