TRANSFORMATION OF EQUATION EDITOR EXPRESSIONS TO LATEX

Jiří Šimek
Bachelor Degree Programme (3), FIT BUT
E-mail: xsimek01@stud.fit.vutbr.cz

Supervised by: Zbyněk Křivka
E-mail: krivka@fit.vutbr.cz

ABSTRACT

This paper discusses the transformation of Microsoft Word Equation Editor expressions to LaTeX. First, Equation Editor expressions are described. As the main topic, the translator is designed along with some solutions of the consequent problems.

1 ÚVOD

Microsoft Word jako součást kancelářského balíku Microsoft Office je bezesporu jedním z nejrozšířenějších nástrojů pro vytváření a editaci textových dokumentů. Existují však i alternativní způsoby tvorby textových dokumentů. Jedním z nich, který je rozšířen zejména v akademické a vědecké sféře, je sázecí systém LaTeX (více viz [1]).

Systém TeX a jeho nadstavba LaTeX primárně vznikly jako nástroje pro sázení matematických textů. Microsoft Word také umožňuje psát matematické výrazy pomocí nástroje Editor rovnic (Equation Editor). Každý z těchto nástrojů ale používá pro reprezentaci matematických výrazů různé způsoby a odlišné formáty. Navíc nelze přímočaře převést jeden formát na druhý.

Mým cílem je usnadnit převod matematických výrazů z Editoru rovnic do LaTeXu a vytvořit převodník, který by tuto činnost prováděl automaticky.

Podobným problémem se zabývají tři nástroje, které z Microsoft Word do LaTeXu nepřevádí jen matematické výrazy, ale celé dokumenty. Prvním nástrojem je *Wv*, který umí převádět dokumenty pouze z Microsoft Word 2000 a nižších verzí. Druhým je *Word-to-LaTeX*, který ale matematické výrazy vytvořené v Microsoft Word 2007 převádí velice neuspokojivě, a posledním nástrojem je komerční *GrindEQ*.

2 MATEMATICKÉ VÝRAZY V EDITORU ROVNIC V MICROSOFT WORD 2007

Převodník jsem se rozhodl implementovat jen pro Microsoft Word 2007. Vedly mě k tomu následující důvody. Word 2007 poprvé umožňuje přímý přístup k matematickým výrazům přes API, což usnadňuje implementaci převodníku při získávání matematických výrazů z dokumentu Wordu. Navíc Word 2007 obsahuje novější verzi Editoru rovnic, která není plně kompatibilní s předchozími verzemi. To znamená, že matematické výrazy vytvořené ve starších verzích Wordu je možné editovat i ve Wordu 2007, ale ne naopak, kdy jsou výrazy uloženy jako obrázky.

2.1 ZPŮSOB ULOŽENÍ MATEMATICKÝCH VÝRAZŮ

Obecně celý kancelářský balík Office 2007 používá nový formát souborů Open XML. Toto platí i pro matematické výrazy, pro které je v rámci Open XML definován jazyk Office Math Markup Language (OMML).

Pro zadávání výrazů poskytuje Editor rovnic dvě možnosti. Buď editovat výraz pomocí grafického uživatelského rozhraní, nebo výraz zapsat v takzvaném lineárním formátu. *Lineární formát* je způsob zápisu výrazu v textovém tvaru, který je podobný zápisu výrazů v LaTeXu. Např. výraz $\frac{a^2}{b}$ je v lineárním formátu v Editoru rovnic zapsán `a^2/b`. Editor rovnic obsahuje dvě funkce *Math AutoCorrect* a *Formula Autobuildup*, která vysází výrazy zapsané v lineárním formátu do formátu, jak má výraz skutečně vypadat, tedy $\frac{a^2}{b}$. Více o lineárním formátu výrazů v Editoru rovnic lze nalézt v [2].

Hlavním rozdílem lineárního formátu oproti LaTeXu je to, že lineární formát místo jednotlivých maker určujících, jaký symbol se má zobrazit, obsahuje Unicode znaky přímo reprezentující daný symbol. Podobně jako v LaTeXu je sice možné psát ve výrazech entity jako např. `\int` pro integrál, ale *Math AutoCorrect* tuto entitu nahradí Unicode znakem reprezentujícím integrál (U+222B).

3 VYTVOŘENÍ PŘEVODNÍKU

Při návrhu převodníku jsem vycházel z lineárního formátu matematických výrazů v Editoru rovnic. Matematický výraz v lineárním formátu tedy bude vstupním řetězcem převodníku a výstupem bude řetězec maker LaTeXu reprezentující stejný matematický výraz.

Základem převodníku je gramatika popisující matematické výrazy Editoru rovnic. Při tvorbě gramatiky jsem vycházel zejména z [2] zabývající se způsobem zápisu matematických výrazů v lineárním formátu v Editoru rovnic. V tomto dokumentu se sice nachází gramatika popisující výrazy Editoru rovnic, ale její použití pro převodník není možné. Tato gramatika totiž popisuje výrazy před jejich korekcí funkcí *Math AutoCorrect* a navíc je nejednoznačná. Proto jsem byl nucen vytvořit vlastní gramatiku. Syntaxi matematických výrazů jsem zjišťoval také přímo praktickým psaním výrazů v Editoru rovnic.

Pro vytvoření samotného převodníku z gramatiky jsem chtěl původně využít nástroje Flex (viz [3]) a Bison, ale Flex neumí vytvořit lexikální analyzátor zpracovávající vstupní řetězec v Unicode kódování. Nakonec jsem místo nástrojů Flex a Bison využil nástroj ANTLR (ANother Tool for Language Recognition, viz [4]), který nabízí podobnou funkčnost a umí zpracovat Unicode řetězce.

3.1 PROBLÉMY PŘEVODU

Hlavní problém při převodu matematických výrazů z Editoru rovnic do LaTeXu vyplývá už ze samotného formátu výrazů v Editoru rovnic, který využívá Unicode kódování. Pro každý Unicode znak neexistuje odpovídající makro LaTeXu. Částečným řešením je použít balík `ucs` obsahující makro `\unichar`, které dokáže zobrazit podstatnou podmnožinu Unicode znaků.

V Editoru rovnic je možné psát text několika druhy písem. Znaky těchto písem jsou přímo součástí Unicode (konkrétně rozsah U+1D400 – U+1D7FF). Lexikální analyzátor vytvořený pomocí nástroje ANTLR dokáže zpracovat znaky pouze z rozsahu U+0000 - U+FFFF. Jedním

z řešení tohoto problému je vytvořit preprocesor, který znaky s hodnotou větší než U+FFFF nahradí odpovídajícím makrem LaTeXu, a teprve takto upravený matematický výraz bude vstupem lexikálního analyzátoru.

3.2 PŘÍKLAD PŘEVODU

Následující rovnice je vytvořená v Editoru rovnic a pod ní se nachází tatáž rovnice v lineárním formátu.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{x^2-1}} dx = \ln\left(x+\sqrt{x^2+1}\right)+c$$

$$\int \_(-\infty)\text{^}\infty \text{▒}[\![1/\sqrt{(x\text{^}2-1)}\ dx]\!] = \ln\left(x+\sqrt{(x\text{^}2+1)}\right)+c$$

Následující rovnice byla vygenerována vytvořeným převodníkem z lineárního formátu uvedeného na předchozím obrázku a pod ní je její zdrojový kód.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{x^2-1}} dx = \ln\left(x+\sqrt{x^2+1}\right)+c$$

```
\int _{-\infty }^{\infty }{{\frac{1}{\sqrt{x^{2} -1}} dx}}=
          \ln {\left(x+\sqrt{x^{2} +1}\right)} +c
```

4 ZÁVĚR A BUDOUCÍ VÝVOJ

V současné době jsem vytvořil první verzi převodníku matematických výrazů z Editoru rovnic do LaTeXu ve formě konzolové aplikace, která při spuštění načte textový soubor s výrazem v lineárním formátu. Tato verze slouží pro ověření správnosti vytvořené gramatiky, která se dosavadními testy potvrdila. I když je převodník stále ve fázi vývoje, lze jej pro převod většiny matematických výrazů použít již nyní. V budoucích verzích převodníku plánuji jednak vylepšit některé nedostatky převodu nynější verze, např. různé styly zobrazení zlomků v Microsoft Word a jim odpovídající styly v LaTeXu nebo umístění horních/dolních indexů u některých symbolů nad/pod nebo v pravém horním/dolním rohu. Dále plánuji doplnit preprocesor na zpracování zatím nepodporovaných Unicode znaků. V neposlední řadě také buď integrovat vytvořený převodník do prostředí Microsoft Word, nebo umožnit načíst dokument Microsoft Word a převést matematické výrazy automaticky.

REFERENCE

[1] RYBIČKA, J.: LaTeX pro začátečníky. Konvoj, 2003, ISBN 80-7302-049-1.

[2] SARGENT, M.: Unicode Nearly Plain-Text Encoding of Mathematics. Technická zpráva, Once Authoring Services, Microsoft Corporation, 2006-08-28.

[3] Flex: The Fast Lexical Analyzer. 2008 [cit. 2009-02-26]. Dostupný z WWW: <http://flex.sourceforge.net/>.

[4] ANTLR Parser Generator. [cit. 2009-02-26]. Dostupný z WWW: <http://www.antlr.org/>.