Detail projektu

Řešitelský kolektiv | Finanční požadavky | Finanční výhled na další roky | Rozšiřující informace | Přílohy

Základní informace o projektu č. 1572314

Český název projektu:	Modelování závislostní syntaxe napříč jazyky
Anglický název projektu:	Modelling dependency syntax across languages
Aktuální řešitel:	Mgr. Rudolf Rosa rur@seznam.cz
První žadatel:	Rudolf Rosa
Studium:	Matematicko-fyzikální fakulta Program: Informatika Obor: Matematická lingvistika Typ studia: doktorské studium
Rok založení projektu:	2014
Délka řešení projektu:	3
Sekce oborové rady:	Společenské vědy - Informatika
Pracoviště	ÚFAL MFF UK
Historie stavu:	13. 11. 2013 - **podaný**

Legenda:

Zpět

Řešitelský kolektiv

Doporučení vedoucího: doporučeno

Číslo osoby	Role	Celé jméno	Typ odměny	Rok 2014	Činnosti
30151175	Řešitel	Mgr. Rudolf Rosa	Stipendia	60	
30987185	Vedoucí	doc. Ing. Zdeněk Žabokrtský Ph.D. zabokrtsky@ufal.mff.cuni.cz	Osobní náklady (mzdy a odvody)	20	
75606749	Spoluřešitel	Bc. Jan Mašek	Stipendia	40	

Charakteristika řešitelského kolektivu - rok 2014:

Hlavní řešitel, Mgr. Rudolf Rosa, je studentem prvního ročníku doktorského studia Matematické lingvistiky na Ústavu formální a aplikované lingvistiky MFF UK v Praze, v červnu 2013 dokončil navazující magisterské studium tamtéž. Po dobu studia se podílel na několika výzkumných projektech, zaměřených na zlepšování kvality strojového překladu, a

je spoluautorem řady článků prezentovaných na mezinárodních konferencích. Téma grantového projektu bude součástí jeho disertace.
Životopis a publikace řešitele se nacházejí v příloze.

Školitel doc. Ing. Zdeněk Žabokrtský Ph. D. je docentem na Ústavu formální a aplikované lingvistiky. Dlouhodobě se zabývá parsingem, závislostní syntaxí, tektogramatickými strukturami, valencí sloves, zdroji lingvistických dat a strojovým překladem. Podílel se na projektu HamleDT, jehož cílem byla konverze různých závislostních korpusů do společného formátu a jejich částečná harmonizace, a na nějž navazuje tento grantový projekt. Řešitelům poskytne metodické vedení při výzkumných pracích a přípravě prezentací výsledků.
Životopis a publikace školitele se nacházejí v přílohách.

Spoluřešitel Bc. Jan Mašek je studentem druhého ročníku magisterského studia Matematické lingvistiky na Ústavu formální a aplikované lingvistiky MFF UK v Praze, v září 2012 absolvoval bakalářské studium Mezikulturní komunikace - angličtina - čeština a Obecné jazykovědy na Filozofické fakultě UK. Podílel se jako anotátor na projektech Prague English Dependency Treebank 1.0 a 2.0 a od ledna 2013 se podílí na projektu SEANCe, analýzy sentimentu v češtině. Téma grantového projektu odpovídá tématu jeho diplomové práce. Pro jeho lingvistické vzdělání a zkušenosti se bude podílet zejména na harmonizaci zdrojových syntakticky anotovaných korpusů a na návrhu jazykově univerzálního anotačního schématu.
Životopis spoluřešitele se nachází v příloze.

Finanční požadavky

Položky	**Rok 2014**
Ostatní neinvestiční náklady	10
Cestovné	90
Doplňkové náklady	33
Osobní náklady (mzdy) a stipendia	120
Celkem	253

Struktura finančních prostředků - rok 2014:

Za prostředky na ostatní neinvestiční náklady bude pořízen nezbytně nutný hardware, zejména pevné disky.

Plánované konference a pobyty:
* LREC 2014, Rejkjavík (poplatek 8000 Kč, ubytování a doprava 35000 Kč) – nejvýznamnější konference zaměřená na lingvistické datové zdroje
* ACL 2014, Baltimore (poplatek 15000 Kč, ubytování a doprava 30000 Kč) – nejvýznamnější konference v oblasti komputační lingvistiky
* TLT 2014 (poplatek 3500 Kč, ubytování a doprava 20000 Kč) – konference specializovaná na syntakticky anotované korpusy
Náklady na pobyty jsou určeny přibližně na základě minulých let.

Částky na stipendia a mzdy jsou navrženy v souladu s požadavky Grantové agentury UK.

Finanční výhled na další roky

Položka	Rok 2015	Rok 2016
Finance celkem	250	250

Rozšiřující informace

Anotace:

V grantovém projektu budeme zkoumat vzájemné podobnosti přirozených jazyků a získané poznatky využijeme pro dva typy úloh komputační lingvistiky, řešících aktuální problémy zpracování jazyka na úrovni syntaxe.
Prvním typem úloh budou technologie mezijazyčné projekce, kdy model jednoho jazyka využijeme pro přibližné modelování jazyka podobného, pro který nemáme dostatečné jazykové zdroje.
Druhým typem úloh bude přenositelnost jednojazyčných technologií, kdy nástroje a postupy vyvinuté pro práci s jedním či několika málo jazyky zobecníme tak, aby umožňovaly zpracování téměř či zcela libovolného jazyka, pro který jsou k dipozici odpovídající datové zdroje.

Přestože existují rozsáhlé jazykové zdroje pro mnoho jazyků, v praxi se často ukazuje, že je obtížné tyto úlohy úspěšně řešit. Dostupné zdroje jsou totiž obvykle silně heterogenní, používají rozdílná anotační schémata a jsou vystavěny na základě odlišných lingvistických tradic a konvencí. Nutným mezikrokem pro uskutečnění hlavních cílů projektu je proto shromáždění a harmonizace existujících syntakticky anotovaných jazykových korpusů.

Anotace v anglickém jazyce:

In this grant project, we will explore mutual similarities of natural languages, and we will use our findings for two types of computational linguistics tasks, dealing with current problems of natural language processing on syntax level.
The first task type will be cross-lingual projection technologies, where a model of one language will be used to approximately model a similar language for which sufficient language resources are not available.
The second task type will focus on portability of monolingual technologies, where tools and procedures developed for working with one or a few languages will be generalized so that they can be used to process any or nearly any language for which sufficient data are available.

Although there exist vast language resources for a number of languages, practice often shows that it is hard to successfully solve the aforementioned tasks. This is due to the fact that the available resources are usually very heterogeneous, are using different annotation schemes and are built on the basis of different linguistic traditions and conventions. A necessary by-step in reaching the main goals of the project is therefore to collect and harmonize existing syntactically annotated language corpora.

Současný stav poznání:

Syntaktická analýza jazyka (parsing) s pomocí syntakticky anotovaných korpusů (treebanků) je již zavedeným a stále se rozvíjejícím směrem, na čemž má velký podíl existence velkých treebanků (Marcus et al. 1993, Böhmová et al. 2003), a také v minulosti organizované soutěže v parsingu (Nilsson et al. 2007). V dnešní době jsou k dispozici desítky treebanků pro mnoho světových jazyků, anotovaných v různých anotačních stylech (Zeman et al. 2012), a mnoho různých parserů (např. McDonald et al 2005b, Nivre at al. 2006), které je možné na těchto treebancích natrénovat a poté použít pro analýzu vět daného jazyka. (Pod pojmem anotační styl rozumíme soubor pravidel a konvencí, s použitím kterých byl daný datový zdroj lingvisticky anotován.)

Jedním z velkých témat současné komputační lingvistiky je multilingualita. Ukazuje se, že nástroje při analýze různých jazyků dosahují různých úspěšností, což odkazuje jednak na typologickou odlišnost jazyků, ale také na odlišnosti v anotačních stylech jednotlivých treebanků. Spolehlivé porovnání úspěšnosti parserů na různých jazycích je proto obtížné a je obvykle nutné parsery více či méně upravovat pro natrénování nad dalším jazykem; vlastnosti některých treebanků dokonce znemožňují nad nimi úspěšně natrénovat určité druhy parserů, například neprojektivní konstrukce v češtině jsou překážkou pro projektivní parsery (McDonald et al. 2005a).
Na významu také získává zaměření na jazyky, pro které je dostupné pouze malé nebo žádné množství potřebných datových zdrojů pro natrénování parseru standardním způsobem. Používají se proto přibližné techniky delexikalizovaného parsingu a mezijazyčné projekce (McDonald et al. 2011), kdy se parser natrénovaný na existujícím treebanku pro jeden jazyk použije pro analýzu jiného podobného jazyka, pro nějž nejsou k dispozici dostatečné datové zdroje. Příbuzným odvětvím je neřízený parsing (Klein a Manning 2004), kdy se parser trénuje nad velkými daty bez syntaktické anotace, pouze na základě definování pravděpodobnostních požadavků na výsledné závislostní stromy. Pro vyhodnocení úspěšnosti obou těchto metod se používají existující treebanky, což velmi znesnadňuje spolehlivé vyhodnocení jejich úspěšnosti (Mareček 2012), neboť odlišnosti v lingvistických tradicích a konvencích vedou k heterogennosti jednotlivých treebanků.

Již nějakou dobu se tedy objevuje myšlenka sjednocení anotačních stylů treebanků, tak aby nenastávaly výše popsané problémy. Prvním velkým projektem tohoto typu byl HamleDT (Zeman 2012), kolekce 29 treebanků pro různé jazyky sjednocených (harmonizovaných) do pražského anotačního stylu, navazující na postupný vznik několika treebanků anotovaných v tomto stylu (Böhmová et al. 2003, Hajič et al. 2004, Čmejrek et al. 2004, Džeroski et al. 2006, Ramasamy a Žaboktský 2012). Značky slovních druhů a morfologických rysů (tagy) byly konvertovány do Intersetu (Zeman 2008), který je pokusem o vytvoření jakési nadmnožiny všech takových značek (kromě těch, které jsou příliš jazykově specifické). Závislostní struktury byly konvertovány do pražského stylu PDT (Böhmová et al. 2003) zejména v případě koordinací, kde se pražský styl ukázal být dostatečně expresivním pro zachycení většiny koordinačních struktur jednotlivých jazyků, mnohé další odlišnosti ale zůstaly ponechány (například anotace složených sloves). Značky závislostních vztahů (deprely) byly namapovány na sadu analytických funkcí definovaných pro PDT, která umožňuje správně zachytit nejdůležitější role jako podmět, předmět či přísudek, ale některé další role zachytit neumí (například členy či negativní částice) a dochází tak ke ztrátě informace při konverzi. Navíc nepodporuje podspecifikovanost deprelů, takže v případě že

zdrojový treebank neobsahuje dostatek informací pro rozlišení jednotlivých deprelů, musejí být použity heuristiky.

Druhým velkým projektem v oblasti vytváření velké kolekce harmonizovaných treebanků je projekt společnosti Google s názvem Universal Dependency Treebanks (McDonald et al. 2013). Ten na rozdíl od HamleDTa nejde cestou konverze existujících treebanků, ale rozhodl se pro vytváření nových treebanků, což umožňuje zaručit skutečně vysokou jednotnost anotace, ale na druhé straně jde o zdlouhavou a finančně náročnou práci - v současné době proto tato kolekce obsahuje pouze šest treebanků, a to poměrně malé velikosti. Slovní druhy jsou reprezentovány pomocí Universal part-of-speech tagset (Petrov et al. 2012), který umožňuje zachytit pouze 12 slovních druhů bez dalších morfologických informací, což je pro mnohé aplikace nedostatečné. Anotace závislostních struktur a deprelů vychází ze Stanford Typed Dependencies (De Marneffe a Manning 2008). Jejich sada seprelů má hierarchickou strukturu, což umožňuje použití podspecifikovaných deprelů, tj. například místo konkrétního druhu slovesného doplnění lze použít obecnější typ deprelu - to je velmi užitečné pro zachycení různé potřebné granularity deprelů v různých jazycích. Výzkumníci Google adaptovali Stanfordské deprely tak, aby byly jazykově nezávislé, zejména pomocí rozšíření definice některých deprelů a spojení více podobných deprelů do jednoho. Kolekce je ale stále ve vývoji, anotace proto dosud není napříč jednotlivými treebanky zcela konzistentní a sada deprelů také ještě není ustálená.

Zdroje:

BÖHMOVÁ, Alena, et al. The Prague dependency treebank. In: Treebanks. Springer Netherlands, 2003. p. 103-127.

ČMEJREK, Martin; HAJIČ, Jan; KUBOŇ, Vladislav. Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation. In: In Proceedings of EAMT 10th Annual Conference. 2004.

DŽEROSKI, Sašo, et al. Towards a Slovene dependency treebank. In: Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation (LREC). 2006.

HAJIČ, Jan, et al. Prague Arabic dependency treebank: Development in data and tools. In: Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools. 2004. p. 110-117.

KLEIN, Dan; MANNING, Christopher D. Corpus-based induction of syntactic structure: Models of dependency and constituency. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004. p. 478.

MAREČEK, David. Unsupervised Dependency Parsing. Praha, 2012. Dizertace. MFF UK.

MARCUS, Mitchell P.; MARCINKIEWICZ, Mary Ann; SANTORINI, Beatrice. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 1993, 19.2: 313-330.

DE MARNEFFE, Marie-Catherine; MANNING, Christopher D. The Stanford typed dependencies representation. In: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation. Association for Computational Linguistics, 2008. p. 1-8.

MCDONALD, Ryan; CRAMMER, Koby; PEREIRA, Fernando. Online large-margin training of dependency parsers. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005. p. 91-98.

MCDONALD, Ryan, et al. Non-projective dependency parsing using spanning tree algorithms. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005. p. 523-530.

MCDONALD, Ryan; PETROV, Slav; HALL, Keith. Multi-source transfer of delexicalized dependency parsers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011. p. 62-72.

MCDONALD, Ryan, et al. Universal dependency annotation for multilingual parsing. Proceedings of ACL, Sofia, Bulgaria, 2013.

NILSSON, Jens; RIEDEL, Sebastian; YURET, Deniz. The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL. sn, 2007. p. 915-932.

NIVRE, Joakim; HALL, Johan; NILSSON, Jens. Maltparser: A data-driven parser-generator for dependency parsing. In: Proceedings of LREC. 2006. p. 2216-2219.

PETROV, Slav; DAS, Dipanjan; MCDONALD, Ryan. A universal part-of-speech tagset. In: Proceedings of LREC. 2012.

RAMASAMY, Loganathan; ŽABOKRTSKÝ, Zdenek. Prague Dependency Style Treebank for Tamil. In: LREC. 2012. p. 1888-1894.

ZEMAN, Daniel. Reusable Tagset Conversion Using Tagset Drivers. In: LREC. 2008.

ZEMAN, Daniel, et al. HamleDT: To Parse or Not to Parse?. In: LREC. 2012. p. 2735-2741.

Přínos projektu k rozvoji fakulty / VŠ:

První část projektu - vytvoření kolekce syntakticky anotovaných korpusů (treebanků) s jednotným anotačním schématem - přímo navazuje na projekt HamleDT (Zeman et al. 2012). Výstupy této práce umožní Ústavu formální a aplikované lingvistiky (ÚFAL) udržet si prestiž a úroveň světové špičky v oblasti treebankingu. ÚFAL by se díky této kolekci například mohl pokusit o zorganizování soutěže v parsingu, navazující na úspěšné soutěže tohoto typu v minulosti (Nilsson et al. 2007).

Zejména díky svému velkému rozsahu, jakož i existenci mnoha nástrojů vyvinutých na

ÚFALu pro práci s daty tohoto typu, se kolekce stane cenným zdrojem pro studenty doktorského studia lingvistiky, kterým usnadní jejich výzkum – studenti ÚFALu se věnují například neřízenému parsingu (Mareček a Straka 2013) či jazykovým projekcím. Kolekce také umožní vytváření dalších odvozených datových zdrojů: ÚFAL je aktivní například v anotaci hloubkových jazykových struktur (Böhmová et al. 2003), valence (Urešová 2009), koreference (Nedoluzhko et al. 2009) či sentimentu (Veselovská 2012).

Vytvořený datový zdroj bude využitelný i ve výuce některých magisterských předmětů garantovaných ÚFALem, jako jsou Zdroje lingvistických dat, Technologie zpracování přirozeného jazyka či Pražský závislostní korpus.

Potřebu masivně paralelního zpracovávání velkých dat při práci s kolekcí lze využít v předmětech zabývajících se oblastmi data-intesive computing a paralelizací; práci s kolekcí pravděpodobně bude možné využít jako benchmark pro úlohy tohoto typu.

Poznatky získané experimenty s modelováním syntaxe napříč jazyky mohou být přínosné například pro systémy strojového překladu založené na syntaxi, jako je ÚFALem vyvíjené TectoMT (Žabokrtský et al. 2008), a s ním související mezinárodní projekt QTLeap, na němž se ústav podílí. ÚFALu by se tak mohly otevřít dveře k syntaktickému překladu mezi jinými jazykovými páry, než je jediný v současnosti podporovaný pár angličtina-čeština.

Na experimenty samotné pak mohou navázat další diplomové či dizertační práce, rozšiřující a prohlubující tyto experimenty a přinášející nové experimenty podobného typu.

Zdroje:

BÖHMOVÁ, Alena, et al. The Prague dependency treebank. In: Treebanks. Springer Netherlands, 2003. p. 103-127.

MAREČEK, David; STRAKA, Milan. Stop-probability estimates computed on a large corpus improve Unsupervised Dependency Parsing. In: In Annual Meeting of the Association for Computational Linguistics (ACL'13), 2013.

NEDOLUZHKO, Anna, et al. Extended coreferential relations and bridging anaphora in the prague dependency treebank. In: Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India. 2009. p. 1-16.

NILSSON, Jens; RIEDEL, Sebastian; YURET, Deniz. The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL. sn, 2007. p. 915-932.

UREŠOVÁ, Zdeňka. Building the PDT-VALLEX valency lexicon. In: On-line proceedings of the fifth Corpus Linguistics Conference. University of Liverpool. 2009.

VESELOVSKÁ, Kateřina. Sentence-level sentiment analysis in Czech. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. ACM, 2012. p. 65.

ZEMAN, Daniel, et al. HamleDT: To Parse or Not to Parse?. In: LREC. 2012. p. 2735-2741.

ŽABOKRTSKÝ, Zdeněk; PTÁČEK, Jan; PAJAS, Petr. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In: Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008. p. 167-170.

Materiální zajištění projektu:

Tento projekt bude vyžadovat především hardware a software umožňující efektivní zpracovávání, ukládání a verzování velkého množství dat. Ústav formální a aplikované lingvistiky již tímto vybavením disponuje, s výjimkou dostatečné diskové kapacity. Prostředky grantu budou proto použity na nákup pevných disků potřebné kapacity.

Cíle řešení projektu:

Hlavním cílem projektu je dosáhnout světové úrovně v úlohách modelování syntaxe napříč jazyky.

Dosažení snadné přenositelnosti jazykových technologií z jednoho jazyka na jiný jazyk, k němuž existují odpovídající datové zdroje, umožní výzkumníkům vymanit se z omezování se na jeden či několik málo jazyků: vyvinuté nástroje bude možné jednoduše aplikovat přinejmenším na desítky světových jazyků, a tak i snadno porovnat jejich úspěšnost s úspěšností nástrojů vyvinutých jinými výzkumníky. Věříme, že to přispěje k usnadnění a urychlení vývoje v komputační i formální lingvistice.

Úspěšné zvládnutí techniky mezijazyčné projekce pak umožní pracovat i s takovými jazyky, pro které dostatečně datové zdroje nejsou k dispozici. Počet jazyků, kterými lidé mluví, se odhaduje na několik tisíc, zatímco zdroje potřebné velikosti a kvality jsou dostupné pouze pro několik desítek z nich. Technologie pro práci s jazyky s omezenými zdroji umožňují použití nástrojů komputační lingvistiky i na tyto jazyky, bez nutnosti nejprve vytvořit potřebná data, což je časově i finančně náročné.

Dílčím cílem projektu je vytvoření velké multilinguální kolekce existujících syntakticky anotovaných korpusů (treebanků), harmonizovaných do jednotného anotačního schématu.

Možností využití této datové sady v komputační lingvistice se nabízí celá řada, zejména jako zdroje trénovacích dat pro parsing včetně jeho variant, jako je například delexikalizovaný parsing. Může také posloužit jako testovací data pro neřízenou závislostní analýzu jazyka, kde vynikne jednotnost jejího anotačního schématu, která umožní srovnání výsledků pro jednotlivé jazyky s velkou vypovídací hodnotou.

Zároveň půjde o cenný zdroj i pro formální lingvisty, kterým umožní snadno zkoumat všechny jazyky obsažené v kolekci, bez nutnosti seznamovat se pro každý jazyk s jeho anotačním schématem, neboť schéma bude pro všechny jazyky společné. Zejména ale zásadním způsobem usnadní práci na vzájemném porovnávání jednotlivých jazyků.

Způsob řešení:

Východiskem pro práci na tomto projektu se stane existující kolekce syntakticky anotovaných korpusů (treebanků) HamleDT (Zeman et al. 2012). Naším cílem bude vylepšit kvalitu této kolekce pomocí opravy chyb a nepřesností v konverzích zdrojových treebanků tak, aby bylo správně zachováno co nejvíce původních informací. Bude také nutné harmonizovat odlišně anotované závislostní struktury, které v rámci projektu HamleDT harmonizovány nebyly – například složená slovesa a podřadící spojky.

Pro odhalení chyb a nepravidelností ve výstupech konverzí budeme využívat jak pravidlových metod, které umožní odhalit přímé rozpory s anotačním schématem, tak metod pravděpodobnostního modelování a strojového učení. Ty nám umožní podchytit jevy pravidly nezachytitelné, jako jsou nepravidelnosti v rozložení jednotlivých značek přiřazených hranám (na základě různých kritérií, zejména slovních druhů slov spojených danou hranou), rozložení počtů potomků jednotlivých rodičovských uzlů, a podobně.

Je možné, že v některých případech se ukáže jako výhodnější použít jinou verzi zdrojového treebanku – HamleDT jako zdroj obvykle používá data ze sad CoNLL (Nilsson et al. 2007), která často již prošla nějakou automatickou konverzí, během které se mohly mnohé informace ztratit. V některých případech byla použita ne zcela kvalitní závislostní konverze původně složkového treebanku, v takových případech může být vhodné implementovat konverzi přímo z původního složkového treebanku.

Jedním z výstupů výše uvedených úprav bude i úprava stávajícího anotačního schématu tak, aby umožňoval vhodným způsobem zachytit všechny informace, které jsou obsaženy v podstatné části treebanků, ale anotační schéma PDT (Böhmová et al. 2003) je zachycuje nedostatečně nebo vůbec, neboť se v českém jazyce běžně nevyskytují – jde například o negativní částice a členy. Bude zváženo, zda místo úpravy existující sady analytických funkcí nezvolit přechod na jinou sadu značek závislostních vztahů, inspirovanou například Stanford Typed Dependencies (De Marneffe a Manning 2008).

Dalším z podúkolů projektu bude zmapování dalších existujících treebanků, které nejsou součástí sbírky HamleDT, a jejich zapojení do projektu. V kolekci dosud chybí některé velké treebanky, jako například treebanky čínštiny, francouzštiny či jeden z německých treebanků. Dále budou přidány i některé menší existující treebanky, například pro polštinu a hebrejštinu.

V druhé fázi projektu se zaměříme na využití vytvořené kolekce pro aktuální úlohy syntaktické analýzy jazyka.

V úloze mezijazyčné projekce se pokusíme vyvinout úspěšnou metodu pro natrénování syntaktického parseru na treebancích pro jeden nebo několik jazyků a jeho následné použití na analýzu jazyka jiného. Jednou z metod, které je možné využít, je tzv. delexikalizovaný parsing (McDonald et al. 2011), kdy se parser natrénuje na treebanku, v němž byla jednotlivá slova nahrazena jejich tagy. Přitom velmi záleží na tom, jak vysokou granularitu tagů použijeme – zda budou zachycovat pouze slovní druhy slov, nebo i některé jejich morfologické rysy, apod. Tato úloha je užitečná pro analýzu jazyků, pro něž nejsou k dispozici dostatečné datové zdroje pro natrénování parseru standardním způsobem. Očekáváme, že pro analýzu daného jazyka bude nejvhodnější natrénování parseru na jednom nebo několika nejpodobnějších jazycích. Naším cílem je dosáhnout s námi

vyvinutou metodou úspěšnosti srovnatelné s nejlepšími světovými systémy.

V úloze přenositelnosti jednojazyčných technologií se zejména pokusíme sestrojit závislostní parser, založený na některém z nejlepších současných parserů (např. Nivre et al. 2006, McDonald et al. 2005), který bude dosahovat vysoké úspěšnosti na všech jazycích s dostatečnými datovými zdroji, přičemž bude stačit jej natrénovat na treebanku tohoto jazyka, bez nutnosti jej navíc ručně ladit na daný jazyk. V současnosti používané parsery je totiž obvykle nutné pro každý jazyk vyladit, tj. nalézt vhodné hodnoty jejich parametrů, tak aby dosahovaly vysoké úspěšnosti. Věříme, že díky harmonizaci treebanků se nám podaří nalézt takovou sadu parametrů, aby ladění na jednotlivé jazyky nebylo nutné. Může se stát, že toto nebude možné, pak bude naším cílem rozdělit jazyky do jednotlivých typologicky odlišných skupin, a vyladit parser na každou takovou skupinu zvlášť.

Práce na projektu bude probíhat na platformě Treex, která poskytuje mnoho nástrojů pro zpracování jazyka, a nad níž byl vystavěn projekt HamleDT. Použity budou odpovídající moderní technologie - výpočetní cluster, programovací model MapReduce, kódování Unicode, a podobně.

Zdroje:

BÖHMOVÁ, Alena, et al. The Prague dependency treebank. In: Treebanks. Springer Netherlands, 2003. p. 103-127.

DE MARNEFFE, Marie-Catherine; MANNING, Christopher D. The Stanford typed dependencies representation. In: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation. Association for Computational Linguistics, 2008. p. 1-8.

MCDONALD, Ryan, et al. Non-projective dependency parsing using spanning tree algorithms. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005. p. 523-530.

MCDONALD, Ryan; PETROV, Slav; HALL, Keith. Multi-source transfer of delexicalized dependency parsers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011. p. 62-72.

NILSSON, Jens; RIEDEL, Sebastian; YURET, Deniz. The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL. sn, 2007. p. 915-932.

NIVRE, Joakim; HALL, Johan; NILSSON, Jens. Maltparser: A data-driven parser-generator for dependency parsing. In: Proceedings of LREC. 2006. p. 2216-2219.

ZEMAN, Daniel, et al. HamleDT: To Parse or Not to Parse?. In: LREC. 2012. p. 2735-2741.

Prezentace výsledků:

Výsledky budou průběžně prezentovány na seminářích Ústavu formální a aplikované lingvistiky a na WDS.
Budeme publikovat příspěvky na mezinárodních konferencích – pokusíme se o přijetí článku na
konferenci LREC, ACL, a/nebo TLT – a plánujeme i publikaci v odborném časopise, například PBML.
Průběžné výsledky budeme popisovat také v technických zprávách.
Vytvořený software bude průběžně zveřejňován na webových stránkách pod svobodnou licencí.

U všech publikací, včetně disertační práce, bude uvedeno, že byly finančně podporovány Grantovou agenturou Univerzity Karlovy.

Přílohy

☐	Vybrané publikace - Zdeněk Žabokrtský	53519 B	13.11.2013 01:30:35	
☐	Životopis a publikace - Rudolf Rosa	200417 B	11.11.2013 21:41:47	
☐	Životopis - Jan Mašek	50022 B	13.11.2013 11:39:59	
☐	Životopis - Zdeněk Žabokrtský	64591 B	13.11.2013 01:28:53	