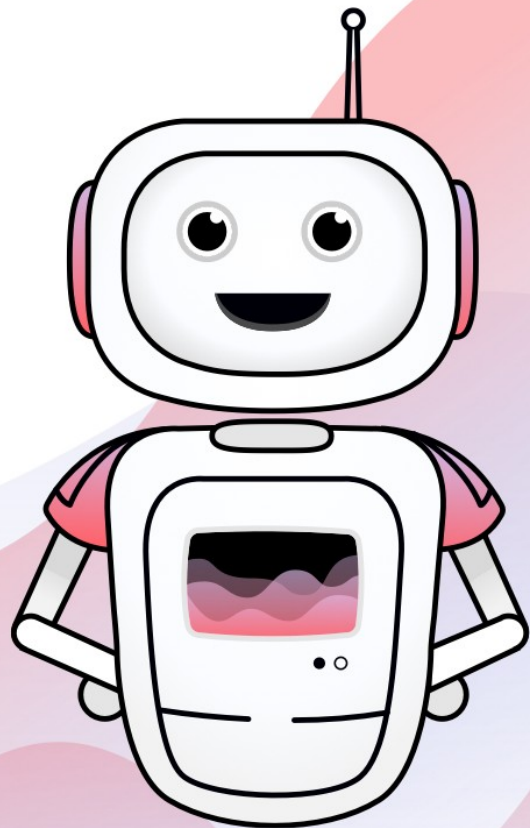


Common Voice

moz://a

aneb

Naučme počítače mluvenou češtinu



Jindřich Dítě
jdite@mozilla.cz

OpenAlt 2020

Proč je Common Voice?

- Google Assistant?
- Alexa?
- Cortana?
- Siri?
- Uzavřené!

Proč je Common Voice?

- Alternativy?
 - Mycroft ...běží nad Common Voice
- Jiné alternativy?
 - NLP, logika ...lze, pomůžou např. Wikidata
 - Rozpoznání řeči (!!!)

Proč je rozpoznávání řeči problém?

- Málo trénovacích dat
 - Wikimedia Commons?
 - VYSTADIAL?

Co je Common Voice?

- Databáze hlasových dat
- Volné dílo
- Crowdsourcing
- commonvoice.mozilla.com



Jak Common Voice funguje?

- Nahrávky celých vět
- Věta do 10 sekund
- Přesně čtené

Jak to probíhá?

- Sběr vět
- Nahrávání vět
- Ověřování
- Vydání 2x za rok

Sběr vět

- Nástroj pro sběr vět
- Wikipedia, Europarl dataset
- Specifické importy – konzultovat!

Nástroj pro sběr vět

- <https://commonvoice.mozilla.org/sentence-collector/>
- Nově předěláno
- 7.11. end-of-day deadline migrace!

Welcome to the Common Voice Sentence Collector

This is a website where we collect and review sentences for [Common Voice](#).

Home

How-to

Add

Review

Rejected Sentences

Statistics

Login

[Discourse](#)

[Report Bugs \(GitHub\)](#)

[Report copyright issues](#)

[Privacy](#)

[Terms](#)

[Cookies](#)



commonvoice.mozilla.org

[Log in with email](#)

jdite@mozilla.cz



Enter



Log in with Firefox



Log in with GitHub



Log in with Google

[Legal](#)

[Privacy](#)

[Need help?](#)

Profile: jdite@mozilla.cz

Migrate your stats and profile settings to your new account now. To do so, use our [migration form](#). We will remove that form on November 7th.

Your languages:

You have not added any languages yet, please add at least one below.

Add a language you want to contribute to

Settings

Experimental: There are two different tools with which you can review sentences. The normal tool lists 5 sentences per page and has an approval and rejection button each. The Swiping tool displays one card at a time where you can swipe right or left to approve and reject sentences. Both work on Desktop, for touch interfaces we would suggest to try out the swiping tool.

[Home](#)

[How-to](#)

[Add](#)

[Review](#)

[Rejected Sentences](#)

[Statistics](#)

[Profile](#)

[Migrate Account](#)

Statistics

Last Update: 3. 11. 2020 17:39:25

The Common Voice Sentence Collector has collected 1975739 sentences in 105 languages!

čeština (Czech)

- 9068 total sentences.
- 725 sentences in review.
- 725 sentences left for you to review. [Review now!](#)
- 7797 validated sentences.
- 546 rejected sentences.

jdite@mozilla.cz

[Logout](#)

[Discourse](#)

[Report Bugs \(GitHub\)](#)

[Report copyright issues](#)

[Privacy](#)

[Terms](#)

[Cookies](#)

Add Sentences

Select Language

Czech (čeština) ▾

Add public domain sentences

Pozdravy divákům mé přednášky!
Letos bohužel ne z Brna.
Dá-li bůh, tak snad napřesrok.

Where are these public domain sentences from?

Own creation

I confirm that these sentences are public domain and I have permission to upload them.

Submit

[Home](#)[How-to](#)[Add](#)[Review](#)[Rejected Sentences](#)[Statistics](#)[Profile](#)[Migrate Account](#)jdite@mozilla.cz[Logout](#)[Discourse](#)[Report Bugs \(GitHub\)](#)[Report copyright issues](#)[Privacy](#)[Terms](#)[Cookies](#)

Review Sentences

Czech (čeština) ▾

[① Review Criteria](#)

Mnoho v nich dokonce nověnarůstá.

Source: PD old 70, R,U.R., taken from wikisources



Amy Schumerová dokáže být velice vtipná.

Source: mozilla.cz sentence collector (people had to confirm that these sentences are public domain)



Nejsou tarantinovské a nesnaží se být tarantinovské.

Source: mozilla.cz sentence collector (people had to confirm that these sentences are public domain)



Předvedená varianta

Source: mozilla.cz sentence collector (people had to confirm that these sentences are public domain)



Navržená varianta

Source: mozilla.cz sentence collector (people had to confirm that these sentences are public domain)

[Finish Review](#)

1

<

1

>

145

Požadavky přidávání vět

- **CC0**
- Číslice, zkratky – NE
- Speciální znaky - .,!? ano, jinak ne
- Max 14 slov, kratší ⇒ lepší

Požadavky ověřování vět

- **CC0**
- Bez pravopisných, gramatických chyb
- Snadno vyslovitelné
- Nevíš \Rightarrow NE

Nahrávání vět

- commonvoice.mozilla.org
- Potřeba:
 - “moderní” prohlížeč
 - Libovolný mikrofon
- Sady po pěti větách

Mluvte

Darujte svůj hlas



Poslouchejte

Pomozte nám ověřovat nahrávky



Projekt Common Voice je iniciativa Mozilly, která pomáhá strojům učit se, jak mluví skuteční lidé.

Hlas je přirozený a lidský. Proto nás tolik zajímá tvorba použitelné hlasové technologie pro naše zařízení. Aby ji ale vývojáři mohli vytvořit, potřebují spoustu hlasových dat.

Většina dat používaných velkými společnostmi nejsou dostupná pro většinu lidí. My si ale myslíme, že to jen zdržuje inovace. Proto jsme spustili projekt Common Voice, projekt, který udělá rozpoznávání hlasu dostupné pro všechny.

[PŘEČÍST SI VÍCE](#)

Hodin nahráno

Hodin ověřeno

VŠE



Aktivní hlasy

VŠE



Děkujeme za potvrzení vašeho účtu, nyní si sestavte svůj profil.

Tím, že nám o sobě poskytnete nějaké informace budou data, která odešlete do Common Voice, více užitečná pro systémy pro prozopznávání řeči, které tato data využívají ke zvýšení přesnosti.

Proč na tomhle záleží? ▾

Uživatelské jméno jdite_mozillac2	Viditelnost v žebříčku Viditelný ▾
Věk 19 - 29 ▾	Pohlaví Muž ▾
Mateřský jazyk Čeština ▾	Príзвиuk ▾
Další jazyk English ▾	Príзвиuk ▾
Další jazyk ▾	Príзвиuk ▾

Přidat jazyk +

E-mail
jdite@mozilla.cz

- Připojte se do elektronické konference Common Voice**
Dostáváte e-maily jako jsou připomínky k nedosaženým cílům či vývám, informace o postupu nebo novinky o projektu Common Voice.
- Zásady ochrany osobních údajů**
Souhlasím se zpracováním těchto informací jak je popsáno v zásadách Mozilly pro [ochranu osobních údajů](#)

[Četli jste naše podmínky?](#)




Mluvte

Poslouchejte

1/5 nahrávek

SPUSTIT NAHRÁVÁNÍ

1

Klikněte  a přečtete nahlas větu

2


3

4

5

Závodník musí překonat
všechny překážky.



 Zkratky

 Hlášení

Přeskočit >>

ODESLAT

Požadavky nahrávání

- **Přesná** výslovnost
- Hluk na pozadí \Rightarrow ano
- Řeč na pozadí \Rightarrow ne
- Problém s větou \Rightarrow Nahlásit



Mluva

Poslouchejte

1/5 nahrávek

SPUSTIT NAHRÁVÁNÍ



2

3

4

5

Nahlásit



Jaké máte potíže s touto větou?

Urážky

Věta obsahuje urážlivé nebo sprosté výrazy.

Gramatické chyby

Věta obsahuje gramatické chyby nebo překlepy.

Jiný jazyk

Věta je v jiném jazyce než jaký mám nastaven.

Obtížná výslovnost

Věta obsahuje těžko čitelná a vyslovitelná slova nebo fráze.

Ostatní

HLÁŠENÍ ↑

Zkratky

Hlášení

dit >>

ODESLAT

Ověřování vět

- commonvoice.mozilla.org
- Sady po pěti větách
- Potřeba
 - “moderní” prohlížeč
 - reproduktor




Mluvte

Poslouchejte

1/5 nahrávek



1

Klikněte  byla věta nahrána přesně?

2

3

4

5


Pod křížem se nachází
vavřínový věnec.




ANO



NE

 Zkratky

 Hlášení

Přeskočit 

Požadavky ověřování

- **Přesná** výslovnost?
- Tip: Napřed poslechnout, poté přečíst

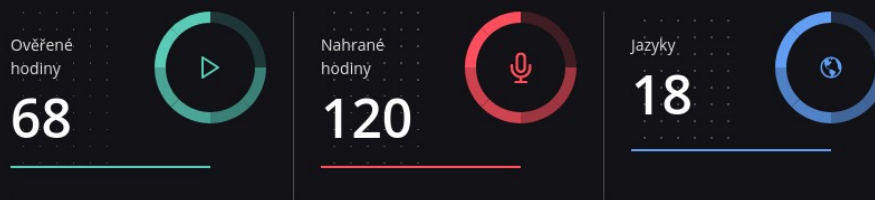
Pravidelná vydání

- 2x do roka
- Červen, Prosinec



Download the Single Word Target Segment

This is a use case driven segment containing data to power spoken digit recognition, yes / no detection, and wakeword testing data for [Firefox Voice](#).



Pro stažení zadejte svou e-mailovou adresu



Proč e-mail? Je možné, že vás budeme potřebovat v budoucnu kontaktovat ohledně změn v datech.

Vytváříme

otevřenou mnohojazyčnou databázi hlasových záznamů, kterou může kdokoli použít k trénování svých hlasových aplikací.

Věříme, že velké, veřejně dostupné hlasové datové soubory podpoří inovace a zdravou konkurenci firem a technologií pro rozpoznávání řeči pomocí strojového učení.



Jazyk	Čeština
VELIKOST	774 MB
VERZE	cs_29h_2020-06-22
CELKEM VALIDOVANÝCH HODIN	26
CELKOVÝ POČET HODIN	29
LICENCE	CC-0

Cílový segment

- Slova Ano/Ne, 0 – 9
- Čeština od čtvrtka



Děkuji za pozornost

- Pozadí prezentace převzáno z repozitáře Common Voice, licence MPL v2.0, <https://github.com/mozilla/common-voice/tree/main/web/img>
- Ikona Public Domain převzata z creativecommons.org