CKL

Centrum komputační lingvistiky

Projekt MŠMT LC536 (LC05)

Univerzita Karlova v Praze, ÚFAL MFF
Západočeská univerzita Plzeň, KKY FAV
Masarykova Univerzita Brno, FI
Ústav pro jazyk český AV ČR Praha

http://www.centrumkomputacnilingvistiky.cz

Program kontroly na pracovišti 12.5.2009 MFF UK, Malostranské nám. 25 Posluchárna S9, 1. patro

- 8:30 - 9:20 Prezentace Centra (J. Hajič)
- 9:20 - 9:50 Diskuse
- 9:50 - 10:15 Prohlídka pracoviště a výpočetních kapacit
- 10:15 - 10:30 Zápis a závěr

Projekt Centra

- Cíl:
  - integrovat statisticko-matematickou, počítačovou a lingvistickou složku výzkumu
  - integrovat výzkum mluvené řeči a zpracování jazyka
  - vytvořit anotovaná data
  - vytvořit nástroje (významové) analýzy a syntézy
  - integrovat lexikální zdroje vč. software

Kontext vzniku Centra

- Dříve: Centrum komputační lingvistiky (program MŠMT LN), 2000-2004
  - UK, ÚJČ, ZČU
- Nyní: Centrum komputační lingvistiky
  - (opět) základní výzkum, program MŠMT LC05
  - k existujícím partnerům přibyla Fakulta informatiky MU Brno (Laboratoř NLP)

Centrum komputační lingvistiky: personální obsazení

- Rozpočet a doba řešení
  - 53,6 mil. Kč, 2005-2009 (4 roky 9 měs.)
- Personální obsazení (2009):
  - 1 řešitel-koordinátor (prof.)
  - Dále: 7 řešitelů a garantů (3x prof., 4x doc.)
  - 9 mladých věd. pracovníků (Ph.D.)
    - Z toho 8 obhájilo v době zaměstnání v CKL
  - 24 doktorandů (Mgr., Ing., RNDr.)

Pracoviště (1)

- UK Praha (ÚFAL MFF)
  - vytváření jazykových dat
  - teoretický výzkum jazyka i metod zpracování
  - SW nástroje pro analýzu a syntézu
- ZČU Plzeň, KKY FAV
  - analýza a syntéza mluvené řeči, zejm. akustika
  - tvorba dat (transkripce, anotace)

Pracoviště (2)

- MU Brno, FI, NLP laboratoř
  - lexikální nástroje
    - lexikální databáze: definice, správa, využití
- ÚJČ AV ČR
  - elektronizace historických lexikálních dat
  - spolupráce na definici lexikální databáze

Koordinace a komunikace

- Neformální i formální schůzky
- Rada Centra
  - Jednou ročně
- Představení projektu pro širší veřejnost
  - Naposledy 22.5.2008, Praha
- Spolupráce v rámci projektů EU / USA
- Obhajoby doktorských prací, rigorózní zk.

Rok 2005

- Zahájení práce v Centru
  - (Až) 1.4.2005 - překlenutí ze zdrojů instituce
  - Změna názvu (tradice, návaznost)
  - Redukovaný rozpočet na cca ¾: 7,0 mil. Kč
  - Pořízení investic (výpočetní stroje - budoucí výpočetní cluster) - cca 1,7 mil. Kč
  - Rozdělení studentů a prací mezi projekty
  - Žádosti o evropský projekt (několik)

Rok 2006

- První ucelený rok práce
  - Dokončení projektu PDT 2.0 (UK, vyd. v USA)
  - Projekt „Rekonstrukce řeči" (UK, specifikace)
  - Práce na slovnících (UK, MU)
  - Mluvená řeč – analýza, syntéza (hl. ZČU)
  - IR – CLEF testovací kolekce, mez. soutěž, 1. část
  - Digitalizace hist. zdrojů (ÚJČ)
  - Spol. mez. projekt EU (IP): UK, ZČU
  - Další zahr. spolupráce: EU, USA
  - 40 výsledků v RIV

Rok 2007

- Polovina projektu
  - Lexikální zdroje (UK, MU + ÚJČ)
  - Důraz na češtinu i angličtinu (noví pracovníci v týmu)
    - Specifikace anotace, anotační software, anotace dat na všech rovinách (ZČU, UK)
  - Integrace mluvené řeči a zpracování jazyka (UK, ZČU)
  - Nástroje pro aut. analýzu a syntézu jazyka
    - Mluvená řeč i psaný jazyk (UK, MU, ZČU)
  - Pokračování mez. spolupráce
    - EU (3 projekty 6.RP: UK, UK+ZČU), USA (UK, UK+ZČU)
  - Organizace celosvět. konference ACL – 1000 úč. (UK)
  - 66 výsledků v RIV (16 čas., 39 sb., 5 SW/data atd.)

Rok 2008

- Upravené cíle
  - Lexikální zdroje (MU, UK, ÚJČ)
    - Softwarové lexikální nástroje
  - Sémantika (rozp. plagiátů: MU),
    - analýza morf., synt., sém. (UK, MU), generování (UK)
  - Nové algoritmy rozpoznávání řeči
    - Prozodie, jaz. modelování, rekonstrukce řeči
  - Získávání dalších jazykových dat, korpusové nástroje
  - Anotace dat a výzkum pro strojový překlad
    - Platforma TectoMT pro strojový překlad
  - Teoretická formální lingvistika, užití jazyka
- Výsledky (RIV): 64, 13 čl., 32 sb., 5 knih, 5 SW aj.

UK v Praze – MFF, ÚFAL

- Rozvoj PDT 2.0
  - Formalizace obsahu sdělení (teoret. výzkum + anotace)
  - teoreticko-empirický výzkum (př.: diskurs)
- Anotace mluvených dat (rekonstrukce řeči)
- Nové metody morf. disambiguace, anotace ČNK
- Generování češtiny, angličtiny z formálního zápisu
- Dialogové systémy – integrace porozumění jazyku
  - Výsledky budou využity pro EU IP „Companions"
- „Information retrieval" – data a aut. Zpracování
- Strojový překlad - nástroje a data

ZČU Plzeň – FAV, Kat. kybernetiky

- Rozpoznávání řeči
  - Parametrizace signálu
  - Akustické a jazykové modely
- Syntéza řeči
  - Prozodické chrakteristiky (ARTIC)
    - Data-driven (statistické) metody modelování
- IR (mluvená data)
  - Vývoj testovací kolekce
- Spolupráce s UK
  - (vč. projektů Companions, Malach)

MU Brno – FI, Laboratoř zpracování přirozeného jazyka

- Lexikální nástroje a zdroje (spol. s ÚJČ)
  - Platforma DEB II
    - Lexikografická stanice Praled, s ÚJČ
  - Verbalex
  - WordNet – rozšíření (29 tis. položek)
- Analýza češtiny
  - Morfologie (derivace), desambiguace
  - SYNT (synt. analyzátor), anafora
  - Sémantické vztahy

ÚJČ AV ČR

- Lexikální zdroje
  - Vývoj lexikograf. stanice Praled (s MU Brno)
  - dokončeno 3,500 položek databáze
- Digitalizace archívu (s UK)
  - Skenování, „identifikace" excerpt
  - cca 4,000,000 celkem (do konce projektu)
  - Nyní: 1,2 mil. identifikováno

Dosažené výsledky
Souhrn

- RIV 2005-2007
  - 138 (unikátních) záznamů
    - 166 celkem (-> spol. publikace)
  - 64 v roce 2008
- Většina: články ve sbornících konferencí
  - Obvyklé schéma v oboru komputační lingvistiky
    - workshop (specializované) / konference (obecnější)
- Některé časopisy (původní, ale spíše souhrnné výsledky)
  - LNCS, IEEE Transactions, LRE
- Software a data: důraz na „open source"

Nejcennější výsledky - publikace

- Články
  - Semi-supervised POS tagging (EACL 2009)
    - Nejlepší dosud dosažené výsledky i pro angličtinu
  - Extension of HVS Semantic Parser by Allowing Left-Right Branching (ICASSP 2008)
    - Nový výsledek navazující na práci S. Younga
  - Large-scale Semantic Networks: Annotation and Evaluation
    - NAACL 2009; výsledek spolupráce s Google Research, švýc.
- Kniha
  - Valenční slovník českých sloves (Karolinum)
    - Elektronická verze k dispozici

Nejcennější výsledky - data

- Korpusy (jazykové databáze, vydané celosvětově)
  - Prague Dependency Treebank 2.0, Linguistic Data Consortium 2006
  - Czech Wordnet 1.0 (ELRA, 2008)
  - Sign Language, Audiovisual (ELRA, 2008)
- Testovací kolekce
  - CLEF 2006, 2007
    - Multilingual cross-langauge search competitions
  - Machine Translation Open Competition - EuroMatrix 2006-8
    - Czech-English, German, French, Italian, Hungarian, Spanish
  - CoNLL Shared Task 2007, 2009, koordinace v r. 2009
    - Dependency parsing, semantic role labeling (čeština)

Nejcennější výsledky - software

- Software
  - Korpusový manažer Bonito/Manatee
    - Celosvětové použití: ČNK, SNK; Hu, Hr, GB
  - Word Sketch Engine
    - Komerční využití, spol. s Lexical Computing
  - ComPOST
    - State-of-the-art POS tagger (Cz, Sk, En, ...)
  - Syntaktický parser „MST" (čeština)
    - Ve spolupráci s Univ. of Pennsylvania

Vliv vzniku Centra na spolupracující organizace

- Využití účelové podpory
  - 3/4 nákladů: osobní náklady
  - Cestovné, technické zabezpečení
  - Investice (rok 1 a 2 - 2005 a 2006)
    - Výpočetní technika, statistické výpočty
  - Malé doplňkové náklady (režie - do 12%)
- „nehmotný dopad" - nejdůležitější:
  - Vytvoření perspektivních týmů
    - Mgr./Ph.D. studenti

Plnění podmínek programu a smlouvy I

- Zaměstnávání a ***vedení*** doktorandů (škol. prac.)
  - Nyní na všech 4 pracovištích (2009: 10/4/9/1)
    - Podmínka: min. 1 pracoviště → Splněno
- Účast studentů (Bc./Mgr./Ph.D.)
  - Celkem prošlo CKL 35 studentů → Splněno
  - 5 národností
- Uplatnění v komerční sféře
  - Petr Němec (UK): TextKernel, Hol.; Kiril Ribarov (UK): ČEZ
  - Jan Romportl, Aleš Pražák: SpeechTech (spinoff, ZČU)
  - Vladimír Kadlec (MU Brno): Acision (GB)
  - Býv. CKL (LN): M. Čmejrek, J. Cuřín (UK): IBM Research

Plnění podmínek programu a smlouvy II

- Podmínka: zapojení do evr. výzkumného prostoru
- 6 projektů EU, 6. a 7. RP
  - Všechny typy: IP, STREP, NoE; SSA, Dig. Libraries
    - Companions (IP) - ZČU, UK; EuroMatrix, EuroMatrixPlus (STREP) - UK
    - Clarin (SSA) - UK, MU, ÚJČ; KYOTO (Dig. Libraries) - MU
- USA
  - Malach (do 2007; UK, ZČU): USC, JHU, IBM, UMD
  - PIRE: rozpoznávání řeči a strojový překlad (UK, nepřímo ZČU): JHU, Brown Univ.
  - Treebanking: Univ. of Colorado → Splněno

EU Project „Companions“

- Cíl
  - Inteligentní společník pro konverzaci
    - nad fotografiemi, „how was your day“
- Technologie
  - Plná ASR, emocionální TTS
  - Porozumění přirozenému jazyku, generování
  - Přirozenost dialogu: „user studies“ / „evaluation“
- CKL
  - UK/ZČU: ASR, TTS, NLU, NLG, částečně dialog

Prodloužení projektu CKL

2010(-2011?)

Účel prodloužení

- Zachovat stabilizovaný tým
  - Experimentální / týmový charakter oboru
    - Společný postup pro evropské projekty
  - Řada studentů před obhajobou (2010-11)
  - Překlenout období 2010-11
    - Skončené programy (2009) - nové neexistují
    - VZ nelze rozšířit, není pro studenty, a je izolovaný (instituce)
- Pokračovat v aktuálních tématech
  - Rozšíření anotovaných korpusů
  - Vytvoření softwarových nástrojů na jejich základě
    - Analýza jazyka od ASR po sémantickou analýzu

Očekávaný přínos

- Obecně
  - Rozvoj statistických metod
    - zpracování přirozeného jazyka a mluvené řeči
  - Tvorba dalších zdrojů, integrace stávajících
    - Lexikální i korpusové (pro statistické metody)
- Pro pracoviště a instituce
  - Jednotlivě
    - udržení nadaných studentů v akademickém prostředí
    - „kritický objem" studentů a mladých věd. pracovníků (týmy)
  - Společně: získání dalších projektů EU (+ spol. s USA)
- Rozpočet: cca ve stávající výši (13-13.5 mil. Kč)
  - Navýšení vzhledem k získání kvalifikace u 7 pracovníků

Cíle prodloužení

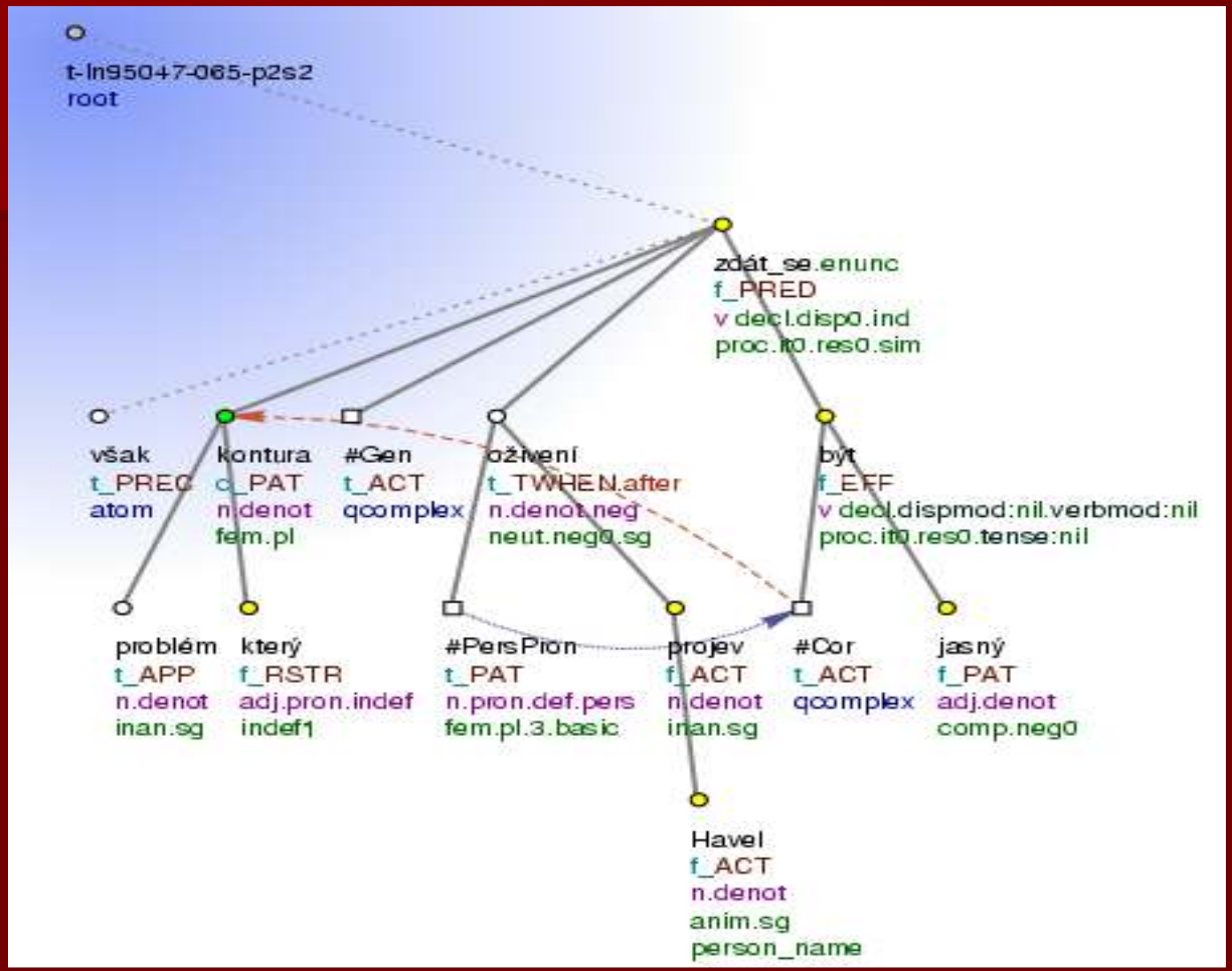
- Čeština, angličtina (jednotlivě, strojový překlad)
  - Data, anotace; analýza, generování
- Mluvený jazyk, integrace
  - ASR, TTS: zdokonalování
  - Rekonstrukce řeči … dialogové systémy
  - Information Retrieval (mluvená jazyková data)
- Lexikální zdroje - budování a začlenění do násstrojů NLP
  - Slovníky několika typů – rozšiřování, dokončení
  - Lexikální databáze češtiny
- Mezinárodní spolupráce (EU, USA)

Očekávané výsledky 2010-2011

- Pokračování publikační činnosti - výstupy:
  - Články ve sbornících a časopisech
  - Anotovaná data (celosvět. publikace)
  - Software - open source nebo licence
- Dokončení Ph.D. studia
  - 5-7 doktorandů
- Zapojení do dalších EU (USA) projektů
  - 3 žádosti podány ve 4. Call (překlad)
- Příprava nových projektů
  - EU, Technologická agentura(?)

Ukázky projektů CKL

Významová anotace věty (UK)



Některé kontury problému se však po oživení Havlovým projevem zdají být jasnější.

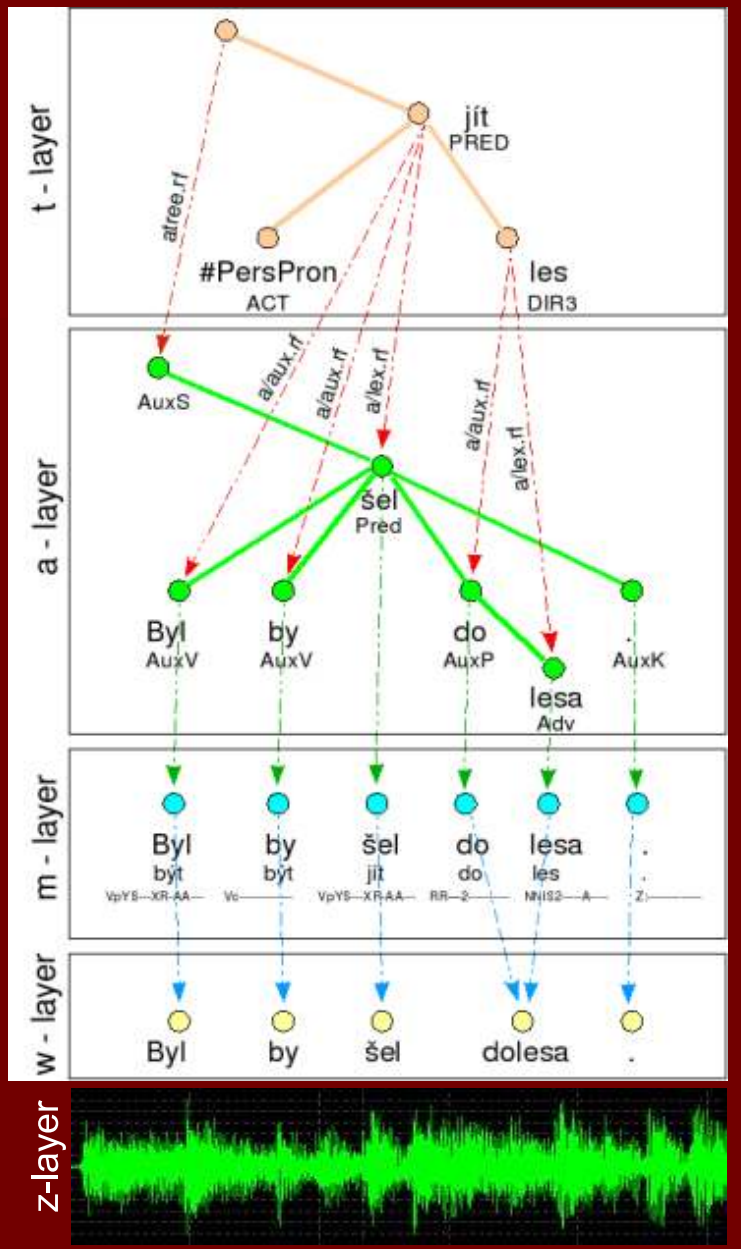
PDT 2.0: Anotační vrstvy

Příklad: věta „Byl by šel do lesa“

Propojení mezi rovinami

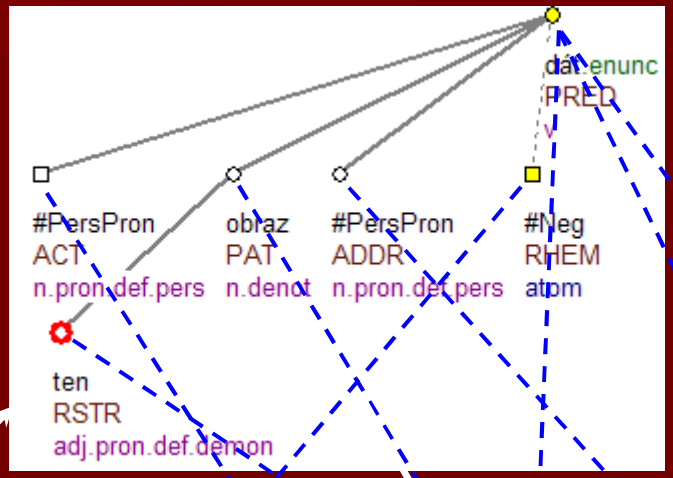
Stand-off anotace

Schéma (Relax NG)



„Rekonstrukce" řeči (UK, ZČU)

- Nyní: anotace
- „Překlad"



Ten obraz jsem jim nemohl dát.

Generation

?

SEM NEMOH SEM TO JIM DÁT TEN VOBRAZ

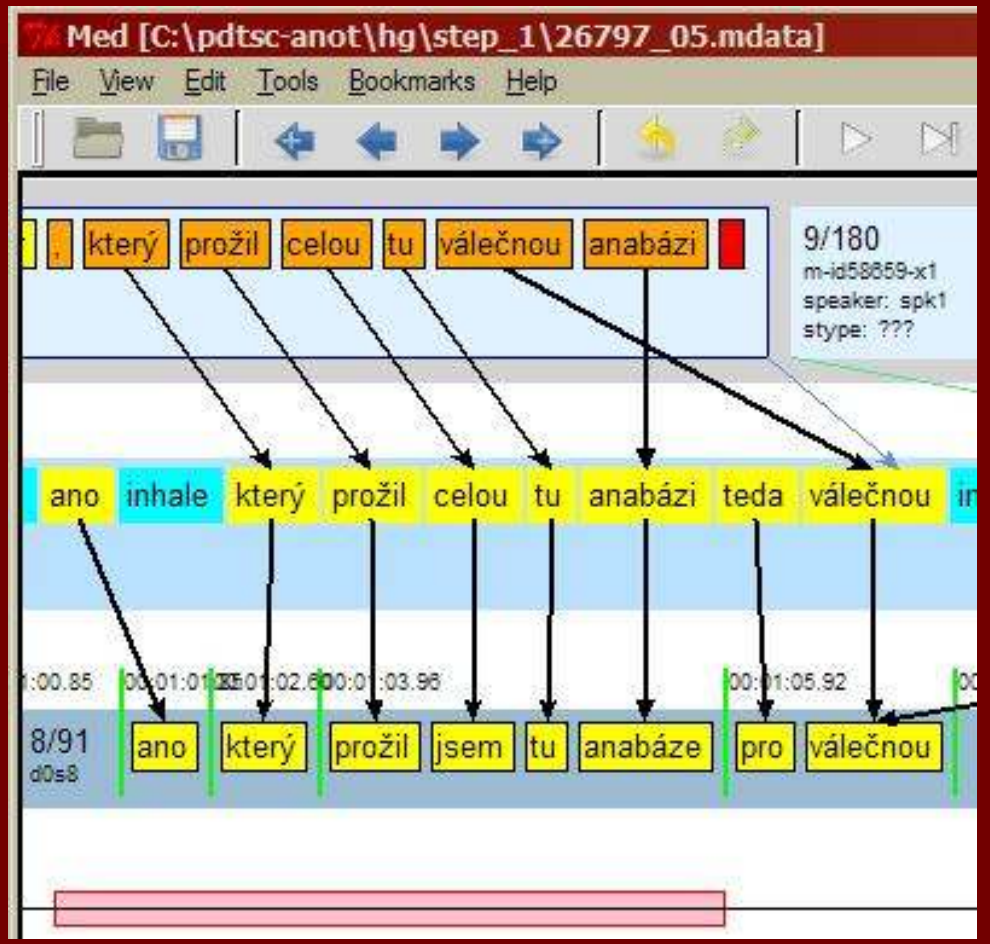
'm couldn't 'm that them give the paintin'

Ten obraz jsem jim nemohl dát.

I could not give them the painting.

“Rekonstrukce” řeči

- Spisovná varianta promluvy
  - „editované interview“
  - Manuální anotace
  - Automatické nástroje, propojení se syntaxí (v budoucnu)



Akustické modelování mezislovního kontextu (ZČU)

- **Využití:** Automatické titulkování televizních pořadů (např. zápasů ledního hokeje) v reálném čase

Automatický překlad čeština -> znaková řeč:

- **Znakovaná čeština**
  - umělý jazykový systém
  - komunikace mezi slyšícími a neslyšícími
  - podobná češtině
- **Český znakový jazyk**
  - mateřský a přirozený jazyk neslyšících
  - komunikace mezi neslyšícími navzájem
  - odlišná od češtiny:
    - simultánnost – např. jeden znak pro „člověk-běží"
    - užití prostoru – do prostoru jsou umístěny objekty komunikace, na něž se mluvčí odkazuje
    - roli intonace zastupuje mimika obličeje
- **Člověk ovládající znakový jazyk nemusí umět česky (ani rozumět textu)**