PARSOVANIE A VIACSLOVNÉ VÝRAZY
Úvodné pracovné stretnutie partnerov COST, akcia IC1207 PARSEME

Daniela Majchráková – Radoslav Brída
Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV, Panská 26
811 01 Bratislava, e-maily: danam@korpus.sk – brida@korpus.sk

COST (European Cooperation in Science and Technology) je medzinárodná platforma na kooperáciu európskych vedeckých pracovísk v rámci rôznych vedných odvetví. Spolupráca je podporovaná a financovaná Európskou úniou. Patrí sem lingvisticky a zároveň počítačovo orientovaná COST akcia *IC1207* PARSEME: *PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing* vznikla s cieľom spojiť odborníkov v oblasti lexikografického a počítačového spracovania viacslovných výrazov a umožniť ich vzájomnú spoluprácu. Do projektu je v súčasnosti zapojených viac ako 100 vedeckovýskumných pracovníkov z 29 štátov vrátane Slovenska, ktoré je v PARSEME zastúpené od júna 2013 prostredníctvom Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV. Spolupráca jednotlivých členských krajín sa má uskutočňovať prostredníctvom štyroch pracovných skupín (Working Groups, ďalej WG), ktorých pracovná náplň je rozdelená podľa úrovní a spôsobov spracovania viacslovných výrazov.

V dňoch 16. – 18. septembra 2013 sa vo Varšave uskutočnilo prvé z pracovných stretnutí PARSEME, ktoré sa majú konať pravidelne dvakrát ročne v priebehu štyroch rokov. Jednotlivé pracovné skupiny zložené z teoretických lingvistov a expertov na počítačové spracovanie prirodzeného jazyka predstavili na podujatí aktuálny stav v oblasti výskumu viacslovných výrazov a navrhli jeho ďalšie smerovanie.

Prvá pracovná skupina (WG1) sa venuje lexikálnym databázam viacslovných výrazov, kolokačným a valenčným slovníkom a ich jednojazyčným a viacjazyčným variantom. Zdôraznila lexikálnu reprezentáciu a anotáciu viacslovných spojení (manuálnu alebo automatickú) a štandardizáciu zozbieraných dát pre algoritmy počítačového spracovania jazyka. Jednotliví účastníci prvej pracovnej skupiny na stretnutí predstavili lexikálne databázy s rozličnými spôsobmi spracovania lexikálnej, syntaktickej a morfologickej roviny.

Nevyhnutným predpokladom formálnej reprezentácie viacslovných výrazov je ich presné lingvistické uchopenie. S tým súvisí najmä deskripcia lexikálnych a syntaktických vlastností viacslovných výrazov a ich odlíšenie od pravidelných syntaktických konštrukcií. Cieľom pracovnej skupiny WG1 je pomocou kontrastívnej analýzy viacslovných výrazov v jednotlivých jazykoch vytvoriť spoločný abstraktný

model ustáleného spojenia slov. Tento model by mohol byť automaticky aplikovaný v rôznych jazykoch a slúžiť napríklad na vytvorenie viacjazyčných paralelných databáz (lexikónov), ktoré by sa mohli využiť v ďalšom počítačovom spracovaní, predovšetkým na parsovanie.

Predmetom záujmu prvej pracovnej skupiny je nielen vytváranie nových lingvistických zdrojov, ale aj rozširovanie a obohacovanie existujúcich lexikálnych a valenčných databáz o viacslovné výrazy. Pozornosť sa zameriava na riešenie otázky, ako zachytiť všetky existujúce slovné spojenia vyskytujúce sa v jazyku a do akej miery môžu byť nástroje automatickej extrakcie slovných spojení efektívne pri zostavovaní lexikónov.

Tematickou náplňou druhej pracovnej skupiny (WG2) je parsovanie ustálených viacslovných výrazov, čo je v súčasnosti jedna z najväčších výziev v oblasti počítačového spracovania prirodzeného jazyka. Parsovanie je proces zameraný na určenie syntaktických štruktúr slov, fráz a viet. Tento proces sa tradične zakladá na lexikónoch a gramatikách, ktoré zachytávajú všeobecné vlastnosti slov a vzájomné vzťahy slov a štruktúr vo vetách. Rôzne formálne gramatiky ponúkajú rozdielne štruktúry a operácie na vytvorenie gramatických pravidiel predovšetkým pravidelných jazykových štruktúr, ktoré sú pomerne ľahko uchopiteľné. Problém nastáva pri vytváraní modelov nepravidelných syntaktických štruktúr, t. j. ustálených viacslovných výrazov, čo môže byť spôsobené ich syntaktickou zrastenosťou, rôznymi syntaktickými variáciami, morfologickými či lexikálnymi anomáliami a podobne. Východiskom je predpoklad, že zohľadnenie týchto vlastností v gramatických modeloch môže viesť k väčšej presnosti parsovacieho procesu.

Jednou z hlavných úloh druhej pracovnej skupiny je zefektívnenie parsovania ustálených viacslovných výrazov pomocou ich integrácie do gramatických pravidiel. Ďalšou výzvou bude začlenenie sémantickej reprezentácie viacslovných výrazov do výsledných štruktúr. Výsledkom spolupráce tejto pracovnej skupiny by malo byť vytvorenie abstraktných modelov viacslovných výrazov, to znamená gramatických pravidiel, ktoré by sa mohli automaticky aplikovať do rôznych gramatických formalizmov, čím by sa znížili náklady spojené s produkciou čiastkových gramatík rôznych jazykov.

Takzvanému hybridnému parsovaniu sa venuje tretia pracovná skupina (WG3). Pri syntaktickej analýze viet je častým problémom správna identifikácia viacslovných výrazov a zakomponovanie tejto informácie do samotnej analýzy textu. Úlohou pracovnej skupiny je výskum v tejto oblasti a snaha o zlepšenie existujúcich alebo vytvorenie nových postupov na spracovanie viacslovných výrazov.

Zložitosť uchopenia ustálených slovných spojení môže spočívať v ich sémantickej nerozložiteľnosti alebo v tom, že komponenty viacslovných spojení sa v textoch nevyskytujú vždy bezprostredne vedľa seba. V diskusii a prednáškach sa preto

účastníci najviac venovali rôznym stratégiám na identifikáciu viacslovných výrazov v korpusoch a ich spracovanie pri syntaktickej analýze textu. Jednou z možností je identifikácia ustálených spojení ako samostatných jednotiek ešte pred analýzou textu. Používajú sa na to automatické nástroje natrénované na ručne značkovaných syntaktických štruktúrach. Takýto postup môže viesť k protichodným výsledkom: buď parser na základe tejto informácie nebude vedieť text spracovať, alebo, naopak, analýza textu sa tým zjednoduší. Hybridnou stratégiou je hľadanie viacslovných výrazov počas analýzy, pričom sú vlastnosti viacslovných výrazov zakomponované do gramatických pravidiel a ich identifikácia prebieha súčasne s morfologickou anotáciou s využitím ďalších zdrojov štatistických údajov, ako sú lexikóny, zoznamy n-gramov a podobne. Tento prístup sa ukazuje pracovnej skupine ako perspektívny, s otvorenými možnosťami na jeho zlepšovanie.

Štvrtá pracovná skupina (WG4) je zameraná na reprezentáciu viacslovných výrazov v treebankoch (syntaktických stromoch). Treebanky predstavujú dôležitý zdroj štatistických informácií o lingvistických jednotkách či štruktúrach, ako aj o ich kontexte, preto sú tieto dáta prínosné pre hybridné parsovacie metódy súvisiace s náplňou práce WG3. Cieľom štvrtej pracovnej skupiny je vytvorenie spoločných pravidiel, podľa ktorých sa môžu označovať viacslovné výrazy v treebankoch (syntaktických stromoch). Tieto treebanky sa dajú ďalej využiť na trénovanie systémov schopných analyzovať vety či extrahovať z nich viacslovné výrazy, ktoré môžu byť po manuálnej oprave zaradené do lexikálnych databáz.

Spoločné zameranie a cieľ COST akcie PARSEME vytvára priestor na kooperáciu jednotlivých pracovných skupín a umožňuje prepojenie rôznych úrovní lingvistického spracovania viacslovných spojení. Prínos projektu má byť predovšetkým v kontrastívnej metodológii, pomocou ktorej sa budú môcť skúmať viacslovné výrazy v kontexte rôznych metodologických východísk. Medzi očakávané výsledky patrí obohatenie súčasných lingvistických zdrojov (lexikálne databázy, kolokačné slovníky, valenčné slovníky), vyhovujúca formálna reprezentácia a automatické generovanie viacslovných spojení.

Prezentácie z prvého stretnutia sú dostupné na internetovej adrese:
http://typo.uni-konstanz.de/parseme/index.php/event/meetings/12-cost-action-ic1207-parseme-meeting-16-18-september-2013-warsaw

Druhé pracovné stretnutie COST akcie PARSEME sa bude konať 10. – 11. marca 2014 v Aténach.