1. Výsledky dosažené v roce 2006

Ve druhém roce řešení projektu jsme pokračovali ve výzkumu ve dvou paralelních větvích. První se zabývá studiem pojmenovaných entit v češtině, druhá jednotným formátem pro lingvistická data. Výsledky byly prezentovány na mezinárodních konferencích LREC, TSD a TLT 2006 a na několika dalších tuzemských seminářích a workshopech. Významná byla prezentace na tutoriálu Prague Treebanking for Everyone v rámci Vilem Mathesius Lecture Series 21, kterého se zúčastnilo několik desítek posluchačů z řad českých i zahraničních odborníků.

A. Pojmenované entity

Data vzniklá anotací v roce 2005 byla dále ručně pročišťována. Byla upravena sada použitých značek - nově povolena podspecifikace, zavedení "šablonových" značek pro kontejnerové pojmenované entity.
Data vzniklá v roce 2005 byla převedena do formátu XML.
Úloha vyhledávání a klasifikace pojmenovaných entit v češtině byla formulována jako "task proposal" pro SemEval-2007 (4th International Workshop on Semantic Evaluations). Úloha byla recenzenty přijata.
Anotační schéma vyvinuté pro anotaci především pojmenovaných entit substantivní povahy bylo rozšířeno i na číselné výrazy. Anotační pokyny byly ověřovány na pilotní sadě 100 vět náhodně vybraných z Českého národního korpusu.
Místo anotace 10000 vět z PDT, která byla plánována pro rok 2006, byla provedena anotace 4000 vět vybraných z ČNK. Důvodem byla skutečnost, že tímto způsobem bylo i při menším množství vět možné dosáhnout daleko větší pestrosti zpracovaného materiálu, a tudíž lze očekávat i vyšší robustnost automatických metod, k jejichž trénování budou tato data použita.
Vlastnosti pojmenovaných entit byly využity ve dvou českých parserech. V prvním případě šlo o parser založený na ručně psaných pravidlech (parser byl publikován na konferenci TSD'06 v níže uvedeném článku). Ve druhém případě šlo o přidání rysů souvisejících s pojmenovanými entitami do statistického parseru založeného na algorimtu minimální kostry.
Výsledky práce na pojmenovaných entitách byly shrnuty v technické zprávě, která je připravena k publikaci - vyjde na pracovišti ÚFAL na začátku roku 2007.

B. Jednotný datový formát PML

Do datového formátu PML byla přidána řada vlastností navrhovaných v závěru technické zprávy TR-2005-29, zejména verzování a podpora modularizace schémat a rozšiřování přejatých datových typů. Formát byl dále rozšířen o kompletní sadu jednoduchých datových typů definovaných jazykem XML Schema.
Byl vytvořen samostatný nástroj sloužící k předzpracování modularizovaných PML schémat do podoby umožňující jejich zpracování jednoduchými transformačními prostředky, jako je např. XSLT 1.0.
Byl vytvořen nástroj automatizující validaci PML datových souborů prostřednictvím validačního jazyka Relax NG. V této souvislosti byly provedeny experimenty vyjádřením dodatečných omezujících podmínek v jazyce ISO Schematron.
Byla vytvořena referenční implementace načítání a ukládání formátu PML v jazyce Perl. Ta byla dále integrována do anotačního nástroje TrEd, jenž byl současně doplněn o řadu nových vlastností. Nová implementace umožňuje mj. transparentní převod z cizích XML formátů do PML a zpět prostřednictvím XSLT šablon.
Byly zdokonaleny nástroje pro převod z formátu CSTS, starších datových formátů užívaných v PDT 1.0, a vytvořeny nové nástroje pro převod textového formátu Penn Treebank, XML formátu Alpino Treebank, formátu Slovene Dependency Treebank založeného na XML TEI P5 a tabulkového formátu CoNLL-X.
Na integraci dalších datových zdrojů se pracuje.

Formát PML byl prostřednictvím anotačního nástroje TrEd nasazen v rámci nizozemského projektu Alpino Treebank (Algorithms for Linguistic Processing - NWO PIONIER Research Project).
Byly zahájeny experimenty s možností uložení a indexace strukturně anotovaných dat v nativní XML databázi (SleepyCat dbXML a eXists). Taktéž byly zahájeny experimenty s ukládáním slovníkových dat ve formátech založených na PML.
Nový formát PML byl použit na datech nově vydaného datového CD PDT 2.0, které vyšlo v LDC (viz položka č. 2 v odstavci Seznam publikací).

Cestovní prostředky byly využity na prezentaci výsledků na domácích i zahraničních konferencích a seminářích, zejména na 5. mezinárodní konferenci LREC 2006 (Language Resources and Evaluation) v italském Janově a na 9. mezinárodní konferenci TSD 2006 (TEXT, SPEECH and DIALOGUE) v Brně. Další významná konference, na které jsme prezentovali své výsledky, (TLT 2006) byla v Praze, tudíž nevyžadovala cestovní náklady. Celkem bylo na cestování vynaloženo 95128 Kč. Zbylých 5 tisíc Kč bylo převedeno do Fondu účelově určených prostředků (FÚUP) a je předpoklad jejich faktického čerpání v příštím roce.

Řešitelský tým doznal na MFF UK v roce 2006 určitých změn, které byly nahlášeny a zdůvodněny ve Změnovém listu 27. června 2006. Na rok 2007 plánujeme stejný řešitelský tým jako v roce minulém, ale změní se pracovní kapacita některých pracovníků, což odráží jejich účast na dalších projektech řešených na pracovišti MFF UK.
Spolupráce mezi oběma institucemi, tj. MFF UK a ÚJČ AV ČR, probíhala ve stejném duchu jako v loňském roce. ÚJČ se soustředí na teoretické otázky, týkající se homonymie vlastních jmen se jmény obecnými, zatímco řešitelé na MFF se zabývají spíše praktickými aplikacemi.

Prostředky na **mzdy** byly čerpány podle plánu. Vzhledem k závěrečným výběrům dovolených se však nepodařilo vyčerpat částku na mzdy úplně, zbylo 10 tisíc Kč, které byly převedeny do Fondu účelově určených prostředků (FÚUP) a budou fakticky vyčerpány v přístím roce.

Celková částka pro převod do FÚUP (15 tis. Kč) byla připočítána podle pokynů k vyplňování webových formulářů tak, aby byla vykázána jako vyčerpaná.

Prostředky naplánované na **investice** byly v roce 2006 čerpány na následující položky:
* chassis PowerEdge 1955,
* 25% účast na servrovém vybavení blade serveru PE 1955 5130.

OON se čerpaly na ruční anotační práce a na pomocné programátorské práce, prováděné převážně studenty MFF a FF UK.

Nejpodstatnější částky z kolonky **Věcné náklady** byly vyčerpány na režii a pojištění z mezd. Zbytek potom tvořil nákup literatury a drobného hmotného majetku.

2. Návrh postupu prací na rok 2007

V roce 2007 plánujeme uskutečnit tyto cíle (opět členěno podle paralelních větví):
A. Pojmenované entity
* další vývoj systému pro automatické rozpoznávání entit,
* implementace prostředí pro vizualizaci a anotaci pojmenovaných entit na tektogramatické rovině v editoru

stromů TrEd, případně i pilotní ruční anotace na této rovině,
* zahájení systematického shromažďování lexikálních zdrojů (rejstříky geografických názvů atd.),
* ruční anotace dalších vět obsahujících pojmenované entity.

B. Jednotný formát
* další experimenty a vyhodncení možností zpracování dat ve formátu PML ve velkých objemech,
* experimenty s PML ve vztahu k relační databázi,
* rozšíření PML formátu o slovníkové a další relační role,
* převod existujících anotačních slovníků do PML,
* návrh a testování aplikačního rozhraní pro zpracování integrovaných datových zdrojů,
* práce na PML podpoře v nástrojích (zejména MEdit; anotační nástroj TrEd s podporou maker již s PML pracuje).

Přidělené prostředky budeme čerpat přibližně ve stejném složení jako dosud.
Konkrétně částku naplánovanou na investice hodláme použít na nákup dalších výpočetních a datových servrů, včetně rozšíření zálohovací, datové a řídicí infrastruktury.
Cestovné využijeme opět na prezentaci našich výsledků na konferencích. Hodláme se zúčastnit konference TSD 2007, dále potom mezinárodní konference ACL 2007, která se letos koná v Praze, konference ITAT a Slovko, které pořádají naši slovenští kolegové, a dalších evropských konferencí.

Seznam publikací

1. Džeroski Sašo, Erjavec Tomaž, Ledinek Nina, Pajas Petr, Žabokrtský Zdeněk, Žele Andreja. Towards a Slovene Dependency Treebank. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006) . Paris, France: 2006, pp. 1388-1391.

2. Hajič Jan, Panevová Jarmila, Hajičová Eva, Sgall Petr, Pajas Petr, Štěpánek Jan, Havelka Jiří, Mikulová Marie, Žabokrtský Zdeněk, Ševčíková-Razímová Magda: Prague Dependency Treebank 2.0, Linguistic Data Consortium, 2006.

3. Hlaváčová Jaroslava. New Approach to Frequency Dictionaries - Czech Example. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). 2006, pp. 373-378.

4. Holan Tomáš, Žabokrtský Zdeněk. Combining Czech Dependency Parsers. In Lecture Notes in Computer Science: Proceedings of the 9th International Conference, TSD 2006. Springer-Verlag Berlin Heidelberg, 2006, pp. 95-102.

5. Mikulová Marie, Bémová Alevtina, Hajič Jan, Hajičová Eva, Havelka Jiří, Kolářová Veronika, Kučová Lucie, Lopatková Markéta, Pajas Petr, Panevová Jarmila, Razímová Magda, Sgall Petr, Štěpánek Jan, Urešová Zdeňka, Veselá Kateřina, Žabokrtský Zdeněk. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Tech. Report 30 ÚFAL MFF UK, 2006, 1287 pp.

6. Mikulová Marie, Bémová Alevtina, Hajič Jan, Hajičová Eva, Havelka Jiří, Kolářová Veronika, Kučová Lucie, Lopatková Markéta, Pajas Petr, Panevová Jarmila, Ševčíková Magda, Sgall Petr, Štěpánek Jan, Urešová Zdeňka, Veselá Kateřina, Žabokrtský Zdeněk. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Reference book. Tech. Report 32 ÚFAL MFF UK, 2006, 193 pp.

7. Mikulová Marie, Bémová Alevtina, Hajič Jan, Hajičová Eva, Havelka Jiří, Kolářová Veronika, Kučová Lucie, Lopatková Markéta, Pajas Petr, Panevová Jarmila, Ševčíková Magda, Sgall Petr, Štěpánek Jan, Urešová Zdeňka, Veselá Kateřina, Žabokrtský Zdeněk. Anotace na tektogramatické rovině Pražského závislostního korpusu. Referenční příručka. Tech. Report 31 ÚFAL MFF UK, 2006, 183 pp.

8. Pajas Petr, Štěpánek Jan. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006). Genova, Italy: 2006, pp. 40-47.

9. Ptáček Jan, Žabokrtský Zdeněk. Synthesis of Czech Sentences from Tectogrammatical Trees. In Lecture Notes in Computer Science: Proceedings of the 9th International Conference, TSD 2006. Springer-Verlag Berlin Heidelberg, 2006, pp. 221-228.

10. Razímová Magda: System of Pronominal Words in Czech with Respect to German and English. In Abstracts of 39th Annual Meeting of Societas Linguistica Europaea. University of Bremen, 2006, pp. 53-54.

11. Razímová Magda, Žabokrtský Zdeněk. Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In Proceedings of the LREC Workshop on Annotation Science. Genova, Italy, May 27: 2006, pp. 12-19.

12. Ševčíková-Razímová Magda, Žabokrtský Zdeněk. Systematic Parameterized Description of Pro-forms in the Prague Dependency Treebank 2.0. In Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT). Prague, Czech Republic, Dec. 1-2: 2006, pp. 175-186.

13. Štěpánek Jan. Post-annotation checking of Prague Dependency Treebank 2.0 data, Prague Bulletin of Mathematical Linguistics, 2006, pp. 23-33.

14. Štěpánek Jan. Post-annotation Checking of Prague Dependency Treebank 2.0 Data. In Lecture Notes in Computer Science: Proceedings of the 9th International Conference, TSD 2006. Springer-Verlag Berlin Heidelberg, 2006, pp. 277-284.

15. Štěpánek Jan. Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat). Ph.D. thesis., 2006, 95 pp.

16. Giger, Markus - Štěpán, Pavel: Česká deverbální příjmení a problém jejich homonymie v elektronických korpusech. Acta onomastica 47, s. 185-196.