

# Rekonstrukce standardizovaného textu z mluvené řeči

Pavel Češka  
*ceska@ufal.mff.cuni.cz*

24. dubna 2009

# Ukázka

## **Spoken text:**

"uh what used to be i know i r- remember were fairly simple uh tasks uh mathematical is what i'm saying uh i- it's like i sort of have forgotten uh how to do those things. so for me the things that i used to know about algebra that's all i can really think of at the moment as an example it really is a strain to get back and do."

## **Reconstructed text:**

"i have forgotten how to do what used to be fairly simple mathematical tasks. it really is a strain for me to get back and do the things that i used to know about algebra."

– *Verbatim, and reconstructed, text taken from the Fisher Conversational Telephone Speech corpus and Spontaneous Speech Reconstruction corpus*

# Proč rekonstruovat?

- Rekonstrovaná řeč je jednodušší a přesnější pro další zpracování přirozeného jazyka (tagging, parsing)
- Rekonstrovaná řeč je vhodnější pro automatický překlad, Q&A systémy, sumarizaci textu, apod.
- Rekonstruovaná řeč je pro lidi čitelnější

# Druhy chyb mluvčího

- Mluvčí opravuje svou předchozí promluvu
- Věta promluvy je negramatická (neshoda v pádu, rodu, čísle...)
- Nedokončené výpovědi
- Nepřesné rozdělení promluvy do větných celků
- Neřečové události (zakašlání...)

# Moje práce

- Identifikace a klasifikace chyb mluvčího
- Rekonstrukce věty dle zamýšleného obsahu promluvy
- Náhled na problém jako překlad mezi dvěma podobnými jazyky (používám nástroj Moses)
- Vyhledávání míst z chybami mluvčího (za použití statistického jazykového modelu)