PHONEXIA

phonexia.com

Anotační manuál

Transkripce mluvené češtiny

PHONEXIA

Anotační manuál

Lidé hledají informace na internetu, Phonexia hledá v lidké řeči. Ať už si chce člověk či počítač osvojit jakoukoli znalost, potřebuje dostatek kvalitních dat. V našem případě audií. Samotný záznam promluvy však nestačí k tomu, abychom naučili systém rozpoznávat řeč. Systém musí vědět, jak řeč "číst".

Z toho důvodu existují anotace. Jejich úkolem je doslovně přepsat řeč, nic nevynechávat, nic nepřidávat. Ani neřečové události, kterých spontánní mluvený projev obsahuje velké množství. Počítač si pak "naposlouchá" jednotlivé zvuky a přiřadí jim hodnotu – grafickou reprezentaci.

Pokud se učíme po hláskách jako v tomto případě, je skutečně důležité přepisovat slova tak, jak je slyšíme. Ne podle jejich významu. Například slovní spojení *okey dokey* přepsané jako *ok* by následně zapříčinilo mylnou interpretci jak na úrovni fonetické, tak lexikální.

Ač se tvorba anotací nezdá být složitá, snadná není. Předem děkujeme, milí anotátoři, za vaši trpělivost a přesnou ruku. I detaily jsou důležité, díky nim vzniknou kvalitní data a fungující systém.

Děkujeme za spolupráci, vážíme si jí!

Nezbytnosti k anotování

- stolní počítač/notebook (Windows, příp. Linux)
- program Transcriber
- kvalitní sluchátka překrývající uši (ne "pecky")
- tiché prostředí
- anotační manuál

Doporučení

- trpělivý a pozorný přístup k anotaci
- anotátor na úrovni rodilého mluvčího v daném jazyce
- odpočinek po přibližně 2 hodinách poslechu, v čase mezi anotováním volit činnosti, které nezaměstnávají sluch

PHONEXIA

Nahrávání

- přístup k hlasovým nahrávkám bude omezený
- data budou primárně určena pro technologii SID (identifikace mluvčího), z toho důvodu je nutné získat od jednoho mluvčího minimálně **3 nahrávky**
- délka jedné nahrávky: cca **3 minuty** (minimálně minuta čisté řeči)
- spontánní projev
- různá témata

mluvčí

- starší 18 let
- různorodost (věk, vzdělání, nářeční oblast, …)
- dobrá vyjadřovací schopnost (plynulý projev, bez dlouhých odmlk)
- získání min. 3 nahrávek pro jednoho mluvčího

prostředky

- **mobilní telefon** – aplikace CSipSimple (Android)
    - other country providers: Odorik
    - nastavit fungování i mimo wifi; 3 G sítě
    - telefonní číslo a heslo na vyžádání
    - vytáčení telefonního čísla přes **číselník v aplikaci**
    - po ukončení hovoru je nahrávka zaslána na e-mail, ze kterého bude předán k distribuci mezi anotátory
    - pův. stereo již bude rozděleno na mono nahrávky
    - ke snadné distribuci je nutné poznačit si **čas pořízení hovoru**
    - v případě nemožnosti stažení aplikace (absence Androidu) je zde možnost zapůjčení mobilního telefonu
    - pokud je účet **neaktivní**, zkontrolujte si prosím připojení k internetu a nastavení viz výše, v aplikaci pak zvolte sekci "účty" (klíč vlevo dole) a stiskněte "telefon" u účtu Odorik
- **diktafon**
- **VoIP** kanál (typově Skype)

Mobilní aplikace nahrává hovor do dvou kanálů, takže i když si mluvčí skočí do řeči, po rozdělení na mono nahrávky uslyšíme vždy pouze jednoho mluvčího.

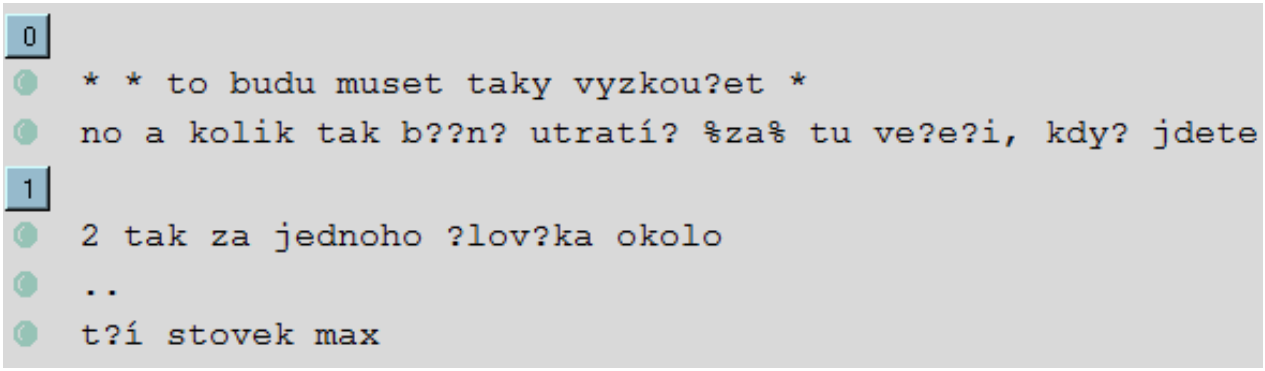
V případě **diktafonů** a **hovorů přes internet** (VoIP) je záznam zvuku obvykle v jedné stopě – v mono nahrávce. Pokud dojde k překryvu řeči v těchto případech, je potřeba úsek vydělit do samostatného segmentu (se značkou "x:") a dále jej neanotovat, čímž nahrávku ukrátíme. Pokud se vyskytnou dva mluvčí, musí se v anotaci důsledně označit jejich střídání segment a značkami (viz níže). Nejsnáze se anotuje monolog, kdy v nahrávce vystupuje pouze jeden mluvčí.

Program Transcriber (verze 1.5.1)

První věc, kterou je důležité provést před samotným anotováním, je **nastavení kódování**.

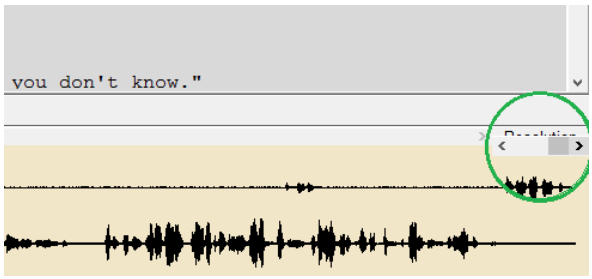
Options	>	General	>	Encoding: Unicode (UTF-8)	>	Ok	>	Options	>	Save configuration

V jiném případě může anotace přijít o zásadní informace, kterými jsou některé znaky, tím i o celá slova. To znamená přijít o anotaci.



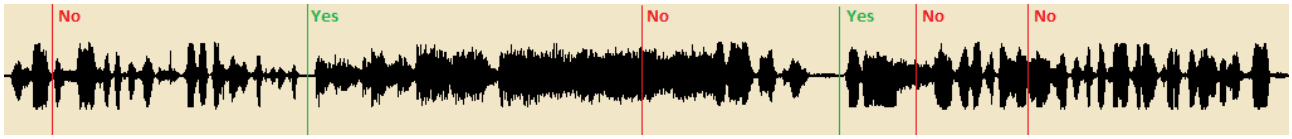
Ovládání

Potřeba	Provedení
Vytvoření nového segmentu	**kliknout myší** do vhodného místa (objeví se červená linka kolmá k signálu), zmáčknout **Enter**
Přehrání/pozastavení zvuku	tabulátor
Přehrání jednoho segmentu	Shift + tabulátor
Oprava hranic segmentu (zkrácení/prodloužení délky segmentu)	podržet **prostřední tlačítko** (kolečko) na myši a posunout hranice segmentu na požadované místo
Smazání segmentu	Ctrl + Backspace

Zvýraznění zvukového signálu	Signal > Control panel > Vertical zoom > posun na škále
Roztažení/shluknutí signálu	

Anotační pravidla

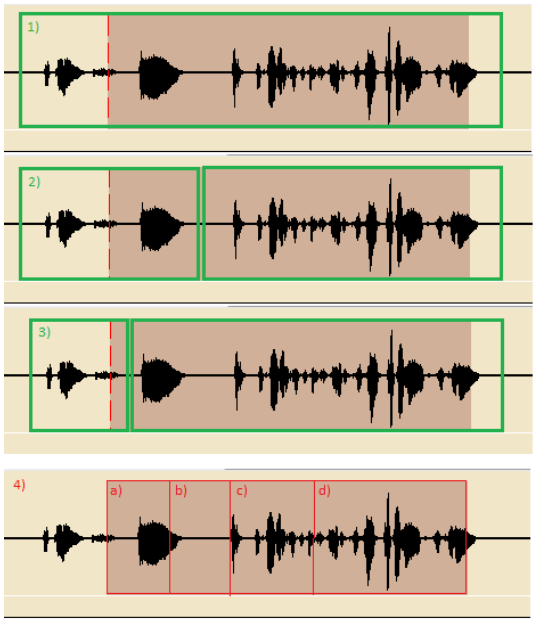
Segmentace



Nahrávku je potřeba rozdělit na menší segmentu (důvody jsou přehlednost a zpracování). Hranice segmentů zpravidla odpovídají jednomu výpovědnímu celku daného řečníka v dialogu, metriku však nelze zakládat na individuálních aspektech. Vycházíme proto z rozmezí **3–8 vteřin**, což je přibližně **5 slov**.

Některé (ne)řečové události vyžadují vlastní segment nehledě na délku jejich trvání. Jedná se například o **problémový segment**, při kterém dochází k překryvu dvou či více řečníků, úsek je nepřepsatelný, a **ticho**. V těchto segmentech je pouze značka dané události, nic dalšího.

Vhodné místo k umístění hranic segmentu vyčteme ze zvukového signálu v programu Transcriber. Pokud signál ukazuje "nahuštěnou" promluvu, nepřetíná se. Ideální je **co nejvodorovnější místo** v signálu. Před řečovou částí signálu a za ní se musí vyskytnout krátké ticho – hranice segmentu se jí tímto nedotýká.



a) b) useknutí řečového signálu
c) "přilepení" hranice segmentu k řečovému signálu - je potřeba dodržovat malý prostor, který obsahuje ticho, na začátku i na konci segmentu
d) hranice segmentu nesmí přetnout řečový signál, ani vyřčené slovo

Informační soubor

Kromě přepisu, anotací zvukového záznamu, je důležité doplnit podstatné informace o nahrávce, které ji pomohou lépe klasifikovat. Šablona ukazuje kategorie, za rovnítkem jsou možnosti jejich doplnění, které anotátor promění v odpovědi (pokud je má).

audio = cstime (formát "jazyk-rok-měsíc-den-hodina-minuta", tj. čas pořízení nahrávky)
annotator_email = ...@...
language = cs
channel = {mobil, diktafon, VoIp}
codec = {G711/G729/G723.1/...}
gender1 = {M/F/C} # C - Child, M - Male, F – Female
age1 = {CHILD/YOUNG/ADULT/OLD}
(gender2 = {M/F/C} # C - Child, M - Male, F - Female)
(age2 = {CHILD/YOUNG/ADULT/OLD})
signal_quality = {GOOD/BAD}
environment_quality = {CLEAN/NOISY}
comment = informace o mluvčím či nahrávce

Anotační značky

Jev	Použití	Značka	Příklad
Střídání mluvčích V nahrávkách s více než jedním mluvčím.	Mluvčí se značí intuitivně **číslicemi**, které udávají jejich pořadí v nahrávce. Za číslicí následuje dvojtečka a přepsaný obsah segmentu. Tato značka se objevuje **na každém řádku**.	1: 2:	1: tady Michal , kdo volá ? 2: Vojta , ahoj . chtěl bych se s tebou domluvit na ten fotbal .
Problematický segment Dva lidé hovoří ve stejnou chvíli, dochází k překryvu řeči. Nesrozumitelné úseky, porušený zvukový signál, rozpoznatelné hlasy v pozadí.	Tyto segmenty obsahují **pouze značku**, dále se neanotuje.	x:	x: 1: ahoj , hele , byls tam včera ? 2: čau ! kde jako x: 2: jo , jo , už chápu , díky , žes to vyřešil .

Ticho Málokdy se vyskytne ticho „laboratorních podmínek", pokud je v místnosti přirozený konzistentní šum (bez hlasů a výrazných ruchů), pak jej považujeme za ticho.	Tyto segmenty obsahují **pouze značku**, dále se neanotuje.	..	1: máš pro mě ty informace prosím ? 2: ano , hned pro ně doběhnu , chviličku prosím ! .. 2: tak jsem zpět , mohu diktovat ?
Psaní velkých písmen	**Začátek věty** zpravidla začíná **malým** počátečním **písmenem**. **Ve vlastních jménech** (jména osob, měst, ulic, firem či produktů) se odráží pravidla pravopisu – **velká písmena**.		dobrý den , je dnes v kanceláři pan //**B**laník byls v **L**ondýně ? pojďme koupit babičce **J**ohance nové auto , zrovna představili novou **T**eslu .
Interpunkce	Psaní čárek vychází z pravidel užívání interpunkce. Tečka, otazník, vykřičník – reflektují **intonaci**.	. , ? ! '	Ač je to zvláštní , mezi slovem a interpunkčním znaménkem se píše mezera . Jako fakt !
Číslovky	Číslovky jsou psány jako jednotlivá slova.		tři sta čtyřicet osm tři sta štyrycet osum
Slang, dialekt, neologismy, neznámá a cizí slova	Zlaté pravidlo anotací: *„píšu to, co slyším".*		
	V případě nejasnosti: a) *Nevím, jak přepsat slovo:* co slyším, napíši do závorky.	( )	(fejsbuk)

	b) *V promluvě je cizí slovo, ale znám jeho správný/oficiální zápis*: za to, co slyším, napíši bezprostředně rovnítko a pokračuji ofiálním zápisem slova. To celé je v závorce. Slovní spojení jsou psána zvlášť.	( )	(fejsbuk=Facebook) (ouky=oukey) (douky=dokey)
Nezřetelná slova	*V segmentu je slovo, kterému nerozumím, nejsem schopen/schopna jej přepsat:* použiji značku pro každé nepřepsa(tel)né slovo v segmentu.	??	They are due to meet next Thursday for a European ?? summit They are due to meet next Thursday for a ?? ?? summit
Hláskování **a** **zkratky**	Jednotlivé hlásky jsou v závorce přiřazovány své grafické interpretaci. (Mezinárodně) užívané zkratky mívají různou výslovnost. Pokud nejsou vysloveny „spojitě“, pak ani v přepisu nejsou značeny jako jeden tvar.	( )	(en=n) (ej=a) (em=m) (i=e) (es=s) (ajbíem=IBM) (ybr=UBER) (ubr=UBER) (uber=UBER) (jubr=UBER)
Nedokončená a „nezačatá“ slova	U nedokončených slov za vysloveným úsekem bezprostředně následuje značka. Pokud mluvčí “polkne” začátek slova, pak značka předchází zbytku slova.	+	některá slova nejsou vždy řečen+ úplně , někdy se nám to prostě +povede .

Hezitační zvuky, zvuky souhlasné a nesouhlasné		2	2 nejsem si jistá ona pracuje v 2 2 nemohu si teď vzpomenout na název firmy
Ruchy	Ruchy řečníka *(nádech/výdech, smích, zakašlání, zívnutí aj.)* a prostředí, pozadí nahrávky *(projíždějící auto, klepání, …)*.	*	
	Jednorázový ruch (jedna značka pro danou událost)	*	byl jsem tam, takže jsem to slyšel * (na konci promluvy se vyskytne nádech/smích/…)
	Déletrvající hluk (párová značka, která ohraničuje začátek a konec ruchu)	*	byl jsem tam, takže * jsem to slyšel * (část promluvy *„jsem to slyšel“* probíhá současně s ruchem na pozadí)
	Konzistentní hluk: objevuje se ve stejné intenzitě po celou dobu trvání nahrávky, je takřka zanedbatelný.	neznačí se	
Citlivé informace	Jméno (v kombinaci s příjmením), adresa, telefonní číslo a další osobní údaje.	// před každou touto informací	jmenuji se Jára //Cimrman moje číslo je //čtyřicet //dva a ač jsem se narodil ve Vídni, jsem hrdý Čech

V případě jakýchkoli otázek k manuálu prosím využijte následující kontakt: **lucie.brychtova@phonexia.com**.