

**ČESKÉ VYSOKÉ  
UČENÍ TECHNICKÉ  
V PRAZE**

**FAKULTA  
ELEKTROTECHNICKÁ**



**DIPLOMOVÁ PRÁCE**

**2024**

**BC. KAROLÍNA BENDOVÁ**

**2024**

**Bc. Karolína Bendová**

**Diplomová práce**

České vysoké učení technické v Praze

Fakulta elektrotechnická

Katedra teorie obvodů



Diplomová práce

Analýza promluv pacientů se schizofrenií

Language Analysis of Patients with Schizophrenia

Autor: Bc. Karolína Bendová

Vedoucí práce: prof. Ing. Roman Čmejla, CSc.

Vedoucí práce – specialista: Mgr. Bc. Juraj Jonáš

Studijní program: Lékařská elektronika a bioinformatika

Specializace: Zpracování signálů

Praha 2024

## I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Bendová** Jméno: **Karolína** Osobní číslo: **483460**  
Fakulta/ústav: **Fakulta elektrotechnická**  
Zadávající katedra/ústav: **Katedra teorie obvodů**  
Studijní program: **Lékařská elektronika a bioinformatika**  
Specializace: **Zpracování signálů**

## II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

**Analýza promluv pacientů se schizofrenií**

Název diplomové práce anglicky:

**Language Analysis of Patients with Schizophrenia**

Pokyny pro vypracování:

Cílem diplomové práce je nalezení biomarkerů schizofrenie v promluvách pacientů.

- Proveďte rešerši literatury týkající se projevů schizofrenie v řečovém signálu se zaměřením na sémantické charakteristiky.
- Pro experimenty použijte řečové signály poskytnuté v NÚDZ v Klecanech, případně databázi doplňte o signály z veřejných zdrojů, např. YouTube.
- Vybrané charakteristiky analyzujte v programovém prostředí MATLAB.
- Vyhodnoťte možnosti detekce onemocnění schizofrenií pro český jazyk.

Seznam doporučené literatury:

- [1] De Boer, J. N., Brederoo, S. G., Voppel, A. E., & Sommer, I. E. (2020). Anomalies in language as a biomarker for schizophrenia. *Current opinion in psychiatry*, 33(3), 212-218.
- [2] Stassen, H. H., Albers, M., Püschel, J., Scharfetter, C. H., Tewesmeier, M., & Woggon, B. (1995). Speaking behavior and voice sound characteristics associated with negative schizophrenia. *Journal of Psychiatric Research*, 29(4), 277-296.
- [3] Hrynchenko, S., Analýza hlasu a řeči u pacientů se schizofrenií. *Bakalářská práce, ČVUT, FEL, 2023.*

Jméno a pracoviště vedoucí(ho) diplomové práce:

**prof. Ing. Roman Čmejla, CSc. katedra teorie obvodů FEL**

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

**Mgr. Bc. Juraj Jonáš Národní ústav duševního zdraví, Klecany**

Datum zadání diplomové práce: **26.01.2023**

Termín odevzdání diplomové práce: \_\_\_\_\_

Platnost zadání diplomové práce: **22.09.2024**

\_\_\_\_\_  
prof. Ing. Roman Čmejla, CSc.  
podpis vedoucí(ho) práce

\_\_\_\_\_  
doc. Ing. Radoslav Bortel, Ph.D.  
podpis vedoucí(ho) ústavu/katedry

\_\_\_\_\_  
prof. Mgr. Petr Páta, Ph.D.  
podpis děkana(ky)

## III. PŘEVZETÍ ZADÁNÍ

Diplomantka bere na vědomí, že je povinna vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

\_\_\_\_\_  
Datum převzetí zadání

\_\_\_\_\_  
Podpis studentky

## **Prohlášení**

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne .....

.....

Podpis autora práce

## **Poděkování**

Ráda bych tímto poděkovala prof. Ing. Romanu Čmejlovi, CSc. a panu Mgr. Bc. Juraji Jonášovi za vedení této diplomové práce, za cenné rady, ochotu a vstřícný přístup. Mé poděkování patří rovněž celému týmu Národního ústavu duševního zdraví za zprostředkování výzkumu. Chtěla bych také poděkovat mé rodině a přátelům, kteří byli po celou dobu studia mojí oporou.

Bc. Karolína Bendová

## Abstrakt

Schizofrenie je psychotická porucha, mezi jejíž hlavní symptomy patří poruchy řeči. Touto nemocí trpí zhruba 1 % populace. K úspěšné léčbě schizofrenie výrazně pomáhá správná a včasná diagnostika. Cílem této práce je navrhnout a otestovat potenciální jazykové biomarkery, které by mohly napomoci k diagnostice schizofrenie.

Pro tuto práci byly použity záznamy řeči šesti pacientů se schizofrenií a k nim věkově odpovídajících kontrolních subjektů. Jejich řečové záznamy byly pořízeny ve čtyřech různých typech úloh zaměřených na spontánní promluvu a čtení. Na základě charakteristických projevů schizofrenie v řeči byly provedeny analýzy zaměřené na lexikální sémantiku, zpracování přirozeného jazyka, latentní sémantickou analýzu a prozódické parametry.

Statisticky signifikantní výsledky byly vyhodnoceny pro zkoumané parametry valence a arousal emoční analýzy a analýzy sentimentu. V rámci kvantitativní lingvistiky byly potvrzeny odlišnosti v relativní četnosti spojek a frekventovanosti použitých slov. V prozódických parametrech byly signifikantní rozdíly nalezeny v tempu řeči a v průměrné délce pauz. Z výsledných hodnot také vyplývá, že úlohy založené na spontánní promluvě, při kterých je pravděpodobná vyšší kognitivní aktivita, jsou pro diagnostiku schizofrenie vhodnější než úloha založená na prostém čtení.

Bylo potvrzeno, že sémantické parametry mohou v budoucnu napomoci k diagnostice schizofrenie z řeči.

**Klíčová slova:** schizofrenie, řeč, sémantika, emoce, zpracování přirozeného jazyka

## **Abstract**

Schizophrenia is a psychotic disorder whose main symptoms include speech disorders. With approximately 1% prevalence in the world population. Successful treatment of schizophrenia is greatly aided by correct and early diagnosis. The aim of this work is to propose and test potential language biomarkers that could aid in the diagnosis of schizophrenia.

Speech recordings of six schizophrenia patients and age-matched healthy controls were used in this work. Their speech was recorded during four different types of tasks focusing on spontaneous speech and reading. Analyses were performed based on the characteristic speech manifestations of schizophrenia using lexical semantics, natural language processing, latent semantic analysis and prosodic parameters.

Statistically significant results were evaluated for the investigated valence parameters and arousal emotional analysis and sentiment analysis. Differences in the relative frequency of conjunctions and the frequency of words used were confirmed within the quantitative linguistics framework. In prosodic parameters, significant differences were found in speech rate and in the average length of pauses. The results also suggest that tasks based on spontaneous speech, which are likely to involve higher cognitive activity, are more appropriate for the diagnosis of schizophrenia than tasks based on simple reading.

It was confirmed that semantic parameters may be helpful in the future diagnosis of schizophrenia from speech.

**Keywords:** schizophrenia, speech, semantics, emotions, natural language processing

# Obsah

<b>1. Úvod .....</b>	<b>11</b>
1.1. Schizofrenie .....	11
1.2. Motivace a cíle práce .....	14
<b>2. Teoretická část .....</b>	<b>15</b>
2.1. Charakteristika jazyka a promluvy pacientů se schizofrenií .....	15
2.2. Analýza sentimentu a emocí .....	16
2.3. Zpracování přirozeného jazyka .....	19
2.3.1. Latentní sémantická analýza .....	19
2.4. Prozódie .....	20
<b>3. Metodika .....</b>	<b>21</b>
3.1. Databáze promluv .....	21
3.2. Předzpracování řečových nahrávek .....	22
3.3. Vývojová prostředí pro analýzy .....	23
3.4. Statistický test .....	24
3.5. Lexikální sémantika .....	25
3.6. Latentní sémantická analýza .....	29
3.7. Prozódie .....	31
<b>4. Výsledky .....</b>	<b>35</b>
4.1. Lexikální sémantika .....	35
4.2. Latentní sémantická analýza .....	46
4.3. Prozódie .....	49
<b>5. Diskuse .....</b>	<b>57</b>
<b>6. Závěr .....</b>	<b>61</b>
<b>Použitá literatura .....</b>	<b>62</b>
<b>Příloha A: Doplnující ilustrace .....</b>	<b>67</b>
<b>Příloha B: Protokol pro výzkum pacientů se schizofrenií .....</b>	<b>68</b>
<b>Příloha C: Souhrn hypotéz .....</b>	<b>70</b>



## Seznam obrázků

Obrázek 1: Dimenzionální 2D model emocí dělený na čtyři kvadranty. ....	18
Obrázek 2: Ilustrace postupu získání transkriptů zvukových nahrávek řeči. ....	22
Obrázek 3: Ilustrace procesu lemmatizace transkriptů. ....	23
Obrázek 4: Vybraný úsek řeči vyobrazený v PRAAT. ....	32
Obrázek 5: Výsledné krabicové grafy relativní četnosti vycpávkových slov. ....	35
Obrázek 6: Krabicové grafy relativní četnosti spojek v použití pro úlohy založené na spontánní promluvě. ....	37
Obrázek 7: Krabicové grafy frekvencovanosti slov pro úlohy založené na spontánní promluvě. ....	38
Obrázek 8: Krabicové grafy hodnocení bohatosti slovní zásoby pro úlohy založené na spontánní promluvě. ....	39
Obrázek 9: Krabicové grafy hodnocení arousal slov pro parametry medián a interkvartilový interval. ....	41
Obrázek 10: Krabicové grafy hodnocení valence slov pro parametry medián a interkvartilový interval. ....	42
Obrázek 11: Krabicové grafy procentuálního zastoupení slov se sentimentem a grafy procentuálního zastoupení slov s pozitivním či negativním sentimentem v řeči pro úlohu popisu obrázku založené na vyvolání emocí. ....	44
Obrázek 12: Krabicové grafy procentuálního zastoupení slov vyjadřujících čtyři vybrané emoce – hněv, smutek, očekávání a důvěra. ....	45
Obrázek 13: Krabicové grafy koherence prvního řádu pro úlohy založené na spontánní řeči. ..	47
Obrázek 14: Krabicové grafy koherence druhého řádu pro parametry minimum, medián a interkvartilový interval pro úlohy založené na spontánní promluvě. ....	49
Obrázek 15: Krabicové grafy porovnání celkových délek nahrávek pro všechny úlohy protokolu. ....	50
Obrázek 16: Krabicové grafy procentuálního zastoupení pauz v nahrávce. ....	51
Obrázek 17: Krabicové grafy průměrné délky pauzy pro všechny úlohy protokolu. ....	53
Obrázek 18: Krabicové grafy tempa řeči pro všechny úlohy protokolu. ....	54
Obrázek 19: Krabicové grafy artikulačního tempa pro všechny úlohy protokolu. ....	56
Obrázek 20: Latentní sémantická analýza. Ilustrativní reprezentace vybraných slov v sémantickém prostoru. ....	67
Obrázek 21: Tématický apercepční test. ....	69

# Seznam tabulek

Tabulka 1: Tabulka dělení příznaků schizofrenie na pozitivní a negativní. ....	12
Tabulka 2: Tabulka přehledu typů schizofrenie s typickými projevy. ....	13
Tabulka 3: Přehled vybraných slov a příslušných parametrů arousal a valence. ....	17
Tabulka 4: Přehled úloh protokolu k výzkumu pacientů se schizofrenií. ....	21
Tabulka 5: Tabulka vyhodnocení relativní četnosti výskytu vycpávkových slov v promluvách subjektů pro úlohy 2, 3 a 4. ....	36
Tabulka 6: Tabulka relativní četnosti spojek v promluvách pro úlohy 2, 3 a 4. ....	36
Tabulka 7: Tabulka frekvence použitých slov pro úlohy založené na spontánní promluvě. ....	37
Tabulka 8: Tabulka o bohatosti slovní zásoby pro všechny úlohy založené na spontánní promluvě. ....	38
Tabulka 9: Tabulka výsledných hodnot analýzy arousal slov z úloh založených na spontánní promluvě. ....	40
Tabulka 10: Tabulka výsledných hodnot analýzy valence slov z úloh založených na spontánní promluvě. ....	42
Tabulka 11: Tabulka výsledných hodnot pro hodnocení sentimentu v promluvě určené k vyvolání emocí, resp. úloha 4 popis obrázku. ....	43
Tabulka 12: Tabulka výsledných hodnot procentuálního zastoupení slov, z celkového počtu slov nesoucích emoce, čtyř vybraných základních emocí – hněv, smutek, očekávání a důvěra. ....	46
Tabulka 13: Tabulka výsledných hodnot koherence prvního řádu pro úlohy založené na spontánní promluvě. ....	47
Tabulka 14: Tabulka výsledných hodnot koherence druhého řádu. ....	48
Tabulka 15: Porovnání celkových délek nahrávek pro jednotlivé úlohy. ....	50
Tabulka 16: Tabulka procentuálního zastoupení pauz v nahrávce. ....	52
Tabulka 17: Tabulka průměrných délek pauz. ....	52
Tabulka 18: Tabulka výsledných hodnot tempa řeči. ....	54
Tabulka 19: Tabulka výsledných hodnot artikulačního tempa. ....	55

# 1. Úvod

---

## 1.1. Schizofrenie

### Definice a charakteristiky

Schizofrenie je psychotická porucha, mezi jejíž hlavní symptomy patří poruchy řeči [1], [2]. Tato porucha postihuje skoro 1 % světové populace [3]. Po století studia schizofrenie zůstává její přímá příčina neznámá. Mezi dosud identifikované mírné vlivy prostředí se řadí několik prenatalních a perinatálních faktorů a těmi jsou podvýživa matky během těhotenství, infekce ve druhém trimestru těhotenství, perinatální poranění či expozice cytokinům [1]. Dalším možným neurovývojovým modelem schizofrenie je přítomnost spouštěče a následné latentní periody, při které stále dochází k vývoji mozku a tato kombinace v budoucnu způsobí vznik psychózy v pozdní adolescenci nebo rané dospělosti. Jedním ze scénářů je přítomnost léze v raném stádiu neuronálního vývoje, kdy následující změny v rámci kompenzace léze již nestačí. Druhým, vzájemně se nevylučujícím, scénářem je, že vývojová léze ovlivňuje dráhu nebo regulační proces například u excitačních a inhibičních synapsí v prefrontální kůře, které mohou mít jen jemné účinky, dokud není vyžadována absolutní a přesná rovnováha, která je nezbytná v průběhu dospívání. [1]

### Příznaky a projevy

Schizofrenie je charakterizovaná souborem symptomů neznámé etiologie, převážně definovaná pozorovanými známkami psychózy [1]. Ve své nejběžnější formě se schizofrenie projevuje paranoidními bludy a sluchovými halucinacemi zejména v dospívání nebo v rané dospělosti [1]. Pro okolí se projevy schizofrenie dají pozorovat například právě kvůli narušené řeči [2]. Obecně se příznaky schizofrenie dělí na pozitivní a negativní (viz Tabulka 1). Pod pozitivním příznakem, resp. patologickým excesem, si můžeme představit stav, kdy oproti běžnému stavu člověku něco přebývá. Těmito příznaky jsou například bludy, sluchové, zrakové či jiné halucinace, zmatené myšlení a řeč nebo neúčelné jednání. Naopak mezi negativní příznaky, resp. patologické deficity, se řadí ty příznaky, které výrazně zhoršují kvalitu života a zároveň se také hůře léčí. Mezi patologické deficity se řadí abulie, alogie, anhedonie, asocialita, otupělost či oploštělé emoce (termíny jsou blíže vysvětlené v Tabulka 1). [4], [5]

**Tabulka 1:** Tabulka dělení příznaků schizofrenie na pozitivní a negativní.

Inspirováno podle [4], [5].

Dělení příznaků	Projevy	Příznaky
Pozitivní příznaky	Člověku oproti běžnému stavu něco přebývá	Bludy (mylná, neviditelná přesvědčení)
		Sluchové, zrakové a další halucinace
		Zmatené myšlení a řeč
		Neúčelné jednání
Negativní příznaky	Výrazně zhoršují kvalitu života	Abulie (nedostatek vůle, ztráta motivace)
		Alogie (ochuzení řeči)
		Anhedonie (neschopnost prožívat radost)
	Hůře se léčí	Asocialita (neschopnost navazovat vztahy)
		Otupělost
		Oploštělé emoce (plochý afekt)

## Epidemiologie, diagnostika a léčba

Léčba schizofrenie je zejména farmakologická. Někteří pacienti ovšem na léčbu nereagují a čelí chronickým problémům, třeba z důvodu nepřesné diagnostiky, kde by právě mohly pomoci řečové biomarkery. Ke schizofrenii se tedy v posledních letech začalo přistupovat jako k neurovývojové poruše s psychózou jako pozdním stádiem onemocnění, kterému lze předcházet. Toto přehodnocení může přinést novou naději na prevenci a léčbu. [1]

Pozdní stádium nemoci, resp. psychóza, se téměř vždy projevuje v pozdní adolescenci nebo v rané dospělosti a vrcholí mezi 18. – 25. rokem, kdy se prefrontální kortex stále vyvíjí [1], [2]. Neurozobrazovací metody dokazují změněné frontotemporální sémantické a fonologické sítě u pacientů trpících schizofrenií [2]. Toto zahrnuje také abnormality ve struktuře Brokova centra, které je v dnešní době spíše spojováno se zodpovědností za gramatiku, morfologii a syntaxi jazyka [6], [7], ve struktuře Wernickeho centra, které ovlivňuje a zpracovává spíše lexikální a sémantické jevy [8], a ve struktuře dalších frontotemporálních regionů [2]. Lze pozorovat abnormality ve traktech bílé [2] a šedé [9] hmoty mozkové a změněné aktivační vzorce v fMRI (funkční magnetická rezonance) v různých jazykových úkolech [1], [2].

## Typy schizofrenie

Podle příznaků (viz Tabulka 1) se obecný pojem schizofrenie dále rozlišuje na šest podtypů (viz Tabulka 2). Prvním a nejčastějším typem je paranoidní schizofrenie, která se často projevuje bludy a halucinacemi. Hebefrenní schizofrenie se projevuje zejména u osob od 15.-25. roku věku. Často se její příznaky připisují a zároveň přirovnávají k opožděné pubertě. Tento typ se projevuje nezodpovědným chováním, hloupým vtipkáním, používáním vulgarismů a zanedbáváním svých povinností. Vzácným typem schizofrenie je katatonní schizofrenie, jejímž příznakem jsou výrazné psychomotorické poruchy. Simplexní schizofrenie vede k uzavřenosti, ke ztrátě zájmu o svět a k neschopnosti prožívání radosti, ke ztrátě vůle. Chronickou formou doprovázenou atakami a zhoršujícími se negativními příznaky je reziduální schizofrenie. Poslední typem schizofrenie je nediferencovaná schizofrenie, která se projevuje kombinací příznaků různých typů. [10]

**Tabulka 2:** Tabulka přehledu typů schizofrenie s typickými projevy.

Inspirováno podle [10].

Typ schizofrenie	Projevy
Paranoidní schizofrenie	Bludy (představy o pronásledování či ohrožování) Halucinace (zejména sluchové vjemy, které mají charakter hlasů)
Hebefrenní schizofrenie	Zejména u osob 15.-25. roku věku Lze přirovnat k opožděné pubertě – nezodpovědné chování, hloupé vtipkání, použití vulgarismů, zanedbávání svých povinností
Katatonní schizofrenie	Vzácná forma Porucha psychomotoriky
Simplexní schizofrenie	Vede k uzavřenosti, ztráta zájmu o svět, neschopnost prožívat radost, ztráta vůle
Reziduální schizofrenie	Chronická forma doprovázená atakami a zhoršujícími se negativními příznaky
Nediferencovaná schizofrenie	Kombinace příznaků různých typů

## 1.2. Motivace a cíle práce

Je potřeba najít objektivní kvantitativní biomarker, který by pomohl diagnostickému nebo prognostickému procesu. Krevní a neurozobrazovací biomarkery nedosahují klinicky použitelné úrovně [2]. Jejich diagnostická přesnost se pohybuje mezi 60-90 % [2]. Nedávné pokroky v počítačové lingvistice dovolují lékařům obrátit se na jazykový výstup jako na nový biomarker [2]. Výhody řeči jako biomarkeru jsou časová efektivnost, neinvazivnost a nízká cena celého procesu. Další výhodou mluveného jazyka jako biomarkeru, oproti klasickým přístupům jako jsou krevní či zobrazovací testy, je jeho reprodukovatelná kvantifikovatelnost [2]. Metody počítačové lingvistiky mohou poskytnout doplňující informace během klinického rozhodování, protože psychiatři nejsou vždy schopni zaznamenat všechny jazykové rozdíly přítomné v mluveném projevu pacienta [11]. Zlepšení schopnosti předvídat psychózu u vysoce rizikových populací by mělo důležité důsledky pro včasnou diagnostiku, což by potenciálně mohlo změnit život lidí s nově se objevujícími psychotickými poruchami [12].

Detailní cíle této diplomové práce byly popsány v zadání diplomové práce a jsou následující:

- I. Provést rešerši literatury týkající se projevů schizofrenie v řečovém signálu se zaměřením na sémantické charakteristiky.*
- II. Pro experimenty použít řečové signály pacientů trpících schizofrenií poskytnuté Národním ústavem duševního zdraví v Klecanech.*
- III. Analyzovat vybrané charakteristiky v programovém prostředí Matlab.*
- IV. Vyhodnotit možnosti detekce onemocnění schizofrenií pro český jazyk.*

## 2. Teoretická část

---

### 2.1. Charakteristika jazyka a promluvy pacientů se schizofrenií

Motivací pro analýzu sémantiky jazyka schizofrenických pacientů je výskyt narušené jazykové komunikace [12], [13]. Obecně pacienti se schizofrenií vykazují velkou škálu sémantických poruch zpracování jazyka [2], [14]. Potíže pacientů se schizofrenií se objevují i při rozhovoru či vyprávění, kdy mají problém s koherencí projevu i s rozhodováním [2], [13], s vyhledáváním slov a s tvořením asociací [2]. Při rozhovoru či vyprávění schizofrenici typicky zapomenou zmínit podstatné informace k pochopení sdělení, nebo naopak zmatou celkové vyznění [2]. U schizofrenických pacientů lze také pozorovat zhoršenou verbální fluenci [13], [15]. Podle studie v článku [16] dělali schizofreničtí pacienti více morfologických chyb než zdravé kontroly. Promluvy pacientů se mohou projevit chudostí obsahu [17], nižší sémantickou hustotou, kdy se v řeči objevují buďto samá slova bez obsahu (např. vycpávková slova), nebo opakováním slov či velkým výskytem pauz [15]. Z hlediska syntaxe či gramatiky můžeme u schizofreniků pozorovat redukovanou syntaktickou složitost [2], resp. méně složitá souvětí s méně závislými a vnořenými větami, což souvisí s použitím méně spojovacích prostředků, resp. spojek.

Řeč pacientů se schizofrenií obvykle není plynulá [15]. Schizofrenie se v řeči může projevovat i hezitací [15], která se projeví v počtu a délce pauz [9]. Dále lze také u pacientů se schizofrenií pozorovat sníženou rychlost řeči a artikulace [18].

Jedním ze symptomů schizofrenie je afektivní zploštění, definované jako redukce rozsahu a intenzity emocionálního projevu [19], [20]. Toto snížení emocionálního projevu se projevuje ve výrazu obličeje, tónu hlasu, řeči těla, ale i v očním kontaktu [9]. U pacientů se schizofrenií lze také pozorovat zhoršenou identifikaci emocí a jejich zpracovávání [20]. V některých případech se můžeme setkat s jevy, kdy samotná mluva se jeví klasicky, ovšem obsah či intenzita neodpovídá adekvátní reakci [20]. Setkáváme se i s definicí abnormální myšlenky či emoce v normální řeči [5].

V následujících kapitolách se budeme teoreticky věnovat tématům jako je lexikální sémantika, komputační a korpusová lingvistika, analýza emocí a sentimentu, zpracování přirozeného jazyka, latentní sémantická analýza nebo prozódie, a jejich souvislostmi s možným využitím k diagnostice schizofrenie.

Zpracování přirozeného jazyka (Natural Language Processing) je název pro interdisciplinární obor, který propojuje informatiku a lingvistiku. Zpracováním přirozeného jazyka se zabývá vědní obor komputační lingvistika [21]. Komputační lingvistika kombinuje metody zpracování přirozeného jazyka s korpusovou lingvistikou a prostředky jako je logika, algebra, statistika a informatika [21]. Dalším často využívaným typem jazykové analýzy je kvantitativní lingvistika, která se soustřeďuje na četnost jevů v řeči [22].

## 2.2. Analýza sentimentu a emocí

Analýzu sentimentu i emocí je možné realizovat třemi možnými technikami: použitím jazykových korpusů, pomocí strojového učení a pomocí hlubokého učení. V následujících kapitolách bude blíže popsána technika založená na použití rozsáhlých jazykových korpusů, která byla zvolena pro tuto práci.

### Sentiment

Analýza sentimentu je definována jako proces získávání smysluplných informací a sémantiky z textu k určení postoje pisatele nebo mluvčího, jehož postoj může být pozitivní, negativní či neutrální [23]. Sentiment je možné hodnotit běžnou kategorizací na pozitivní/negativní (případně neutrální), ale také méně obvyklými škálami, kdy jsou slova rozdělena do kategorií souhlas/nesouhlas, dobré/špatné, anebo mohou být rozdělena například do pěti kategorií: velmi dobré, dobré, uspokojivé, špatné, velmi špatné [24].

Jednou z možností analýzy sentimentu je analýza založená na použití korpusů jazyka. V této metodě používáme rozsáhlé korpusy, ve kterých je slovům přiřazena polarita, např. pozitivnost/negativnost. Sentiment věty, textu nebo dokumentu se dále určí jako suma nebo průměr z detekovaných polarit. [23] Pro český jazyk byl například na Matematicko-fyzikální fakultě Karlovy univerzity v Ústavu formální a aplikované lingvistiky vytvořen lexikon SubLex 1.0 [25], který obsahuje 4625 slov s přiřazenou polaritou – pozitivní/negativní. Tento lexikon kromě informace o polaritě také informuje například o negaci slova nebo o jeho základním tvaru – lemma.



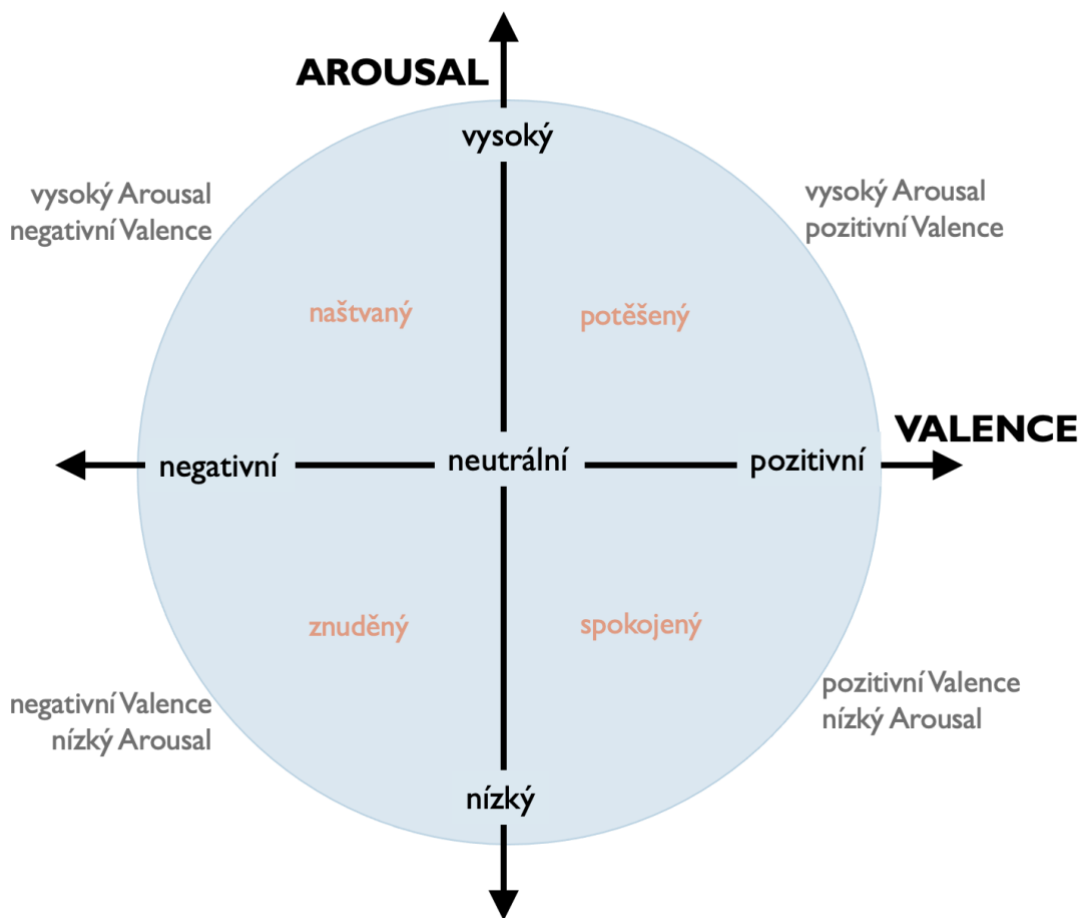
## Emoce

Pro analýzu emocí byl vybrán korpusový přístup. Dvěma nejvíce užívanými lexikony pro detekci emocí jsou WordNet-Affect [26], který se skládá z afektivních slov a emocí k nim přiřazených, a NRC word-emotion lexicon [27], ve kterém je ke každému slovu přiřazena polarita a emoce. Emoce bývají nejčastěji kategorizovány do osmi základních emocí, kterými jsou hněv, očekávání, znechucení, strach, radost, smutek, překvapení a důvěra. Oba zmíněné lexikony WordNet-Affect a NRC word-emotion lexikony jsou kategorické, ovšem zanedbávají informaci o intenzitě emocí, která daná slova vyjadřují. Proto jako rozšíření jednoho z nich byl vytvořen lexikon NRC-Valence arousal dominance [28], který je dimenzionálním lexikonem a doplňuje informaci o valence, arousal a dominanci emoce obsažené ve slovech. Tento lexikon byl přeložen do 108 jazyků včetně češtiny. Obsahuje v sobě hodnoty pro dimenze valence, arousal a dominance pro 20 000 slov. Hodnoty pro tato tři kritéria jsou mezi 0-1 a každé slovo je definováno pomocí číselné hodnoty v každé dimenzi. Pro češtinu obsahují největší informaci kritéria valence a arousal, resp. polarita a intenzita. Ukázkou českých slov a jim příslušné hodnoty z NRC Valence arousal dominance lexikonu je možné vidět v Tabulka 3. Vyobrazení jednotlivých dimenzí lexikonu je vidět na Obrázek 1.

**Tabulka 3:** Přehled vybraných slov a příslušných parametrů arousal a valence.

Hodnoty arousal a valence vzaty z NRC VAD lexikonu [28]. Inspirováno v [28].

<b>Dimenze</b>	<b>slovo</b>	<b>nejvyšší skóre</b>	<b>slovo</b>	<b>nejnižší skóre</b>
<b>Arousal</b>	únos	0.990	měkkost	0.086
	exorcismus	0.980	gauč	0.082
	noční můra	0.942	siesta	0.046
<b>Valence</b>	láska	1.000	nebezpečný	0.020
	radost	0.980	bolest	0.012
	svoboda	0.969	noční můra	0.005



**Obrázek 1:** Dimenzionální 2D model emocí dělený na čtyři kvadranty.

Tento model reprezentuje emoce v závislosti na dvou parametrech – arousal a valence. Uprostřed každého kvadrantu je umístěna emoce, která tuto pozici reprezentuje. Inspirováno v [29].

## 2.3. Zpracování přirozeného jazyka

### 2.3.1. Latentní sémantická analýza

Sémantika je obor, který se zabývá výzkumem významu slov a vyšších jednotek jazyka, resp. slovních spojení, frází, vět [30]. Sémantika může být zkoumána pomocí modelů sémantického prostoru. Jedním z nich je Latentní sémantická analýza (dále jen LSA) [31]. LSA je vysokodimenzionální asociativní analýza sémantické struktury řeči [31]. LSA umožňuje vypočítat podobnosti mezi prvky jazyka na základě vzorců společného výskytu slov ve velkých korpusech [32]. Cílem LSA je zachytit význam slova pomocí vektoru v sémantickém prostoru. Vzdálenost mezi slovy v sémantickém prostoru udává, jaký vztah mají daná slova mezi sebou, jak jsou si příbuzná. [32]

V rámci LSA je možné hodnotit koherenci slov, frází či textů. V praxi to znamená, že například věta, která má nízkou koherenci bude složená ze slov, která reflektují více oddělené pojmy. [2] Sémantické a zároveň i syntaktické informace dokáží z textu zachytit například modely a metody jako Word2vec [33]. Word2vec je populární metodou využívající neuronové sítě k vytváření distribuovaných reprezentací slov (tzv. word embeddings), která k procesu učení používá velké korpusy dat. Po natrénování je Word2vec schopno vytvořit vektorovou reprezentaci slova v prostoru na základě jeho významu a sémantické podobnosti. Tato metoda vznikla v roce 2013 a jejím autorem je Tomáš Mikolov ve spolupráci se společností Google [34]. Pro český jazyk je možné použít metodu fastText, která je rozšířením metody Word2vec, a vznikla pod záštitou Facebook AI Research [35]. FastText je schopen vytvořit vektory pro slova, která se přímo v korpusu (v trénovací množině) nevyskytují [36]. Vektory pro tato slova jsou vytvořené z vektorů podřetězců znaků, které obsahují [36]. Kromě vlastního natrénování těchto modelů na vlastních datech je také možné použít již předtrénované verze, ke kterým jsou používány rozsáhlé korpusy, např. Wikipedie. Dalšími možnými metodami a modely jako je Word2vec jsou například GloVe (Global Vectors of Word Representations), BERT (Bidirectional Encoder Representations from Transformers) nebo ELMo (Embeddings from Language Models) [11], [37]. Každý z těchto zmíněných modelů nebo metod je vhodný pro jiné specifické použití.

Pro diagnostiku schizofrenie jsou metody využívající LSA často doplňovány o parametry kvantitativní lingvistiky. Například v článku [2] byl vytvořen klasifikátor, který k odlišení skupiny pacientů a kontrol používá kombinaci hodnocení koherence spontánní promluvy, normalizované četnosti výskytu determinantů (např. to, co, cokoli, který, jakýkoli) a maximální délky fráze spontánní promluvy. V článku [13] bylo zjištěno, že kosinová podobnost mezi po sobě následujícími slovy při testech verbální fluence je nižší pro pacienty se schizofrenií než pro kontroly. V práci [38] byl nalezen rozdíl mezi pacienty se schizofrenií a kontrolními subjekty

ve snížené sémantické koherenci, větší varianci v této koherenci a sníženém používání přivlastňovacích zájmen.

## 2.4. Prozódie

Dalším možným hodnocením projevů schizofrenie jsou prozódické parametry. Alogie, neplynulost řeči a hezitace patří mezi negativní symptomy schizofrenie. Dalším projevem může být i snížení rychlosti řeči, případně artikulace. V článku [15] bylo zmíněno, že klinicky vyhodnocené faktory alogie a plochého afektu silně korelují s délkou pauz a hezitací. V jiné práci [19] byla nalezena souvislost mezi neplynulostí řeči a morfologickými chybami v řeči. Také v článku [9] bylo demonstrováno, že délka pauz v promluvách a jejich průměrná délka je vyšší u pacientů než u zdravých kontrol a počet pauz dosahuje vyšších hodnot pro skupinu pacientů oproti kontrolám.

V této práci byly jako možné parametry k rozlišení skupiny zdravých kontrol a pacientů navržené počet a délka detekovaných pauz v promluvách (okno bez detekce řeči delší než 30 ms), celková délka nahrávek a řečové a artikulační tempo. Tempo řeči je definováno jako počet slabik za jednotku času a zahrnuje i pauzy a hezitace. Artikulační tempo definuje počet slabik za jednotku času stejně jako tempo řeči, ovšem bez zahrnutí délek pauz a hezitací. Tyto parametry mohou být z řečových záznamů vyhodnoceny manuálně, ale takový postup je pro analýzu více nahrávek časově velice náročný. Proto lze k analýze těchto parametrů použít řadu sofistikovaných algoritmů k tomu určených. Jedním z nich je také algoritmus použitý v této práci, který je blíže popsán v Kapitola 3.7.

## 3. Metodika

---

### 3.1. Databáze promluv

Databáze pro tuto práci vzniká ve spolupráci s Národním ústavem duševního zdraví v Klecanech a obsahuje záznamy šesti pacientů se schizofrenií (pět mužů a jedna žena) a šesti kontrolních subjektů (tři muži a tři ženy). Tato databáze vzniká od roku 2022. Sběr dat byl schválen Etickou komisí Národního ústavu duševního zdraví a každý subjekt podepsal informovaný souhlas s účastí na výzkumu. Mateřským jazykem všech osob byla čeština. Všechny provedené analýzy byly vybrány a následně provedeny nezávisle na pohlaví z důvodu nerovnoměrného zastoupení mužů a žen ve skupinách kontrol a pacientů. Databáze kontrolních subjektů dodaná z Národního ústavu duševního byla mnou doplněná o nahrávky dalších tří kontrolních subjektů, aby byl počet subjektů stejný jako ve skupině pacientů, tedy šest pacientů a šest kontrol.

Nahrávání probíhalo v tiché místnosti podle schváleného protokolu pro tento výzkum (viz Příloha B). Při nahrávání byl použit kondenzátorový hlavový mikrofon Bayerdynamic Opus 55 v 16bitovém režimu se vzorkovací frekvencí 48kHz.

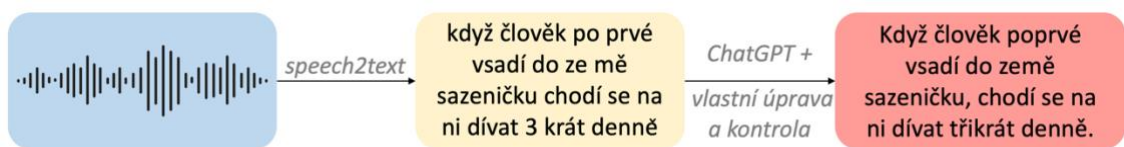
**Tabulka 4:** Přehled úloh protokolu k výzkumu pacientů se schizofrenií.

Kompletní přehled protokolu uveden v Příloha B.

<b>Úloha 1</b>	<b>Čtený text (akustická analýza)</b> - hlasité čtení emočně neutrálního textu
<b>Úloha 2</b>	<b>Spontánní promluva (prozódická informace)</b> - převyprávění pohádky o Popelce dle předloženého textu, cca 90 s
<b>Úloha 3</b>	<b>Spontánní promluva (emočně neutrální)</b> - odpověď na otázku - „Popište byt nebo dům, ve kterém bydlíte.“, cca 90 s
<b>Úloha 4</b>	<b>Spontánní promluva (vyvolání emocí)</b> - popis obrázku - vytvoření příběhu a detailní popis obrázku, který pochází ze standardizovaného projektivního psychologického testu TAT (Tematický apercepční test), cca 90 s

## 3.2. Předzpracování řečových nahrávek

Pro analýzu sémantiky a emocí bylo potřeba získat přepis promluv. Nahrávky pacientů a kontrol byly pomocí algoritmu společnosti Google - speech2text - převedeny do psané formy. Pomocí ChatGPT byl text rozdělen na věty a byla doplněna interpunkce. Jako poslední krok byla provedena manuální kontrola s nahrávkami. Dále bylo sjednoceno označení čísel, aby každé číslo bylo psané jako slovo, a ne pomocí číslovky. Pokud některý ze subjektů použil zkratku místo celého názvu, zůstala použita zkratka přesně tak, jak ji subjekt řekl. Upravení bylo inspirováno podle [39]. Ilustrace procesu transkripce je zobrazena v Obrázek 2.



**Obrázek 2:** Ilustrace postupu získání transkriptů zvukových nahrávek řeči.

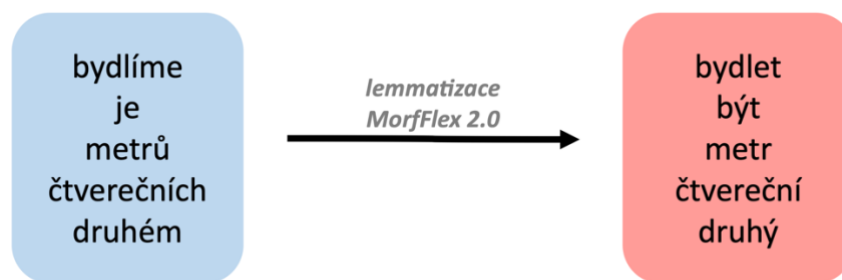
V prvním kroku byl použit algoritmus společnosti Google, speech2text. Dále byl použit ChatGPT k doplnění interpunkce a rozdělení textu na věty a zároveň byla provedena vlastní kontrola se zvukovou nahrávkou.

Dále pro lexikální analýzu sentimentu bylo potřeba provést lemmatizaci transkriptů. Ilustrační ukázka procesu viz. Obrázek 3. Pro tento krok byl použit volně dostupný korpus MorfFlex CZ 2.0 [40] vytvořený Ústavem formální a aplikované lingvistiky na Karlově univerzitě v Praze. Tento korpus obsahuje více než 125 milionů slov a jejich tvarů, proto byla lemmatizace textu provedena pomocí MetaVO<sup>1</sup>, virtuální organizace české Národní Gridové Iniciativy, kde všichni akademičtí pracovníci a studenti registrovaní na této stránce mají možnost bezplatného využití výpočetní a úložní kapacity a řady aplikačních programů, např. Matlab.

U zvukových nahrávek byly odstraněny neřečové části na začátku a konci každého souboru, aby bylo zamezeno artefaktům a falešným detekcím.

---

<sup>1</sup> <https://metavo.metacentrum.cz>



**Obrázek 3:** Ilustrace procesu lemmatizace transkriptů.

Lemmatizace byla provedena pomocí morfologického slovníku MorfFlex 2.0. [40]

### 3.3. Vývojová prostředí pro analýzy

#### Matlab

Analýzy v implementační části této práce byly provedeny v prostředí Matlab. Matlab je programovací jazyk a interaktivní prostředí vhodné pro numerické výpočty [41]. Matlab je také vhodný pro tvorbu algoritmů a vizualizaci dat [41].

#### PRAAT a Parselmouth

K analýze dat v oblasti fonetiky se používá PRAAT, rozsáhlý softwarový balík, který fonetici využívají běžně ke každodenní práci [42]. PRAAT pracuje například jako v našem případě se soubory s příponou *.wav*. O souboru můžeme získat jakékoliv akustické nebo artikulační parametry.

Jako doplněk PRAAT byla vytvořena open-source knihovna jazyka Python, Parselmouth [42]. Díky této knihovně je možné získat hlavní funkce PRAAT v jazyce Python. Tato kombinace je benefitem z důvodu možnosti kombinace sofistikovaných analýz díky PRAAT a rozsáhlých implementačních možností jazyka Python.

### 3.4. Statistický test

Ke statistickému vyhodnocení analýz byl v této práci použit nepárový neparametrický Mann-Whitney test [43]. Tento test byl zvolen z důvodu nesplnění normálního rozdělení dat ve skupinách subjektů v tomto výzkumu. Nepárový test byl zvolen kvůli složení datasetu, který neobsahuje opakovaná měření jedné úlohy, a tedy data ve skupinách pacientů a kontrol jsou nezávislá. Pro tento test byla v Matlabu použita funkce *ranksum*. Hladina významnosti byla standardně zvolena 5 %.

V následujících kapitolách je ke každému testování uvedena nulová hypotéza  $H_0$  definovaná za předpokladu, že distribuce kontrolní skupiny a skupiny pacientů se rovnají. Kompletní přehled nulových hypotéz definovaných pro tuto práci je uveden v Příloha C.

$H_0$ : *distribuce v obou skupinách jsou identické*

Případně při zamítnutí nulové hypotézy je formulována alternativní hypotéza.

$H_A$ : *distribuce v obou skupinách nejsou identické*

Výsledky budou v následujících kapitolách uváděné ve tvaru:

*střední hodnota  $\pm$  směrodatná odchylka (medián).*

K zobrazení výsledných distribucí pro testy skupin pacientů a kontrol budou použity krabicové grafy.



## 3.5. Lexikální sémantika

### Frekvence výskytu vycpávkových slov

Analýza frekvence výskytu vycpávkových slov byla provedena za předpokladu, že pacienti trpící schizofrenií čelí problémům s nalezením správných slov [2], často mají chudší aktivní slovní zásobu [5], a proto si mohou napomáhat vycpávkovými slovy [44]. Jejich časté použití v promluvách může indikovat nízkou sémantickou hustotu řeči [11]. Po rešerši byla jako relevantní vybrána tato vycpávková slova: *jakoby, vlastně, prostě, jako, takový/á/é, jaksi, takže*. Tato slova byla vyhledána ve výpovědích subjektů a následně byla analyzována jejich relativní četnost v promluvách. Byly vygenerovány boxploty a proveden Mann-Whitney U-test.

Pro tuto analýzu byly vybrány úlohy založené na spontánní řeči, tedy úlohy 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Pacienti i kontrolní skupina používají vycpávková slova se srovnatelnou četností.*

### Frekvence použití spojek

Analýza frekvence použití spojek a předložek proběhla za předpokladu, že pacienti trpící schizofrenií mají potíže s tvořením komplikovanějších vět, používají méně závislých a vnořených vět [2], [44]. V jejich promluvách pozorujeme sníženou syntaktickou složitost [2], [44]. Oproti kontrolním subjektům tedy existuje předpoklad nižšího užití spojovacích prostředků – spojek [2], [44]. V textu byly proto vyhledány výskyty spojek podle seznamu spojek, který byl vytvořen pomocí dokumentace k Prague Dependency TreeBank 2.0, jež je volně dostupný online [45]. Opět byla vyhodnocena relativní četnost použití tohoto slovního druhu v promluvách. Byly vygenerovány boxploty proveden Mann-Whitney U-test.

Pro tuto analýzu byly vybrány úlohy založené na spontánní řeči, tedy úlohy 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Syntaktická složitost promluvy pacientů a kontrolní skupiny je stejná.*

## Frekventovanost použitých slov

Tato analýza proběhla s cílem zjistit, jak četná slova účastníci se subjekty volí ve svých promluvách, zda používají běžná slova, či používají slova více raritní, neběžná. Předpokladem tohoto testu byl fakt, že schizofrenní pacienti používají méně častá, méně frekventovaná slova [2]. K této analýze byl použit korpus českého jazyka SYN2015, volně dostupný online [46], kde jsou slova seřazena podle frekventovanosti použití v řeči v populaci. K analýze byla použita jedinečná slova (seznam slov bez opakování). Každé jedinečné slovo z promluv pacientů bylo tedy nalezeno v korpusu SYN2015 a byla uložena pozice, na které se v korpusu nachází. Z takových dat byl vypočten průměrný index pro každý subjekt a každou úlohu zvlášť. Z těchto výsledných hodnot byly vygenerovány boxploty a proveden Mann-Whitney U-test.

Pro tuto analýzu byly vybrány úlohy založené na spontánní řeči, tedy úlohy 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Skupina pacientů a kontrol v řeči používá srovnatelně frekventovaná slova.*

## Bohatost jazyka

Tento test proběhl na základě předpokladu o vlivu schizofrenie na slovní zásobu pacientů, kdy se s vlivem působící nemoci ukazuje efekt i na slovní zásobě jedinců [11], [14]. Bohatost byla vyhodnocena v Matlabu a byl analyzován počet slov, kdy bylo do seznamu pro každou promluvu uloženo každé slovo pouze jednou a všechny další výskyty jsou zanedbány, protože nám seznam slov neobohacují. Tento test byl vyhodnocen spočítáním poměru unikátních slov ku celkovému počtu slov v promluvě a tento poměr byl přepočten na procenta. Z těchto dat byly vygenerovány boxploty a proveden Mann-Whitney U-test.

Pro tuto analýzu byly vybrány úlohy založené na spontánní řeči, tedy úlohy 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Skupina pacientů a skupina kontrol používá v řeči stejné procentuální zastoupení unikátních slov.*

## **Emoční analýza**

V této části byla provedena analýza sentimentu a emocí pro český jazyk. Jako hlavní byl pro tuto analýzu vybrán NRC VAD Lexicon [47], který v sobě nese informace pro Valence a Arousal slov, které se pro český jazyk vyhodnocují. Každé slovo v tomto lexikonu obsahuje hodnotu pro Valence, tedy pozitivnost nebo negativnost, a pro Arousal, tedy jak je emoce silně vyjádřená tím daným slovem. Hodnocení slov vyšlo z teorie sémantického diferenciálu [48] a výsledné hodnoty pro Arousal a Valence jsou naškálovány od 0-1, kde hodnota blízká 1 označuje silný výraz emoce či pozitivnosti/negativnosti a hodnota blízká 0 naopak slabou emoci či pozitivnost/negativnost (viz Tabulka 3).

Pro hodnocení Valence byl přidán lexikon SubLex 1.0 [25], kde je kromě jiných hodnotících parametrů také každé slovo v tomto lexikonu označené jako pozitivní či negativní.

Stejný postup byl zvolen i pro hodnocení Arousal slov v transkriptech. K lexikonu NRC VAD byl přidán lexikon NRC Affect Intensities [49], kde je seznam slov a k nim přidaná emoce, kterou vyjadřují. V tomto lexikonu jsou slova rozdělena do 8 základních emocí, kterými jsou hněv, očekávání, znechucení, strach, radost, smutek, překvapení a důvěra.

## **Arousal**

Předpokladem pro tento experiment je zaznamenaná oploštělost vyjádření emocí, objevující se u pacientů trpících schizofrenií [20]. Tedy jedna z hypotéz je, že kontroly budou oproti pacientům vykazovat silnější vyjádření emocí, budou používat slova s větší intenzitou a variabilita jejich emočního vyjadřování bude vyšší. Naopak u pacientů bychom měli pozorovat spíše neutrální vyjádření emocí. Pro tuto analýzu byl použit parametr Arousal z NRC VAD lexikonu [47].

Arousal slov byl hodnocen pro každou úlohu zvlášť. Pro každé slovo z transkriptů, resp. lemma, které v sobě nese emoci, byla nalezena jeho příslušná hodnota v lexikonu a byl vyhodnocen medián hodnot arousal pro každý subjekt a každou úlohu. Dále byla vyhodnocen interkvartilový interval hodnot arousal pro každý subjekt i úlohu. Výsledné hodnoty byly vyobrazené v krabicových grafech a statistické vyhodnocení bylo provedeno prostřednictvím Mann-Whitney U-testu.

Pro tuto analýzu byly vybrány úlohy založené na spontánní řeči, tedy úlohy 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Parametr arousal slov v promluvách pacientů a kontrol je shodný.*

## Valence

Dále byly emoce hodnocené z hlediska celkové polarity emocí. Tedy zda se obě skupiny od sebe odlišují v parametru Valence uvedeného v NRC VAD lexikonu [47].

Valence slov byla vyhodnocena opět pro každou úlohu a pro každý subjekt zvlášť. Slova, resp. lemma, byla nalezena v lexikonu a byla uložena příslušná hodnota. Pro každý subjekt a úlohu byla vyhodnocena mediánová valence a také interkvartilový interval. Výsledky byly zobrazeny pomocí krabicových grafů a statisticky pomocí Mann-Whitney U-testu.

Pro tuto analýzu byly vybrány úlohy založené na spontánní řeči, tedy úlohy 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Parametr valence slov v promluvách pacientů a kontrol je shodný.*

## Kvalitativní zastoupení emocí v úloze vyvolání emocí

Samostatnou kapitolou při kvalitativním hodnocení emocí je analýza pro úlohu 4 – popis obrázku (viz Příloha B), která byla do protokolu umístěna se záměrem vyvolání emocí. Stanovená hypotéza tedy testuje, zda u popisu obrázku, který měl vyvolat spíše negativní emoce, výpověď pacientů odpovídá a zda jejich emoce z hlediska lexikálního hodnocení odpovídají předpokladům a zda se odlišuje reakce skupiny pacientů a kontrol na tento obrázek.

Kvalitativní zastoupení emocí a sentimentu v úloze popisu obrázku bylo provedeno pomocí dvou lexikonů, a to lexikonu SubLex 1.0 [25] a NRC Affect Intensities [49].

Nejprve byla v každém transkriptu identifikována slova se sentimentem, resp. lemma nacházející se v SubLex 1.0. A bylo vyhodnoceno jejich procentuální zastoupení v řeči ku celkovému počtu slov v transkriptu. Tímto byla získána informace o sémantická hustotě každé výpovědi, protože slova označující polaritu výpovědi jsou zároveň i nosiči informací [17]. Dále byl pro každé slovo, resp. lemma, přímo určen sentiment pomocí lexikonu SubLex 1.0. Bylo vyhodnoceno procentuální zastoupení pozitivních či negativních slov ve slovech se sentimentem. Vyhodnocení bylo provedeno pomocí krabicových grafů a Mann-Whitney U-testu.

*H<sub>0</sub>: Procentuální zastoupení slov se sentimentem v promluvách je stejné pro skupinu pacientů a kontrol.*

*H<sub>0</sub>: Procentuální zastoupení slov s pozitivním či negativním sentimentem je shodné pro skupinu pacientů a kontrol.*

Dále byla pomocí lexikonu NRC Affect Intensities přiřazena každému slovu přímo emoce a bylo analyzováno procentuální zastoupení 8 základních emocí (viz Kapitola 2.2) pouze ve slovech, která emoci vyjadřují. Cílem bylo zanalyzovat poměrové zastoupení daných emocí v promluvě.

*H<sub>0</sub>: Procentuální zastoupení jednotlivých emocí v promluvách pacientů a kontrol je shodné.*

## 3.6. Latentní sémantická analýza

### Koherenční model

Sémantika jazyka se dá hodnotit na základě koherence. Předpokladem pro tuto analýzu byla ztráta cíle či pointy v řeči [12], kdy tento jev lze pozorovat u pacientů se schizofrenií jako častý. Dalším znakem, který hodnocení koherence může zachytit je tzv. vykojení [11], což je jev, kdy se objevují přerušování v toku myšlení a subjekt skáče od jednoho tématu k druhému. Daná témata spolu z většiny nesouvisí a nemají žádnou zjevnou spojitost [50].

Pro samotnou analýzu koherence bylo nejprve potřeba upravit podobu transkriptů. Transkripty výpovědí byly rozdělené na věty a dále na slova. Každé slovo bylo potřeba převést na vektorovou reprezentaci. (viz Kapitola 2.3.1). Tento krok proběhl pomocí lexikonu fastText [35], kde ke každému slovu náležel vektor o dimenzi 1x300 a použita byla funkce v Matlabu *word2vec*. Lexikon s předtrénovanými vektory slov je volně dostupný online na webu fastText<sup>2</sup>. Tento předtrénovaný model byl zvolen z důvodu malé velikosti datasetu pro tento výzkum, protože by nebylo možné natrénovat vlastní validní model. Z vektorů slov získaných pomocí fastText byl vytvořen jeden průměrný vektor, který reprezentuje význam celé věty. [12] Ilustrovaný příklad této vektorové reprezentace redukováný do třídímenzionálního prostoru byl vytvořen pomocí statistické metody t-SNE a následně byla slova rozdělena do tří klastrů pomocí metody k-means (viz Přílohy Obrázek 20).

---

<sup>2</sup> <https://fasttext.cc/docs/en/crawl-vectors.html>

## Koherence prvního řádu

Prvním typem hodnocení koherence je koherence prvního řádu (resp. First order coherence – dále jen *FOC*). Pomocí *FOC* se mezi sebou porovnávají věty za sebou následující [12]. Dle vzorce (1) z článku [12] byla tato koherence analyzována.

$$FOC = \langle \cos (V_i, V_{i+1}) \rangle, \quad (1)$$

kde  $V_i$  označuje průměrný vektor reprezentující větu a  $V_{i+1}$  označuje průměrný vektor věty následující.

Pro samotný výpočet koherence byla vybrána kosinová podobnost vypočítaná podle vzorce (2) [11], [12].

$$\text{kosinová podobnost} = \frac{V_i \cdot V_{i+1}}{\|V_i\| \cdot \|V_{i+1}\|} \quad (2)$$

Z výsledných hodnot *FOC* byly analyzovány parametry jako minimum, medián a interkvartilový interval koherencí v každé promluvě každého subjektu. Tyto parametry byly vybrány vzhledem k předpokladům o nižší koherenci a větší varianci koherence v promluvě pacientů se schizofrenií [11], [12], [38]. Pro každý parametr byl proveden U-test a vyobrazení pomocí krabicových grafů.

Pro tuto analýzu byly vybrány úlohy 2, 3 a 4 (viz Příloha B) z důvodu nutnosti spontánní promluvy pro analýzu tohoto typu.

*H<sub>0</sub>: Koherence prvního řádu dosahuje srovnatelných hodnot ve zkoumaných parametrech minima, mediánu a interkvartilového intervalu pro skupinu pacientů a kontrol.*

## Koherence druhého řádu

Druhým typem hodnocení koherence promluvy je koherence druhého řádu (resp. Second order coherence – dále jen *SOC*). Pomocí *SOC* se porovnává podobnost významu vět, které odděluje jedna věta mezi nimi. Dle vzorce (3) z článku [12] byla tato koherence analyzována.

$$SOC = \langle \cos (V_i, V_{i+2}) \rangle, \quad (3)$$

kde  $V_i$  označuje průměrný vektor reprezentující větu a  $V_{i+2}$  označuje průměrný vektor věty o dva indexy v textu dále.

Pro výpočet *SOC* dvou vět byla použita kosinová koherence podle vzorce (2) upravená pro dva vektory vět, které odděluje jedna věta mezi nimi [12]. Z výsledných hodnot *SOC* byly

analyzovány parametry minimum, medián a interkvartilový interval koherencí v každé promluvě každého subjektu. Pro každý parametr byl proveden U-test a vyobrazení pomocí krabicových grafů.

Pro tuto analýzu byly vybrány úlohy 2, 3 a 4 (viz Příloha B) z důvodu nutnosti spontánní promluvy pro analýzu tohoto typu.

*H<sub>0</sub>: Koherence druhého řádu dosahuje srovnatelných hodnot ve zkoumaných parametrech minima, mediánu a interkvartilového intervalu pro skupinu pacientů a kontrol.*

### 3.7. Prozódie

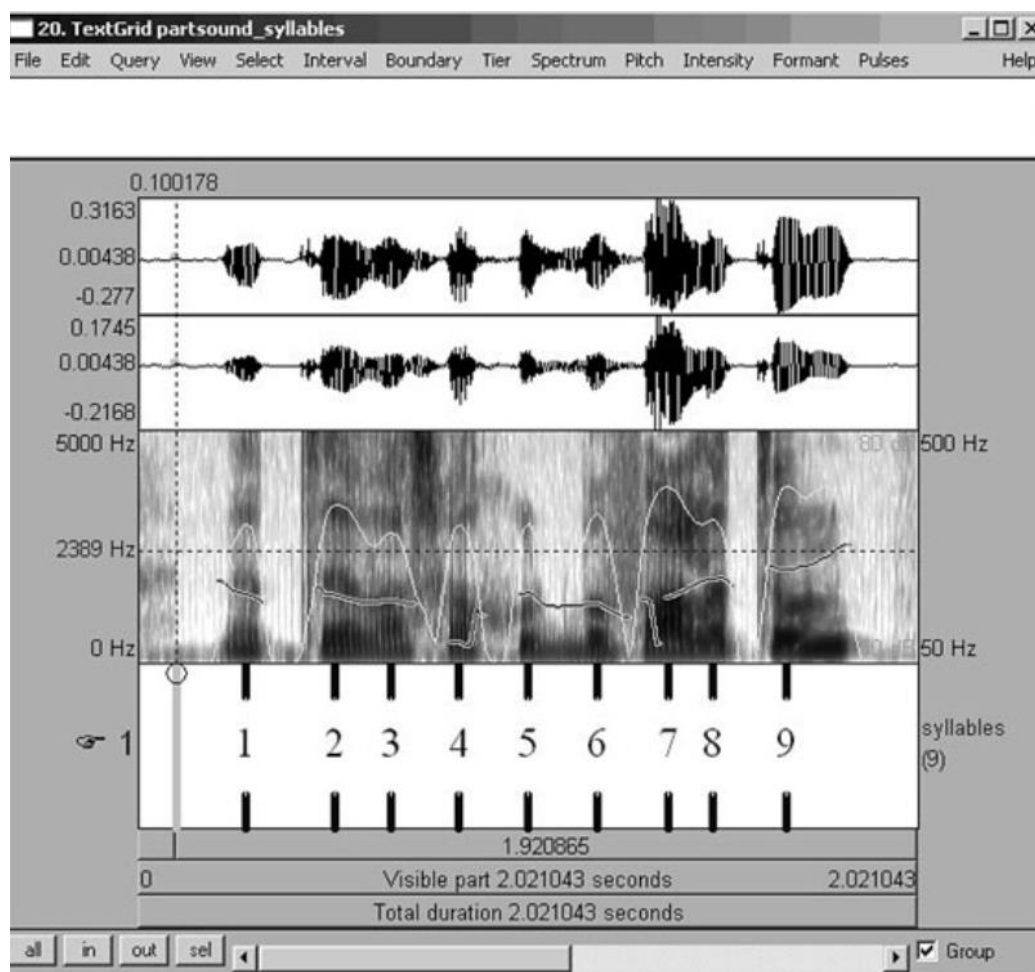
Tato akustická analýza byla provedena za předpokladu, že u pacientů se schizofrenií se v řeči objevují delší pauzy oproti kontrolním subjektům, proto tedy prvním parametrem byla zvolena průměrná délka pauz [9], [15]. Dále byla k analýze vybraná celková délka promluv a poměr celkové délky pauz ku celkové délce nahrávky, tempo řeči a artikulační tempo [9], [15], [19]. K získání těchto parametrů bylo potřeba odlišit řečové a neřečové části nahrávek.

Pro zmíněné akustické analýzy byla použita open-source knihovna jazyka Python, Parselmouth. Tato knihovna umožňuje využívat hlavní funkce, které má aplikace a knihovna PRAAT, která je nejznámějším a nejvyužívanějším nástrojem pro analýzu řečových záznamů [42].

K akustické analýze řečových nahrávek byl použit volně dostupný kód [51], který byl využit k odlišení řečových a neřečových částí nahrávek. Cílem tohoto skriptu je detekovat jádra slabik. Jádra slabik většinou tvoří samohlásky. V signálu jsou hledány znělé úseky, které jsou charakterizované vyšší intenzitou oproti zbytku signálu a je také přítomná základní hlasivková frekvence. Z tohoto důvodu je v tomto algoritmu po segmentech zkoumaný průběh intenzity a jsou hledána lokální maxima překračující práh. Práh je definovaný jako medián všech intenzit. Aby bylo zamezeno falešným detekcím v rámci jedné slabiky, jsou vyloučena všechna lokální maxima, kterým nepředchází pokles intenzity minimálně o 2 dB. Dále je použita znělost, zjištěná pomocí základní hlasivkové frekvence, k vyloučení detekovaných vrcholů, které jsou v neznělých částech signálu. Zbylé detekované vrcholy jsou považovány za jádra slabik a byly uloženy pro další analýzu. [51]

Vzorová ukázka detekovaných jader slabik je vidět na Obrázek 4.

Pomocí tohoto kódu [51] byl automaticky detekován počet pauz v promluvách. Za pauzu byl považovaný každý úsek promluvy, kde nebyla detekovaná řeč a který byl delší než 30 ms. Dýchání je v tomto provedení bráno jako podmnožina pauz.



**Obrázek 4:** Vybraný úsek řeči vyobrazený v PRAAT.

Na ilustraci je zobrazena intenzitou (bílá čára ve spektru) a výškou tónu (tmavá čára ve spektru). Ve spodní části obrázku je vyobrazeno umístění devíti detekovaných jader slabik. Převzato z [51].

Dalším automaticky získaným parametrem byla celková délka řeči, fonace, v nahrávkách. Tento parametr byl použit k získání celkové délky pauz odečtením celkové délky řeči od celkové délky nahrávky.

Protože výše zmíněné parametry celkové délky pauz a počtu pauz nejsou objektivně hodnotitelné pro různé délky nahrávek, byly použity objektivní metriky jako procentuální zastoupení pauz v nahrávce a průměrná délka pauzy, které jsou blíže popsány v následujících kapitolách. Kód pro detekci jader slabik byl tedy následně doplněn o vlastní výpočty parametrů. Výsledné hodnoty byly uloženy pro každou nahrávku a extrahované do prostředí Matlab, kde bylo provedené statistické vyhodnocení.



## **Celková délka nahrávek**

Zkoumaným parametrem byla celková délka nahrávek. Hodnoty celkové délky nahrávek byly získané pomocí algoritmu [51].

Analýza byla provedená pro všechny úlohy – 1, 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Délky nahrávek pro každou úlohu samostatně jsou pro skupiny pacientů a kontrol shodné.*

## **Procentuální zastoupení pauz v nahrávce**

S předchozí kapitolou celkové délky nahrávek souvisí procentuální zastoupení detekovaných pauz v nahrávce. Pomocí algoritmu [51] popsaného výše Kapitola 3.7 byla zjištěna celková délka pauz. Protože pro každý subjekt má nahrávka jinou délku, byla zvolena obecná míra, a to procentuální zastoupení pauz v nahrávce.

Analýza byla provedena pro všechny úlohy – 1, 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Procentuální zastoupení pauz v nahrávce je shodné pro skupiny pacientů a kontrol pro každou úlohu.*

## **Průměrná délka pauzy**

Pomocí algoritmu [51] popsaného v Kapitola 3.7 byla vypočítána průměrná délka pauzy. A to pomocí celkové délky pauz v nahrávce vydělené počtem detekovaných pauz.

Analýza byla provedena pro všechny úlohy – 1, 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Průměrná délka pauzy je pro pacienty a kontrolní skupinu stejná.*

## Tempo řeči

Dalším zkoumaným akustickým parametrem je tempo řeči. Tempo řeči se vypočítá jako počet vyslovených slabik ku celkové délce nahrávky. Parametr počtu slabik byl zjištěn pomocí algoritmu [51] popsaného v Kapitola 3.7.

Analýza byla provedena pro všechny úlohy – 1, 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Tempo řeči je pro pacienty a kontrolní skupinu stejné.*

## Artikulační tempo

Artikulační tempo bylo zvoleno jako další vhodný akustický parametr. K výpočtu artikulačního tempa se používá počet slabik a délka fonace, resp. délka nahrávky bez zahrnutí délky pauz. Data byla opět získána pomocí algoritmu [51] popsaného v Kapitola 3.7.

Analýza byla provedena pro všechny úlohy – 1, 2, 3 a 4 (viz Příloha B).

*H<sub>0</sub>: Artikulační tempo je pro pacienty a kontrolní skupinu stejné.*

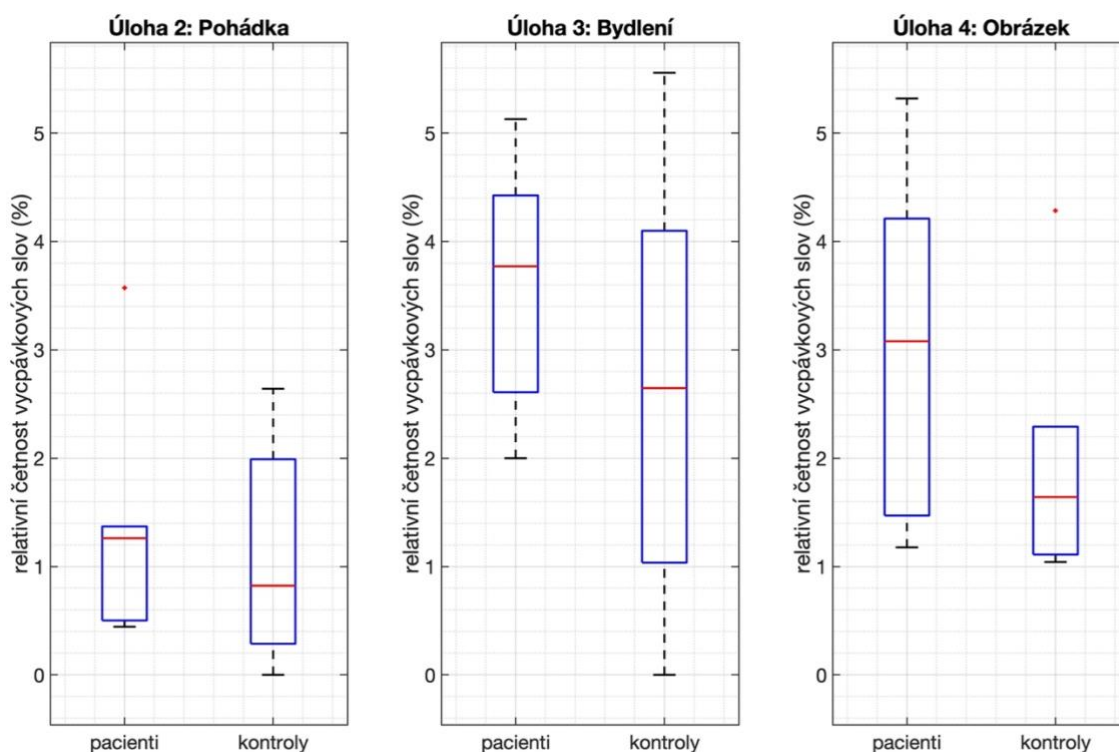
## 4. Výsledky

### 4.1. Lexikální sémantika

#### Relativní četnost vycpávkových slov

Dle kapitoly 3.5 byla vyhodnocena četnost použití vycpávkových slov pro úlohy 2, 3 a 4. Dle výsledných krabicových grafů (viz Obrázek 5) a výsledků U-testu (viz Tabulka 5) lze vidět, že relativní četnost vycpávkových slov se mění v závislosti na úloze. V úloze 4, která měla vyvolat emoční reakci, použili pacienti více vycpávkových slov než kontroly. V úlohách 2 a 3, emočně neutrálních, použily kontrolní subjekty opět v průměru méně vycpávkových slov než pacienti. Žádná výsledná p-hodnota U-testu nepotvrzuje signifikanci těchto trendů ( $p > 0.05$ , U-test).

Nulové hypotézy o shodnosti relativní četnosti v promluvách nebyly zamítnuty.



**Obrázek 5:** Výsledné krabicové grafy relativní četnosti vycpávkových slov.

Pomocí U-testu nebyly mezi distribucemi pro úlohy založené na spontánní promluvě nalezeny signifikantní rozdíly ( $p > 0.05$ , U-test). Průměrné hodnoty i medián relativní četnosti vycpávkových slov je ve všech úlohách vyšší pro skupinu pacientů než pro zdravé kontroly.

**Tabulka 5:** Tabulka vyhodnocení relativní četnosti výskytu vycpávkových slov v promluvách subjektů pro úlohy 2, 3 a 4.

Výsledné hodnoty jsou uvedeny v procentech. Ve všech třech testech vyšla výsledná p-hodnota U-testu nad hranicí 5 % a mezi jednotlivými distribucemi nebyl nalezen signifikantní rozdíl. Průměrné hodnoty relativní četnosti i mediány jsou obecně vyšší pro skupinu pacientů oproti kontrolní skupině.

	<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test (p-hodnota)
<b>Úloha 2:</b> pohádka	1.4 ± 1.1 (1.3)	1.1 ± 1.0 (0.8)	0.5887
<b>Úloha 3:</b> bydlení	3.6 ± 1.2 (3.8)	2.7 ± 2.1 (2.6)	0.3939
<b>Úloha 4:</b> obrázek	3.1 ± 1.6 (2.1)	2.0 ± 1.2 (1.6)	0.3095

## Relativní četnost spojek

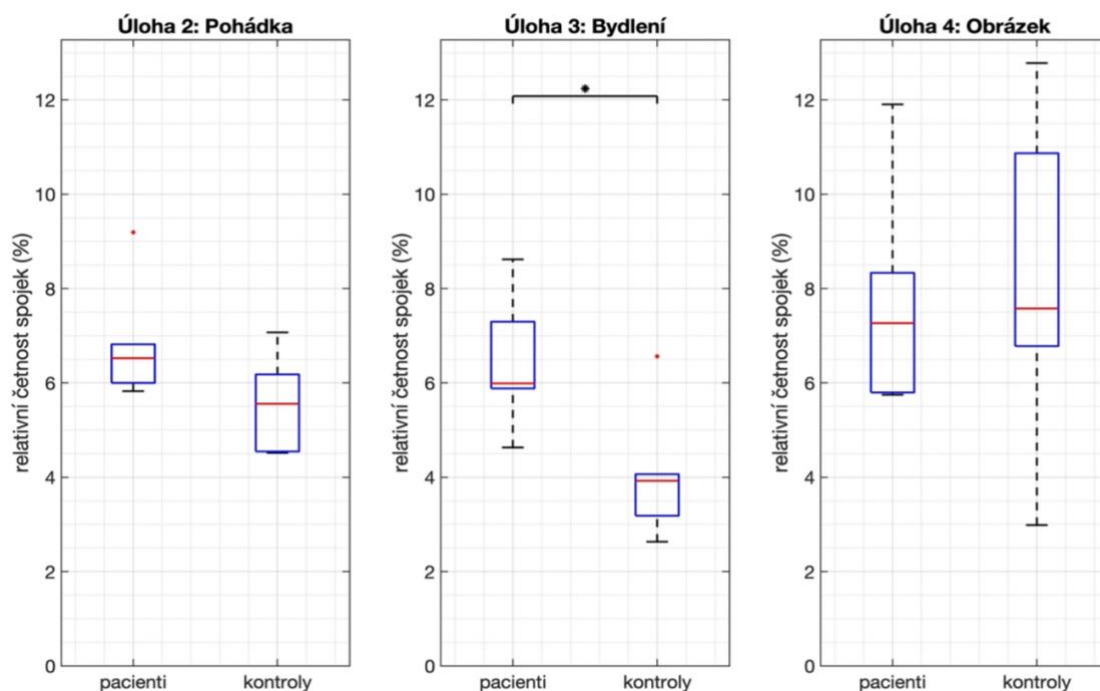
Dle Kapitoly 3.5 byla vyhodnocena relativní četnost spojek. Pro úlohu 2 – pohádku a úlohu 3 – bydlení lze (viz Obrázek 6) pozorovat trend, kdy u distribuce pacientů vidíme obecně vyšší procento spojek v řeči než u kontrol. U úlohy 3, popisu bydlení, je výsledná hodnota U-testu (viz Obrázek 6 a Tabulka 6) signifikantní ( $p < 0.05$ , U-test), a tedy lze pozorovat signifikantní rozdíl mezi skupinou pacientů a kontrol. Pro úlohu 4 – popis obrázku je medián i průměr velmi srovnatelný.

Hypotézy shodnosti relativní četnosti spojek nebyly zamítnuty pro úlohy 2 a 4 a nulová hypotéza byla zamítnuta pro úlohu 3 na 5% hladině významnosti.

**Tabulka 6:** Tabulka relativní četnosti spojek v promluvách pro úlohy 2, 3 a 4.

Výsledné hodnoty jsou uvedeny v procentech. V úloze bydlení je vidět signifikantně vyšší relativní četnost spojek pro pacienty než pro kontrolní skupinu ( $p < 0.05$ , U-test). V testech pro úlohu pohádky a obrázku jsou vidět vzájemně opačné trendy, mezi kterými nebyl nalezen statisticky signifikantní rozdíl ( $p > 0.05$ , U-test). (hvězdička označuje signifikanci hodnot:  $*p < 0.05$ )

	<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test (p-hodnota)
<b>Úloha 2:</b> pohádka	6.8 ± 1.2 (6.5)	5.6 ± 1.0 (5.6)	0.0931
<b>Úloha 3:</b> bydlení	6.4 ± 1.4 (6.0)	4.0 ± 1.4 (4.0)	0.0238*
<b>Úloha 4:</b> obrázek	7.7 ± 2.3 (7.3)	8.1 ± 3.4 (7.6)	0.9372



**Obrázek 6:** Krabicové grafy relativní četnosti spojek v použití pro úlohy založené na spontánní promluvě. V úloze o bydlení vidíme signifikantní rozdíl mezi distribucemi ( $p < 0.05$ , U-test). Ve zbylých dvou úlohách jsou trendy odlišné ve všech zkoumaných parametrech (průměr, směrodatná odchylka a medián).

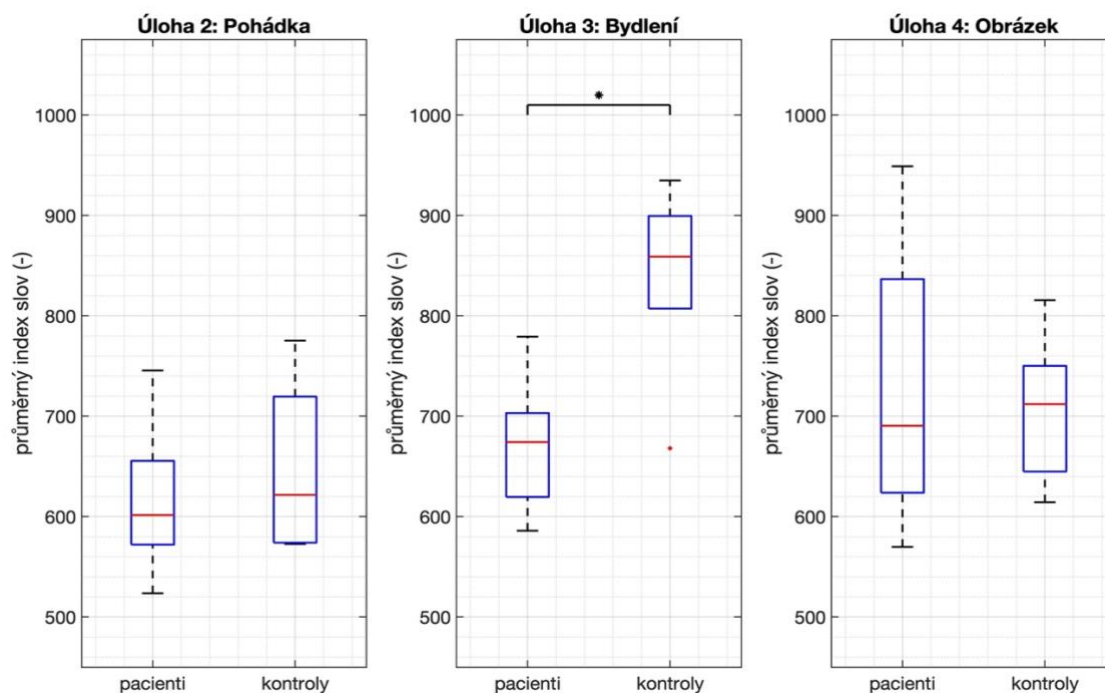
## Frekventovanost použitých slov

Frekventovanost použitých slov byla hodnocena pro všechny transkripty spontánní promluvy, tedy pro úlohy 2, 3 a 4. Mezi distribucemi úloh 2 a 4 nebyl nalezen signifikantní rozdíl ( $p > 0.05$ , U-test) (viz Tabulka 7). U úlohy 3 lze vidět možný trend (viz Obrázek 7), který je statisticky významný na 5% hladině významnosti při použití Mann-Whitney U-testu.

Hypotézy o shodných distribucích frekventovanosti použitých slov nebyly zamítnuty pro úlohy 2 a 4 a hypotéza byla zamítnuta na 5% hladině významnosti pro úlohu 3.

**Tabulka 7:** Tabulka frekventovanosti použitých slov pro úlohy založené na spontánní promluvě. Pro úlohu 3 byla zamítnuta nulová hypotéza o shodnosti distribucí ( $p < 0.05$ , U-test). Zdravé kontroly mají mediány frekventovanosti použitých slov vyšší ve všech úlohách. Výsledné hodnoty jsou uvedeny v indexech pozice slov v lexikonu. (hvězdička označuje signifikanci hodnot:  $*p < 0.05$ )

	Pacienti	Kontroly	Mann-Whitney
	$\mu \pm \sigma$ (median)	$\mu \pm \sigma$ (median)	U-test (p-hodnota)
Úloha 2: pohádka	616.7 $\pm$ 78.6 (601.6)	647.5 $\pm$ 82.9 (621.7)	0.4848
Úloha 3: bydlení	672.4 $\pm$ 68.4 (674.3)	837.9 $\pm$ 95.3 (859.0)	0.0158*
Úloha 4: obrázek	726.7 $\pm$ 142.6 (690.6)	708.2 $\pm$ 72.5 (712.0)	0.9372



**Obrázek 7:** Krabicové grafy frekventovanosti slov pro úlohy založené na spontánní promluvě. Mediány frekventovanosti použitých slov jsou vyšší u kontrolní skupiny pro všechny úlohy. Signifikantní rozdíl mezi distribucemi byl nalezen pro úlohu bydlení ( $p < 0.05$ , U-test).

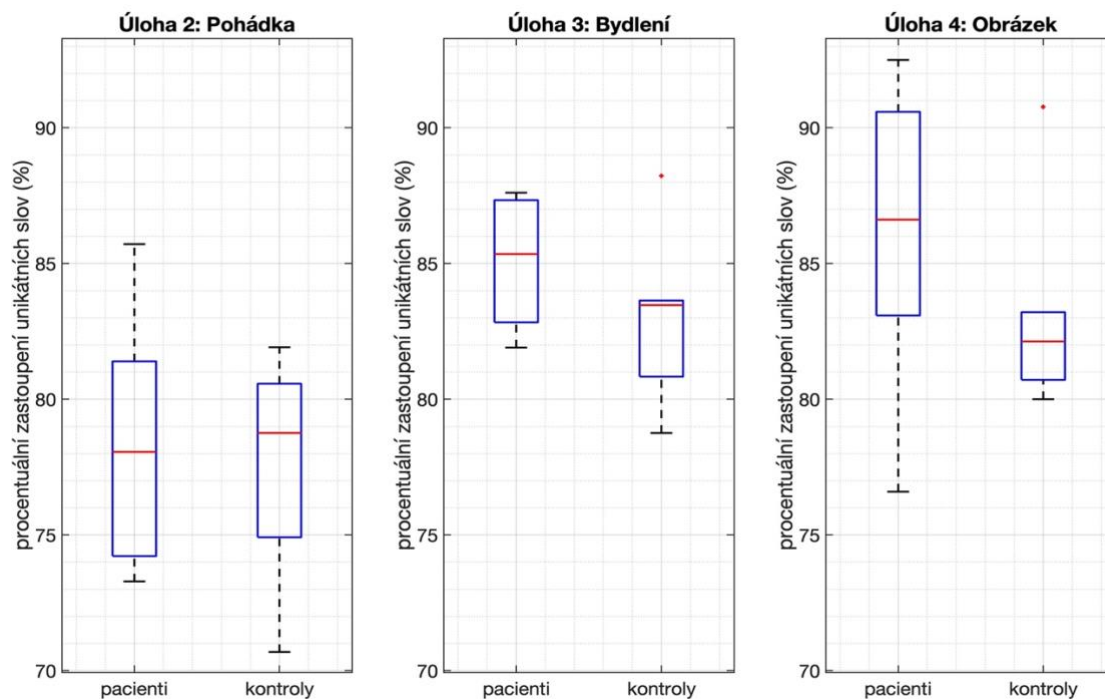
## Bohatost slovní zásoby

Bohatost slovní zásoby byla hodnocena postupem popsáním v Kapitola 3.5. Na výsledných grafech (viz Obrázek 8) a tabulce (viz Tabulka 8) lze vidět průměrně vyšší zastoupení unikátních slov u pacientů než u kontrol ve všech úlohách. U kontrol zejména u úloh 3 a 4 lze pozorovat, že kontrolní subjekty více opakovaly slova, a tedy jejich procentuální zastoupení jedinečných slov bylo nižší než u pacientů. Výsledné p-hodnoty Mann-Whitney U-testu (viz Tabulka 8) ukazují, že skupiny dat nejsou statisticky odlišné ( $p > 0.05$ , U-test).

**Tabulka 8:** Tabulka o bohatosti slovní zásoby pro všechny úlohy založené na spontánní promluvě. Výsledné hodnoty jsou uvedeny v jednotkách procentuálního zastoupení unikátních slov v celkové promluvě. Mezi distribucemi nebyl ani pro jednu úlohu nalezen signifikantní rozdíl ( $p > 0.05$ , U-test).

	Pacienti	Kontroly	Mann-Whitney
	$\mu \pm \sigma$ (median)	$\mu \pm \sigma$ (median)	U-test (p-hodnota)
Úloha 2: pohádka	78.5 $\pm$ 4.6 (78.1)	77.6 $\pm$ 4.2 (78.8)	0.9372
Úloha 3: bydlení	85.1 $\pm$ 2.4 (85.4)	83.1 $\pm$ 3.2 (83.5)	0.3939
Úloha 4: obrázek	86.0 $\pm$ 5.9 (86.6)	83.2 $\pm$ 3.9 (82.1)	0.3095

Nulové hypotézy o shodnosti distribucí v bohatosti slovní zásoby se nepodařilo zamítnout.



**Obrázek 8:** Krabicové grafy hodnocení bohatosti slovní zásoby pro úlohy založené na spontánní promluvě. Mezi jednotlivými distribucemi je možné pozorovat určité trendy, ovšem statistickým U-testem nebyl nalezen signifikantní rozdíl ( $p > 0.05$ , U-test).

## Emoční analýza

V následujících podkapitolách jsou shrnuty výsledky lexikální analýzy emocí (viz Kapitola 3.5) pro úlohy založené na spontánní promluvě.

### Arousal

Dle Kapitola 3.5 byl hodnocen arousal, resp. intenzita, slov. Pro každou úlohu spontánní promluvy byly vyhodnoceny dva parametry – medián a interkvartilový interval hodnot arousal všech slov vyjadřujících emoci. Hodnoty byly získané z NRC VAD Lexikonu [47]. U výsledných boxplotů (viz Obrázek 9) lze vidět trend zejména pro medián úlohy 4 a také p-hodnota U-testu (viz Tabulka 9) ukazuje signifikanci tohoto parametru pro úlohu založenou na spontánní promluvě vybrané za účelem vyvolání emocí ( $p < 0.01$ , U-test). V úlohách vyprávění pohádky a popisu bydlení nebyly mezi distribucemi nalezeny signifikantní rozdíly.

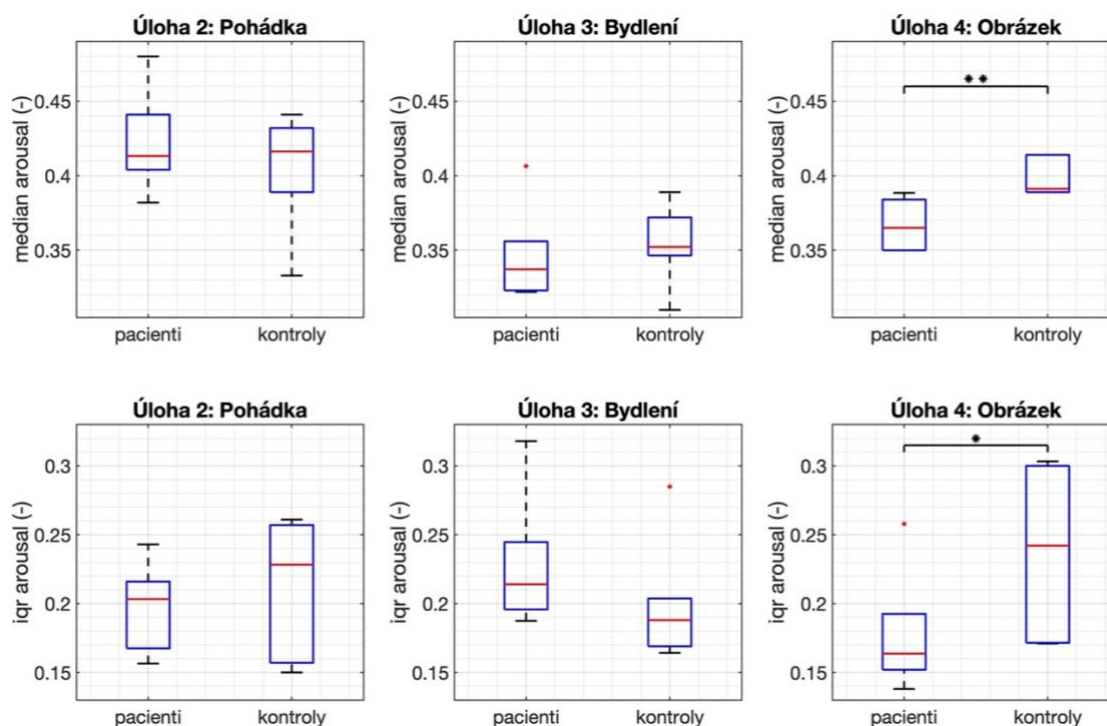
Hodnoty interkvartilového intervalu vykazují vyšší hodnoty pro kontrolní skupinu než pro pacienty v úlohách číslo 2 a 4. Opačný výsledek pozorujeme pro úlohu 3 – popis bydlení, kdy pacienti dosahují vyšších hodnot v hodnocení interkvartilového intervalu síly slov než kontroly. V Tabulka 9 a Obrázek 9 můžeme vidět, že rozdílné distribuce interkvartilového intervalu byly U-testem vyhodnoceny pro úlohu popisu obrázku ( $p < 0.05$ , U-test). Pro ostatní úlohy jsou distribuce hodnot shodné ( $p > 0.05$ , U-test).

**Tabulka 9:** Tabulka výsledných hodnot analýzy arousal slov z úloh založených na spontánní promluvě. Výsledné hodnoty emoční intenzity jsou škálovány v intervalu 0-1. Signifikantní rozdíl byl nalezen pro parametr medián ( $p < 0.01$ , U-test) a interkvartilový interval ( $p < 0.05$ , U-test) pro úlohu popisu obrázku. Pro zbylé dvě úlohy nebyl mezi úlohami nalezen signifikantní rozdíl. (hvězdička označuje signifikanci hodnot: \* $p < 0.05$ , \*\*  $p < 0.01$ )

		Pacienti $\mu \pm \sigma$ ( <i>median</i> )	Kontroly $\mu \pm \sigma$ ( <i>median</i> )	Mann-Whitney U-test (p-hodnota)
<b>Medián</b>	Ú2: pohádka	0.42 $\pm$ 0.03 (0.41)	0.40 $\pm$ 0.04 (0.42)	0.7338
	Ú3: bydlení	0.35 $\pm$ 0.03 (0.34)	0.35 $\pm$ 0.03 (0.35)	0.5887
	Ú4: obrázek	0.37 $\pm$ 0.02 (0.37)	0.40 $\pm$ 0.01 (0.39)	0.0022**
<b>IQR</b>	Ú2: pohádka	0.20 $\pm$ 0.03 (0.20)	0.21 $\pm$ 0.05 (0.23)	0.4848
	Ú3: bydlení	0.23 $\pm$ 0.05 (0.21)	0.20 $\pm$ 0.04 (0.19)	0.2403
	Ú4: obrázek	0.18 $\pm$ 0.04 (0.16)	0.24 $\pm$ 0.06 (0.24)	0.0411*



Hypotézy o shodnosti distribucí parametru arousal slov nebyly zamítnuty pro úlohy 2 a 3, ale hypotéza byla zamítnuta pro úlohu 4 pro medián na 1% a pro interkvartilový interval na 5% hladině významnosti.



**Obrázek 9:** Krabicové grafy hodnocení arousal slov pro parametry medián a interkvartilový interval.

Pro úlohu popisu obrázku byly pro oba parametry nalezeny signifikantní rozdíly mezi distribucemi (medián:  $p < 0.01$ , iqr:  $p < 0.05$ , U-test). Pro úlohu vyprávění pohádky a popisu bydlení nebyly mezi distribucemi nalezeny signifikantní rozdíly a nulové hypotézy o shodnosti distribucí nebyly zamítnuty.

## Valence

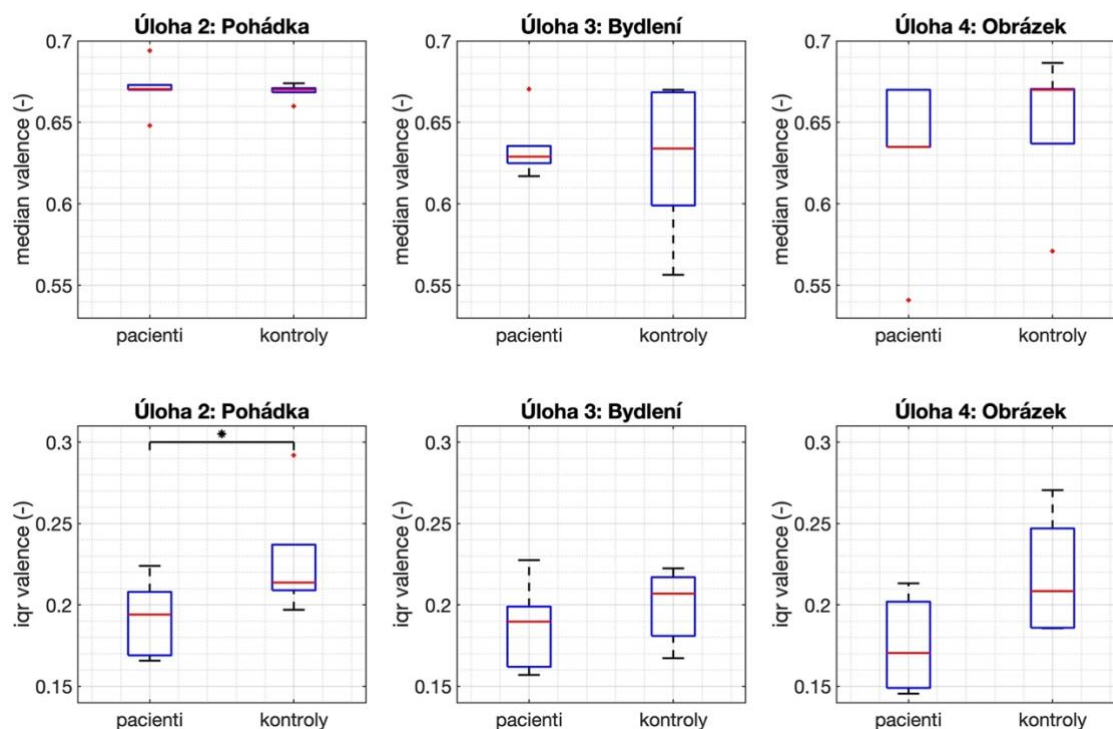
Valence, resp. pozitivnost/negativnost, promluv založených na spontánní promluvě byla vyhodnocena pomocí dvou parametrů mediánu a interkvartilového intervalu. Výsledné hodnoty pro medián nevyšly signifikantně ani pro jednu úlohu (viz Tabulka 10) a mezi distribucemi nebyl nalezen žádný rozdíl ( $p > 0.05$ , U-test) (viz Obrázek 10).

Dále byl vyhodnocen interkvartilový interval hodnot valence použitých slov. Tato analýza vyšla signifikantně pro úlohu založenou na převyprávění pohádky (viz Tabulka 10) ( $p < 0.05$ , U-test), ovšem pro zbylé dvě úlohy U-test signifikanci nenašel. U úlohy popisu obrázku můžeme pozorovat určitý trend, kdy variabilita promluvy kontrolní skupiny je vyšší než u pacientů, ale výsledná p-hodnota není pod hranicí významnosti ( $p > 0.05$ , U-test).

Hypotézy o shodnosti distribucí parametru valence slov nebyly zamítnuty pro úlohy 2, 3 a 4 v parametru mediánu stejně tak i pro interkvartilový interval s výjimkou, kdy hypotéza byla zamítnuta pro úlohu 2 pro interkvartilový interval na 5% hladině významnosti.

**Tabulka 10:** Tabulka výsledných hodnot analýzy valence slov z úloh založených na spontánní promluvě. Výsledné hodnoty polarity emocí jsou škálovány v intervalu 0-1. Signifikantní rozdíl byl nalezen pro parametr interkvartilový interval ( $p < 0.05$ , U-test) pro úlohu vyprávění pohádky. Pro zbylé dvě úlohy nebyl mezi úlohami ani pro jeden parametr nalezen signifikantní rozdíl. (hvězdička označuje signifikanci hodnot:  $*p < 0.05$ )

		Pacienti $\mu \pm \sigma$ (median)	Kontroly $\mu \pm \sigma$ (median)	Mann-Whitney U-test (p-hodnota)
<b>Medián</b>	Ú2: pohádka	0.67 $\pm$ 0.01 (0.67)	0.67 $\pm$ 0.00 (0.67)	0.6190
	Ú3: bydlení	0.63 $\pm$ 0.02 (0.63)	0.63 $\pm$ 0.04 (0.63)	0.9372
	Ú4: obrázek	0.63 $\pm$ 0.05 (0.64)	0.65 $\pm$ 0.04 (0.67)	0.1645
<b>IQR</b>	Ú2: pohádka	0.19 $\pm$ 0.02 (0.19)	0.23 $\pm$ 0.03 (0.21)	0.0411*
	Ú3: bydlení	0.19 $\pm$ 0.03 (0.19)	0.20 $\pm$ 0.02 (0.21)	0.3939
	Ú4: obrázek	0.18 $\pm$ 0.03 (0.17)	0.22 $\pm$ 0.04 (0.21)	0.0649



**Obrázek 10:** Krabicové grafy hodnocení valence slov pro parametry medián a interkvartilový interval. Pro úlohu vyprávění pohádky byl pro parametr interkvartilový interval nalezen signifikantní rozdíl mezi distribucemi ( $p < 0.05$ , U-test). Pro parametr medián nebyl nalezen signifikantní rozdíl v žádné z úloh. Nullovou hypotézu o shodnosti distribucí se podařilo zamítnout pouze pro úlohu vyprávění pohádky.

## Kvalitativní zastoupení emocí v úloze vyvolání emocí

V této části byl dle Kapitola 3.5 vyhodnocen sentiment úlohy číslo 4, která byla do analýzy zařazena z důvodu vyvolání emocí. Dle Obrázek 11 a výsledných hodnot v Tabulka 11 lze vidět, že u kontrol se v promluvách nachází poměrně více slov nesoucích sentiment než u skupiny pacientů. U zastoupení slov s negativním sentimentem vidíme signifikantní rozdíl potvrzený statistickým testem ( $p < 0.01$ , U-test) (viz Obrázek 11 a Tabulka 11). U procentuálního zastoupení slov s pozitivním sentimentem nebyl statistickým testem nalezen rozdíl mezi distribucemi obou skupin ( $p > 0.05$ , U-test).

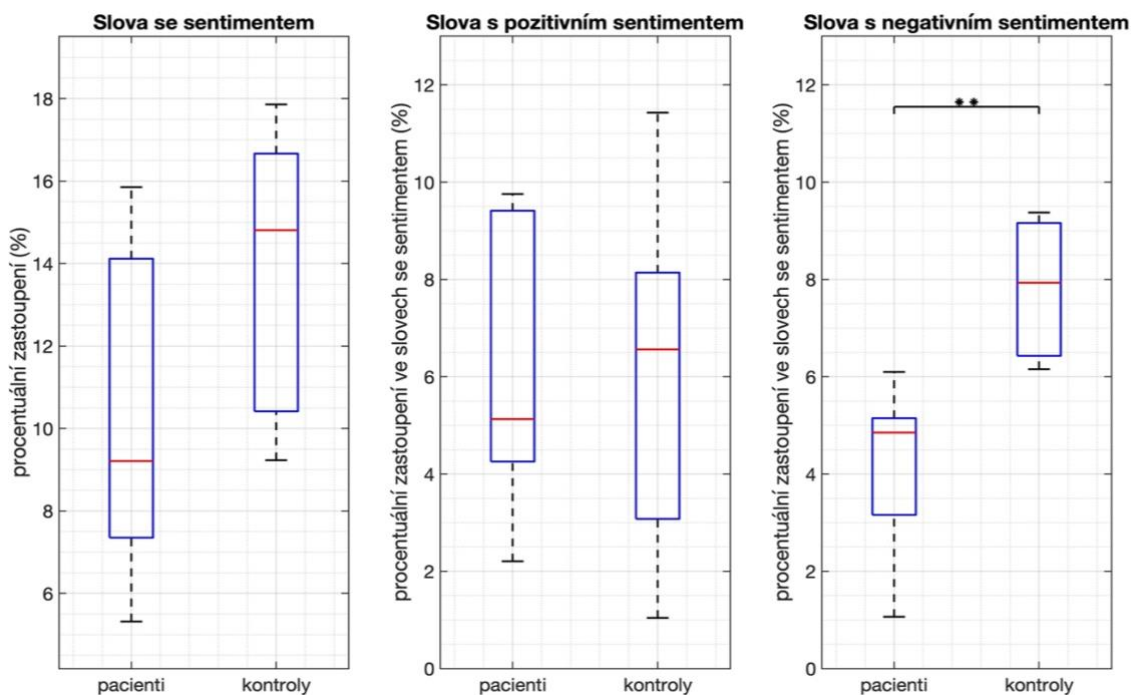
Hypotéza o shodném procentuálním zastoupení slov se sentimentem v promluvách nebyla zamítnuta.

Hypotéza o shodném procentuální zastoupení slov s pozitivním sentimentem nebyla zamítnuta, ale hypotéza o negativním sentimentu byla zamítnuta na 1% hladině významnosti.

**Tabulka 11:** Tabulka výsledných hodnot pro hodnocení sentimentu v promluvě určené k vyvolání emocí, resp. úloha 4 popis obrázku.

Výsledné hodnoty jsou v jednotkách procent. U-test objevil signifikanci pro procentuální zastoupení negativních slov ze slov se sentimentem ( $p < 0.01$ , U-test), kdy u kontrolní skupiny bylo použito signifikantně více slov s negativním sentimentem než u skupiny pacientů. Pro ostatní testy se nepodařily nulové hypotézy zamítnout. (hvězdička označuje signifikanci hodnot:  $**p < 0.01$ )

	<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test ( <i>p-hodnota</i> )
Procentuální zastoupení <b>slov nesoucích sentiment</b> v celkové promluvě	10.2 $\pm$ 4.1 (9.2)	14.0 $\pm$ 3.4 (14.8)	0.0931
Procentuální zastoupení <b>pozitivních slov</b> ze slov se sentimentem	6.0 $\pm$ 3.0 (5.1)	6.1 $\pm$ 3.8 (6.6)	0.9372
Procentuální zastoupení <b>negativních slov</b> ze slov se sentimentem	4.2 $\pm$ 1.8 (4.9)	7.8 $\pm$ 1.5 (7.9)	0.0022**



**Obrázek 11:** Krabicové grafy procentuálního zastoupení slov se sentimentem a grafy procentuálního zastoupení slov s pozitivním či negativním sentimentem v řeči pro úlohu popisu obrázku založené na vyvolání emocí.

U kontrolní skupiny se v průměru v promluvě vyskytovalo větší procento slov se sentimentem. Mezi jednotlivými distribucemi nebyl nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). Mezi procentuálním zastoupením pozitivních slov nebyl U-testem nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). Mezi procentuálním zastoupením negativních slov ze slov se sentimentem byl nalezen signifikantní rozdíl ( $p < 0.01$ , U-test). (hvězdička označuje signifikanci hodnot:  $**p < 0.01$ )

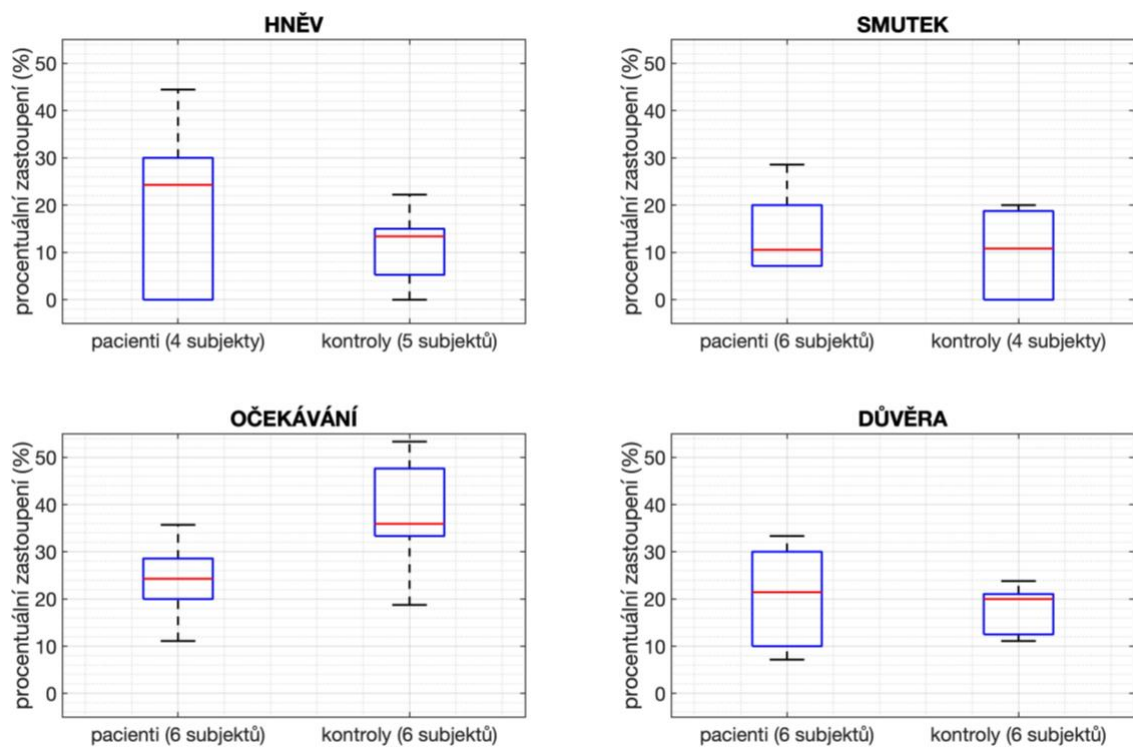
Navazující část byla přímo zaměřena na identifikaci emocí vyjádřených v této úloze. Evaluace byla provedena v rámci osmi základních emocí zmíněných blíže v Kapitola 3.5. A bylo vyhodnoceno procentuální zastoupení slov určité emoce v celkovém počtu slov nesoucích emoce.

Prvotní porovnání bylo provedeno z hlediska počtu subjektů v každé skupině, které použily slova dané emoce. Emoce překvapení byla zastoupena jen u jednoho subjektu v každé skupině, proto je z hodnocení vyřazena. Emoce znechucení a radosti byla ve skupinách použita pouze polovinou subjektů, proto jsou z hodnocení také vyřazeny. Emoce strachu byla z obrázku vyvolaná více u kontrolní skupiny než u pacientů v poměru 4:1, ale i tato emoce byla z důvodu nedostatečného počtu vzorků z analýzy vyřazena.

Naopak emoce očekávání a důvěry byly použity všemi subjekty v obou skupinách. A posledními dvěma ponechanými emocemi je zlost a smutek. Dle výsledné Tabulka 12 a Obrázek 12 není žádný výsledek statistického testu signifikantní ( $p > 0.05$ , U-test), a tedy mezi

jednotlivými distribucemi nebyl nalezen žádný signifikantní rozdíl. Ovšem u emoce očekávání můžeme pozorovat určitý trend, kdy byla tato emoce poměrově mnohem více zastoupena u kontrol než u pacientů. Naopak u hněvu pozorujeme opačný trend, kdy pacienti používali slova emoce hněvu více než kontrolní skupina. Distribuce smutku a důvěry jsou srovnatelné a mezi distribucemi nebyl nalezen žádný rozdíl.

Nulové hypotézy o shodných distribucích slov nesoucích zmíněné emoce nebyly zamítnuty.



**Obrázek 12:** Krabicové grafy procentuálního zastoupení slov vyjadřujících čtyři vybrané emoce – hněv, smutek, očekávání a důvěra.

Mezi jednotlivými distribucemi nebyly pro žádnou z vybraných emocí nalezeny signifikantní rozdíly ( $p > 0.05$ , U-test). U emoce očekávání můžeme pozorovat určitý trend, který ovšem není statisticky významný ( $p = 0.0866$ , U-test).

**Tabulka 12:** Tabulka výsledných hodnot procentuálního zastoupení slov, z celkového počtu slov nesoucích emoce, čtyř vybraných základních emocí – hněv, smutek, očekávání a důvěra.

Mezi skupinami nebyly pro tyto čtyři emoce nalezené signifikantní rozdíly ( $p > 0.05$ , U-test). U skupiny pacientů lze pozorovat vyšší procentuální zastoupení slov vyjadřujících smutek, hněv a důvěru oproti kontrolní skupině. U kontrolní skupiny lze pozorovat vyšší procentuální zastoupení slov vyjadřujících emoci očekávání oproti skupině pacientů.

	<b>Pacienti</b>	<b>Kontroly</b>	<b>Mann-Whitney</b>
	$\mu \pm \sigma$ ( <i>median</i> )	$\mu \pm \sigma$ ( <i>median</i> )	<b>U-test (p-hodnota)</b>
<b>Hněv</b>	20.5 $\pm$ 17.7 (24.3)	11.5 $\pm$ 7.8 (13.4)	0.3874
<b>Smutek</b>	14.0 $\pm$ 8.6 (10.6)	10.1 $\pm$ 8.7 (10.8)	0.6602
<b>Očekávání</b>	24.0 $\pm$ 8.7 (24.3)	37.5 $\pm$ 12.1 (35.9)	0.0866
<b>Důvěra</b>	20.6 $\pm$ 11.4 (21.4)	18.1 $\pm$ 5.1 (20.0)	0.7879

## 4.2. Latentní sémantická analýza

### Koherence prvního řádu

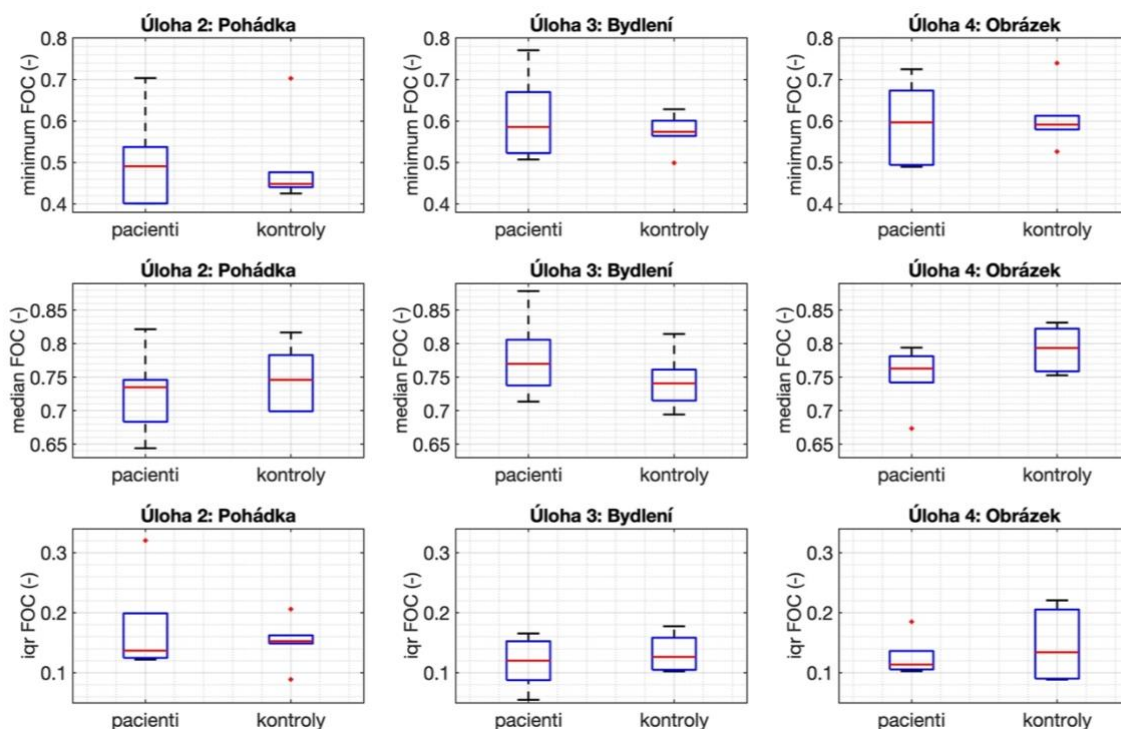
Koherence prvního řádu byla hodnocena postupem popsáním v Kapitola 3.6. Pro koherenci prvního řádu byly vyhodnoceny parametry minimum, medián a interkvartilový interval. Analýza byla provedena pro úlohy založené na spontánní promluvě. Dle tabulky výsledných hodnot (viz Tabulka 13) lze vidět, že mezi jednotlivými distribucemi nebyl nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). U parametru mediánu úlohy 4 popisu obrázku můžeme vidět náznak trendu (viz Obrázek 13), který ovšem nebyl potvrzen pomocí Mann-Whitney statistického testu ( $p = 0.1320$ , U-test).

Hypotézy o shodnosti distribucí skupinu pacientů a kontrolní skupiny pro koherenci prvního řádu nebylo možné zamítnout ani pro jednu z úloh a parametrů.

**Tabulka 13:** Tabulka výsledných hodnot koherence prvního řádu pro úlohy založené na spontánní promluvě.

Mann-Whitney U-test nebyl schopen odhalit signifikantní rozdíly mezi jednotlivými distribucemi ( $p > 0.05$ , U-test). Výsledné hodnoty se pohybují v intervalu 0-1.

		<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test ( <i>p-hodnota</i> )
<b>Minimum</b>	Ú2: pohádka	0.50 ± 0.11 (0.49)	0.49 ± 0.11 (0.45)	0.8182
	Ú3: bydlení	0.61 ± 0.10 (0.59)	0.57 ± 0.04 (0.57)	0.6991
	Ú4: obrázek	0.60 ± 0.10 (0.60)	0.61 ± 0.07 (0.59)	0.8182
<b>Medián</b>	Ú2: pohádka	0.73 ± 0.06 (0.73)	0.75 ± 0.05 (0.75)	0.6991
	Ú3: bydlení	0.78 ± 0.06 (0.77)	0.74 ± 0.04 (0.74)	0.4848
	Ú4: obrázek	0.75 ± 0.04 (0.76)	0.79 ± 0.03 (0.79)	0.1320
<b>IQR</b>	Ú2: pohádka	0.17 ± 0.08 (0.14)	0.15 ± 0.04 (0.15)	0.6991
	Ú3: bydlení	0.12 ± 0.04 (0.12)	0.13 ± 0.03 (0.13)	0.5887
	Ú4: obrázek	0.13 ± 0.03 (0.11)	0.15 ± 0.06 (0.13)	0.6991



**Obrázek 13:** Krabicové grafy koherence prvního řádu pro úlohy založené na spontánní řeči.

Pro hodnocení byly použity parametry minimum, medián a interkvartilový interval. Mezi jednotlivými distribucemi nebyl pomocí U-testu nalezen signifikantní rozdíl ( $p > 0.05$ , U-test).

## Koherence druhého řádu

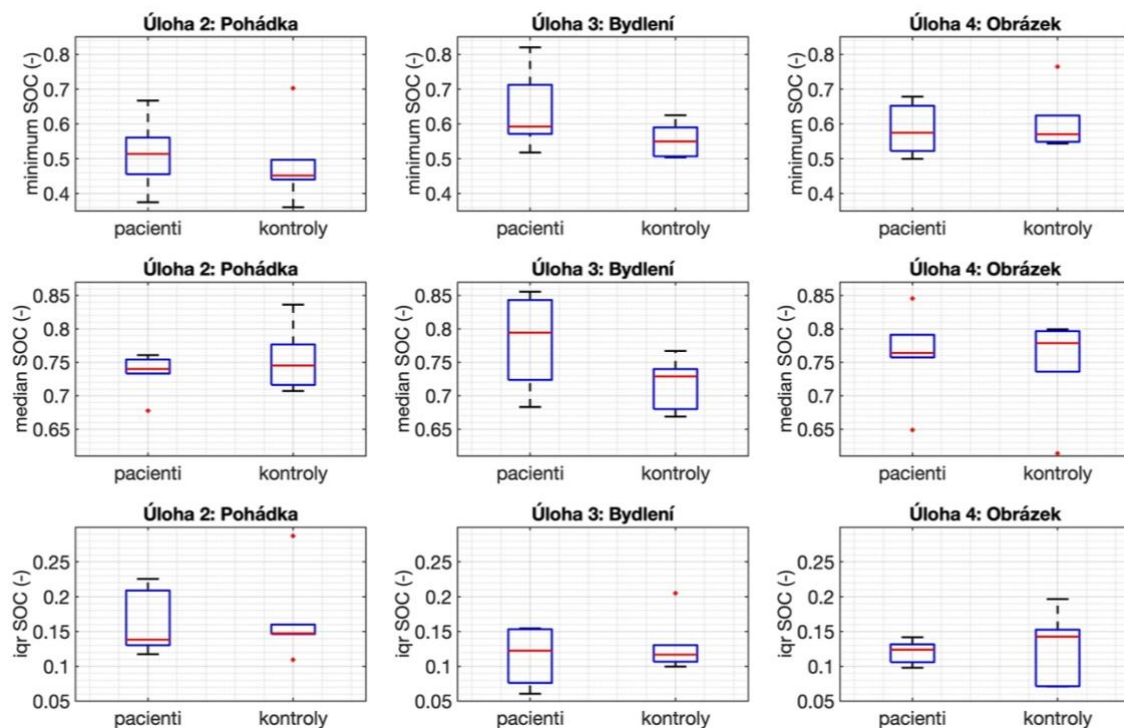
Koherence druhého řádu byla vyhodnocena dle postupu blíže popsaneho v Kapitola 3.6. Pro koherenci druhého řádu byly opět vyhodnoceny parametry minimum, medián a interkvartilový interval. Analýza koherence byla provedena pro úlohy založené na spontánní promluvě. Dle výsledné Tabulka 14 lze vidět, že žádný ze tří zkoumaných parametrů nepotvrdil signifikantní rozdíly mezi distribucemi pacientů a kontrol. U parametru minima a mediánu pro úlohu 3 můžeme pozorovat určité trendy (viz Obrázek 14), které ovšem nebyly U-testem potvrzeny (minimum:  $p = 0.1797$ , medián:  $p = 0.1320$ ).

**Tabulka 14:** Tabulka výsledných hodnot koherence druhého řádu.

Hodnoty U-testu neprokázaly signifikantní rozdíly mezi jednotlivými distribucemi ( $p > 0.05$ , U-test). Průměrné hodnoty minima, mediánu a interkvartilového intervalu koherence se liší vzhledem k úlohám. Výsledné hodnoty se pohybují v intervalu 0-1.

		<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test ( <i>p</i> -hodnota)
<b>Minimum</b>	Ú2: pohádka	0.51 $\pm$ 0.10 (0.51)	0.48 $\pm$ 0.12 (0.45)	0.4848
	Ú3: bydlení	0.63 $\pm$ 0.11 (0.59)	0.55 $\pm$ 0.05 (0.55)	0.1797
	Ú4: obrázek	0.58 $\pm$ 0.08 (0.57)	0.60 $\pm$ 0.08 (0.57)	0.5887
<b>Medián</b>	Ú2: pohádka	0.73 $\pm$ 0.03 (0.74)	0.75 $\pm$ 0.05 (0.75)	0.6991
	Ú3: bydlení	0.78 $\pm$ 0.07 (0.79)	0.72 $\pm$ 0.04 (0.73)	0.1320
	Ú4: obrázek	0.76 $\pm$ 0.06 (0.76)	0.75 $\pm$ 0.07 (0.78)	0.9372
<b>IQR</b>	Ú2: pohádka	0.16 $\pm$ 0.05 (0.14)	0.17 $\pm$ 0.06 (0.15)	0.5887
	Ú3: bydlení	0.11 $\pm$ 0.04 (0.12)	0.13 $\pm$ 0.04 (0.12)	0.9372
	Ú4: obrázek	0.12 $\pm$ 0.02 (0.12)	0.13 $\pm$ 0.05 (0.14)	0.4848





**Obrázek 14:** Krabicové grafy koherence druhého řádu pro parametry minimum, medián a interkvartilový interval pro úlohy založené na spontánní promluvě.

Dle výsledků U-testu není rozdíl mezi skupinami pacientů a kontrol statisticky signifikantní ( $p > 0.05$ , U-test).

### 4.3. Prozódie

#### Celková délka nahrávek

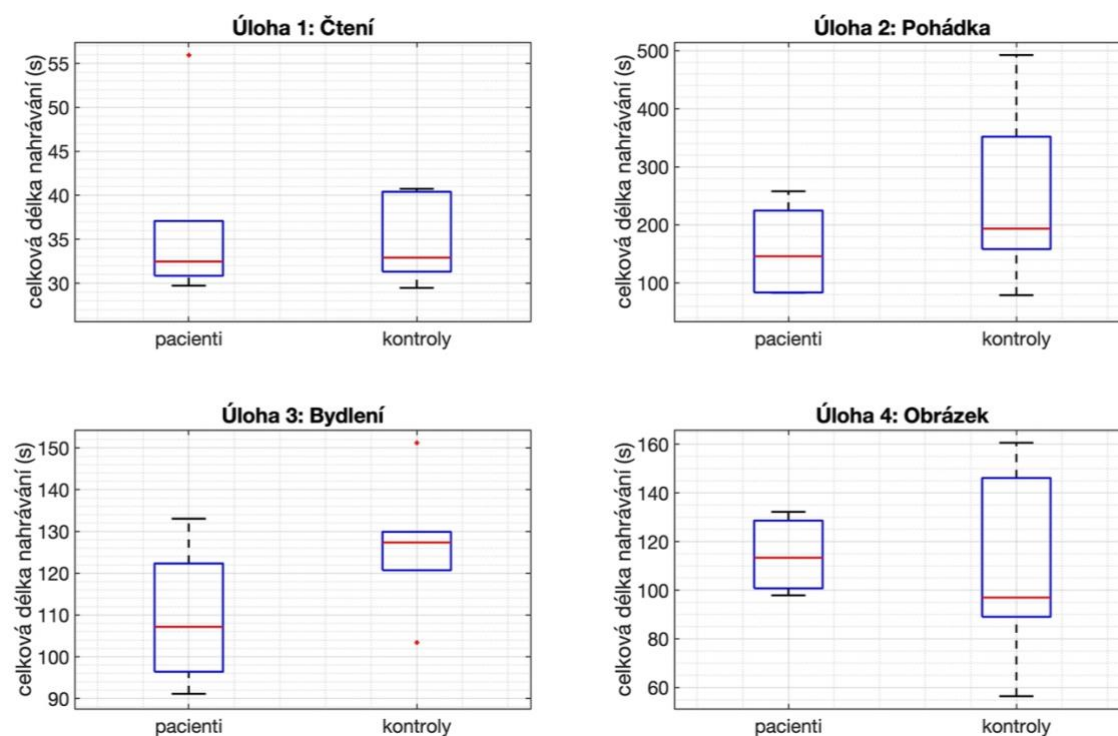
Dle porovnání délky nahrávky (viz Kapitola 3.7) pro každou úlohu zvlášť bylo zjištěno, že tento parametr není signifikantní ani pro jednu úlohu (viz Tabulka 15). Výsledné p-hodnoty U-testu nejsou pod hranicí signifikance ( $p > 0.05$ , U-test). Mezi daty nebyl nalezen signifikantní rozdíl. V délce nahrávek pro úlohu 3 – bydlení můžeme pozorovat trend (viz Obrázek 15), kdy délka nahrávky pro kontrolní skupinu je vyšší než pro pacienty. V úloze vyprávění pohádky můžeme pozorovat podobný trend. U-test ovšem nedokázal zamítnout hypotézu o shodnosti distribucí ( $p = 0.3939$ , U-test).

Hypotéza o vlivu schizofrenie na délku nahrávek nebyla zamítnuta ani pro jednu z úloh.

**Tabulka 15:** Porovnání celkových délek nahrávek pro jednotlivé úlohy.

Výsledné hodnoty jsou v jednotkách sekund. Výsledné hodnoty pro čtení jsou srovnatelné pro obě skupiny ( $p = 0.9372$ , U-test). Pro úlohu vyprávění pohádky a popisu bydlení pozorujeme stejný trend, a to takový, že kontrolní skupina je schopna průměrně déle mluvit na daná témata. Naopak u úlohy popisu obrázku výsledné hodnoty ukazují, že pacienti popisovali průměrně déle než kontroly. Žádná z analýz neprokázala signifikantní rozdíl mezi jednotlivými distribucemi.

	Pacienti $\mu \pm \sigma$ ( <i>median</i> )	Kontroly $\mu \pm \sigma$ ( <i>median</i> )	Mann-Whitney U-test ( <i>p-hodnota</i> )
Úloha 1: čtení	36.4 ± 9.9 (32.5)	34.6 ± 4.8 (32.9)	0.9372
Úloha 2: pohádka	157.0 ± 71.9 (146.1)	244.9 ± 150.3 (193.8)	0.3939
Úloha 3: bydlení	109.5 ± 16.3 (107.2)	126.7 ± 15.5 (127.4)	0.1320
Úloha 4: obrázek	114.4 ± 14.1 (113.3)	107.7 ± 38.7 (97.0)	0.4848



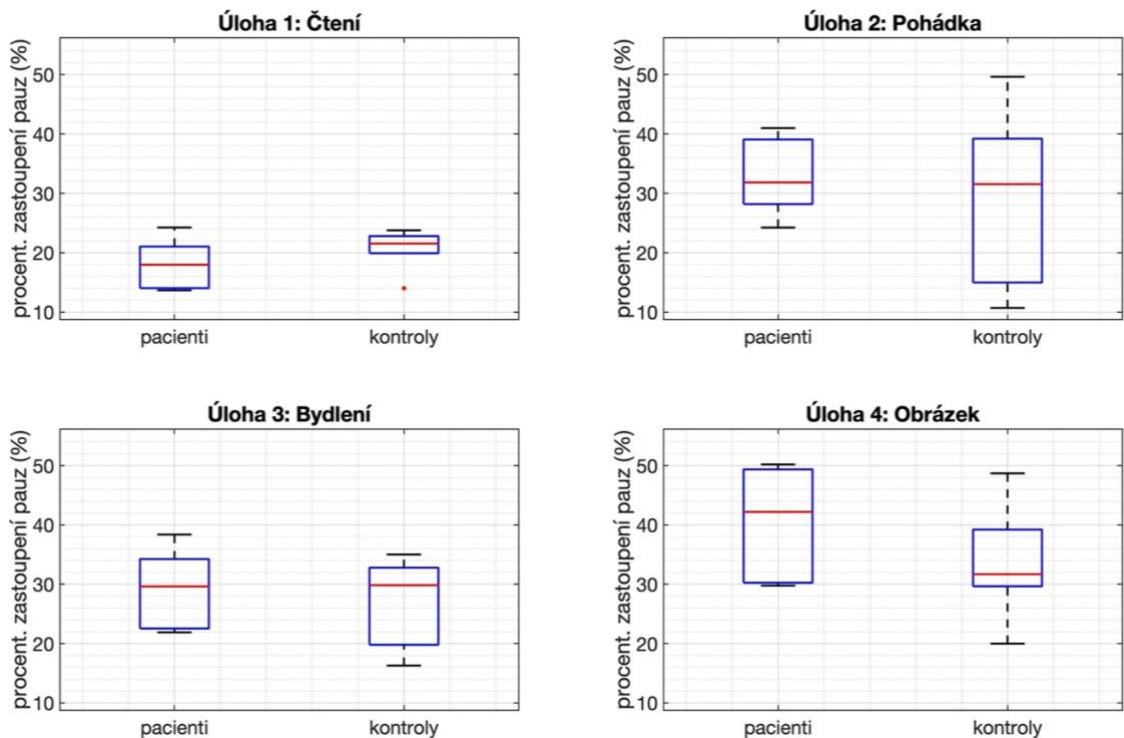
**Obrázek 15:** Krabicové grafy porovnání celkových délek nahrávek pro všechny úlohy protokolu.

Mezi jednotlivými distribucemi nebyl nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). U úlohy vyprávění pohádky a popisu bydlení mluví kontroly průměrně déle než pacienti. U úlohy čtení jsou obě distribuce shodné ve všech parametrech ( $p = 0.9372$ , U-test). U úlohy popisu obrázku je průměrná hodnota a medián délek nahrávání vyšší pro pacienty než pro kontroly.

## Procentuální zastoupení pauz v nahrávce

V této části kapitoly vyhodnocujeme procentuální zastoupení pauz v nahrávce (viz Kapitola 3.7). Dle Tabulka 16 lze vidět, že pro žádnou úlohu se nepodařilo zamítnout nulovou hypotézu o shodnosti jednotlivých distribucí. Pro úlohu vyprávění pohádky a popisu bydlení lze vidět srovnatelné výsledné hodnoty parametrů pro obě skupiny (viz Tabulka 16). U úlohy čtení dosahuje kontrolní skupina vyššího průměrného i mediánového zastoupení pauz v nahrávce oproti skupině pacientů (viz Tabulka 16 a Obrázek 16). V úloze popisu obrázku pozorujeme opačný trend, tedy u skupiny pacientů se vyskytuje průměrně vyšší procentní zastoupení pauz v nahrávce než u kontrolní skupiny (viz Tabulka 16 a Obrázek 16).

Nulové hypotézy o shodnosti distribucí nebyly zamítnuty ( $p > 0.05$ , U-test).



**Obrázek 16:** Krabicové grafy procentuálního zastoupení pauz v nahrávce.

Mezi jednotlivými distribucemi nebyl nalezen signifikantní rozdíl v žádné úloze ( $p > 0.05$ , U-test).

**Tabulka 16:** Tabulka procentuálního zastoupení pauz v nahrávce.

Výsledné hodnoty pro kontroly a pacienty jsou uvedeny v procentech. Mezi jednotlivými distribucemi nebyl nalezen signifikantní rozdíl v žádné úloze ( $p > 0.05$ , U-test). Pro úlohu čtení a popisu obrázku lze pozorovat určité trendy, které ovšem nebyly statisticky potvrzené.

	<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test ( <i>p-hodnota</i> )
<b>Úloha 1:</b> čtení	18.2 ± 4.1 (18.0)	20.6 ± 3.5 (21.5)	0.3095
<b>Úloha 2:</b> pohádka	32.7 ± 6.4 (31.8)	29.6 ± 14.8 (31.6)	0.9372
<b>Úloha 3:</b> bydlení	29.4 ± 6.9 (29.6)	27.3 ± 7.5 (29.8)	0.5887
<b>Úloha 4:</b> obrázek	40.7 ± 8.9 (42.2)	33.5 ± 9.7 (31.7)	0.2403

## Průměrná délka pauzy

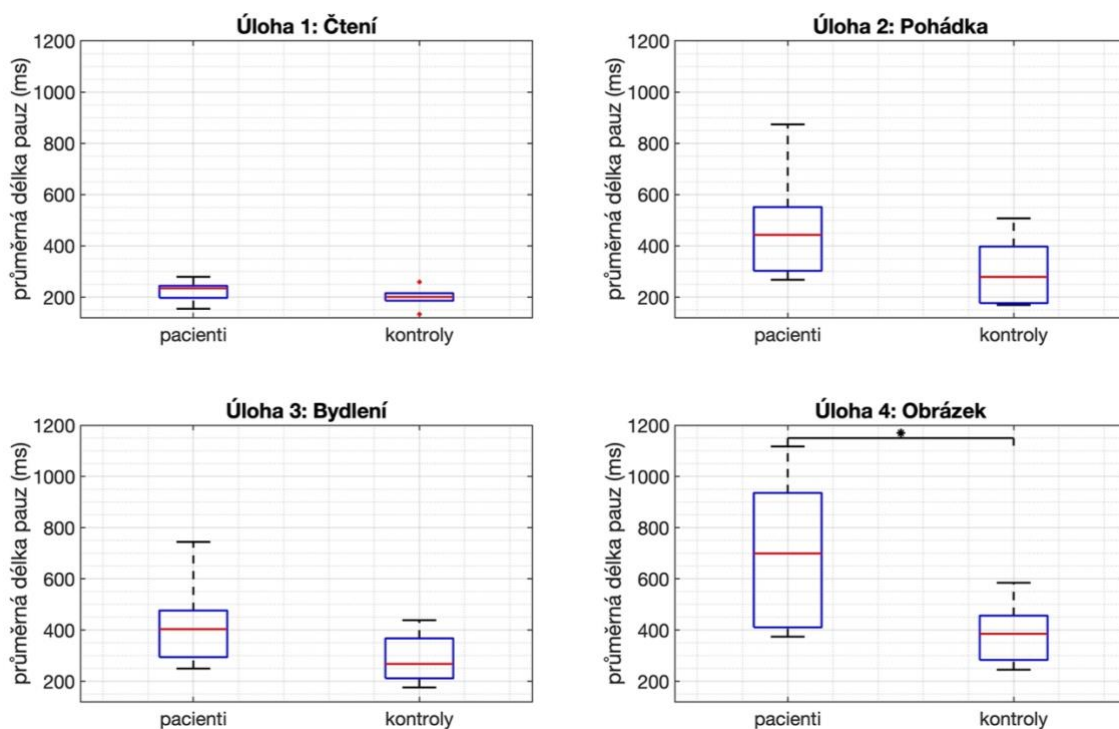
Dalším hodnotícím faktorem je průměrná délka pauzy (viz Kapitola 3.7). Výsledné hodnoty (viz Tabulka 17) následují stejný trend, pro všechny úlohy je průměrná délka pauzy vyšší pro pacienty než pro kontrolní skupinu. Tento trend lze pozorovat ve všech statistických parametrech pro všechny úlohy (viz Tabulka 17 a Obrázek 17). Výsledné hodnoty U-testu prokázaly signifikanci pro úlohu popisu obrázku ( $p < 0.05$ , U-test).

Hypotéza o shodnosti distribucí průměrné délky pauzy pro pacienty a kontroly byla zamítnuta pro úlohu popisu obrázku na 5% hladině významnosti. Pro úlohy čtení, vyprávění pohádky a popisu bydlení se nepodařilo nulové hypotézy zamítnout.

**Tabulka 17:** Tabulka průměrných délek pauz.

Výsledné hodnoty pro skupinu pacientů a kontrol jsou uvedeny v milisekundách. Dle všech statistických parametrů průměrné délky pauz je vidět, že pacienti mají ve všech úlohách detekovanou vyšší průměrnou délku pauz než skupina kontrol. (hvězdička označuje signifikanci hodnot:  $*p < 0.05$ )

	<b>Pacienti</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Kontroly</b> $\mu \pm \sigma$ ( <i>median</i> )	<b>Mann-Whitney</b> U-test ( <i>p-hodnota</i> )
<b>Úloha 1:</b> čtení	223.9 ± 42.8 (234.3)	199.1 ± 41.3 (200.7)	0.3095
<b>Úloha 2:</b> pohádka	480.2 ± 219.3 (442.7)	301.3 ± 136.7 (278.9)	0.1320
<b>Úloha 3:</b> bydlení	428.6 ± 177.6 (403.9)	288.1 ± 102.3 (267.7)	0.1320
<b>Úloha 4:</b> obrázek	706.1 ± 296.3 (699.5)	390.0 ± 122.6 (385.4)	0.0411*



**Obrázek 17:** Krabicové grafy průměrné délky pauzy pro všechny úlohy protokolu.

Pro parametr průměrné délky pauz jsou průměr, medián a směrodatná odchylka vždy vyšší pro skupinu pacientů pro všechny typy úloh než pro kontrolní skupinu. U-test prokázal signifikanci pro úlohu popisu obrázku ( $p = 0.0411$ , U-test).

## Tempo řeči

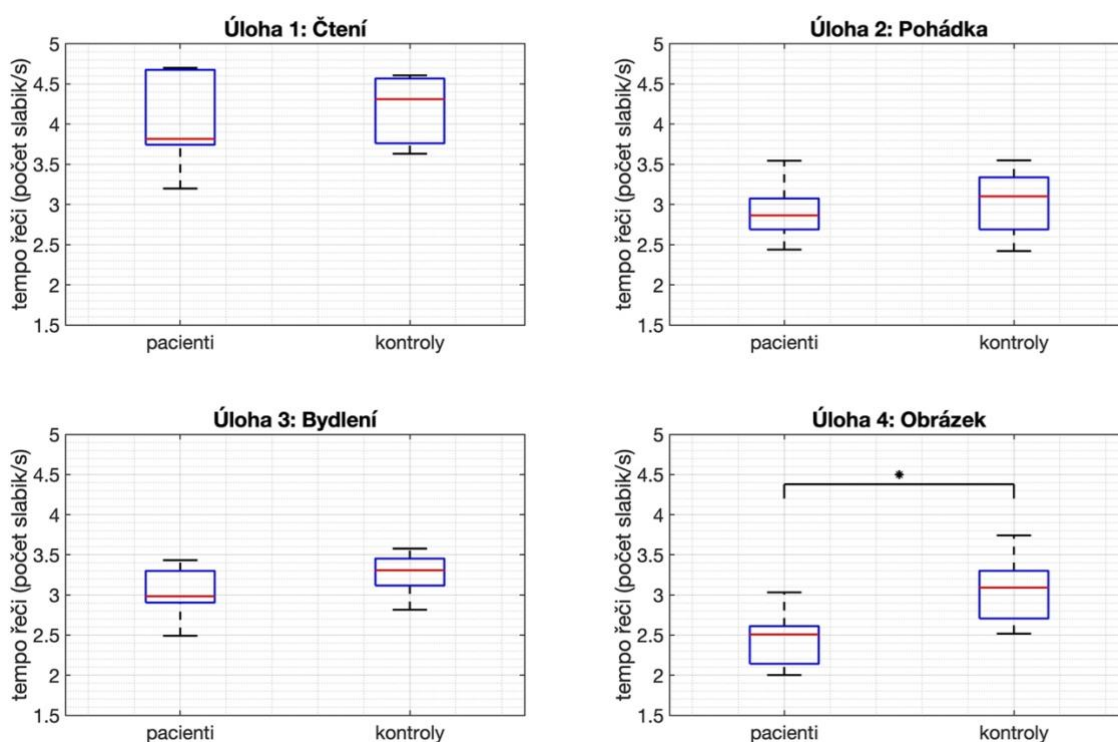
Dalším prozódickým parametrem hodnoceným v této práci je tempo řeči (viz Kapitola 3.7). Dle výsledných hodnot v Tabulka 18 lze vidět, že pro úlohy čtení, vyprávění pohádky a bydlení nebyl pomocí U-testu nalezen signifikantní rozdíl. Pro úlohu popisu obrázku byl nalezen signifikantní rozdíl na 5% hladině významnosti (viz Obrázek 18, viz Tabulka 18). Pro všechny úlohy lze pozorovat shodný trend, kdy je průměrná hodnota i medián tempa řeči vyšší pro kontrolní skupinu než pro pacienty.

Hypotézy o shodném tempu řeči pro skupinu pacientů a kontrol se nepodařilo zamítnout pro úlohy čtení, pohádky a bydlení. Pro úlohu popisu obrázku se podařilo zamítnout nulovou hypotézu na 5% hladině významnosti.

**Tabulka 18:** Tabulka výsledných hodnot tempa řeči.

Výsledné hodnoty jsou v jednotkách počet slabik za sekundu. Mezi jednotlivými distribucemi pro úlohy čtení, pohádka a bydlení nebyl nalezen signifikantní rozdíl. Ve všech úlohách je možné vidět, že kontrolní skupina má průměrně vyšší tempo řeči než skupina pacientů. U-test prokázal signifikanci pro úlohu popisu obrázku ( $p = 0.0411$ , U-test). (hvězdička označuje signifikanci hodnot:  $*p < 0.05$ )

	Pacienti	Kontroly	Mann-Whitney
	$\mu \pm \sigma$ (median)	$\mu \pm \sigma$ (median)	U-test (p-hodnota)
Úloha 1: čtení	3.99 $\pm$ 0.59 (3.82)	4.20 $\pm$ 0.41 (4.31)	0.9372
Úloha 2: pohádka	2.91 $\pm$ 0.38 (2.86)	3.03 $\pm$ 0.43 (3.10)	0.6991
Úloha 3: bydlení	3.02 $\pm$ 0.33 (2.98)	3.26 $\pm$ 0.27 (3.31)	0.2403
Úloha 4: obrázek	2.47 $\pm$ 0.37 (2.51)	3.08 $\pm$ 0.44 (3.09)	0.0411*



**Obrázek 18:** Krabicové grafy tempa řeči pro všechny úlohy protokolu.

Pro úlohu popisu obrázku byl mezi distribucemi nalezen signifikantní rozdíl ( $p < 0.05$ , U-test). Pro úlohu čtení je průměrné tempo řeči rychlejší než pro úlohy založené na spontánní promluvě. (hvězdička označuje signifikanci hodnot:  $*p < 0.05$ )

## Artikulační tempo

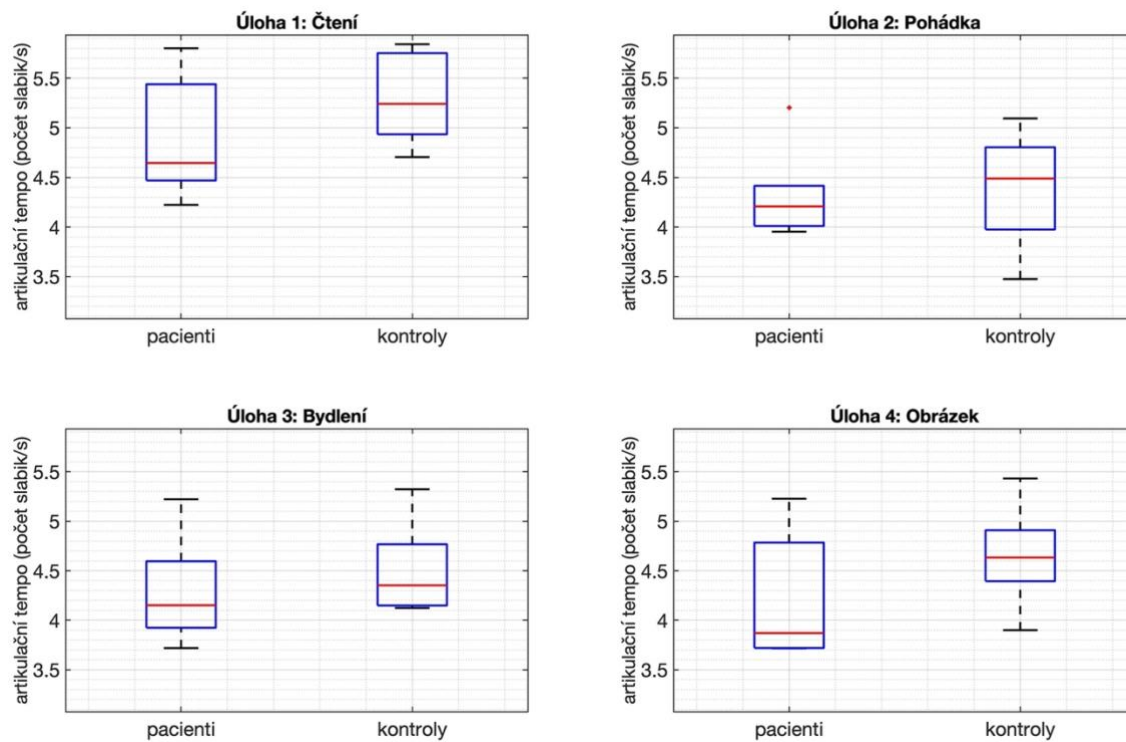
Dalším zkoumaným parametrem je artikulační tempo (viz Kapitola 3.7). Výsledné hodnoty v Tabulka 19 ukazují, že mezi jednotlivými distribucemi nebyl ani pro jednu úlohu nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). Ve všech úlohách lze pozorovat shodný trend, kdy pro kontrolní skupinu je artikulační tempo vyšší v průměrné hodnotě i mediánu než pro skupinu pacientů (viz Obrázek 19, viz Tabulka 19).

Nulové hypotézy o shodnosti distribucí artikulačního tempa pro kontrolní skupinu a skupinu pacientů se nepodařilo zamítnout ani pro jednu z úloh ( $p > 0.05$ , U-test).

**Tabulka 19:** Tabulka výsledných hodnot artikulačního tempa.

Výsledné hodnoty v tabulce jsou v jednotkách počtu slabik za sekundu. Mezi jednotlivými distribucemi nebyl pomocí U-testu nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). Průměrné artikulační tempo i medián jsou ve všech úlohách vyšší pro skupinu kontrol než pro skupinu pacientů.

	<b>Pacienti</b>	<b>Kontroly</b>	<b>Mann-Whitney</b>
	$\mu \pm \sigma$ ( <i>median</i> )	$\mu \pm \sigma$ ( <i>median</i> )	<b>U-test (p-hodnota)</b>
<b>Úloha 1:</b> čtení	4.87 $\pm$ 0.62 (4.65)	5.29 $\pm$ 0.46 (5.24)	0.2403
<b>Úloha 2:</b> pohádka	4.33 $\pm$ 0.46 (4.21)	4.39 $\pm$ 0.58 (4.49)	0.6991
<b>Úloha 3:</b> bydlení	4.29 $\pm$ 0.54 (4.15)	4.51 $\pm$ 0.46 (4.35)	0.3095
<b>Úloha 4:</b> obrázek	4.20 $\pm$ 0.65 (3.87)	4.65 $\pm$ 0.51 (4.63)	0.2403



**Obrázek 19:** Krabicové grafy artikulačního tempa pro všechny úlohy protokolu.

Mezi jednotlivými distribucemi nebyl pomocí U-testu nalezen signifikantní rozdíl ( $p > 0.05$ , U-test). Kontrolní skupina ve všech úlohách vykazuje vyšší průměrné a mediánové hodnoty artikulačního tempa než skupina pacientů.



## 5. Diskuse

---

Tato práce prezentuje pilotní analýzy potenciálních sémantických biomarkerů řeči pro diagnostiku schizofrenie. Cílem pilotního projektu bylo provést rešerši literatury týkající se projevů schizofrenie v řeči se zaměřením na sémantické charakteristiky, vhodné parametry použít k experimentům a vyhodnotit jejich využití v rámci detekce schizofrenie v řeči. Komputační metody hodnocení řečového narušení pro schizofrenii mají velký potenciál, protože jsou neinvazivní, snadno interpretovatelné a finančně nenáročné. Na rozdíl od nynějšího přístupu, kdy posouzení diagnostiky musí provést lékař, je tento přístup počítačového hodnocení schopen vyloučit subjektivní složku a přidává možnost reprodukovatelnosti měření.

V této diplomové práci byly použity promluvy šesti pacientů a šesti věkově odpovídajících kontrolních subjektů. Nahrávky řeči se pro každý subjekt skládaly ze čtyř úloh: čtení, vyprávění pohádky, popisu bydlení a popisu obrázku (viz Příloha B). Pro sémantickou analýzu řeči byly nahrávky založené na spontánní promluvě transkribovány (viz Kapitola 3.2).

Byla vyhodnocena relativní četnost spojek, která se ukázala jako signifikantní pro úlohu popisu bydlení ( $p < 0.05$ , U-test), kdy pacienti používali více spojek než kontroly (pacienti:  $6.4 \pm 1.4$  (%), kontroly:  $4.0 \pm 1.4$  (%)). Schizofrenie má negativní vliv na mozková centra, která jsou zodpovědná za zpracování a tvorbu řeči [2]. Proto je vyšší relativní četnost spojek v řeči pacientů zajímavým jevem, neboť neodpovídá předpokladu o nižší komplikovanosti vět a použití méně spojovacích prostředků [2] pro skupinu pacientů trpících schizofrenií. Zajímavostí je, že relativní četnost spojek se liší v závislosti na zkoumané úloze protokolu výzkumu. Pro úlohu vyprávění pohádky je relativní četnost spojek vyšší pro pacienty o 21 % oproti kontrolám, pro popis bydlení je četnost pro pacienty vyšší o 60 % a pro úlohu popisu obrázku je relativní četnost spojek naopak nižší pro pacienty o 5 % oproti kontrolám. Pro skupinu pacientů se relativní četnost spojek téměř nemění v závislosti na úloze. Pro kontrolní skupinu je ovšem variabilita četnosti spojek v závislosti na úloze výrazná. To může znamenat větší schopnost změny stylu řeči v závislosti na kontextu, kterou pacienti vlivem nemoci ztrácí [2].

Frekventovanost použitých slov, resp. hodnocení, jak frekventovaně je dané slovo populací používáno, je dalším parametrem, který prokázal signifikantní rozdíl mezi pacienty a kontrolami. V úloze popisu bydlení použili pacienti signifikantně běžnější slova než kontrolní skupina ( $p < 0.05$ , U-test). Předpokladem bylo, že pacienti budou trpět zhoršenou verbální fluencí, proto budou používat běžnější slova [13]. Tento předpoklad byl splněn pro úlohy vyprávění pohádky a bydlení. Indexy použitých slov byly podle lexikonu o 5-25 % nižší u pacientů než u kontrolní skupiny. Tento parametr frekventovanosti použitých slov souvisí se slovní zásobou, jejíž velikost je vlivem nemoci negativně ovlivněná [44].

Vyhodnocení analýzy sentimentu a parametrů arousal a valence v kapitole analýzy emocí přineslo pro tuto práci velice optimistické prognózy pro budoucí navazující výzkum. Mezi skupinami byl pozorován markantní trend, který dle předpokladu a stejně jako ve článku [20] potvrdil použití „silnějších“ a více emočně zabarvených slov v kontrolní skupině oproti pacientům (úloha 4:  $p < 0.01$ , U-test). Tento výsledek potvrzuje možné afektivní zploštění, které je jedním ze symptomů pro schizofrenii. Při zaměření pouze na úlohu popisu obrázku určenou k vyvolání emocí, bylo zjištěno, že zdravé kontroly byly schopnější v pochopení emocí a byly výraznější v jejich vyjádření. Kontroly také použily více slov s emocí „očekávání“ oproti skupině pacientů při úloze popisu obrázku. To bychom mohli vysvětlit tím, že kontrolní skupina při vytváření příběhu, inspirovaného obrázkem v úloze 4, častěji vložila optimistický a pozitivní vzhled do příběhu oproti skupině pacientů.

Překvapivé výsledky přinesla analýza koherence prvního a druhého řádu pro úlohy založené na spontánní promluvě. Předpokladem byly potíže pacientů se schizofrenií s alogií, s verbální fluencí, se schopností zůstat u tématu a vyjádřením pointy. Vzhledem k výsledkům v článkách [11]–[13] byl předpoklad signifikance v minimální koherenci, ve varianci a průměrné hodnotě. V rámci této diplomové práce nebyly pozorovány stejné signifikantní výsledky jako v uvedených článkách. Ovšem zajímavé počátky trendů lze pozorovat pro koherenci druhého řádu v úloze popisu bydlení, kde by zvětšení datasetu či použití jiného modelu sémantického prostoru mohlo objevit významné rozdíly.

Z hlediska prozodie byla mimo jiné hodnocena průměrná délka pauz, která se ukázala jako signifikantní faktor ( $p < 0.05$ , U-test) pro úlohu založenou na emočně zabarveném popisu obrázku (pacienti:  $706.1 \pm 296.3$  (ms) a kontroly:  $390.0 \pm 122.6$  (ms),  $p < 0.05$ ). Průměrná délka pauz je ve všech úlohách pro pacienty vyšší (čtení: +12 %, pohádka: +59 %, bydlení: +49 %, obrázek: +81 %) oproti kontrolní skupině. Zvýšená délka pauz může být v případě pacientů se schizofrenií spojená s hezitací, s neschopností se vyjádřit a s problémem nalezení správných slov, které jsou způsobené narušením jazykových center v mozku. Zvýšená délka pauz u pacientů se schizofrenií byla shodně potvrzená i v citované studii [9].

Řečové i artikulační tempo bylo pro všechny úlohy vyšší u kontrolní skupiny než u pacientů. Signifikantní výsledek byl prokázán pouze v případě tempa řeči pro úlohu popisu obrázku, která je založená na vyvolání emocí ( $p < 0.05$ , U-test). Lze tedy diskutovat, zda na tento výsledek může mít, kromě alogie nebo hezitace, vliv také problematika zpracování a vyjádření emocí, na které je úloha 4 popisu obrázku zaměřená.

Nelze zanedbat korelaci dvou prozodických charakteristik, a to průměrné délky pauz a tempa řeči. Z výsledků artikulačního tempa řeči lze vidět, že podstatným faktorem k rozlišení pacientů a kontrolní skupiny je počet a délka pauz v promluvě.

Pro hodnocení sémantiky, sentimentu a emocí se největší rozdíly prokázaly v úlohách popisu bydlení a popisu obrázku. Proto by do budoucna mohl být protokol výzkumu doplněn o další úlohy spontánní promluvy, emočně neutrální úlohu a úlohu určenou k vyvolání emocí. Neboť tyto úlohy se ukazují jako nejvhodnější pro sémantické, ale i prozódické analýzy. Obecně z provedených experimentů vychází, že v prostém čtení pro zkoumané parametry není možné pozorovat patologické rysy.

Výsledky experimentů zaměřených na sémantické a prozódické parametry potvrzují, že diferenciace mezi zdravými kontrolami a pacienty je možná a že řečové biomarkery by v budoucnu mohly být použity k diagnostice schizofrenie.

## **Limitace**

Ve statistickém testování nebyla zahrnuta korekce na mnohonásobné testování z důvodu velikosti datasetu. Právě vzhledem k velikosti datasetu byly v této práci hlavně komentovány trendy, které lze v datech pozorovat.

V lingvistické sekci byla evaluována bohatost jazyka, resp. procentuální zastoupení unikátních slov v řeči. Tento test vycházel z předpokladu, že schizofreničtí pacienti trpí zhoršenou verbální fluencí [52]. Před nahráváním subjektů výzkumu ovšem nebyl proveden inteligenční test nebo nebylo zjištěno jejich dosažené vzdělání. Tyto faktory společně s věkem nebo právě neurodegenerativní nemocí mohou mít na verbální fluenci vliv [52]. Při zvětšování databáze pro tento výzkum by do budoucna mohly být vybírány nejen věkově blízké kontrolní subjekty k pacientům, ale také kontrolní subjekty blízké dosaženým vzděláním. Tím bychom mohli předejít případné limitaci způsobené nevyvážeností datasetu v tomto parametru. Do budoucna by také mohlo být přínosné přímo zahrnout test verbální fluence do protokolu pro tento výzkum podobně jako v těchto studiích [13], [53]. I přes tyto limitace diplomová práce prokázala užitečnost takového přístupu k výzkumu diagnostiky schizofrenie z řeči.

## **Shrnutí dosažených cílů této diplomové práce:**

- I. Provést rešerši literatury týkající se projevů schizofrenie v řečovém signálu se zaměřením na sémantické charakteristiky.*

Provedla jsem rešerši literatury týkající se projevů schizofrenie v řečovém signálu se zaměřením na sémantické charakteristiky. Rešerše uvádím v teoretické části této diplomové práce (viz Kapitola 2.1). Nad rámec této práce jsem přidala rešerši prozódických charakteristik, které se týkají projevů schizofrenie v řeči.

- II. Pro experimenty použít řečové signály poskytnuté Národním ústavem duševního zdraví v Klecanech.*

Pro experimenty provedené v této diplomové práci jsem použila řečové signály pacientů trpících schizofrenií a zdravých kontrol, které mi byly poskytnuté Národním ústavem duševního zdraví v Klecanech. Kontrolní skupinu jsem doplnila řečové signály dalších subjektů, aby byly skupiny pacientů a kontrol početně vyrovnané.

- III. Analyzovat vybrané charakteristiky v programovém prostředí MATLAB.*

Vybrané sémantické a prozódické charakteristiky jsem analyzovala v programovém prostředí Matlab a Python.

- IV. Vyhodnotit možnosti detekce onemocnění schizofrenií pro český jazyk.*

V této diplomové práci jsem použila korpusový přístup detekce schizofrenie pro český jazyk z hlediska lexikální sémantiky a latentní sémantické analýzy. Tento přístup prokázal schopnost budoucího využití sémantických charakteristik v rámci detekce onemocnění schizofrenie.

Vybrané prozódické charakteristiky také prokázaly možné budoucí využití v rámci detekce onemocnění schizofrenie. Tyto charakteristiky svým provedením ovšem nejsou vázány na specifický jazyk.

## 6. Závěr

---

Byla provedena rešerše literatury týkající se projevů schizofrenie v řečovém signálu se zaměřením na sémantické charakteristiky, sentiment, emoce a prozódii. Na základě této rešerše byly vybrány parametry pro analýzu v Matlab a Python. Všechny analýzy byly provedeny na řečových záznamech pacientů pořízených v Národním ústavu duševního zdraví v Klecanech.

Jako potenciální parametry pro detekci schizofrenie byly vyhodnoceny valence a arousal emocí, sentiment promluvy, relativní četnost spojek, frekvencovanost použitých slov, průměrná délka pauz a tempo řeči.

Řečové biomarkery navržené v této práci byly schopné rozlišit mezi zdravými kontrolami a pacienty a dávají naději na budoucí využití v rámci diagnostiky schizofrenie.

### Budoucí výzkum

Do budoucna by mohlo přinést významné výsledky zahrnutí kontextu zejména pro analýzu emocí. Zvláště pro český jazyk, který je specifický různým významem slov v závislosti na slovních spojeních, by mohlo zahrnutí kontextu znamenat velký benefit pro analýzu promluvy.

Lexikální analýza by do budoucna mohla být doplněna o parts-of-speech tagging, tedy o doplnění analýzy slovních druhů.

S vyhlídkou zvětšení databáze řečových nahrávek pacientů a zdravých kontrol by mohl být vytvořen klasifikátor s využitím signifikantních biomarkerů pro diagnostiku schizofrenie.

## Použitá literatura

---

- [1] T. R. Insel, „Rethinking schizophrenia“, *Nature*, roč. 468, č. 7321, s. 187–193, lis. 2010, doi: 10.1038/nature09552.
- [2] J. N. de Boer, S. G. Brederoo, A. E. Voppel, a I. E. C. Sommer, „Anomalies in language as a biomarker for schizophrenia“, *Curr. Opin. Psychiatry*, roč. 33, č. 3, s. 212–218, kvě. 2020, doi: 10.1097/YCO.0000000000000595.
- [3] World Health Organisation, „Schizophrenia“. Viděno: 6. leden 2024. [Online]. Dostupné z: <https://www.who.int/news-room/fact-sheets/detail/schizophrenia>
- [4] S. R. Kay, A. Fiszbein, a L. A. Opler, „The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia“, *Schizophr. Bull.*, roč. 13, č. 2, s. 261–276, led. 1987, doi: 10.1093/schbul/13.2.261.
- [5] H. H. Stassen, M. Albers, J. Püschel, C. Scharfetter, M. Tewesmeier, a B. Woggon, „Speaking behavior and voice sound characteristics associated with negative schizophrenia“, *J. Psychiatr. Res.*, roč. 29, č. 4, s. 277–296, 1995, doi: 10.1016/0022-3956(95)00004-o.
- [6] A. Moro, M. Tettamanti, D. Perani, C. Donati, S. F. Cappa, a F. Fazio, „Syntax and the brain: disentangling grammar by selective anomalies“, *NeuroImage*, roč. 13, č. 1, s. 110–118, led. 2001, doi: 10.1006/nimg.2000.0668.
- [7] P. Indefrey, P. Hagoort, H. Herzog, R. J. Seitz, a C. M. Brown, „Syntactic Processing in Left Prefrontal Cortex Is Independent of Lexical Meaning“, *NeuroImage*, roč. 14, č. 3, s. 546–555, zář. 2001, doi: 10.1006/nimg.2001.0867.
- [8] Y. Harpaz, Y. Levkovitz, a M. Lavidor, „Lexical ambiguity resolution in Wernicke’s area and its right homologue“, *Cortex*, roč. 45, č. 9, s. 1097–1103, říj. 2009, doi: 10.1016/j.cortex.2009.01.002.
- [9] V. Rapcan, S. D’Arcy, S. Yeap, N. Afzal, J. Thakore, a R. B. Reilly, „Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia“, *Med. Eng. Phys.*, roč. 32, č. 9, s. 1074–1079, lis. 2010, doi: 10.1016/j.medengphy.2010.07.013.
- [10] WHO, „Mezinárodní klasifikace nemocí a souvisejících zdravotních problémů“. Viděno: 4. leden 2024. [Online]. Dostupné z: <https://mkn10.uzis.cz/prohlizec/F20.9>
- [11] J. Sarzynska-Wawer *et al.*, „Detecting formal thought disorder by deep contextualized word representations“, *Psychiatry Res.*, roč. 304, s. 114135, říj. 2021, doi: 10.1016/j.psychres.2021.114135.

- [12] G. Bedi *et al.*, „Automated analysis of free speech predicts psychosis onset in high-risk youths“, *NPJ Schizophr.*, roč. 1, s. 15030, 2015, doi: 10.1038/npjSchz.2015.30.
- [13] B. Elvevåg, P. W. Foltz, D. R. Weinberger, a T. E. Goldberg, „Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia“, *Schizophr. Res.*, roč. 93, č. 1–3, s. 304–316, čvc. 2007, doi: 10.1016/j.schres.2007.03.001.
- [14] A. Alqahtani, E. S. Kay, S. Hamidian, M. Compton, a M. Diab, „A Quantitative and Qualitative Analysis of Schizophrenia Language“. arXiv, 25. leden 2022. Viděno: 29. prosinec 2023. [Online]. Dostupné z: <http://arxiv.org/abs/2201.10430>
- [15] M. Alpert, A. Kotsaftis, a R. Pouget, „At Issue: Speech Fluency and Schizophrenic Negative Signs“, *Schizophr. Bull.*, roč. 23, č. 2, 1997.
- [16] M. Walenski, T. W. Weickert, C. J. Maloof, a M. T. Ullman, „Grammatical processing in schizophrenia: evidence from morphology“, *Neuropsychologia*, roč. 48, č. 1, s. 262–269, led. 2010, doi: 10.1016/j.neuropsychologia.2009.09.012.
- [17] N. Rezaei, E. Walker, a P. Wolff, „A machine learning approach to predicting psychosis using semantic density and latent content analysis“, *NPJ Schizophr.*, roč. 5, s. 9, čer. 2019, doi: 10.1038/s41537-019-0077-9.
- [18] M. Alpert, E. R. Pouget, J. Welkowitz, a J. Cohen, „Mapping schizophrenic negative symptoms onto measures of the patient’s speech: Set correlational analysis“, *Psychiatry Res.*, roč. 48, č. 3, s. 181–190, zář. 1993, doi: 10.1016/0165-1781(93)90070-W.
- [19] M. Alpert, S. D. Rosenberg, E. R. Pouget, a R. J. Shaw, „Prosody and lexical accuracy in flat affect schizophrenia“, *Psychiatry Res.*, roč. 97, č. 2, s. 107–118, pro. 2000, doi: 10.1016/S0165-1781(00)00231-6.
- [20] R. E. Gur *et al.*, „Flat Affect in Schizophrenia: Relation to Emotion Processing and Neurocognitive Measures“, *Schizophr. Bull.*, roč. 32, č. 2, s. 279–287, dub. 2006, doi: 10.1093/schbul/sbj041.
- [21] V. Petkevič, „KOMPUTAČNÍ LINGVISTIKA | Nový encyklopedický slovník češtiny“. Viděno: 4. leden 2024. [Online]. Dostupné z: <https://www.czechency.org/slovník/KOMPUTA%C4%8CN%C3%8D%20LINGVISTIKA>
- [22] L. Uhlířová, „KVANTITATIVNÍ LINGVISTIKA | Nový encyklopedický slovník češtiny“. Viděno: 4. leden 2024. [Online]. Dostupné z: <https://www.czechency.org/slovník/KVANTITATIVN%C3%8D%20LINGVISTIKA>

- [23] P. Nandwani a R. Verma, „A review on sentiment analysis and emotion detection from text", *Soc. Netw. Anal. Min.*, roč. 11, č. 81, srp. 2021, doi: 10.1007/s13278-021-00776-6.
- [24] R. Prabowo a M. Thelwall, „Sentiment analysis: A combined approach", *J. Informetr.*, roč. 3, č. 2, s. 143–157, dub. 2009, doi: 10.1016/j.joi.2009.01.003.
- [25] K. Veselovská, „Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification", *K Gajdošová Žáková Eds Nat. Lang. Process. Corpus Linguist. E-Learn.*, s. 279–284, čer. 2013.
- [26] C. Strapparava a A. Valitutti, „WordNet Affect: an Affective Extension of WordNet", *Lrec*, roč. 4, č. 1083–1086, s. 40, kvě. 2004.
- [27] S. M. Mohammad a P. D. Turney, „Crowdsourcing a Word–Emotion Association Lexicon", *Comput. Intell.*, roč. 29, č. 3, s. 436–465, 2013, doi: 10.1111/j.1467-8640.2012.00460.x.
- [28] S. Mohammad, „Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, s. 174–184. doi: 10.18653/v1/P18-1017.
- [29] L. He *et al.*, „Deep learning for depression recognition with audiovisual cues: A review", *Inf. Fusion*, roč. 80, s. 56–86, dub. 2022, doi: 10.1016/j.inffus.2021.10.012.
- [30] K. Kearns, *Semantics*. Bloomsbury Publishing, 2017.
- [31] T. K. Landauer, P. W. Foltz, a D. Laham, „An introduction to latent semantic analysis", *Discourse Process.*, roč. 25, č. 2–3, s. 259–284, led. 1998, doi: 10.1080/01638539809545028.
- [32] R. M. Suleman a I. Korkontzelos, „Extending latent semantic analysis to manage its syntactic blindness", *Expert Syst. Appl.*, roč. 165, s. 114130, bře. 2021, doi: 10.1016/j.eswa.2020.114130.
- [33] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, a J. Dean, „Distributed Representations of Words and Phrases and their Compositionality". arXiv, 16. říjen 2013. doi: 10.48550/arXiv.1310.4546.
- [34] T. Mikolov, K. Chen, G. Corrado, a J. Dean, „Efficient Estimation of Word Representations in Vector Space", prezentováno v International Conference on Learning Representations, led. 2013. Viděno: 3. leden 2024. [Online]. Dostupné z: <https://www.semanticscholar.org/paper/f6b51c8753a871dc94ff32152c00c01e94f90f09>

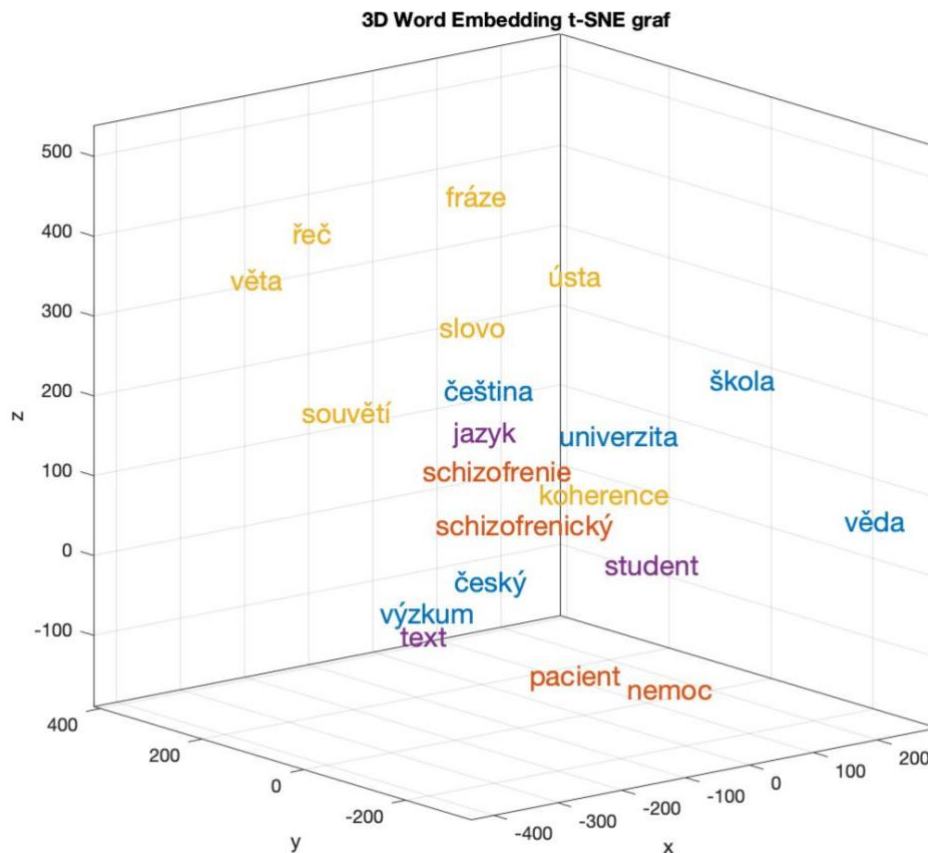


- [35] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, a T. Mikolov, „Learning Word Vectors for 157 Languages". arXiv, 28. březen 2018. doi: 10.48550/arXiv.1802.06893.
- [36] P. Bojanowski, E. Grave, A. Joulin, a T. Mikolov, „Enriching Word Vectors with Subword Information", *Trans. Assoc. Comput. Linguist.*, roč. 5, s. 135–146, 2017, doi: 10.1162/tacl\_a\_00051.
- [37] K. Hořeňovská, „An evaluation of Czech word embeddings", in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, M. Hartmann a B. Plank, Ed., Turku, Finland: Linköping University Electronic Press, zář. 2019, s. 65–75. Viděno: 3. leden 2024. [Online]. Dostupné z: <https://aclanthology.org/W19-6107>
- [38] C. M. Corcoran *et al.*, „Prediction of psychosis across protocols and risk cohorts using automated language analysis", *World Psychiatry Off. J. World Psychiatr. Assoc. WPA*, roč. 17, č. 1, s. 67–75, úno. 2018, doi: 10.1002/wps.20491.
- [39] M. Mikulová *et al.*, „Prague Dependency Treebank of Spoken Czech 2.0 (PDTSC 2.0)", 2017, Viděno: 29. prosinec 2023. [Online]. Dostupné z: <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3189>
- [40] J. Hajič, J. Hlaváčová, M. Mikulová, M. Straka, a B. Štěpánková, „MorFlex CZ 2.0", pro. 2020, Viděno: 29. prosinec 2023. [Online]. Dostupné z: <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3186>
- [41] „MATLAB". Viděno: 29. prosinec 2023. [Online]. Dostupné z: <https://www.mathworks.com/products/matlab.html>
- [42] Y. Jadoul, B. Thompson, a B. de Boer, „Introducing Parselmouth: A Python interface to Praat", *J. Phon.*, roč. 71, s. 1–15, lis. 2018, doi: 10.1016/j.wocn.2018.07.001.
- [43] T. MacFarland a J. Yates, *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. 2016. doi: 10.1007/978-3-319-30634-6.
- [44] K. Hong, A. Nenkova, M. E. March, A. P. Parker, R. Verma, a C. G. Kohler, „Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls", *Psychiatry Res.*, roč. 225, č. 1, s. 40–49, led. 2015, doi: 10.1016/j.psychres.2014.10.002.
- [45] Sgall, Petr *et al.*, „Prague Dependency Treebank 2.0". Linguistic Data Consortium, s. 515072 KB, 21. červenec 2006. doi: 10.35111/E6P0-9S32.

- [46] M. Křen *et al.*, „SYN2015: Representative Corpus of Contemporary Written Czech", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, a S. Piperidis, Ed., Portorož, Slovenia: European Language Resources Association (ELRA), kvě. 2016, s. 2522–2528. Viděno: 29. prosinec 2023. [Online]. Dostupné z: <https://aclanthology.org/L16-1400>
- [47] S. Mohammad, „Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, s. 174–184. doi: 10.18653/v1/P18-1017.
- [48] R. Garland, „A Comparison of Three Forms of the Semantic Differential", *Mark. Bull.*, roč. 1, s. 19–24, 1990.
- [49] S. M. Mohammad, „Word Affect Intensities". arXiv, 15. říjen 2022. doi: 10.48550/arXiv.1704.08798.
- [50] „APA Dictionary of Psychology". Viděno: 29. prosinec 2023. [Online]. Dostupné z: <https://dictionary.apa.org/>
- [51] N. H. de Jong a T. Wempe, „Praat script to detect syllable nuclei and measure speech rate automatically", *Behav. Res. Methods*, roč. 41, č. 2, s. 385–390, kvě. 2009, doi: 10.3758/BRM.41.2.385.
- [52] T. Nikolai *et al.*, „Testy verbální fluence, česká normativní studie pro osoby v...", *Čes. Slov. Neurol. Neurochir.*, roč. 78, č. 111, s. 292–299, úno. 2015, Viděno: 7. leden 2024. [Online]. Dostupné z: <https://www.csn.eu/casopisy/ceska-slovenska-neurologie/2015-3-8/testy-verbalni-fluence-ceska-normativni-studie-pro-osoby-vyssiho-veku-52143>
- [53] A. Docherty, H. Berenbaum, a J. Kerns, „Alogia and formal thought disorder: differential patterns of verbal fluency task performance", *J. Psychiatr. Res.*, roč. 45, č. 10, říj. 2011, doi: 10.1016/j.jpsychires.2011.04.004.

## Příloha A: Doplnující ilustrace

---



**Obrázek 20:** Latentní sémantická analýza. Ilustrativní reprezentace vybraných slov v sémantickém prostoru.

Redukce dimenze pomocí statistické metody t-SNE z dimenze 300 do 3. Následně provedeno shlukování slov, resp. vektorů, do skupin pomocí metody k-means. Barvy sdružují slova přiřazená do stejných klastrů pomocí k-means.

# Příloha B: Protokol pro výzkum pacientů se schizofrenií

---

## Úloha 1: Čtený text – Akustická analýza

Když člověk poprvé vsadí do země sazeničku, chodí se na ni dívat třikrát denně: „Tak co, povyrostla už nebo ne?“ I tají dech, naklání se nad ní, přitlačí trochu půdu u jejich kořínků, načechrává jí lístky a vůbec ji obtěžuje různým konáním, které považuje za užitečnou péči. A když se sazenička přesto ujme a roste jako z vody, tu člověk žasne nad tímto divem přírody, má pocit čehosi jako zázraku a považuje to za jeden ze svých největších úspěchů.

## Úloha 2: Spontánní promluva – vyprávění pohádky – Prozódická informace (cca 90 s)

Popelka žije v domečku se svými dvěma línými sestrami Kateřinou a Dorou. Obě sestry se zlou macechou nutí Popelku k neustálé práci. Jednou se Popelčin hodný tatínek vydá do města. Kateřina s Dorou po něm chtějí drahé šaty a šperky, jen Popelce stačí to, co mu cestou zavadí o kabát. Když se tedy otec vrátí, předá Popelce větvíčku se třemi oříšky, která mu cestou zpět zavádila o kabát. Svě ženě také řekne novinu, že princ se chystá oženit. A skutečně ze zámku se ozve hlas, který oznamuje, že se na zámku budou po tři dny konat plesy, kde si princ Mojmír vybere svoji budoucí ženu. Na ples jsou zvány všechny dívky z království – chudé i bohaté, urozené i neurozené. V den, kdy se koná ples je macecha se svými dcerami v plné parádě. Než se ale vydají na ples, připraví Popelce práci. Dají jí za úkol přebrat popel od hrachu. Popelka se smutně pustí do práce, když najednou se na okně objeví holoubek, který Popelce poradí, aby rozlouskla oříšek a o práci, že se postará on a jeho bratři. Holoubek jí zároveň varuje, že do půlnoci musí být zpátky. Popelka rozlouskne oříšek, obleče si modré šaty a odjede na ples.

Princ Mojmír se ženit nechce a z hledání nevěsty je otráven. Vše se ale změní, když do sálu vstoupí Popelka. Princ se do ní ihned zamiluje a protancuje s ní celý večer. Jenomže ještě než odbije půlnoc, Popelka se z plesu vytratí. Druhý den musí Popelka opět přebírat popel od čocky. Ale i tentokrát jí přispěchají na pomoc holoubci. Popelka rozlouskla druhý oříšek a najde stříbrné šaty. Princ už na ní na plese čeká a znovu s ní stráví celý večer. Na třetí den je princ celý nedočkavý, a tak jeho otec a jeho rádce vymyslí malou lest. Natřeli schody lepidlem, aby se dívka, až bude odcházet, přilepila. Popelka tentokrát najde v oříšku zlaté šaty a zase spěchá na ples. Princ na ni již čeká a požádá ji o ruku. Nicméně když odbije půlnoc, Popelka uteče pryč. Na schodech se jí přilepí střevíček, který tam musí nechat. Princovi tak po jeho neznámé dívce zbude střevíček. Nakonec dojde až do chaloupky, kde bydlí Popelka. Její sestry se pokouší střevíček obout, ale střevíček je moc malý. Princ už skoro ztrácí naději, když do chaloupky vejde

Popelčin otec a řekne princ, že má ještě jednu dceru – Popelku. Popelka střevíček zkusí a padne jí jako ulitý. Princ pozná, že je to ta dívka, do které se zamiloval a odveze si ji i s jejím otcem na zámek.

**Úloha 3: Spontánní promluva – Odpověď na emočně neutrální otázku (cca 90 s)**

Popište dům nebo byt, ve kterém bydlíte.

**Úloha 4: Spontánní promluva – popis obrázku – založené na vyvolání emocí (cca 90 s)**



**Obrázek 21:** Tématický apercepční test.

Obrázek volně dostupný na <https://www.psychestudy.com/general/personality/detailed-procedure-thematic-procedure-test>.

## Příloha C: Souhrn hypotéz

---

### Lexikální sémantika:

*H<sub>0</sub>: Pacienti i kontrolní skupina používají vycpávková slova se srovnatelnou četností.*

*H<sub>0</sub>: Syntaktická složitost promluvy pacientů a kontrolní skupiny je stejná.*

*H<sub>0</sub>: Skupina pacientů a kontrol v řeči používá srovnatelně frekventovaná slova.*

*H<sub>0</sub>: Skupina pacientů a skupina kontrol používá v řeči stejné procentuální zastoupení unikátních slov.*

*H<sub>0</sub>: Parametr arousal slov v promluvách pacientů a kontrol je shodný.*

*H<sub>0</sub>: Parametr valence slov v promluvách pacientů a kontrol je shodný.*

*H<sub>0</sub>: Procentuální zastoupení slov se sentimentem v promluvách je stejné pro skupinu pacientů a kontrol.*

*H<sub>0</sub>: Procentuální zastoupení slov s pozitivním či negativním sentimentem je shodné pro skupinu pacientů a kontrol.*

*H<sub>0</sub>: Procentuální zastoupení jednotlivých emocí v promluvách pacientů a kontrol je shodné.*

### Latentní sémantická analýza:

*H<sub>0</sub>: Koherence prvního řádu dosahuje srovnatelných hodnot ve zkoumaných parametrech minima, mediánu a interkvartilového intervalu pro skupinu pacientů a kontrol.*

*H<sub>0</sub>: Koherence druhého řádu dosahuje srovnatelných hodnot ve zkoumaných parametrech minima, mediánu a interkvartilového intervalu pro skupinu pacientů a kontrol.*

### Akustické parametry:

*H<sub>0</sub>: Délky nahrávek pro každou úlohu samostatně jsou pro skupiny pacientů a kontrol shodné.*

*H<sub>0</sub>: Procentuální zastoupení pauz v nahrávce je shodné pro skupiny pacientů a kontrol pro každou úlohu.*

*H<sub>0</sub>: Průměrná délka pauzy je pro pacienty a kontrolní skupinu stejná.*

*H<sub>0</sub>: Tempo řeči je pro pacienty a kontrolní skupinu stejné.*

*H<sub>0</sub>: Artikulační tempo je pro pacienty a kontrolní skupinu stejné.*