Jazyky a jazykové technológie na Slovensku

prof. Ing. Jozef Juhár, CSc.
Katedra elektroniky a multimediálnych telekomunikácií
FEI TU v Košiciach
http://kemt.fei.tuke.sk

ELRC Workshop, Bratislava, 14.3.2016

KEMT
Katedra elektroniky
a multimediálnych telekomunikácií

Obsah

1. Úvod
2. Základné princípy jazykových technológií a súčasný stav v ich výskume na Slovensku
3. Niektoré podporné nástroje na jazykové zdroje na výskum a vývoj jazykových technológií na TUKE (a ÚI SAV)
4. Aplikačné výstupy výskumu a vývoja rečových technológií na TUKE (a ÚI SAV)
5. Záver

Úvod

- V histórii každej civilizácie a každého národa hrá jazyk významnú rolu, tak kultúrnu, politickú ako aj ekonomickú.

- V histórii Slovenska zohral aj významnú emancipačnú úlohu.

- Zatiaľ čo v minulosti zaostávanie resp. plnohodnotný vývoj jazyka bol výsledkom ekonomicko-politicko-vojenských okolností.

- V súčasnosti napredovanie ale aj úpadok jazyka môže byť výsledkom technologického zaostávania.

Úvod

Krátke historické pripomenutie

- Reč a písmo – dve podoby jazyka, nesúce informáciu
- Hovorená forma jazyka má dve základné nedostatky – časové a priestorové obmedzenie
- Písaná forma – objav písma – môže byť považovaná za vôbec prvú jazykovú technológiu – jej nedostatkom bola stále priestorová obmedzenosť – časové oneskorenie, ktorú úplne neriešil ani objav kníhtlače.
- Prvé komplexné riešenie výmeny informácie hovorenou rečou priniesol objav telefónu, ktorý umožnil výmenu informácie v reálnom čase a na ľubovoľnú vzdialenosť.
- Ďalšiu revolučnú zmenu v spracovaní a prenose informácie v písanej forme prinesli osobný počítač a počítačová sieť

Úvod

Ako to vyzerá dnes?

- bezdrôtový prenos a mobilný telefón
  - nulové časovo-priestorové obmedzenie v komunikácii
- mobilné telefónne siete prepojené s internetom
  - mobilný komunikačný terminál – kombinácia telefónu a počítačového terminálu
- mobilný telefón + internet + sociálna sieť
  - nová forma „technologicky“ podporovanej skupinovej „diaľkovej“ komunikácie v reálnom čase
- V takých podmienkach sa význam jazykových technológií umocňuje.

Témy jazykových technológií v programoch EÚ

- Automatizovaný preklad
- Tvorba a manažovanie obsahu na multilingválnej báze (Multilingual content authoring and management)
- Rečové technológie a interaktívne služby
- Analytika obsahu (content analytics)
  - Textová
  - Rečová
- Jazykové zdroje
- Kolaboratívne platformy.

Jazykové technológie na podporu slovenčiny

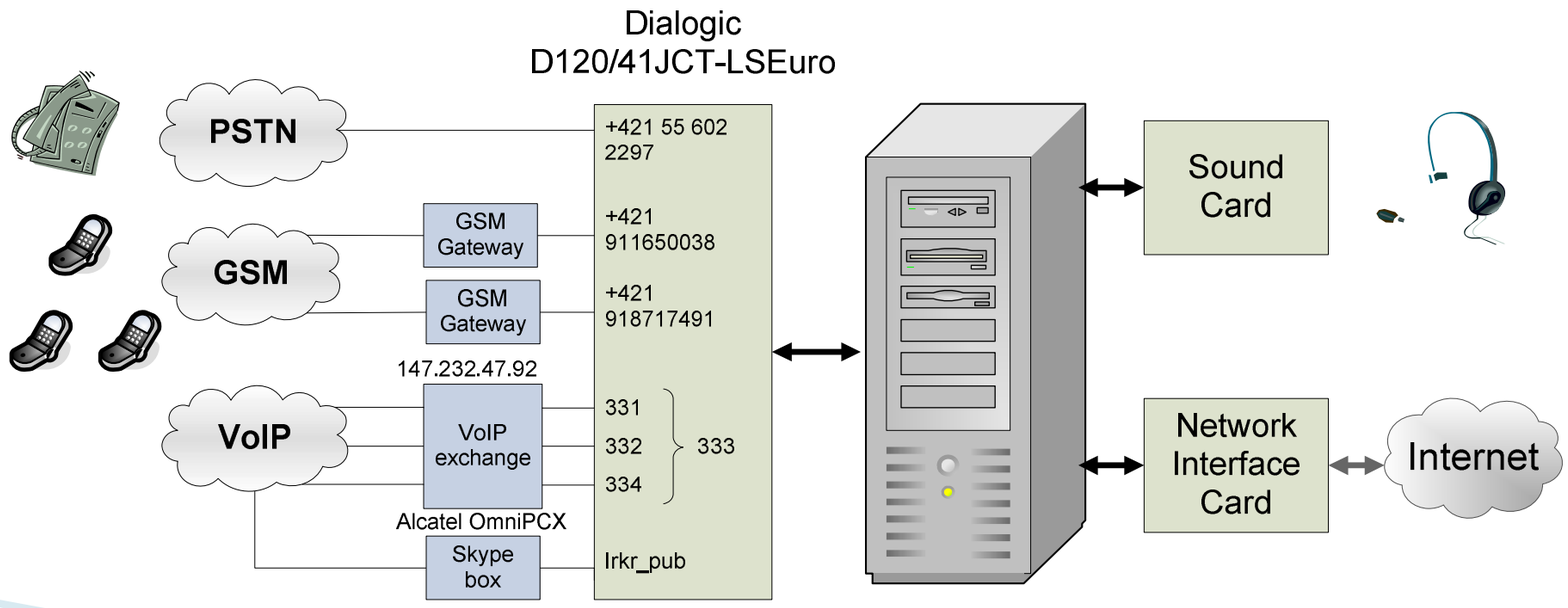
- K začiatku prebiehajúcej dekády sú pomerne dobre zmapované v publikácii:
  - Mária Šimková et al., Slovenský jazyk v digitálnom veku/The Slovak Language in the Digital Age. Springer-Verlag, 85p., 2012, ISBN 978-3-642-30369-2.

- Výsledky akademického výskumu a vývoja jazykových technológií sú naďalej viditeľné najmä na pracoviskách:
  - SAV (JULS, ÚI)
  - „klasických“ univerzít (UK, PU, UPJŠ)
  - technických univerzít (STUBA, TUKE, ŽU)

- Charakteristika akademického výskumu:
  - veľa malých tímov, riešiacich malé partikulárne témy
  - spôsob financovanie nepodporujúci spoluprácu medzi pracoviskami

Inteligentné rečové komunikačné rozhranie (2003 – 2006)

- Pozitívny príklad zmysluplného financovania
- Spolupráca TUKE, ÚI SAV, STUBA a ŽU
- Pokračujúca spolupráca TUKE a ÚI SAV



Slovenské telefónne rečové korpusy

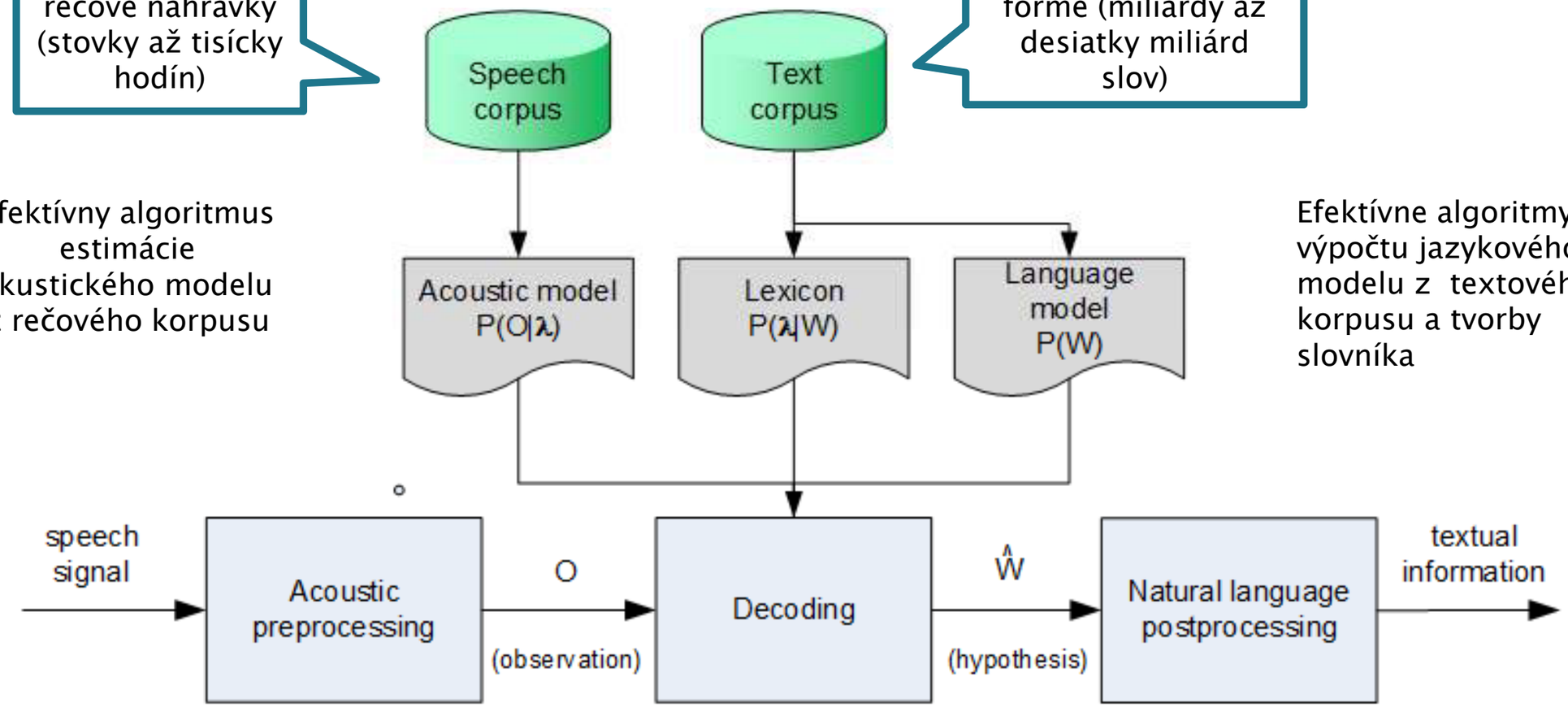
- SpeechDat-E Slovak
  - Jediná „regulérna“ databáza telefónnej reči slovenčine komerčne ponúkaná v sieti ELRA
  - 1000 hovoriacich / PSTN
- MobilDat-Sk
  - Vytvorená podľa vzoru SpeechDat-E v rámci projektu IRKR
  - >1000 hovoriacich / GSM

Spolupráca TUKE a ÚI SAV (2006 –2016)

Rozpoznávanie plynulej slovenčiny

Anotované rečové nahrávky (stovky až tisícky hodín)

Textové dokumenty v elektronickej forme (miliardy až desiatky miliárd slov)



Efektívny algoritmus estimácie akustického modelu z rečového korpusu

Efektívne algoritmy výpočtu jazykového modelu z textového korpusu a tvorby slovníka

Rýchly, presný a spoľahlivý prehľadávací algorimus

Anotované rečové korpusy pre slovenský systém rozpoznávania plynulej slovenčiny

- Parlamenté debaty (cca. 120 hod.)
- Diktovaná reč z domény súdnictva APD1 (cca. 100 hod.)
- Diktovaná reč z domény súdnictva APD2 (cca. 150 hod.)
- Spravodajské relácie Bnews1 a 2 (cca. 200 hod.)
- Korpus „Súdna sieň“ (cca. 80 hod.)
- Spravodajské relácie Bnews-Multichannel (20 hod.)
- Korpus prednášok „Tedex“ (12 hodín)

Podporné nástroje na tvorbu jazykových modelov

Agenty na zber textu z internetu

- Dolovanie slovenských web stránok a extrahovanie textov z HTML kódu
  - Všeobecný crawler
  - Špecializované

- Ukladanie textov a meta-informácií (kľúčové slová, kategória, dátum, autor, ...) lokálne vo forme databázy

- Zbieranie dát v niekoľkých etapách od roku 2010

- Ostatná etapa od 10/2015 do 03/2016
  - 4.6 GB skomprimovaných dát
  - 5,3 milóna dokumentov
  - 185 miliónov viet
  - 1.82 miliardy tokenov

[1] D. Hládek, J Staš: Text mining and processing for corpora creation in Slovak language, Journal of Computer Science and Control Systems 3 (1) 2010, pp. 65

Podporné nástroje na tvorbu jazykových modelov

Tokenizátor

- Založený na pravidlách (Chomského hierarchia)
- Implementovaný v jazyku Ragel + kompilácia do jazyka C++
- Výsledkom je stavový stroj
- Vykonáva lexikálna analýza
  - Identifikácia hraníc slov a viet
  - Identifikácia čísiel a skratiek
- Súčasťou tokenizátora je pravidlový anotátor
  - Prepis čísloviek do slovnej podoby
  - Indetifikácia pomenovaných entít
  - Korekcia preklepov
- Demo: http://nlp.web.tuke.sk/nlpform
- Zdrojový kód: https://github.com/hladek/slovak-lexer

[1] Ján Staš, Daniel Hládek, Matúš Pleva, Jozef Juhár: Slovak Language Model from Internet Text Data Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues Volume 6456 of the series Lecture Notes in Computer Science pp. 340-346

Podporné nástroje na tvorbu jazykových modelov

Morfologický značkovač Dagger

- Morfologická analýza a klasifikácia slov v texte
  - Identifikácia koncovky slova
  - Identifikácia gramatickej kategórie slova vzhľadom na kontext

- Demo: http://nlp.web.tuke.sk/nlpform
- Zdrojové kódy: budú dostupné

[1] D. Hládek, J. Staš and J. Juhár : Morphological Analysis of the Slovak Language, Advances in Electrical and Electronic Engineering 13 (4) 2015, pp. 289-294
[2] D. Hládek, J. Staš and J. Juhár, "Dagger: The Slovak morphological classifier," ELMAR, 2012 Proceedings, Zadar, 2012, pp. 195-198

Podporné nástroje na tvorbu jazykových modelov

Automatický korektor preklepov

- Identifikuje chybné slová
  - Pomocou slovníka a pravidiel Hunspell
- Navrhuje kandidátov na opravu chybného slova
- Zoraďuje kandidátov na opravu podľa kontextu a podľa druhu chyby
  - Pomocou jazykového modelu a učiacej sa metriky medzi slovami

[1] D. Hládek, J. Staš and J. Juhár: Unsupervised spelling correction for Slovak, Advances in Electrical and Electronic Engineering 11 (5) 2013

Klasifikácia a kategorizácia dokumentov

- Chceme znížiť perplexitu jazykového modelu a mieru chybovosti systému rozpoznávania reči.
- Jedným z riešení sú doménovo orientované jazykové a akustické modely, adaptované pomocou dokumentov, kategorizovaných podľa témy
- Pretože informácia obsiahnutá v HTML kóde nie vždy postačuje, skúmame ďalšie nástroje na kategorizáciu dokumentov.
- Zhlukovacie a klasifikačné algoritmy, založené na:
  - kľúčových slovách a kľúčových frázach
  - Metódach kontrolovaného učenia
  - Metódach nekontrolovaného učenia

[1] Staš, J., Zlacký, D., Hládek, D., Juhár, J. Categorization of unorganized text corpora for better domain-specific language modeling (2013) Advances in Electrical and Electronic Engineering, 11 (5), pp. 398-403.
[2] Staš, J., Juhár, J., Hládek, D. Classification of heterogeneous text data for robust domain-specific language modeling (2014) Eurasip Journal on Audio, Speech, and Music Processing, 2014, art. no. 14.

Špecializované jazykové (textové) korpusy

- Podporné jazykové nástroje boli použité na prípravu viacerých špecializovaných jazykových korpusov
- Všeobecný webový korpus
  - Korpus webových blogov
  - Korpus webových diskusií
  - Korpus novinových článkov
  - Korpus dokumentov verejnej správy
- Iné špeciálne korpusy
  - Korpus sémantických rolí

Špecializované jazykové (textové) korpusy

Korpus webových diskusií

- Anotovaný ako v korpus novinových článkov
- Doplnený o automatickú anotácie kandidátov na opravu
- Vhodný na štúdium „ledabolo“ písaného prejavu
- Obsahuje veľa slangových a nárečových výrazov
- Dostupnosť : http://nlp.web.tuke.sk/pages/webdiscussion

D. Hládek, J. Staš, J. Juhár: Slovak Web Discussion Corpus, Advances in Natural Language Processing, 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014

Špecializované jazykové (textové) korpusy

Korpus novinových článkov

- Obsahuje 3164 dokumentov – novinových článkov s rôznymi témami
- Korpus je anotovaný (Kategória, autor, morfologický tvar slova, léma, pomenované entity)
- Vhodný na vyhodnocovanie systémov/algoritmov vyhľadávania informácií v texte (80 vzorových dopytov)
- Dostupný (na požiadanie): http://nlp.web.tuke.sk/pages/categorizednews

D. Hládek, J. Staš, J. Juhár: Slovak Categorized News Corpus, LREC 2014, Reykjavik, 2014

Špecializované jazykové (textové) korpusy

Korpus sémantických rolí „Semienko“

- Obsahuje 300 jednoduchých viet v slovenčine.
- Je manuálne anotovaný podľa modifikovanej dvojúrovňovej schémy označovania sémantických rolí.
- Je určený na experimentovanie v oblasti komunikácie človeka so strojom.
- Predbežné výsledky ukazujú na nutnosť rozširovania korpusu !!!

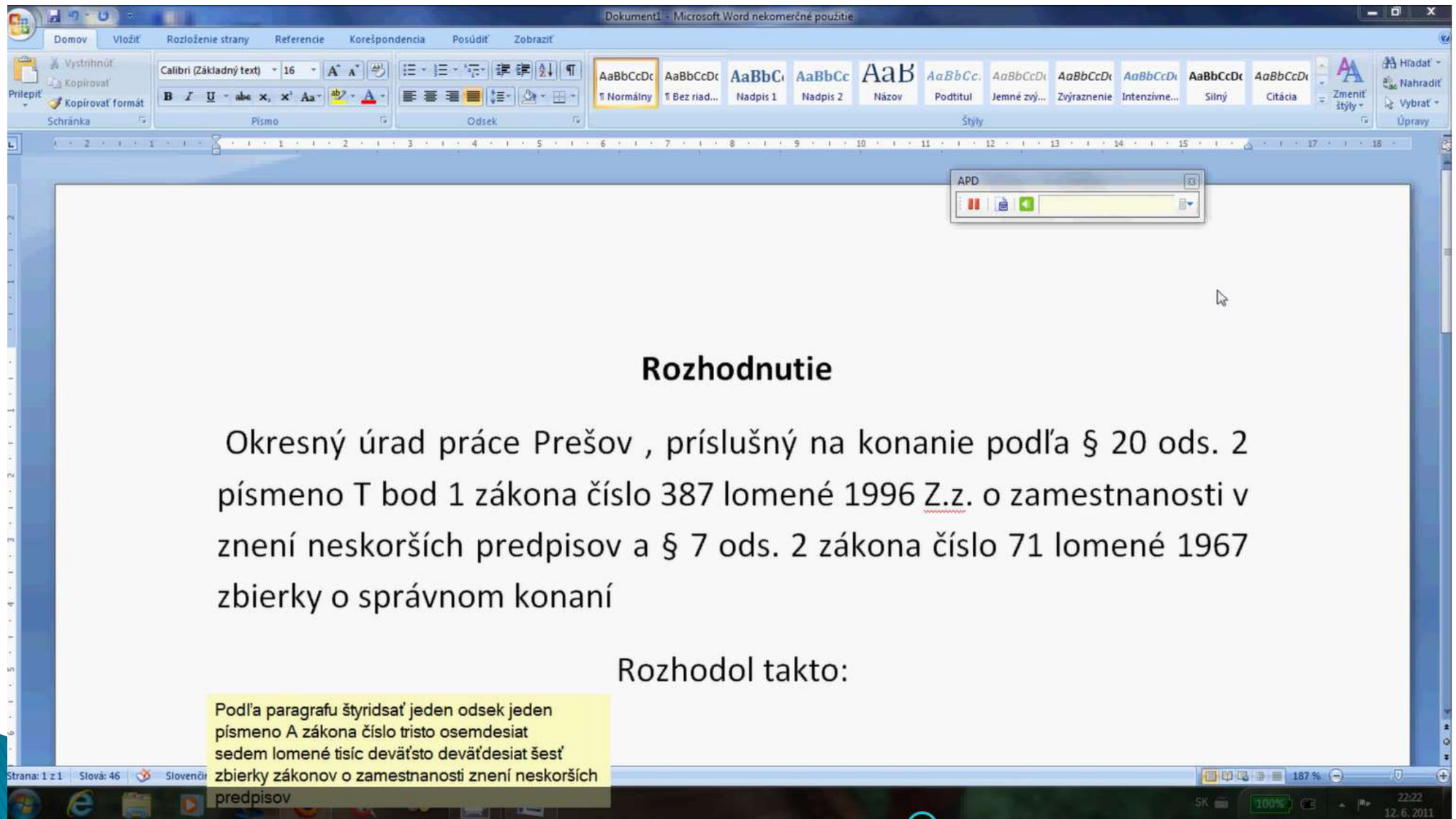
S Ondáš, D Hládek, J. Juhár: Semantic roles labeling system for Slovak sentences. 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom), 2014, 161-166

Vybrané aplikačné výstupy (2008 – 2011, 2013–2015)

Automatický prepis diktátu

Ústav informatiky SAV

Technická univerzita v Košiciach

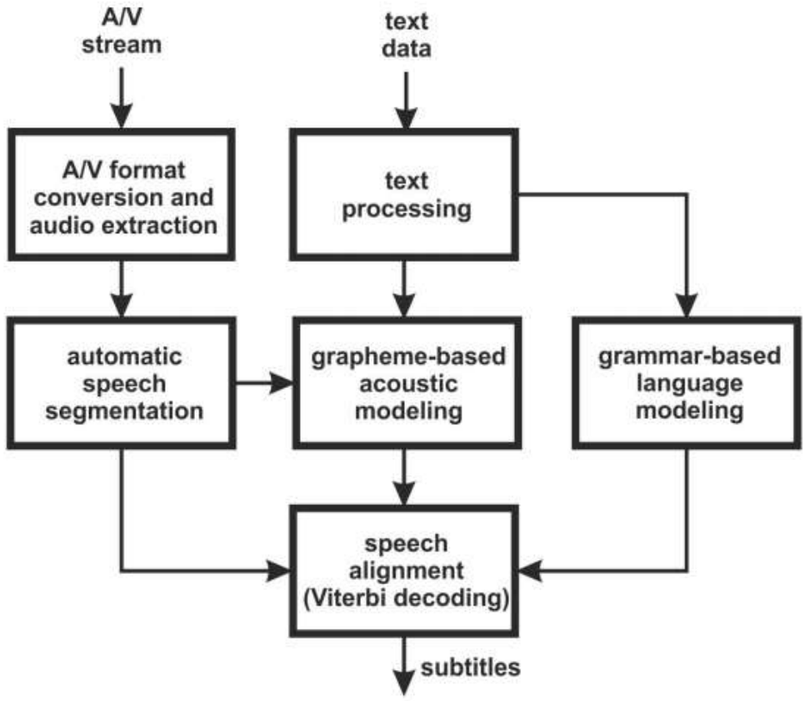


Vybrané aplikačné výstupy (2015)

Automatické titulkovanie pre nepočujúcich

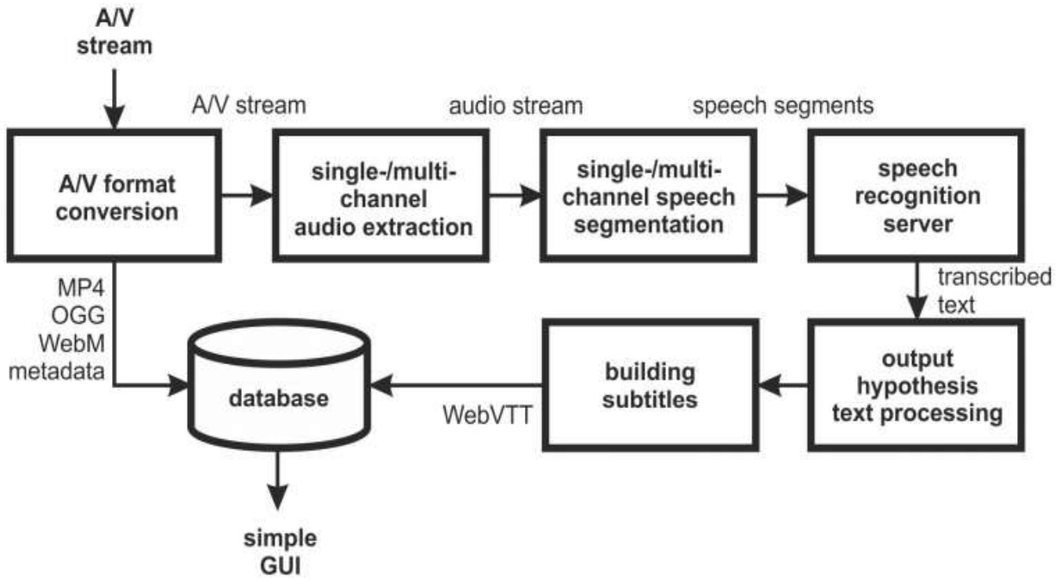
Ústav informatiky SAV
Technická univerzita v Košiciach

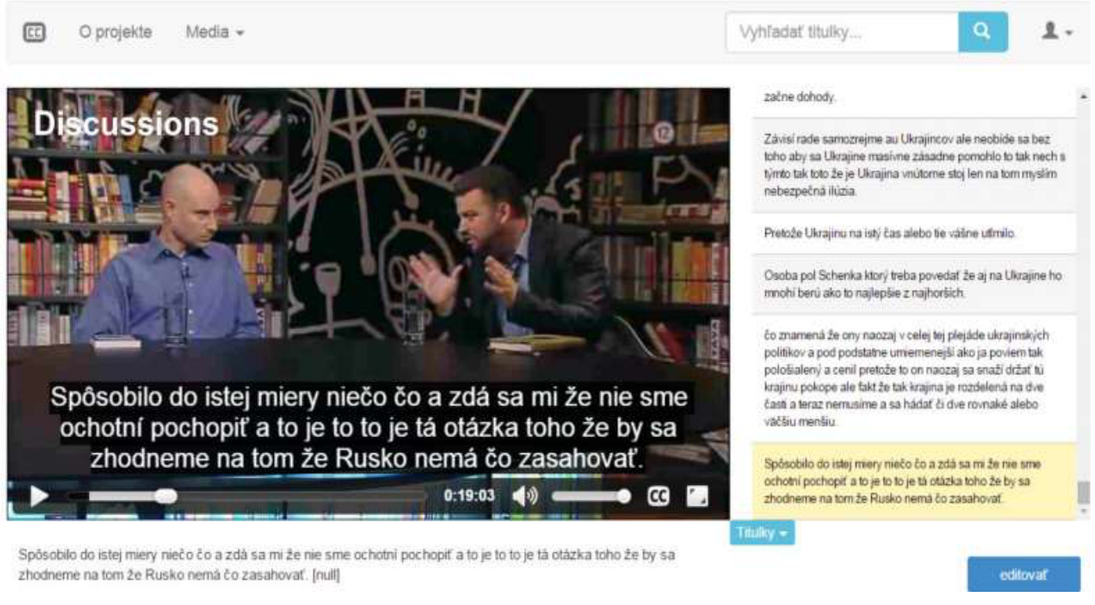
- Spolupráca s TV Markíza
- Systém je založený na automatickom „off-line“ zarovnaní hlasu a textu (na úroveň viet).
- Textovým vstupom sú nespracované scenáre (dabingové alebo nahrávacie)
- Akustickým vstupom je audio stopa nahrávky
- Pracujeme na „on-line“ zarovnávaní „živého“ hlasu a textu na úrovni slov
- Aplikovateľné na titulkovanie živého spravodajstva



Vybrané aplikačné výstupy – prebiehajúci výskum a vývoj

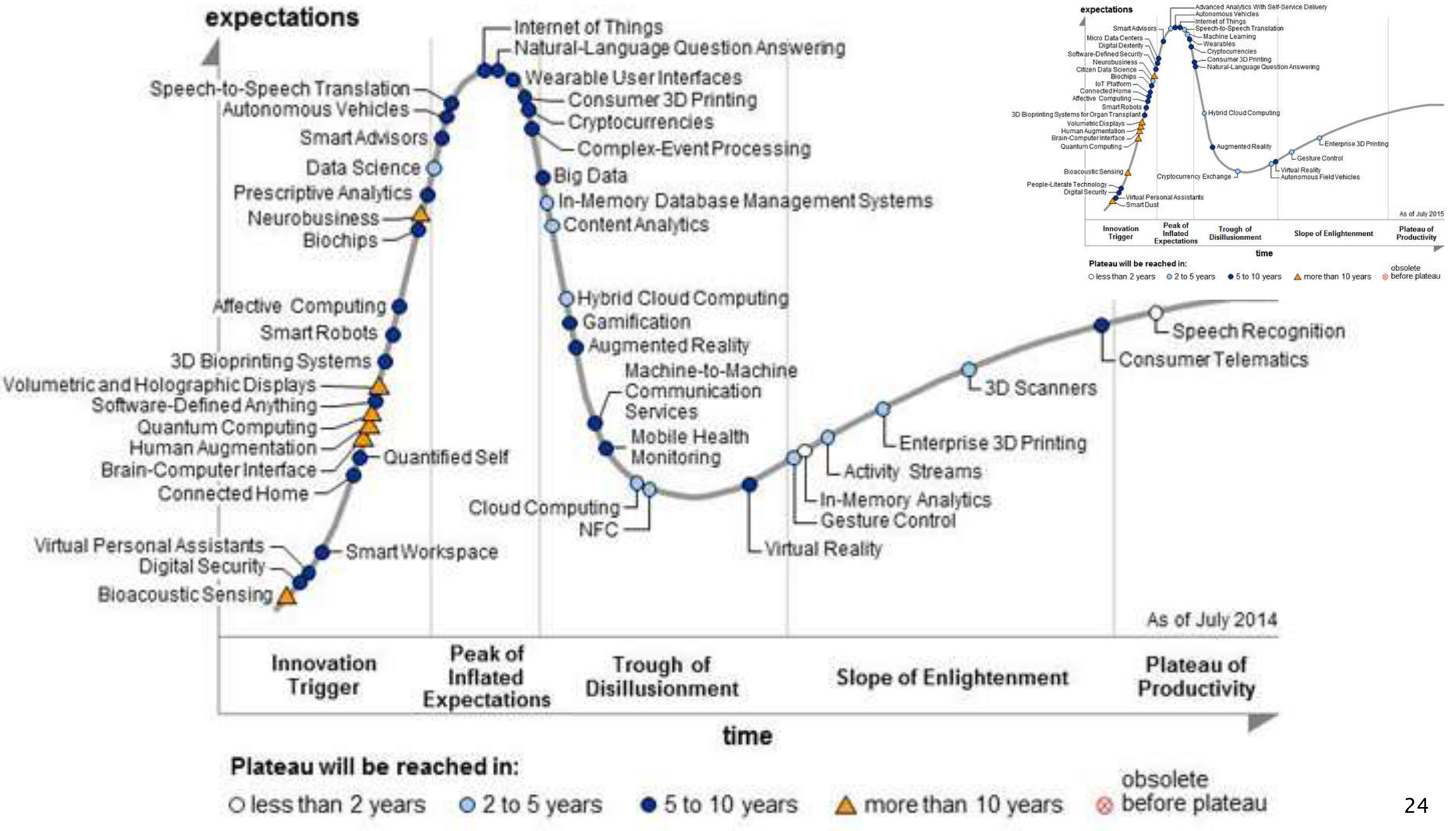
Automatická transkripciu audio-vizuálnych záznamov (s možnosť ou indexovania)





Kam sa uberajú jazykové technológie?

Gartner's Hype Cycle of Emerging Technologies



Ďakujem za pozornosť