Nové EU projekty (od 2010)

Faust (STREP)
T4ME Net (NoE)

Jan Hajič & al.

FAUST

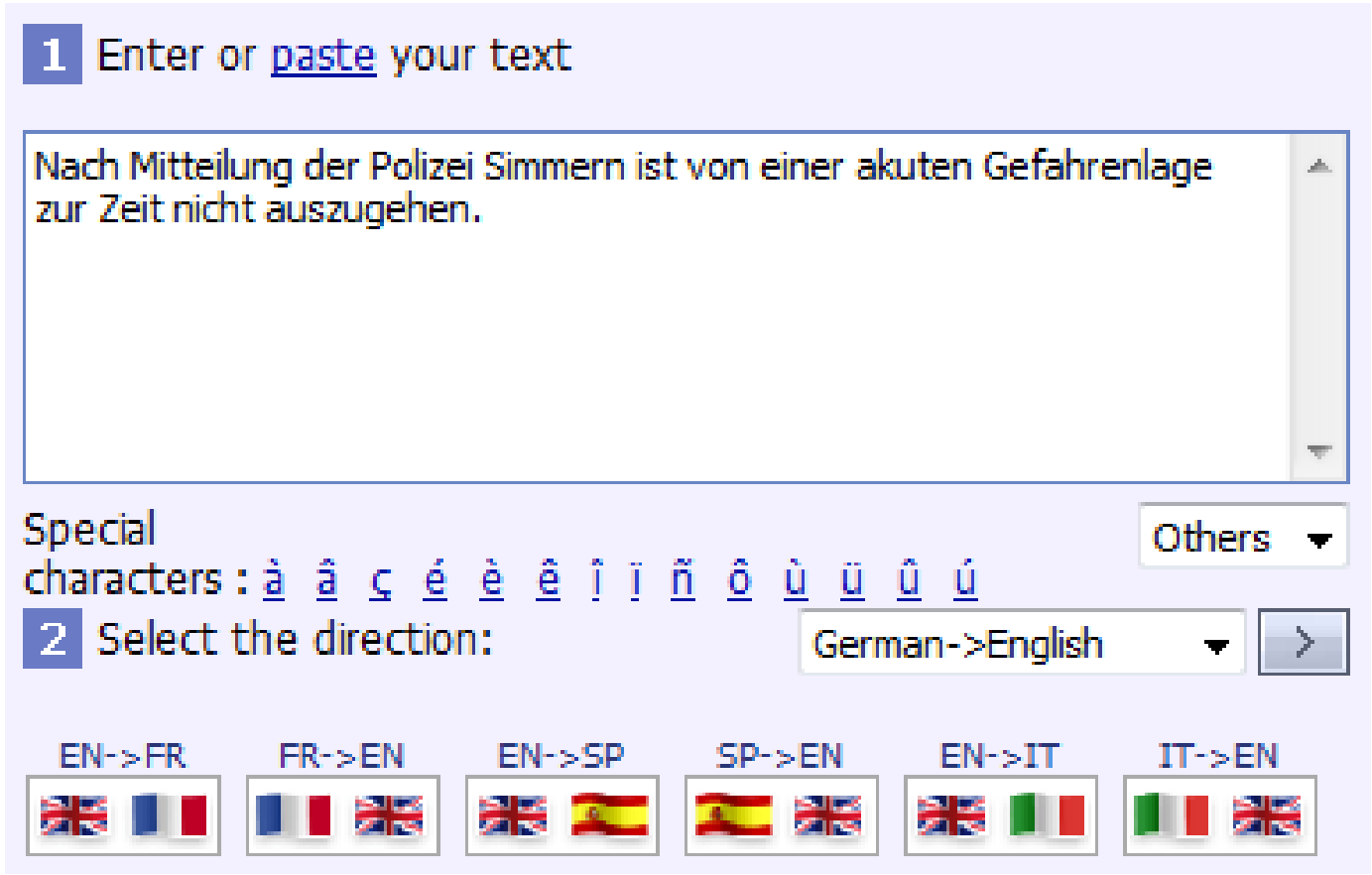
Feedback Analysis for User-adaptive Statistical Translation

- 7. RP, kooperativní, STREP, 2010-13 (únor), GA #247762
- Unit E1: multilingvalita – strojový překlad
  - „Language-based interaction“
  - další: např. Panacea (P. Pecina, DCU)
- Koordinátor: Univ. of Cambridge
  - Bill Byrne, dříve na JHU Baltimore
- Partneři:
  - UCAM, CU, UPC Barcelona (Lluís Márquez), LanguageWeaver USA/Rum. (Daniel Marcu), Softissimo (Theo Hoffenberg)
- Peníze: EUR 772k, z toho 370k na mzdy+poj.
  - Ročně cca 2.2 mil. na hrubé mzdy
  - JH, Jelínek, D. Mareček, J. Ptáček(?); data(?): MM, SC, JM, další

- Jazyky: EN, CZ, ES/CA, RO, FR, AR, ZH)
- Hlavní témata
  - Začlenění zpětné vazby od uživatele do MT
    - Rychle (cíl: „online“)
    - Sběr dat
  - Zlepšení plynulosti překladu
    - Plynulost ~ “gramatičnost” (fluency)
    - Pomocí generování (NLG)
      - První aproximace: n-best z MT, robustní analýza, kombinace, generování (pravidly(?)); cíl: začlenění do celkového modelu, dekodéru
- Další témata
  - Metriky pro strojový překlad

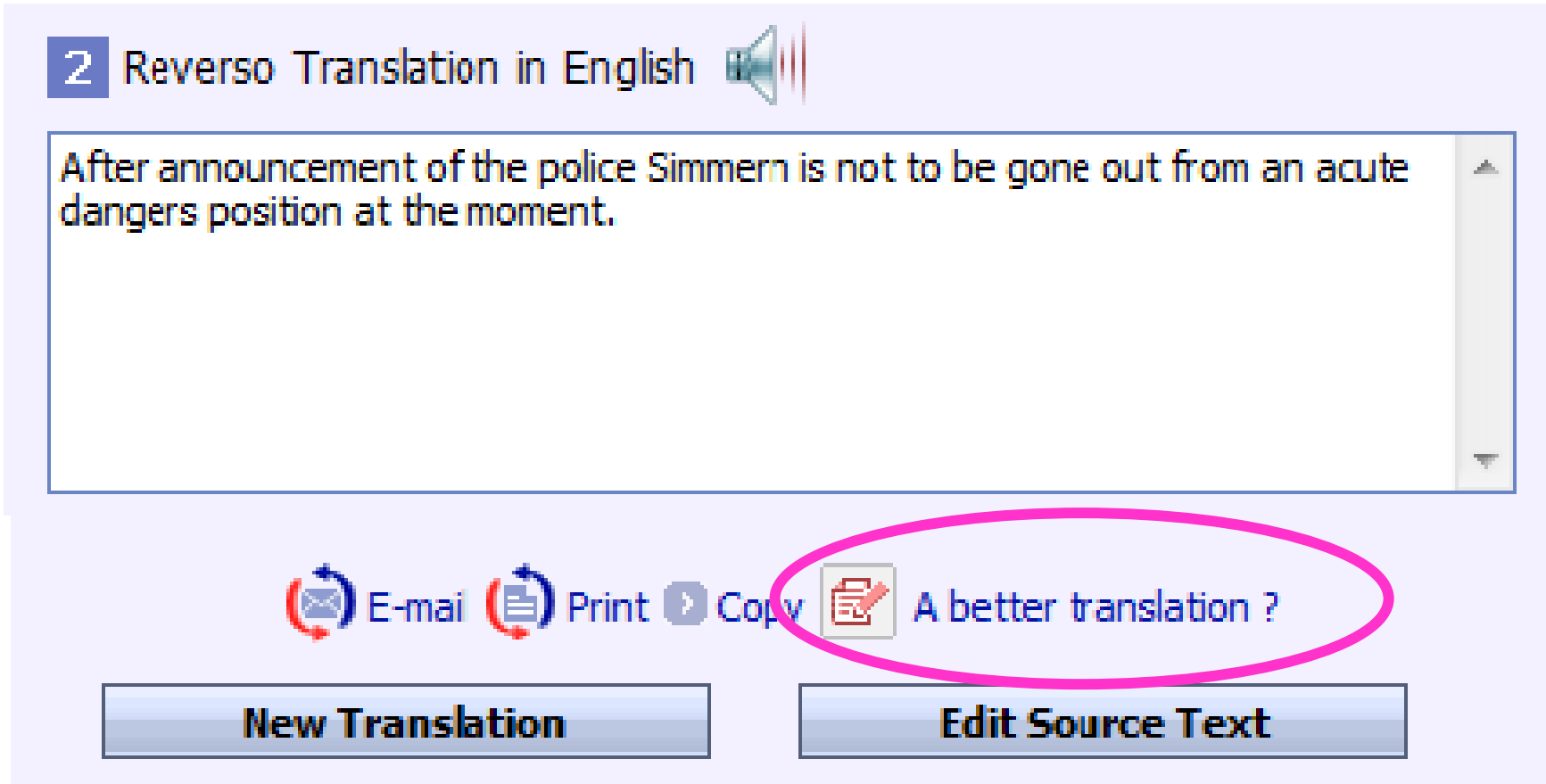
FAUST

- Vstup


1 Enter or paste your text
Nach Mitteilung der Polizei Simmern ist von einer akuten Gefahrenlage zur Zeit nicht auszugehen.
Special characters : à â ç é è ê î ï ñ ô ù ü û ú
Others
2 Select the direction:
German->English
EN->FR
FR->EN
EN->SP
SP->EN
EN->IT
IT->EN

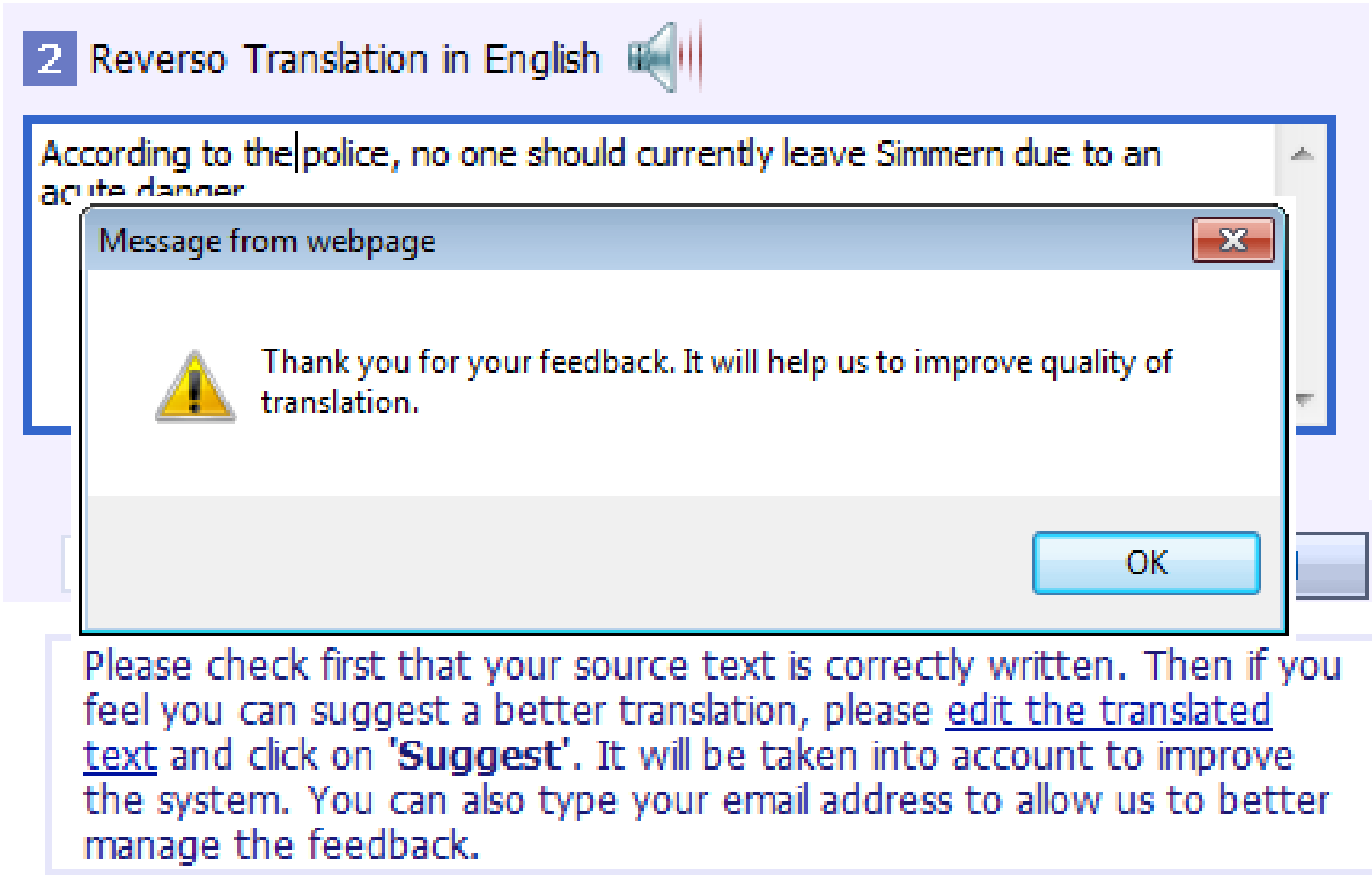
FAUST

- Automatický překlad (LW?)



FAUST

- Oprava uživatele

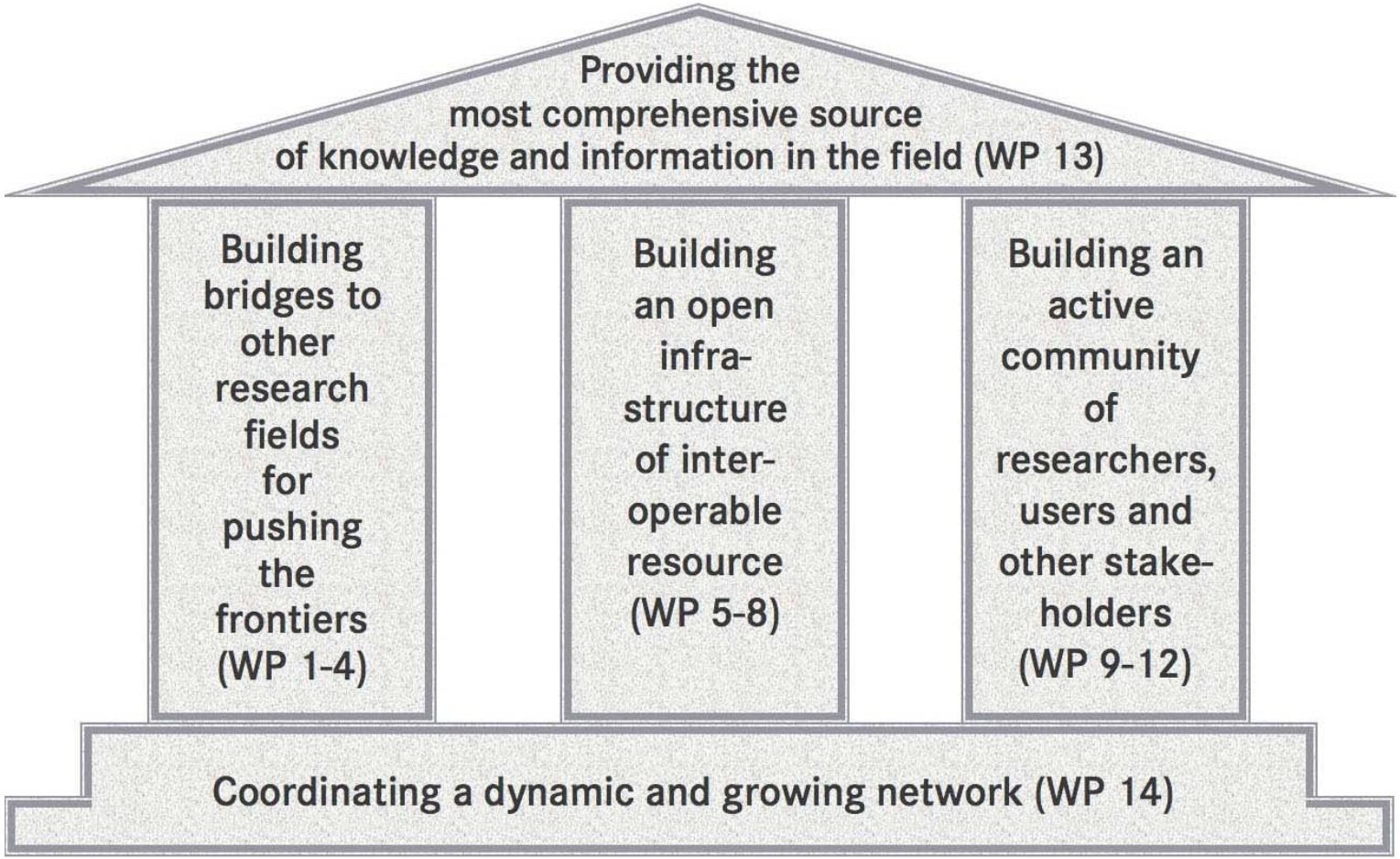


T4ME Net:

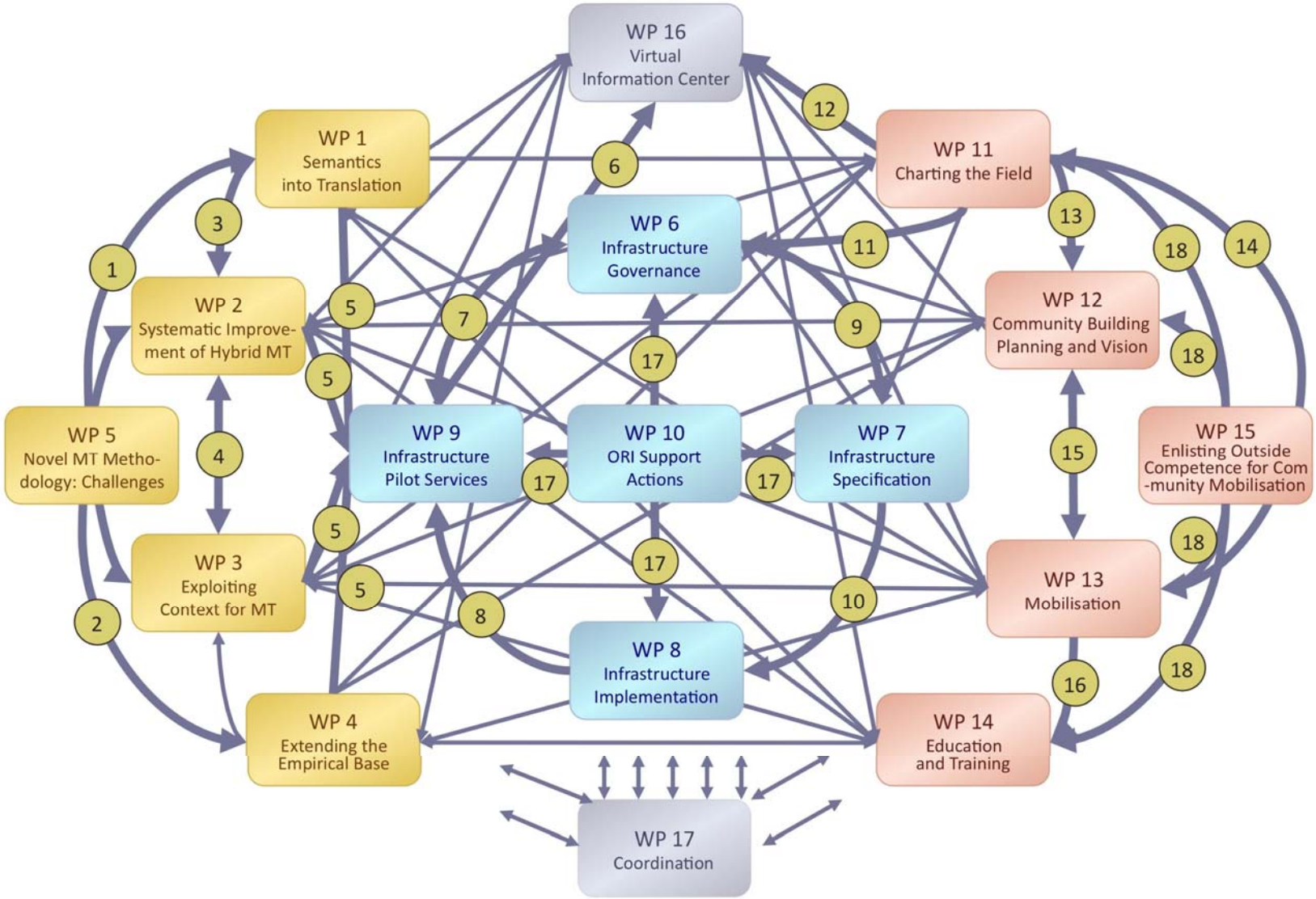
Technologies for the Multilingual European Information Society

- 7. RP, „NoE“, 2010-2013 (únor), GA #249119
- Unit E1: multilingvalita – strojový překlad
  - „Language-based interaction“
- Partneři: DFKI (koord.), CNR (Pisa), DCU, ELRA (celkem 13)
- Peníze: EUR 353k, z toho cca 100k na mzdy+poj.
  - 12 člověkoměsíců na výzkum
  - 28 člověkoměsíců na infrastrukturu (5) a „networking“ (23)
  - JH, P. Straňák, ?

T4ME Net


Providing the most comprehensive source of knowledge and information in the field (WP 13)
Building bridges to other research fields for pushing the frontiers (WP 1-4)
Building an open infra-structure of inter-operable resource (WP 5-8)
Building an active community of researchers, users and other stake-holders (WP 9-12)
Coordinating a dynamic and growing network (WP 14)

T4ME je… network



GA ČR P406/2010/0875

- *název:* **Komputační lingvistika: Explicitní popis jazyka a anotovaná data se zřetelem na češtinu**
- *trvání:* 2010–2013
- *řešitelé:* **prof. J. Panevová**

  M. Ševčíková, J. Hlaváčová, J. Mírovský, V. Kettnerová, V. Kolářová, P. Pajas, S. Cinková, (M. Mikulová)
  - celkem cca 3 úvazky

GA ČR P406/2010/0875

- *celkový finanční objem:* 15 milionů
- *finance na 1. rok řešení:* 3,5 milionu, *z toho*
  - *mzdy* 1 413 tisíc
  - *OON* 288 tisíc
  - *cestovné* 80 tisíc

GA ČR P406/2010/0875

- *náplň projektu:*
  - *(A)* Prohloubení informací na TGTS PDT 2.0 (uvnitř věty i nadvětně), některé aspekty zabudování mimojazykového obsahu
  - *(B)* Metodologické závěry z anotace diskuzních a marginálních jevů
  - *(C)* Instrukce k anotování diskurzu se zřetelem k anotování mluvených projevů
  - *(D)* Zvětšení objemu (syntakticky) anotovaných dat (PDT 3.0, PCEDT, PDTSL)

GA ČR P406/2010/0875

- *očekávané výsledky:*
  - teoretické články a studie o nově zjištěných jevech *(A, B)*
  - pokyny (doplněný manuál, technická zpráva o změnách v anotačním scénáři pro PDT 3.0) *(C)*
  - vydání CD s PDT 3.0 *(D)*

GA ČR P406/2010/0875

- PDT 3.0
  - anotování ATS a TGTS na PDTSC (mluvený korpus po etapě „rekonstrukce“, cca 25 tisíc vět)
    - vstup: zvukový signál, výstup: ATS a TGTS (včetně koreference)
  - výhody – nevýhody

GA ČR P406/2010/0875

- změny v anotačním scénáři (příklady):
  - nové funktory (kvazivalenční: Překážka, Mediátor; subfunktory)
  - nové gramatémy (diatéze rezultativní, recipientní; factmod: asserted, potential, irreal)
  - požadavky na slovník plynoucí z těchto zjištění:
    - promítnutí sémantických diatezí (alternací) do VALLEXu
    - uplatnění rozdílu syntaktické a sémantické derivace u deadjektivních příslovcí
    - další typy koreference a analytické predikáty
  - další typy textové koreference

PDT 2.x

Jan Štěpánek

- Náplň: opravy drobnějších chyb v PDT 2.0
- Lidé:
  - ze svého post-doc GAČRu mohu platit jen sebe
  - spolupracovat může každý, kdo projeví zájem
- Financování:
  - mě platí můj GAČR, na nic jiného peníze nejsou
- Trvání:
  - GAČR je na tři roky
- Výsledky: prvních pár oprav.
  - Je to víc práce, než jsem původně čekal.

PCEDT – cz

Pražský česko-anglický závislostní korpus – česká část

Vstup: texty Wall Street Journal (PTB): 49 208 vět

- překlad textů do češtiny a revize a postrevize překladu
- tektogramatická anotace

Výstup: vydání CD (v LDC) na konci roku 2010

Kdo?

- Jan Hajič
- Marie Mikulová a Jan Štěpánek
- Anotátoři: aktuálně 8

  Ivana Klímová, Olga Hromadová, Martina Otradovcová, Katka Voleková, Alena Kropíková, Lenka Hrejsemnová, Michala Lvová, Jitka Faktorová

Financování:

- Companions, EuromatrixPlus, CKL, GAČR JPa, GAČR JŠt
- a řada dalších (PIRE, MŠMT KONTAKT, GAUK, IS, GAČR)

PCEDT – cz

Pražský česko-anglický závislostní korpus – česká část

Překlad textů do češtiny a několikanásobná revize překladu

- překlad, revize, postrevize
- pokyny k překladu/post/revizi, glosář
- zbývají cca 2 % postrevizí (984 vět, 19107 slov)
- předpokládaný konec: únor 2010

Po postrevizi: *Zisk po zdanění a minoritní podíly, avšak před mimořádnými položkami se zvýšily...*

Správně: *Zisk po započtení daně a menšinových podílů, avšak před započtením mimořádných položek se zvýšil...*

[Originál: *Profit after taxes and minority interests but before extraordinary items increased...*]

PCEDT – cz

Pražský česko-anglický závislostní korpus – česká část

Tektogramatická anotace (1. fáze)

- automatické zpracování m-roviny, a-roviny i předzpracování t-roviny
- anotace struktury a funktorů, valence sloves a substantiv na -ní/tí, zástupné t-lema *#NewNode,* odkazy do a-roviny (10 atributů z 39)
- rozšiřování valenčního slovníku sloves o nové rámce (cca 2 000 nových rámců)
- neanotuje se koreference, aktuální členění, subfunktory, gramatémy, „uvozovky“
- průběžné automatické kontroly správnosti anotace: přes 100 kontrolních skriptů (structure, coord, valency, links, attribute)
- měření mezianotátorské shody, chybovosti, výkonnosti (9,2 věty/hod)
- hotovo: 95 % korpusu (46 890 vět), předpokládaný konec: březen 2010

Plán do konce roku (příprava CD k vydání):

- kontroly a opravy, anotátorské poznámky
- anotace gramatické a textové koreference

Prague English Dependency Treebank

- 2004-2010
- EuroMatrix+, GA P406/2010/0875 (?)
- manuální tektogr. anotace 1M slov z PennTreebanku (WSJ)
- Dosavadní výsledky (cca od semináře v Třeboni):
  - vylepšení automat. předzpracování, kontrolní skripty, urychlení anotace (z 25% v lednu 2009 na 65% v lednu 2010).
- Plán:

  konec 2010: anotace kompletní, příprava k vydání u LDC. Manuál? 😐

- Spolupracovníci *(aktuální):*

Jan Hajič, Silvie Cinková, Eva Fučíková, Kristýna Čermáková, Matěj Korvas, Ema Krejčová, Lucie Mladová, Jan Mašek, Adam Pospíšil, Kateřina Rysová, Magdaléna Rysová, Jana Šindlerová, Kristýna Tomšů, Kateřina Veselá, Kateřina Veselovská

Od struktury věty k textovým vztahům

GA ČR 405/09/0729

Seminář ÚFAL, Hejnice

25. 1. 2010

Charakteristika projektu

- řešitelka prof. Hajičová
- 2009-2011
- celkový rozpočet 3,6 mln. Kč

  (940 – 1260 – 1360)
- souvislost s dalšími projekty (např. GAUK Linh, GAUK L. Mladová, Kontakt E. Hajičová)

Náplň projektu

Vztahy přesahující rámec věty (stromu):

- koreference
- aktuální členění
- textové vztahy (diskurz)

1. Anotace koreference a bridging vztahů na tektogramatické rovině PDT

Lidé: Anja Nedolužko (zodp. osoba), Jiří Mírovský (technická stránka), Radek Ocelák, Jiří Pergler (anotátoři)
trvá od začátku r. 2009, **plánuje se do** konce r. 2010
průběh práce:
anotace rozšířené textové koreference (navazuje na\pokračuje v anotaci pronominální textové koreference prof. Hajičové a Lucie Kučové)
anotace asociační (bridging) anafory
v prosinci 2009 je hotovo cca 50% PDT

2. PlayCoref

Lidé: Pavel Schlesinger, Barbora Vidová Hladká

Výstup na www.lgame.cz

(viz samostatná prezentace)

3. Automatické určování anafory

Lidé: Nguy Giang Linh, Zdeněk Žabokrtský

Výstup: dva systémy pro automatické určování zájmenné anafory:

- klasifikační systém založený na rozhodovacích stromech C5.0
- rankovací systém založený na perceptronu (jednoznačně lepší, f-skóre 79,43 %).

(viz samostatná prezentace)

4. Aktuální členění

Lidé: Šárka Zikánová, Kateřina Rysová (DP, disertace)

Výstup: ověření hypotézy systémového uspořádání z hlediska valence

- hotovo: CAUS, DIR3, LOC, MANN, TWHEN

5. Anotace textových vztahů na tektogramatické rovině PDT

Lidé: Lucie Mladová (zodp. osoba), Eva Hajičová, Šárka Zikánová, Jiří Mírovský (technická stránka), Zuzanna Bedřichová, Pavlína Jínová

Jana Zdeňková, Jana Pěnčíková, Helena Filipová, Veronika Pavlíková (anotátorky)

1. fáze – anotace textových vztahů vyjádřených konektorem

2009: přípravné anotace, školení anotátorů, vypracování základních anotačních instrukcí, vývoj anotačních nástrojů

2010: anotace naostro, manuál, úprava anotačních nástrojů, experimenty

Kontakt

K počítačové analýze struktury textu

grantový projekt česko-americké spolupráce
s University of Pennsylvania

Seminář ÚFALu, Hejnice

25. 1. 2010

Povaha projektu

- Program mezinárodní spolupráce ve výzkumu a vývoji MŠMT
- „Kontakt“
- Agentura AMVIS (Americké vědecké informační středisko) – dvoustranná spolupráce s USA
- Charakterem „cestovní, výměnný“ typ grantu
- Americký partner musí doložit vlastní grantovou podporu pro daný vědecký projekt (NSF)
- Není prvním takovým „podpůrným“ grantem na ÚFALu (JH – MULDA)

Náplň projektu

Spolupráce „diskurzní“ skupiny na ÚFALu s touž skupinou ve Filadelfii, hlavním cílem je kooperace a vědecká výměna při vytváření korpusů zachycujících textové vztahy v typologicky různých jazycích (v Pennu kromě angl. i Hindi)

Náš cíl v Praze: Vytvoření komplexního způsobu anotace textových (významových mezivětných) vztahů pro češtinu a vznik takto zaměřeného korpusu

- Anotační schéma - založeno částečně na systému korpusu Penn Discourse Treebank, částečně na funktorech t-roviny PDT.
- Manuál
- Anotační nástroj
- Anotovaný český korpus
- Vzájemné ověřování postupů zde a v USA, publikace (i společné), srovnání výsledků, komparativní studie…

Projekt zaštiťuje zejména:

- **Vědecký mezinárodní kontakt**, tedy:
  - Cesty do USA (do Filadelfie či na americké konference s účastí UPenn)
  - Discourse Workshop – pro UPenn a ÚFAL (ale i další zájemce), v Praze, 2011
- Mzdy – pro administrativní pracovníky (organizace workshopu aj.)

Kdo se účastní

- Na americké straně:
  - tým prof. Aravinda Joshiho ve Filadelfii

  (pracovníci na projektu Penn Discourse TreeBank)

- Na české straně:
  - Eva Hajičová, Zuzanna Bedřichová, Ondřej Bojar, Pavel Češka, Jiří Mírovský, Lucie Mladová, Šárka Zikánová, studenti (schopní anotátoři)

Trvání a finance

- 3 roky, 2010 – 2012
- Návrh financí 2,3 milionu, zkráceno na 1,7
- Schválený rozpočet po jednotlivých letech: 500 – 600 – 600 tisíc
- Současný stav – spolupráce s prof. Joshim již několik let, nyní začíná být i finančně podporována ☺. Probíhají úpravy rozpočtu dle schválených financí.

PlayCoref

JAZYKOVÁ HRA S KOREFERENCÍ

PlayCoref - financování

- **Barbora Hladká:** 2010 – VZ + GAČR Šárky Zikánové
- **Jiří Mírovský:** 2010 – VZ + GAČR prof. Panevové
- **Pavel Schlesinger:** 2010 – CKL

- **Jan Kohout:** 2009 – OON, provoz (klient)
- **Lenka Studničná:** 2009 – ročníkový projekt (server)
- **Vladimír Rovenský:** 2009 – OON, IS Jarky Hlaváčové (administrace portálu LGame)
- **Helena Pouchová:** 2009 – OON, provoz (grafika)

PlayCoref

http://ufallab2.ms.mff.cuni.cz/lgame/sb/playcoref.php

PlayCoref

Jen některá slova jsou aktivní.

PlayCoref

Hráč označuje koreferenční vztahy.

PlayCoref

Hra končí po vypršení času.

PlayCoref – otevřené problémy

- jmenné fráze
  - šedá ekonomika, česká ekonomika
- pojmenované entity
  - pan Dušín, České Budějovice
- automatické určování koreference
  - pro výpočet skóre
  - výstup projektu

CLARIN – Common Language Resources and Technology Infrastructure

- projekt FP 7-21230
- Hlavní koordinátor University of Utrecht, Nizozemí (S.Krauwer), mimo ÚFAL dalších 31 podporovaných subjektů z EU
- Financování ze zdrojů EU 53.928 EUR, dofinancování z ČR (MŠMT) 26.712 EUR
- 1.1.2008-31.12.2010 , tj. 36 měsíců

Účast ÚFALu

- Řešitelské pracoviště v ČR: Univerzita Karlova
- Tzv. národní kontaktní osoba: Hajičová
- Spoluúčastníci: z UK ještě ÚČNK, dále MUNI (Pala)
- Administrátor: Kotěšovcová
- Dosud se podíleli: Hajič, Pajas, Straňák, Štěpánek ad.

Náplň

- CLARIN x FLaReNet:
- z počátku nejasný vztah, téměř duplikát, jen zaměření CLARINu užší: využití počítačových korpusů a metod v humanitních vědách
- postupně se CLARIN profiluje ve smyslu konkrétních úkolů: standardizace, využití softwarových nástrojů, atd.

Výhledy

- Součást evropské „roadmap“ – tři fáze: přípravná (dnešní CLARIN – do konce r. 2010), konstrukční a „provozní“
- Vytváření velkých infrastruktur výzkumu: návrh projektu LINDAT-CLARIN jako českého uzlu mezinárodní distribuované sítě
- Řešitel-koordinátor: Hajič
- Vědecký tajemník: Straňák
- Administrátor: Kotěšovcová
- Odborný garant a koordinátor spolupráce s CLARIN: Hajičová
- Spoluřešitelská pracoviště: ÚJČ, MUNI, ZČU

Náplň

- sběr jazykových dat (textová, řečová, paralelní)
- anotace dat (včetně anotačních nástrojů)
- zpřístupnění dat a jejich distribuce
- začlenění softwarových nástrojů do poskytovaných webových služeb

Časové rozvržení

- 2010 – příprava Centra
- 2011 – 2013: konstrukční fáze
- 2014 – 2015 (-2020): provoz (tzv. operační fáze)

- Finance: ročně 20 mil. ☺

FLaReNet – Fostering Language Resource Network

- projekt ECP-2007-LANG-617001

  Hlavní koordinátor CNR-ILC, Itálie (N.Calzolari), mimo ÚFAL dalších 27 podporovaných subjektů z EU a 10 dalších, nepodporovaných subjektů z celého světa

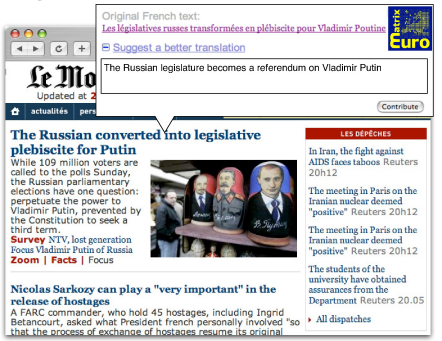
- Financování ze zdrojů EU bez dofinancování z ČR, částka na celé období je 9.000 EUR
  1.9.2008-1.8.2011, tj. 36 měsíců
- Odpovědný řešitel z ÚFALU: Hajičová
- Administrátor: Kotěšovcová
- Tým není přesně specifikován, kdokoli z ÚFALu

Náplň projektu

- tzv. network – nejde o research project
- Podpora kontaktů, workshopů, „forum“
- V čem pro nás užitečný:
  - Dáváme o sobě vědět
  - Dozvídáme se, co dělají ostatní
  - Některé akce projektu organizovány při konferencích, kde máme příspěvky -> financování účasti

EuroMatrixPlus celkově

- Překlad mezi všemi evropskými jazyky
- Zpřístupnění MT pro uživatele



- Témata:
  - Bohatší frázové modely, hierarchické modely
  - Stromové a hloubkově syntaktické modely
  - Hybridní metody (zejm. pravidlový + statistický MT)
  - Nástroje, data (Moses, TectoMT, …) a **komunita**
  - WikiTrans
  - MT pro lokalizaci (Localization Workflow)

Sársko, Edinburk, my, FBK Trento, Le Mans, JHU, DCU + firmy Lucy Software a CEET

EM+ na ÚFALu

- Úkoly:
  - Data: ruční i automatická
  - Software: TectoMT, nástroje k Mojžíšovi, ...
  - Modely: bohatší frázové, různé přes t-rovinu
- Zatřešuje projekty:
  - PEDT
  - PCEDT_CZ
  - TectoMT

Data

- Ruční (zhruba 2009-2010):
  - PEDT (aj): 0,05 mil. vět, 1,2 mil. slov; hotovo 65%
  - PCEDT_CZ: 1,0 mil. slov
    - 98% zrevidován překlad
    - 95% anotace hotova
- Automatická: CzEng 0.9; budou další
  - 8 mil. paralelních vět
  - 93 mil. anglických a 82 mil českých a-uzlů
  - 59 mil. anglických a 59 mil. českých t-uzlů

TectoMT

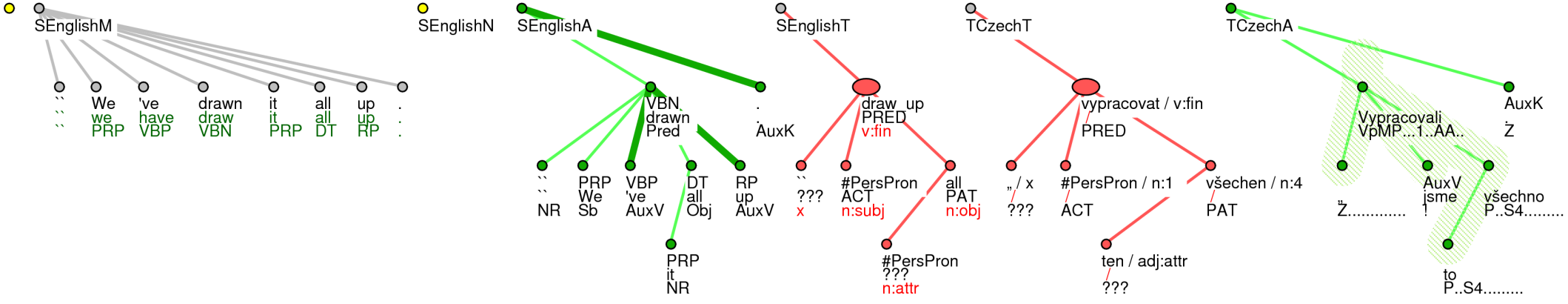
- TectoMT jako platforma pro vývoj NLP
  - Využíváno v mnoha projektech.
  - Předanotace PEDT, PCEDT, ... Anotace CzEngu
  - Dialogový systém Companions, ...
- TectoMT jako experimentální systém MT
  - „Mělká“ t-rovina (formémy místo funktorů, víceméně zachováno povrchové pořadí uzlů).
  - Nově max.ent. model pro transfer t-lemat a formémů.

Aktuální výsledky MT

	BLEU	NIST
Google	0.1741 (17.41±0.53)	5.6476
Ondřejův Moses	0.1566 (15.63±0.51)	5.4042
TectoMT (12.1.)	0.1085 (10.79±0.41)	4.9041

Protipříklad:

SRC	"We've drawn it all up.
REF	"Vypracovali jsme všechno.
Google	"Máme vypracován to všechno nahoru.
Moses	"Máme tady všechno.
TectoMT	"Vypracovali jsme to všechno.



TectoMT

- SW laboratoř pro experimenty a vývoj aplikací v NLP
- struktura dat (roviny) převážně podle PDT
- vlajková aplikace: překlad přes tektogramatickou rovinu
- open-source, veřejně dostupné ze svn
- integrovaná řada existujících NLP nástrojů (hlavně pro češtinu a angličtinu)
  - taggery, závislostní parsery, složkové parsery, rozpoznávače pojm. entit, konvertory formátů...
- kompletní analýza (a syntéza) českých a anglických vět na (z) tektogramatickou rovinu
- společné úfalí vývojářské hřiště, nikoli jeden konkrétní projekt/grant

Projekt JAZZ

Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů

Celkové uznané náklady na řešení projektu ze všech zdrojů financování za všechny příjemce MFF + ÚJČ

Účelové prostředky od AV ČR (v tis. Kč)					
2005	**2006**	**2007**	**2008**	**2009**	**celkem**
3700	4185	4303	4286	4239	**20713**

Celkové uznané náklady na řešení projektu ze všech zdrojů financování jen MFF

Účelové prostředky od AV ČR (v tis. Kč)					
2005	**2006**	**2007**	**2008**	**2009**	**celkem**
2940	3397	3485	3432	3344	**16598**

Hlavní řešitel: Jarka Hlaváčová

Sekce **Jednotný formát**:
Petr Pajas
Honza Štěpánek

Sekce **Pojmenované entity**:
Zdeněk Žabokrtský
Magda Ševčíková
Honza Ptáček
Jana Kravalová
Martin Popel
David Mareček
Oldřich Krůza

Další účastníci:
Milan Fučík
Maruška Křížková
Honza Raab
David Kolovratník

Projekt “Jazz”, část “pojmenované entity”

- Ručně anotovaná data
  - Czech Named Entity Corpus 1.0
    - cca 6000 vět, cca 33000 výskytů p.e.
  - Manual Word Alignment Corpus 0.5
    - česko-anglický korpus, párování na úrovni slov

- Nástroje pro rozpoznávání pojmenovaných entit.
  - rozpoznávač založený na rozhodovacích stromech (O. Krůza)
  - rozpoznávač založený na pravidlech a seznamech geografických názvů (M. Popel)
  - rozpoznávač založený na klasifikátoru SVM (J. Straková)

- Další data
  - CzEng 0.9 - česko-anglický paralelní korpus, cca 80/90 mil.slov

- Další software
  - řada komponent v TectoMT

Projekt “JAZZ” – ukončený k 31.12.2009

Část „jednotný formát“

Datový formát PML

- Specifikace, implementace, převodní nástroje
- Integrace do TrEdu a dalších nástrojů

Nástroje

- TrEd, MEd, jtred, ntred, různé utility
- XML knihovny pro Perl (XML::LibXML, XML::CompactTree::XS)
- „Tisk“ stromů do vektorového formátu SVG

Vyhledávání v datech - PML-TQ

- Dotazovací jazyk PML-TQ
- Dvě implementace vyhledávače (překladem do SQL a v „čistém“ Perlu)
- Grafická rozhraní: webový browser a TrEd

Integrace zdrojů a jejich zpřístupnění pomocí PML-TQ

- PDT 2.0 + PDT ValLex, CAC, PEDT
- Další treebanky: Penn (English), Tiger, Penn Arabic, Penn Chinese, Sinica (složkově a závislostně), Hyderabad, závislostní treebanky z CoNLL2009 ST

Spolupráce se zahraničím (uživatelé TrEdu a PML-TQ)

- Německo, Švýcarsko, Polsko, Japonsko, Korea a další

Typické vzory užívání anglických sloves

a jejich anotace v PEDT

Patrick Hanks

Silvie Cinková

Martin Holub

Záměr projektu

- Zmapovat syntakticko-sémantické vzory užívání sloves s ohledem na
  - významy sloves
  - četnost výskytu vzorů v korpusu
  - normální vs. figurativní (vy)užití vzorů
- Formální popis typických vzorů užívání sloves
  - Pattern Dictionary of English Verbs (PDEV)
- Procedury pro automatickou podporu vytváření slovníku PDEV
- Aplikovatelnost slovníku PDEV

Složení týmu

- Patrick Hanks, Silvie Cinková, Martin Holub, Eva Fučíková
  - MFF UK, Praha
  - financování: Výzkumný záměr, projekt Companions
  - trvání 1 rok
- Pavel Rychlý, Adam Rambousek
  - FI MU, Brno
  - financování: brněnské zdroje
  - dosavadní vývoj cca 4 roky
- Lenka Smejkalová, Jan Popelka
  - studenti MFF UK, Praha
  - financování: zatím 0; přislíbeny drobné peníze na anotační práce z GAČR grantu prof. Panevové

Současný stav projektu

- **Slovník typických vzorů užívání pro 650 anglických sloves**
  - **vytvořeno na základě analýzy výskytů sloves v British National Corpus (BNC)**
  - **pro každé sloveso vzorek označkovaných výskytů v BNC**
  - **pokrytí > 10% všech slovesných výskytů v BNC**

- **Probíhající anotace vybraných sloves**
  - **1000 výskytů v BNC (20 různých sloves)**
  - **1075 výskytů v PEDT (18 různých sloves)**
  - **3-4 anotátoři (PH, SC, MH, studenti)**
  - **čištění slovníku, konsistence, feedback pro autora**
  - **evaluace (zejm. míra mezianotátorské shody)**
  - **„korelace“ vzorů užití s překladovými ekvivalenty**

Vzdálenější cíle

- Rozšíření PDEV
  - větší pokrytí sloves
  - přesnější struktura vzorů užívání
- Automatické rozpoznávání popsaných vzorů
- Automatická detekce ne-popsaných vzorů
  - spíše generování návrhů pro ruční ověření při analýze specifického korpusu
- Aplikace PDEV
  - lexikografické využití
  - automatická syntakticko-sémantická analýza textu
  - hypoteticky podpora strojového překladu

Syntaktická analýza souvětí pro počítačové zpracování češtiny

GAČR 405/08/0681

Syntaktická analýza souvětí pro počítačové zpracování češtiny

trvání: 2008-2010

řešitelka: Markéta Lopatková

zúčastnění: Petr Homola (0,5 úvazku)

Vláďa Kuboň, Martin Plátek, Jarmila Panevová

Tomáš Holan (2008/9)

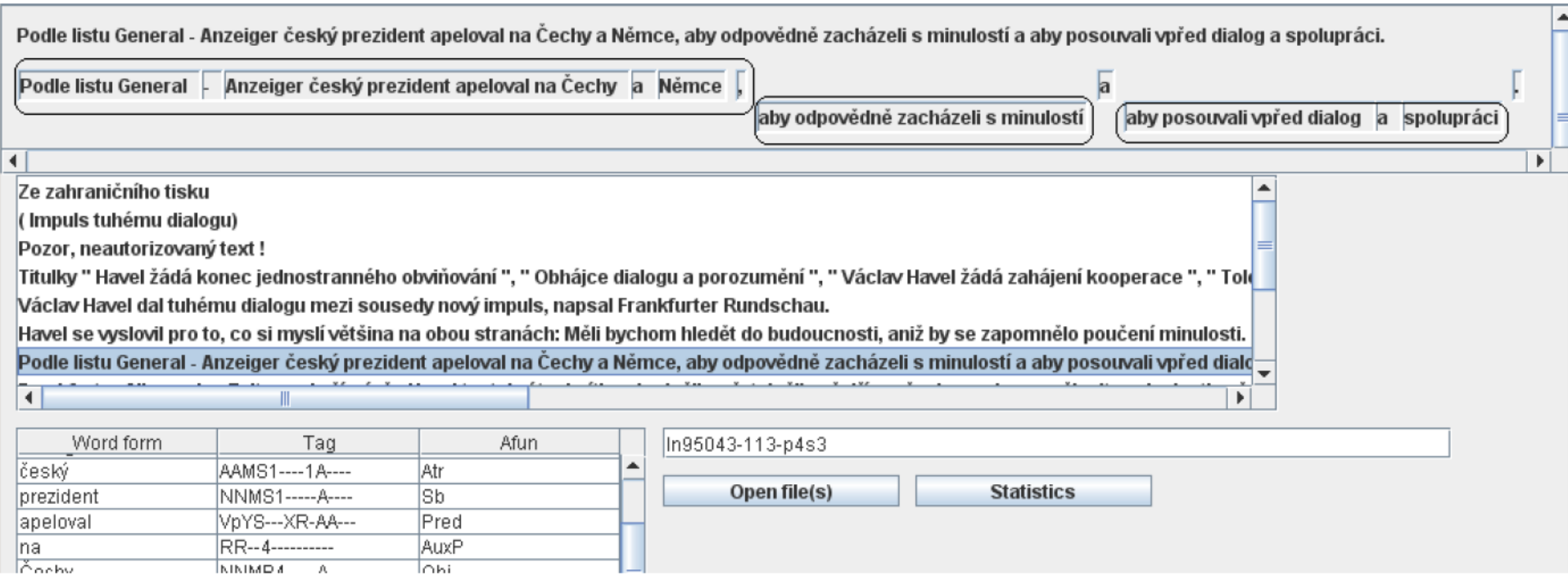
Natalia Klyueva (2008/9), Oldřich Krůza (2008/9)

finance: 1,995 mil.

(531 tis. – 702 tis. – 762 tis.)

Syntaktická analýza souvětí pro počítačové zpracování češtiny

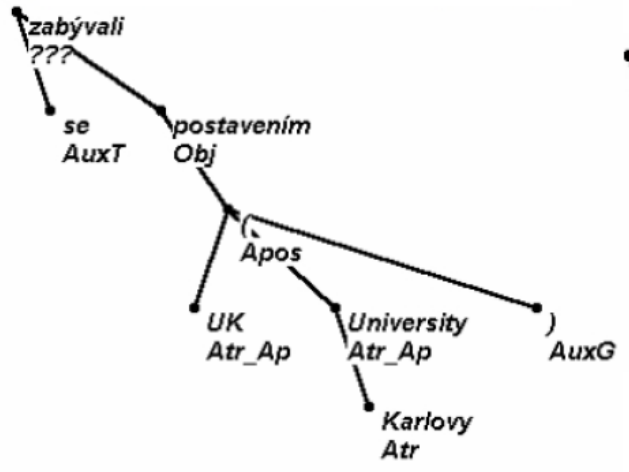
Náplň: vývoj a testování automatické metody pro odhad struktury českých souvětí

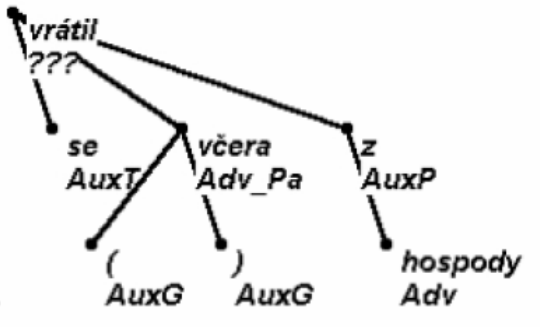


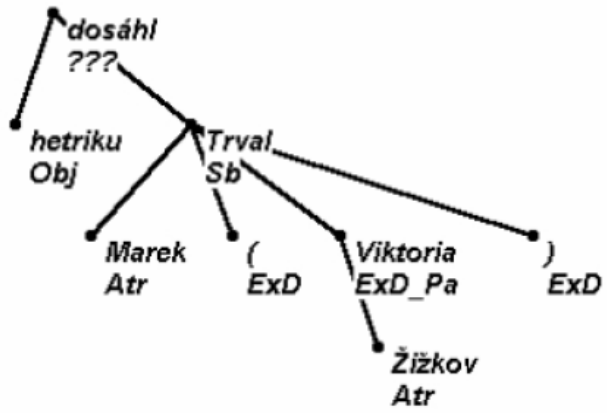
Syntaktická analýza souvětí pro počítačové zpracování češtiny

Náplň: vývoj a testování automatické metody pro odhad struktury českých souvětí

1. Datová struktura segmentů a segmentačních schémat;
2. Sběr lingvistických dat a jejich klasifikace;
3. Příprava a modifikace dat Pražského závislostního korpusu pro získání testovacích dat;







Syntaktická analýza souvětí pro počítačové zpracování češtiny

Náplň: vývoj a testování automatické metody pro odhad struktury českých souvětí

4. Vývoj a implementace automatické procedury pro segmentaci souvětí;
5. Systém značek pro segmenty a návrh pravidel pro spojování segmentů do klauzí;
6. Implementace automatické procedury pro spojování segmentů do klauzí;
7. Vyhodnocení implementovaných procedur;

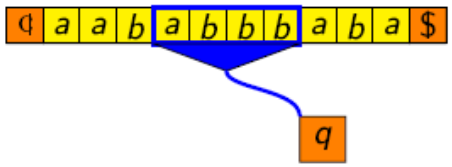
	sentences	agree	%	segments	agree	%	clauses	agree	%
BL	3 443	2 137	62.06	3 775	2 657	70.76	–	–	–
IAA	1 294	1 181	91.27	3 755	3 512	93.53	2 015	1 748	86.75
LH[1]	1 958	1 361	69.51	5 793	4 388	75.75	–	–	–
KK[2]	3 443	2 144	62.27	–	–	–	5 609	4 637	82.67

Syntaktická analýza souvětí pro počítačové zpracování češtiny

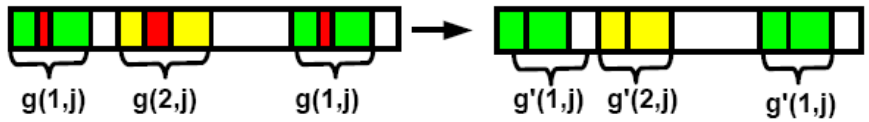
Náplň: vývoj a testování automatické metody pro odhad struktury českých souvětí

8. Aplikace … ČESÍLKO

9. Formální matematický model pro teorii segmentů.



restartovací automat



paralelně komunikující gramatické systémy

Internet jako jazykový korpus

January 15, 2010

Základní údaje

- Název: Internet jako jazykový korpus
- Funding: GAČR standardní projekt
- Doba trvání: 2009-2011
- Peníze: 2.7M celkem (500k 2009, 1M 2010, 1.2M 2011)
- V hlavním rolích: Johanka a Mirek Spoustovi
- Ve vedlejších rolích (special guests): Pavel Pecina, Jan Hajič, Jarmila Panevová

Náplň

- Hlavní cíl: Vytvořit (český) korpus z webu a releasnout sw nástroje potřebné pro jeho výrobu
- Vedlejší cíle: korpus něčím označkovat, tj. trochu si pohrát s taggingem, parsingem..
- Co už máme: pracujeme na nástrojích, hrajeme si s taggingem a parsingem :)
- Hlavní nástroje: Victor - čistič (existuje, chceme vylepšit), Hector - stahovač (Mirek píše), Featurama - Mirkova obecná implemtace ,,morčecího perceptronu“ (existuje, využíváno)

Lexemann

Lexikálně sémantická anotace víceslovných výrazů v PDT

- Ruční anotace víceslovných „frazémů“ a NE
  - stand-off anotace nad tektogramatickou rovinou: s-rovina
  - vlastní slovník SemLex, tvorba anotátorských instrukcí
- Řešitelský kolektiv
  - Pavel Straňák, Eduard Bejček, anotátoři Pavel Šidák, Pavlína Vimmrová, Natalia Kljueva, Eva Šťastná
- 2006? – 2009
- IS J. Hajiče (2005-2009), GAUK P. Straňáka (2009-2010?)
  - placeni 2 řešitelé a 2-3 anotátoři
- Hotová je anotace 99% t-roviny PDT 2.0
- Zbývá pročistit a zkontrolovat data, sjednotit anotátorské slovníky, slévání/přemapování/všívání pro snadné zobrazení a vyhledávání v TrEdu s PML-TQ, vydat data

Projekt ValLink

Prolinkování VALLEXu a PDT-VALLEXu

- Automaticky propojit odpovídající si rámce
  - schválit méně jisté a ručně propojit zbylé
- Řešitelský kolektiv
  - Eduard Bejček, Markéta Lopatková, Zdeňka Urešová, Pavel Straňák, Karel Vandas, (Václava Kettnerová)
- GAUK 2009-2010 (129 tis. Kč na rok 2009)
- Další plán:
  - rozgenerovat vidové dvojice
  - zobrazit v TrEdu a vyhledávat pomocí PML-TQ
    - s odkazy na vidové protějšky a na ekvivalentní rámce

GAUK 19008

- název: Multilingvální zdroj valenčních vlastností sloves
- realita: propojení PDT-VALLEXu a Engvallexu, na úrovni rámců i jednotlivých valenčních pozic
- trvání: 2008-2010
- lidé:
  - 2008, 2009: Jana Šindlerová, Markéta Lopatková, Zdeňka Urešová, Ondřej Bojar, Josef Toman
  - 2010: -JT, +Eva Fučíková, +Kateřina Veselovská
- cíle:
  - komputační: oboustranný překladový elektronický slovník
  - lingvistický: zdroj pro zkoumání shod a rozdílů ve valenci Cz a En sloves
- rozpočet: průměrně 120 tis. ročně na projekt

Kompetence

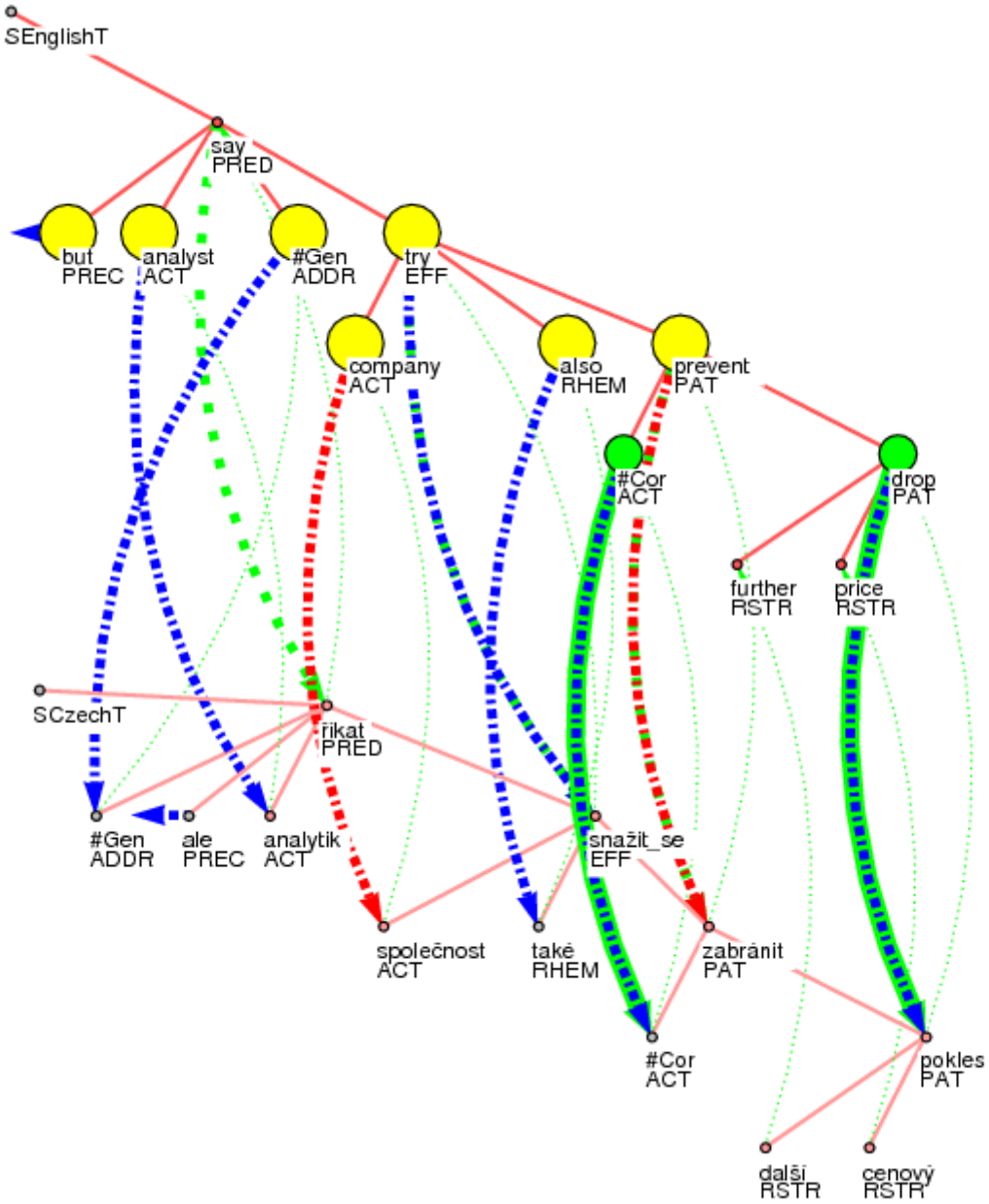
- JŠ:
  - Engvallex: kontrola obsahu a příprava na propojení
  - anotace propojování rámců na korpusových příkladech
- ZU:
  - PDT-VALLEX: kontrola obsahu a příprava na propojení
- JT, EF:
  - Engvallex a anotační prostředí: technická příprava
- OBo:
  - Anotační prostředí, alignment, příprava korpusových dat, nedocenitelné nápady a další práce všeho druhu :)
- ML:
  - valenční poradce, konzultant a „dozor“ :)
- KV
  - anotace propojování rámců na korpusových příkladech

Jak jsme daleko

- Oba slovníky i anotační prostředí připraveny, začínáme anotovat
- Anotace nejprve En->Cz, výsledky ukládány do Engvallexu
  - seznam id odpovídajících rámců u každého rámce Engvallexu
  - u každého id informace o mapování jednotlivých valenčních pozic (uspořádaná dvojice funktorů)

Anotace

- v TrEdu, na zvláštních verzích obou slovníků i PCEDT (updatovány pravidelně)
- na současně zobrazených paralelních stromech, včetně alignmentu uzlů a valenčních pozic
- filelisty pro každou dvojici alignovaných sloves (alignment jiného druhu není brán v úvahu)
- první vhodný výskyt dvojice rámec-rámec anotován ručně, uložený výsledek se ihned automaticky zobrazí u všech dalších výskytů a anotátor pokračuje kontrolou případných konfliktů



- modré šipky: automatický návrh alignmentu dcer slovesa
- červené šipky: návrh anotátora
- zelené plné šipky: alignment potvrzený anotátorem a uložený ve slovníku

Diateze a transformace povrchového vyjádření valenčních doplnění

NÁPLŇ:	výzkum vyjádření slovesného děje jako základu každé věty a možnosti vyjádření jeho argumentů podle požadované diateze
KDO:	Urešová, Hajičová, Šindlerová, Ptáček
FINANCOVÁNÍ:	Grantová agentura UK 2008 - 154K, 2009 - 139K, 2010 - 301K?
JAK DLOUHO:	2008 až 2010
VÝSLEDKY a PLÁNY:	popis transformačních pravidel pro sekundární diateze, Využití: kontrola anotace generování

Hejnice, únor 2010

FGP jako formální překlad: redukční analýza a restartovací automaty

FGP jako formální překlad: restartovací automaty a redukční analýza

finanční podpora:

- IS 1ET100300517

  Methods for Intelligent Systems and Their Applications in Datamining and Natural Language Processing

  řeš. Jiří Šíma, ÚI AV ČR

  2005-2009

- GAČR P202/10/1333

  NoSCoM: Non-Standard Computational Models and Their Applications in Complexity, Linguistics, and Learning

  řeš. Jiří Šíma, ÚI AV ČR

  2010-2013

  MFF: Martin Plátek, Markéta Lopatková (Fr. Mráz, I. Mrázová)

FGP jako formální překlad: restartovací automaty a redukční analýza

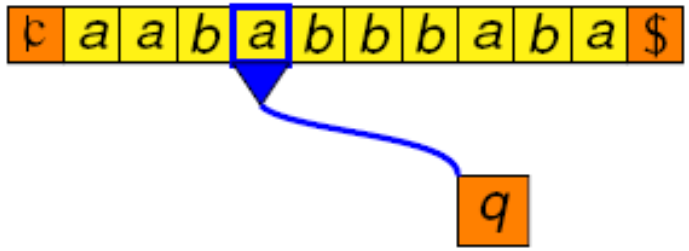
Dlouhodobý cíl:

- formalizace analýzy přirozeného jazyka
- redukční analýza: lingvisticky adekvátní model
  - rekurzivita
  - závislostní $\approx$ valenční syntax
  - lexikalizace $\approx$ slovník jako gramatika
  - nelokální chování
- model restartovacího automatu:
  - jako akceptor
  - jako převodník řetěz $\rightarrow$ řetěz (ITAT 2007)
- zde: zařízení generující stromové struktury

FGP jako formální překlad: restartovací automaty a redukční analýza

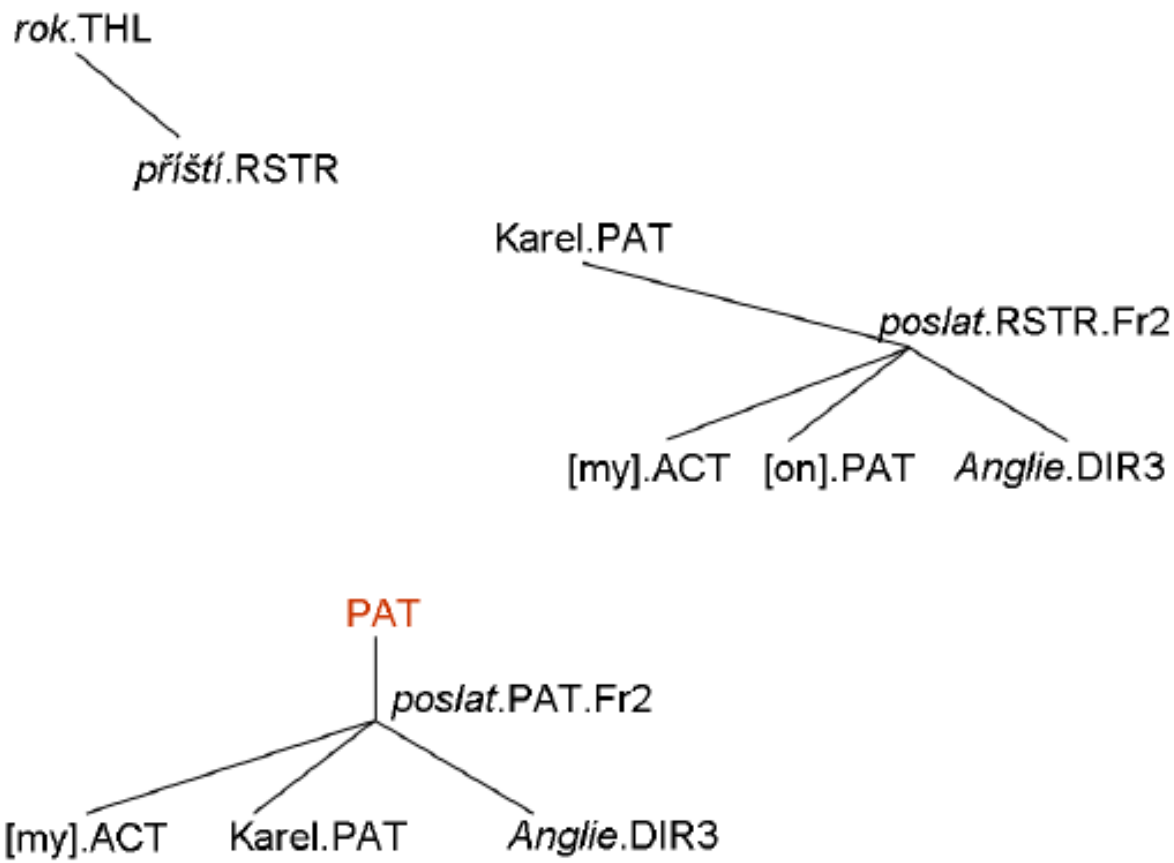
Restartovací automat (převodník):

- nedeterministický stroj
  $M = (Q, \Sigma, \delta, q_0, ¢, \$, \delta)$
- řídící jednotka s kon. množinou stavů Q, kde $q_0 \in Q$
- konečný charakteristický slovník Σ, kde $¢, \$ \in \Sigma$
- pracovní hlava s oknem 1
- instrukce δ:
  - pohyby doleva a doprava
  - vypouštěcí kroky + přepisovací kroky (min 1 políčko, max t políček)
  - restart



+ pokládání "oblázků"

FGP jako formální překlad: restartovací automaty a redukční analýza



FGP jako formální překlad: restartovací automaty a redukční analýza

plán na nejbližší období:

- formální model pro paralelní gramatiky generující stromové struktury
- restartovací automaty-převodníky zpracovávající paralelní stromové strukturu

model reprezentace na analytické a tektogramatické rovině

COMPANIONS
INTELLIGENT, PERSISTENT, PERSONALISED
MULTIMODAL INTERFACES TO THE INTERNET

Projekt PDTSC

Prague Dependency Treebank of Spoken Czech
http://ufal.mff.cuni.cz/pdtsc

Marie Mikulová, Petr Pajas
Nino Peterek,
Jan Hajič, anotátoři, mluvčí, ZČU

Cíl(e)

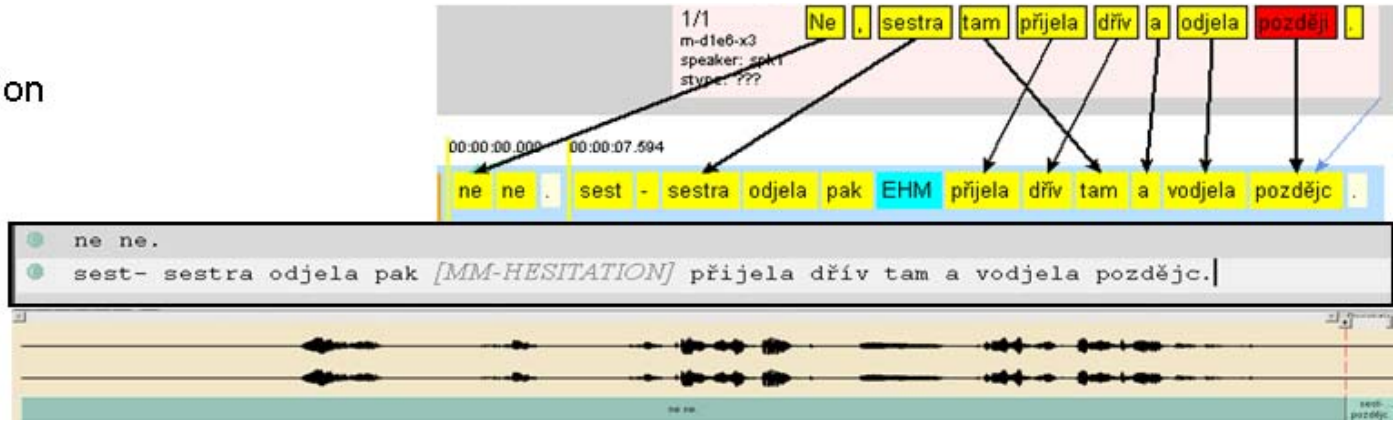
DOUHOBĚ

Krátkodobé:

speech reconstruction

transcription

audio recording



Data

- Malach
  - 30h, ~200 tis. slov
  - 9 tis. vět

- Companions
  - 140h, ~700 tis. slov
  - 36 tis. vět

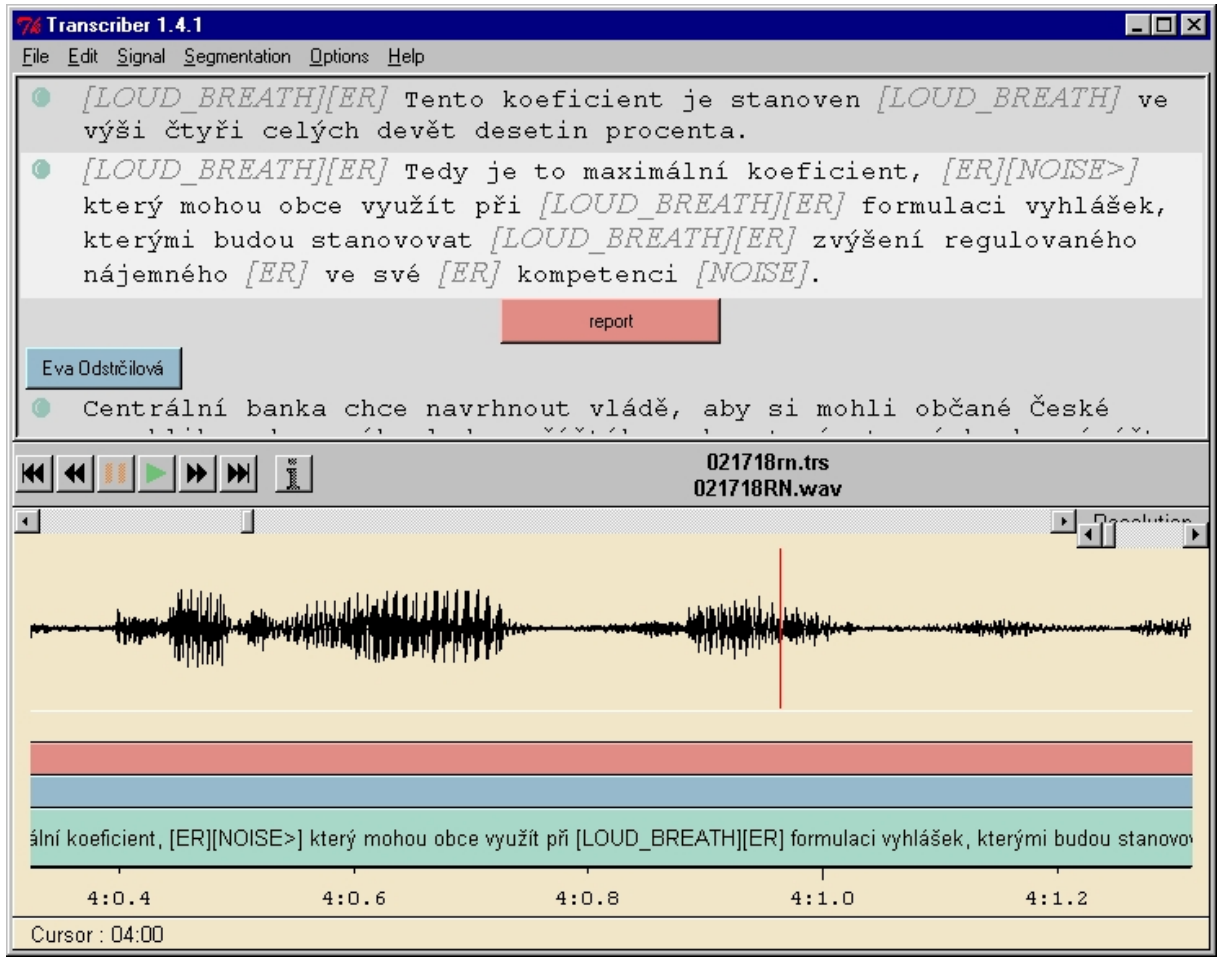
Hotovo: transkripce **80%,** aspoň 1x **71%,** speech reconstruction 2x **63%**

Nahrávání

- Wizard-of-Oz technologie: iluze strojové inteligence
- Moderátor
  - Člověk
  - Skrytý
  - TTS
- Interface
  - Stejné
    - …téměř

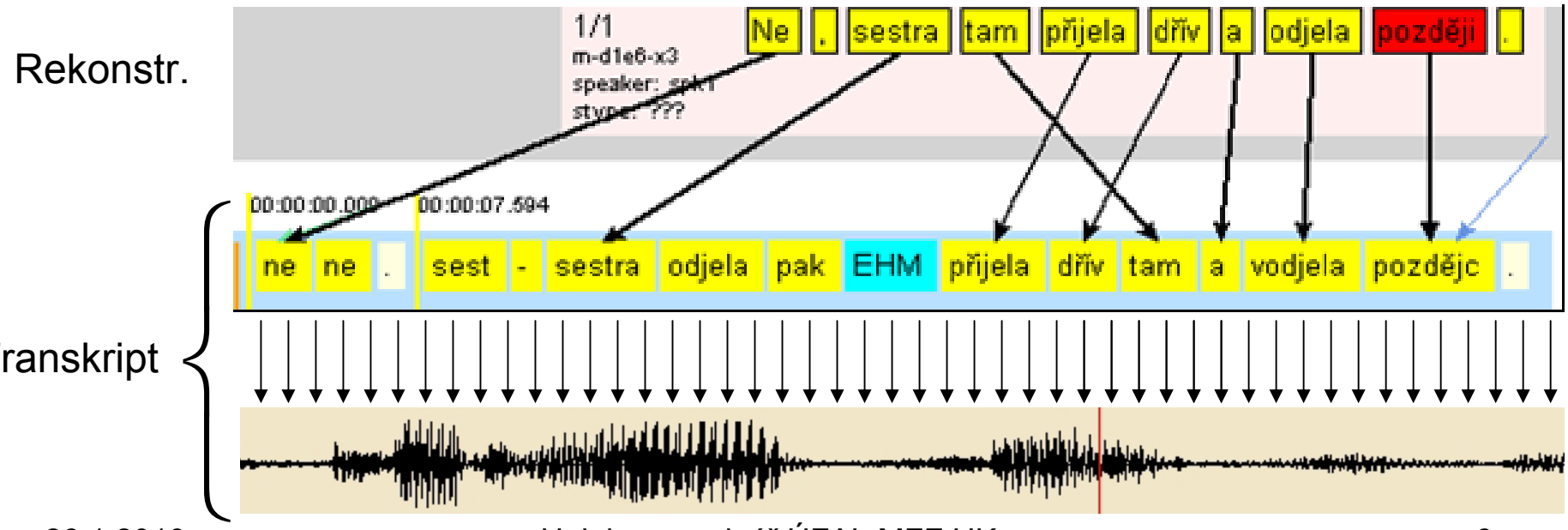
Transkripce

- Transcriber
- Formát
  - XML
  - Synchronizace
- Audio
  - mono
  - 44kHz
  - .ogg



Speech Reconstruction

- Editace jako v redakci novin, časopisů
  - Vše je dovoleno
  - Linky mezi SR a transkripcí, dále až na audio
  - (Re)segmentace do skutečných vět



Organizace

- CKL, EU Companions + co se dá
- Nahrávání a transkripce
  - Systém: Napier Univ. -> Pavel Ircing, ZČU
  - Nino Peterek, ZČU (2008-2009)
- Speech reconstruction
  - Marie Mikulová, Petr Pajas
  - Anotátoři
    - Anna Kapsová, Ludmila Kaplanová, Michaela Luňáčková, Hanka Štěpánková, Petr Míčková (kontroly) + další (ASAP)

Resource-light Morphological Analysis and Tagging

Jiří Hana

- GA ČR P406/10/P328 postdoktorský grant
- 2010-2012, 2.4 mil Kč
- Cíl: Rychlý a levný vývoj taggerů a morf. analyzátorů na základě:
  - zdrojů dostupných pro příbuzný jazyk
  - omezeného množství ručně vytvořených dat

Co už máme

- MA a tagger pro
  - ruštinu (přes češtinu)
  - katalánštinu, portugalštinu (přes španělštinu)
- MA pro češtinu
  - 10K word-list, paradigmata, 20 derivací
  - evaluace na substantivech (tagy, ne lemata): ambiguita 4.0, recall: 96.6%
  - state of the art: 3.8, 98.7%

Jak to děláme

- Ruština přes češtinu – HMM tagger
  - transitions – natrénované na upraveném PDT (ň -> n', nevím -> ne vím, ...)
  - emissions – kombinace
    - uniformního rozdělení výstupu ruské MA
    - emission z PDT přenesené do ruštiny pomocí kognátů
  - MA
    - 1000 slov ve word-listu
    - guesser na základě vzorů
    - automaticky naučený slovník

Co chceme dělat

- Deepening scope:
  - Improve the performance of the morphological analyzer, esp. automatic lexicon acquisition.
  - Explore alternative algorithms for cognate detection and transfer.
  - Incorporate Machine Learning methods.
- Broadening scope: Experiment with other lgs
  - Middle Czech/Sorbian/Belorussian/Russian via Czech.
  - Romanian via French/Spanish/Bulg/Slovene.

Co chceme dělat

- Better insights:
  - Explore the pay-off of various possibilities to improve resources manually and identifying where the manual effort should be directed.
- Practical results: An out-of-the-box toolkit and guides for creating taggers a analyzers for new lgs