Syntaktická analýza vnorených programovacích jazykov

Tomáš Belan Vedúci práce: RNDr. Richard Ostertág, PhD.

MOTIVÁCIA

Existujú tisíce programovacích jazykov, špecializované na rôzne účely. Nie vždy platí, že rôzne jazyky sú uložené v rôznych súboroch – čoraz častejšie sa stretávame so situáciou, že jeden súbor používa viacero jazykov. Medzi praktické príklady patria: SQL dotazy vnorené v PHP/Python/Java kóde, JavaScript udalosti v HTML dokumente, HTML fragmenty v PHP, Assembler v C, shell skript v Makefile a mnohé ďalšie.

Tieto vnorené jazyky sú spravidla spracovávané oddelenými parsermi, ktoré navzájom nekooperujú: PHP parser neoveruje korektnosť SQL dotazov, atď. To sťažuje statickú analýzu a otvára dvere bezpečnostným a iným chybám. Moja diplomová práca ponúka lepší spôsob, ako navrhovať a analyzovať vnorené programovacie jazyky.

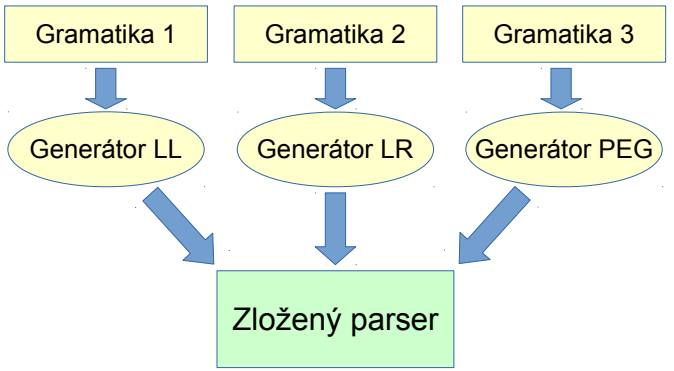
MULTIPARSER

Navrhli a implementovali sme systém **Multiparser**. Ide o generátor parserov, ktorý dokáže spojiť parsery pre jednotlivé vnorené jazyky do jedného celku.

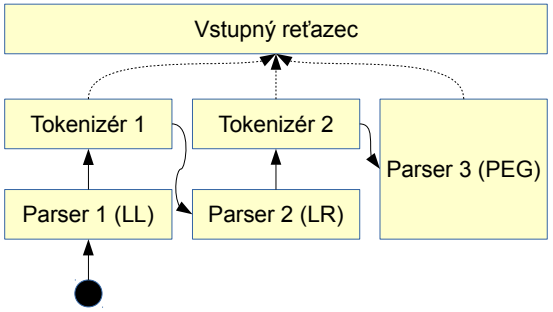
Multiparser podporuje tri klasické modely syntaktickej analýzy – LL, LR a PEG – a rozširuje ich o podporu oddeľovačov, ktoré určujú začiatok a koniec vnoreného jazyka. Parsery pre jednotlivé jazyky potom pracujú s tým istým vstupom, a navzájom si odovzdávajú kontrolu, keď začne alebo skončí úsek napísaný v inom jazyku. Výsledkom je jediný syntaktický strom, ktorý obsahuje aj informáciu o vnorených jazykoch. Tento strom môže byť ďalej použitý na statickú analýzu alebo v ďalších fázach kompilácie.

Každý jazyk je určený samostatnou gramatikou, vďaka čomu môžu jednotlivé jazyky používať aj rôzne pravidlá lexikálnej analýzy – rôzne formáty pre komentáre, reťazcové a číselné literály, povolené operátory atď.

GENERÁTOR ZLOŽENÝCH PARSEROV



BEH ZLOŽENÉHO PARSERA



Automaty pre jednotlivé vnorené jazyky vystupujú ako vzájomne rekurzívne procedúry, ktoré pracujú so spoločným vstupom.