

Vládní opatření, očkování nebo hnutí ANO: wordcloudy nejčastěji se vyskytující slov ve zpravodajství a publicistice o onemocnění COVID-19 (1. 11. 2020 – 30. 9. 2021)

Veronika Macková, Victoria Nainová, Natálie Pančochářová, Alice Němcová Tejkalová, Radek Mařík

Úvod

Jaké zprávy o koronaviru se objevily v českých médiích během září? Je stále referování o onemocnění COVID-19 aktuální téma? V další analýze jsme se rozhodli pomocí nástrojů strojového učení zmapovat výskyt ve zpravodajství za měsíc září a také porovnat delší časové období s první analýzou wordcloudů provedenou v prosinci 2020.

Korpus mediálního monitoringu společnosti Newton Media týkající se zpravodajství o onemocnění COVID-19 obsahoval v období od 1. 11. 2020 do 30. 9. 2021 celkem 716 804 zpráv. Analyzovali jsme jak celek, tak jsme ze všech zdrojů vybrali několik shluků se specifickou charakteristikou mediálního typu (např. komerční rádia, komerční televize nebo veřejnoprávní média).

Rozdělení jsme provedli na základě předvýzkumu a generování wordcloudů pro jednotlivá média z korpusu monitoringu společnosti Newton Media (více o metodě v kapitole Způsob generování wordcloudů a použitá data). Vizualizace pomocí wordcloudů je v tomto případě neefektivnější metodou, jelikož jeden wordcloud dokáže pojmut až 200 slov a zároveň ukázat množství výskytů těchto slov v článcích a příspěvcích prostřednictvím velikosti písma.

Vládní opatření zůstávají. Počet nakažených nahradilo v médiích očkování

Stejně jako v předchozím sledovaném období (viz obrázek 1), tak také od listopadu 2020 do září 2021 dominovalo v celkovém analyzovaném vzorku v českých médiích zpravodajství o onemocnění COVID-19 slovo *opatření* (viz obrázek 2), což je pochopitelné, protože dynamicky se měnící podoba různých typů opatření má přímý vliv na běžný život obyvatel. Ve druhém sledovaném období je velmi výrazné také slovo *vláda*, která připravovala mimo jiné i zmiňovaná opatření.

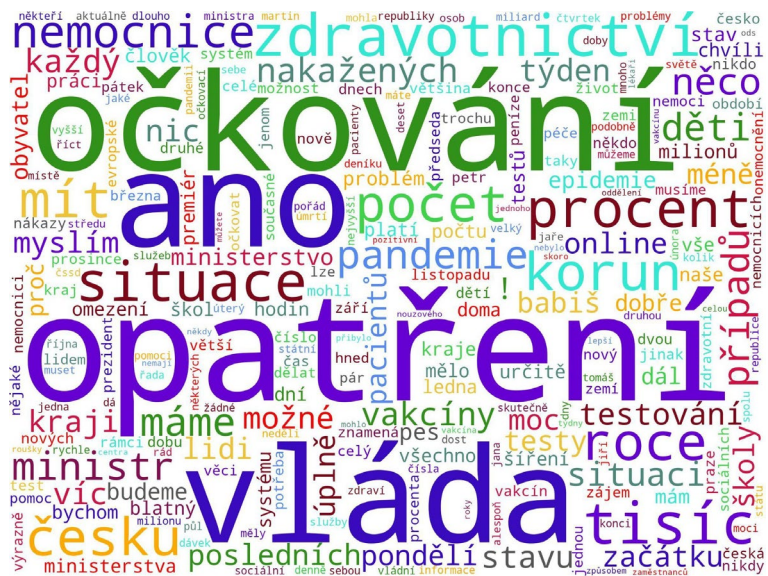
Od začátku listopadu 2020 rostly [počty nakažených](#) s laboratorně prokázaným onemocněním COVID-19, dle hlášení krajských hygienických stanic, velmi razantně. Od května se situace začala zlepšovat

a přírůstek byl naopak velmi mírný, i proto je z obrázku 2 patrné, že slovo *nakažených* nepatří mezi nejvýraznější, jako tomu bylo v prvním sledovaném období (viz obrázek 1). Do popředí vstupuje *očkování* proti onemocnění COVID-19, které bylo zahájeno ve vybraných pražských a brněnských nemocnicích na konci prosince 2020. Právě očkování plnilo titulní stránky i zpravodajské relace během

prvních měsíců roku 2021. V dubnu pak odstartovala i národní informační kampaň, jež měla za cíl především odpovídat na otázky kolem vakcinace.



Obrázek 1. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v období 1. 12. 2019 – 31. 10. 2020



Obrázek 2. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v období 1. 11. 2020 – 30. 9. 2021

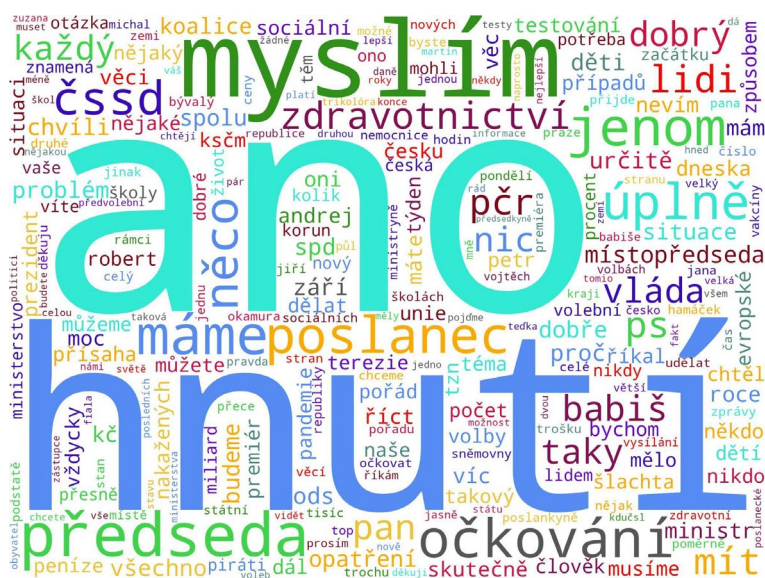
COVID-19 spojený s volbami do Poslanecké sněmovny Parlamentu České republiky

V následující části se věnujeme pěti vybraným skupinám médií, u kterých docházelo z hlediska použitých jazykových prostředků k nejvýrazněji odlišným situacím oproti celku nebo k zajímavým zjištěním.

Vysílání komerčních televizí se věnovalo očkování méně oproti předchozím měsícům. Ve [fotogalerii wordcloudů](#) je tento pokles výskytu viditelný. Když srovnáme wordcloudy za srpen a září 2021 (obrázek 3 a 4) uvidíme, jak předvolební téma v září dominovalo v televizním zpravodajství. Kvůli blížícím se říjnovým volbám se tak nejčastěji vyskytovala slova *ANO* a *hnutí*.



Obrázek 3. Wordcloud ze slov použitých ve zpravodajství komerčních televizí a jejich webu o onemocnění COVID-19 v období 1. 8. – 31. 8. 2021



Obrázek 4. Wordcloud ze slov použitých ve zpravodajství komerčních televizí a jejich webu o onemocnění COVID-19 v období 1. 9. – 30. 9. 2021

Z obrázku 5 vyplývá, že zpravodajské časopisy a jejich weby se v září poměrně vyrovnaným způsobem věnovaly dvěma tématům – volbám do Poslanecké sněmovny Parlamentu České republiky a pandemii onemocnění COVID-19. Nejčastěji se objevovala slova: *procent* (volby se konaly 8. a 9. října a lze tak předpokládat, že procenta se vztahovala k předvolebním průzkumům), název hnutí *ANO*, příjmení premiéra *Babiše*, *koalice* a název koalice *SPOLU*. I v textech o volbách však musel být zmíněn koronavirus. Mezi analyzované reportáže byl totiž text zařazen až poté, co obsahoval slova související s problematikou onemocnění COVID-19 (viz kapitola Způsob generování wordcloudů a použitá data). V pondělí 20. září 2021 se spustila také registrace k přeočkování třetí dávkou, i proto mediálnímu pokrytí tématu pandemie dominují: *očkování*, *opatření*, *tisíc*, *zdravotnictví*, *vacciny* a *pandemie*.

I v celostátních a regionálních denících a jejich webech (obrázek 6) můžeme vidět, že slovo *zář* se vyskytovalo nejčastěji, opět zejména ve spojení se začátkem školního roku a s možnými omezeními ve spojitosti s koronavirem. V těchto denících převažovalo v textech zmiňujících nemoc COVID-19 téma *očkování*. Druhým slovem s nejvyšším počtem výskytu bylo *tisíc*. To souviselo zejména s referováním o počtu provedených testů, očkováných lidí, počtu nakažených, a také s navýšením počtu účastníků klinických studií u vakcíny pro děti od 5 let. Slovo se vyskytovalo i v článcích týkajících se financí, např.

Cílem bulvárních médií je mít co nejvyšší čtenost. Aby jí bulvární deníky a jejich weby dosáhly využívají jednodušší a expresivnější jazyk, kterým připoutají vysokou míru pozornosti. Jak ukazuje wordcloud vytvořený ze všech nejfrekventovanějších slov v hlavních bulvárních médiích a na bulvárních webech (obrázek 7) objevovala se v nich nejvíce slova *očkování*, *opatření*, *test*, *pandemie* a *PCR*. V září nastoupili žáci a studenti do škol a bulvární média tak věnovala pozornost *dětem* a *testování*. Expresivitu jazyka bulvárních médií dokládá slovo *nutné* a *vykřičník* (!), který titulům dodává apelativní význam.

procenta školy začátku období podobně určitě premiéra vojtech
vlebe svět školách testů jedné korunní opatření pandemie
úplně komunikační posledních dětí státu českupremiér babiš
srovnání zdražování sociálních sebe dní studie některých doby krize voličů dost
získo stan programu piráti někdy cen nových stavu dětí státu českupremiér jina
nejvyšší jednoho protitlak každý šíření lidi cen někdy nový dlouhodobě unie rámcí sedm evropské polské
stan programu protitlak každý šíření lidi cen někdy nový dlouhodobě unie rámcí sedm evropské polské
zajíři měn naše týden práci přání dáli škola vlna vláda lidem
práti volby milard bychom všo nimi škola vlna vláda lidem
pořád výrazné volbách zaměstnanci nem finanční celý strana deset týden vláda lidem
dlouho možné otázka člověk kampaň inflace většina zemi státní život koalice všchno
věci český sociální rozdí náky konce případů systém volební mám evropské obyvate
problém epidemii náky konce případů systém volební mám evropské obyvate
dš stran budeme národní děti nížvec dobře myslím volební roky proč bábíš
tísň vzáky lze stranu dona růst testování srpn ministr ministerstva peněz
skutečně řašení milionů vyběh nobil



COVID-19
INFOMORE.CZ



Veřejnoprávní média nebo také média veřejné služby by měla občany vždy informovat o aktuálních událostech, které se v dané zemi odehrávají. Není proto překvapením, že ve wordcloudu složeného ze zářijových nejfrekventovanějších slov veřejnoprávních médií a jejich webů v souvislosti s onemocněním COVID-19, dominují slova: *kraji*, název hnutí *ANO*, *kandidátky* a *lídr*. Méně pozornosti byla věnována pandemii a vládním opatřením s ní spojená: *očkování*, *zdravotnictví*, *testování* a *případů*. Z analýzy vybraných zářijových wordcloudů vyplývá, že volbám do Poslanecké sněmovny Parlamentu České republiky v souvislosti s COVID-19 se nejvíce věnovala veřejnoprávní média (obrázek 8), komerční televize (obrázek 4) a zpravodajské časopisy (obrázek 5).

Způsob generování wordcloudů a použité data

Společnost Newton Media shromáždila od listopadu 2020 do září 2021 více než 716 tisíc mediálních zpráv monitorováním více než tří tisíc vydavatelů a jejich různých zpravodajských kanálů. V první fázi jsme proto využili metod dolování dat aplikovaných na texty zpráv v kombinaci se základními postupy zpracování přirozeného jazyka se zaměřením na specifiku češtiny.

Mediální zprávy jsou firmou Newton Media získávány z mnoha zpravodajských zdrojů. Jsou monitorovány jak tištěné formy deníků a časopisů, které se z jejich různých, např. elektronických, forem převádí manuálně, poloautomatizovaně, či plně automatizovaně do normalizovaného tvaru obsahujícího jak vlastní text, tak i jeho metadata. Zprávy se poté kopírují do speciálního úložiště. Podobně se monitorují i řečové kanály rozhlasových a televizních stanic, které se technikami rozpoznávání řeči převádí opět na text.

Úložiště umožňuje efektivně ukládat násobné výskyty zpráv, které jsou se stejným obsahem publikovány v různých variantách zpravodajských kanálů příslušného vydavatele. Např. některé regionální deníky přebírají základní centrálně vytvořený vzor a doplní na vybraná místa speciální zprávy týkající se daného konkrétního regionu. Tento fakt je potřeba zohlednit při vytváření statistik, proto jsou např. v současném exportu zpráv takové zprávy uvedeny pouze jednou. Je potřeba i podotknout, že ačkoliv firma Newton Media řadu dalších zpravodajských kanálů monitoruje, tak ne všechny zdroje či relace jsou archivovány.

Mezi analyzované texty byla zpráva zařazena, pokud obsahovala alespoň části slov s problematikou onemocnění COVID-19 související. Vzhledem k různým inflexím českých slov se vyhledávají všechny možné varianty přípon. Technicky se takový dotaz řeší pomocí tzv. regulárních výrazů, kdy znak “*” zastupuje libovolný řetězec znaků, v našem případě nějakou příponu slova. Výsledný dotaz je poměrně komplikovaný, ale pro jednoduchost můžeme uvést, že se hledají v principu následující slova: covi*, koron*, coron*, epidem*, pandem*, vir*, onemocnění*, zápal*, plic*, netopý*, sars*, karantén*, wuchan*, infek*, infik*, lékař*, zdravot*, šest*, nákaz*, nakaž*, vakcín*, čín*, lockdown*. Při exportu dat se obsah zpráv doplňuje o metadata a ve formě několika souborů komprimovaných pomocí ZIP techniky tak tvoří vstupní data do výpočetního procesu provádějící vlastní analýzu zaměřenou na problematiku projektu Infodemie.

Analýza dedikovaná hodnocení zaměření obsahu zpráv a jejich průběžných změn během vývoje pandemie pomocí wordcloudů je zahájena vytvořením přehledu, jak příslušný mediální zdroj používal jednotlivá slova v určitém měsíci. Před vytvořením přehledu se text zprávy upraví. Eliminují se všechny speciální znaky, které netvoří slova. Rovněž se v přehledu neuvažují čísla. Do tohoto přehledu nejsou zahrnuta slova zařazená mezi tzv. stopwords. Tato hojně se vyskytující slova typicky přispívají pouze do vytvoření syntaktické struktury věty tím, že vytváří vazby s dalšími slovy, ale hodnocení sémantického obsahu založeného pouze na slovech ovlivňují už mnohem méně.

Tzv. stopwords jsou navíc doplněna vybranými zájmeny a dalšími slovy, která se pouze odkazují na další slova, která však mohou být od takových zájmen poměrně již hodně v textu vzdálená. V rámci analýzy pomocí wordcloud diagramů jsme vyloučili i slova 'koronavirus', 'koronavir', neboť zprávy byly vybrány s ohledem na tato slova, a názvy vybraných médií, např. 'irozhlas', 'blesk', 'seznam', 'právo', 'právu', 'nova', 'tn', 'hn', 'televize', 'já', 'cz'. Tato zřejmě častá slova jsou jinak v diagramech uvedena jako nejčastější, a tudíž největší, a zbytečně proto zabírají místo ostatním slovům. Všechna slova se normalizují na použití pouze malých písmen. Další normalizace českého textu v analýze v současné době neprovádíme.

Podle cíleného diagramu se pak příslušným příkazem z databáze vybere frekvenční tabulka výskytů slov pro zvolenou množinu mediálních zdrojů a analyzované období a konstruují se wordcloudy. V diagramu se přibližně dvěma stovkám nejfrekventovanějších slov podle počtu výskytů slov nebo podle počtu zpráv obsahujících takové slovo přiřadí velikost fontu a barva podle vybrané škály použité barevné palety. Posledním krokem je umístění obarvených slov s přiřazenou velikostí fontu do diagramu. Existuje řada metod, které se liší i kvalitou výsledků. V našem projektu používáme knihovnu Wordcloud implementovanou v jazyce Python.

Reference

- Komenda M et al. COVID-19: Přehled aktuální situace v ČR. *Onemocnění aktuálně* [online]. Praha: Ministerstvo zdravotnictví ČR, 2020 [cit. 04.11.2021]. Dostupné z: <https://onemocneni-aktualne.mzcr.cz/covid-19>. Vývoj: společné pracoviště ÚZIS ČR a IBA LF MU. ISSN 2694-9423.
- Němcová Tejkalová A et al. Jak se v médiích referuje o koronaviru: wordcloudy nejčastěji se vyskytujícími slovy ve zpravodajství a publicistice o onemocnění COVID-19 (1. 12. 2019 – 31. 10. 2020). *COVID-19 infodemie: AI komunikační platforma potlačující infodemii ve vazbě na*

novinářskou a mediální etiku [online]. 2020. [cit. 25.10.2021]. Dostupné z: <https://www.infomore.cz/res/file/analyzy/20201214-jak-se-referuje/20201214-jak-se-referuje-o-koronaviru.pdf>

- Wordcloudy – komerční televize a jejich weby. *COVID-19 infodemie: AI komunikační platforma potlačující infodemii ve vazbě na novinářskou a mediální etiku* [online]. 2020. [cit. 25.10.2021]. Dostupné z: <https://www.infomore.cz/cs/wordcloudy-galerie/wordcloudy-komerzni-televize-weby/>

T A Projekt TL04000176 *Covid-19 infodemie: AI komunikační platforma potlačující*
Č R *infodemii ve vazbě na novinářskou a mediální etiku* je spolufinancován se
státní podporou Technologické agentury ČR v rámci Programu Éta