FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

odb. as. Ing. Kamil Ekštein, Ph.D.
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd Západočeské univerzity
Univerzitní 22, CZ-30614 Plzeň
Tel.: (+420) 377 63 2406
E-mail: kekstein@kiv.zcu.cz

Posudek oponenta diplomové práce

Bc. Oto Šťáva
Automatický překladač jazyka Blason

Diplomová práce Bc. Oto Šťávy se zabývá strojovým překladem zdrojového textu tzv. *blasonu*, což je v heraldice užívaný formální slovní popis vzhledu znaku či erbu, který lze současně použít i pro popis vzhledu vlajek, pečetí atp. Zápis blasonu by měl dovolit zkušenému herladikovi zrekonstruovat podobu znaku (erbu) bez významnějších odchylek od originálu a zároveň různí heraldici by při interpretaci jednoho konkrétního blasonu měli dospět k (téměř) identickým podobám znaků (erbů). Z tohoto pohledu se tedy jedná o v podstatě formální jazyk, jehož gramatiku lze zachytit v Backus-Naurově formě a pro který lze zkonstruovat překladač. O to se také diplomant pokusil a předložená diplomová práce dokumentuje postup jeho prací a také konečný výsledek jeho snahy.

Problém, který diplomant v práci řešil, je bezesporu velmi komplexní, neboť jazyk blasonů je sice formalizovaný, ale stále vystavěný na prakticky neomezené podmnožině nějakého přirozeného etnického jazyka, což z něj činí v zásadě kontextovou gramatiku typu 1 dle Chomského klasifikace, a to ještě velice rozsáhlou. Je tedy zřejmé, že výsledkem nemůže být (a není) nástroj, který by problém řešil v celé jeho komplexnosti, což ale nelze autorovi vyčíst. Naopak, předložená práce je za daných okolností mimořádně dobře odvedená, zejména demonstrované programátorské dovednosti autora hodnotím jako vynikající.

Autorem předložené dílo překvapuje (v porovnání s jinými pracemi tohoho druhu z poslední doby, které měl oponent možnost hodnotit) mimořádně zralým, promyšleným a precizně provedným programátorským řešením. Nejen samotný způsob, styl a úprava zápisu zdrojového kódu, ale i propracovaná koncepce architektury software s řadou efektivních postupů a chytrých nápadů, je na zcela profesionální úrovni a naplňuje oponenta optimismem, že nemusíme ještě nad úrovní absolventů lámat hůl.

Ke zdrojovému kódu díla tedy nelze mít závažnější připomínky, snad jen několik drobných poznámek: (i) umístění zdrojových souborů `.c`/`.h` do podsložky `lib` není zcela obvyklé, oponent by je spíše očekával v `src`; (ii) pro zamezení kruhových či vícenásobných vložení hlavičkových souborů příkazem preprocesoru `#include` používá autor méně bezpečnou variantu s příkazem `#pragma once`; (iii) v některých modulech lze občas narazit na „magická čísla“, tj. konstanty zapsané literálem přímo na místo výskytu v kódu (např. řada výrazů blasonu ve funkci `cbl_token_keyword` `cbl_morpho_tokenizer_word_keyword(·)`), což by šlo vyřešit elegantněji např. inkluzí hlavičkového souboru s definicemi těchto konstant; atp. Zdrojový text je středního rozsahu (5 883 řádek kódu) a autorovi se podařilo udržet jeho vzornou konzistenci.

Software jde bez problémů přeložit a spustit, struktura úložiště zdrojového kódu a komponent je jasná, přehledná a zejména odpovídá zvyklostem. Autor naneštěstí zapomněl v té části dokumentu, která pojednává o sestavení a spuštění programového vybavení, mezi potřebnými nástroji zmínit sestavovací systém *Ninja*, bez kterého software sestavit nelze, neboť při dodržení autorových pokynů generuje *CMake* sestavovací předpis `.ninja` specificky pro tento nástroj.

Po spuštění pracuje software dle očekávání a s omezeními, která jsou dobře popsána v průvodním dokumentu – zejména jde o fakt, že autorem navržená gramatika pro zpracování blasonů nepokrývá (ani dost dobře nemůže) všechny přípustné varianty popisu znaků, a tedy software předložený vstupní blason vyhodnotí jen zčásti. K rychlosti a paměťové náročnosti zpracování nelze mít připomínek.

Předložený průvodní dokument programového díla je na vynikající úrovni. Je napsán výbornou, dobře čitelnou technickou češtinou s minimem chyb. Je dobře logicky sturkturovaný a jednotlivé části rozumně pokrývají předmětné oblasti podmiňujících teoretických poznatků a programového díla samotného. Typografická úroveň je dobrá, text práce je vysázen v LaTeXu a působí harmonicky a přehledně. I přes použití LaTeXu se v dokumentu vyskytují místy drobné odchylky od typografických zvyklostí (např. velmi rušivé odsazení odstavců na začátcích kapitol/sekcí), ale nejedná se o nijak dramatické prohřešky. Také lze nalézt pár překlepů a prohřešků proti jazyku českému („defaultní barvy“), ovšem s ohledem na rozsah textu (71 str. vč. příloh) je jejich množství zanedbatelné.

U výpisů zdrojového kódu chybí čísla řádek, takže není možné se na konkrétní úseky kódu jednoznačně odkazovat – to je docela mrzuté. V práci lze také nalézt tvrzení, se kterými není dost dobře možné se zcela ztotožnit: Na str. 15 „... kdy se ukázalo, že [gramatika popsaná BNF] není zcela odpovídajícím vodítkem pro ruční přepis na analyzátor rekurzivním sestupem“. Autor následně „vynalezl“ lepší formu zápisu gramatiky, čímž znovu objevil kolo již dříve

nazvané „rozšířená Backus-Naurova forma“ (EBNF). Je zřejmé, že autor neprostudoval některé vhodné knihy, např. J.-P. Tremblay, P. Sorenson: The Theory and Practice of Compiler Writing, kde nejenže je perfektně popsaná EBNF, ale také je tam uveden zcela jednoznačný deterministický postup, jak jedno každé pravidlo gramatiky převést na jednu konkrétní proceduru prediktivního LL(k) parseru rekurzivním sestupem.

Autor si také postup řešení zbytečně zkomplikoval tím, že nevyužil standardní mechanismy syntaktické analýzy rekurzivním sestup v podobě akceptačních funkcí `expect(·)` a `match(·)`, ovšem jeho řešení – byť složitější – zase dobře dokumentuje jeho programátorské dovednosti.

Také tvrzení v úvodu práce na str. 8 „Ani velmi pokročilá neuronová síť pravděpodobně nebude nikdy efektivnějším a zároveň přesným řešením logicky jasně vyhraněného problému než vhodně zvolený deterministický algoritmus určený přímo k jeho řešení.“ lze v mnoha případech snadno zpochybnit a ač to tak může na první pohled vypadat, jistě nemá obecnou platnost.

V práci lze místy narazit na výrazy, které nejsou zcela běžné a mohou být zavádějící, ale frekvence jejich výskytu je velmi nízká a kvalitu práce zásadněji negativně neovlivňují.

Autor při řešení vycházel z celkem 21 zdrojů, mezi nimiž postrádám již zmíněnou knihu J.-P. Tremblaye a P. Sorensona nebo nějakou podobně dobrou knihu zabývající se konstrukcí syntaktických analyzátorů. Návrh jakéhokoliv komplexního parseru (což ten pro blason nepochybně je) je teoreticky celkem náročná disciplína, a proto nepřítomnost vlastně jakékoliv literatury na toto téma v přehledu použitých zdrojů považuji za docela překvapivou a alarmující.

Autorem uvedené zdroje se týkají především heraldiky (což je na místě), dále formátů různých deskriptivních souborů a dat (BNF, SVG, Markdown, Doxygen) a potom použitých knihoven *utf8proc* a *MorphoDiTa*, přičemž k poslední uvedené se vážou ještě zdroje popisující morfologii českého textu a její zpracování. Až na zmíněný nedostatek literatury k formálním jazykům a jejich algoritmické analýze lze výběr zdrojů považovat za vhodný a jejich objem na dolní hranici dostatečnosti.

Citace v textu i bibliografie na konci dokumentu splňují nutné podmínky na ně kladené, umožňují jednoznačnou identifikaci zdroje díky přítomnosti všech povinných identifikátorů, a tedy neporušují práva jejich autorů. Nicméně při současném stavu vývoje různých pomocných nástrojů v této oblasti by bylo vhodnější doplnit i některé chybějící identifikátory nepovinné, které by nemělo být zas tak těžké dohledat (např. místo vydání), příp. prostě zaznamenat (datum a čas citace online zdroje). Tak, jak je bibliografie provedena, je hodně minimalistická.

Body zadání 1 a 4 až 6 byly nepochybně splněny v celém rozsahu. Body 2 a 3 byly s největší pravděpodobností splněny také, avšak části textu průvodního dokumentu, které se k nim vztahují, jsou relativně malého rozsahu. Např. k bodu 3 „Seznamte se s problematikou generování vektorových obrázků a formáty pro jejich ukládání“ není uvedena žádná podstatná teorie (jen velmi povrchní všeobecný popis základních principů vektorové grafiky) a zmíněn je jen formát SVG, který je „popsán“ jen třemi krátkými odstavci, což při rozsahu a možnostech tohoto formátu rozhodně nelze považovat za dostatečné. Oponent ovšem nemůže vyloučit, že se autor s vektorovými formáty skutečně v potřebném rozsahu seznámil, ovšem v textu práce to explicitně neuvedl.

Prakticky totéž se týká bodu 2 „Seznamte se s problematikou formálních jazyků a s nástroji pro tvorbu vlastního překladače“. Dvě stránky celkem povrchního popisu obecných skutečností týkajících se gramatik formálních jazyků v práci sice jsou, ovšem chybí zde některé velmi důležité skutečnosti, které s řešeným tématem přímo souvisejí, např. Chomského klasifikace gramatik, která je zásadní proto, aby adaptace přirozené gramatiky blasonu splňovala podmínky, které vyžaduje např. autorem zvolená technika syntaktické analýzy rekurzivním sestupem. Soudě pak podle textu práce, s nástroji pro tvorbu vlastního překladače se autor neseznámil vůbec (očekával bych přinejmenším zmínku o Lex/Yacc, byť by asi jejich použití nebylo vhodné, a dále pak o alespoň některém z generátorů RDP, např. ParserTongue, Parsley, pgen, rdp, . . .).

Výše uvedené výtky nesnižují hodnotu práce jako vynikajícího programátorského díla, které obsahuje množství invence a užití řady profesionálních pokročilých technik. Autor bezesporu prokázal schopnost řešit inženýrským způsobem zadané problémy. Proto práci i přes uvedené drobné pochybnosti týkající se úplného splnění všech bodů zadání rozhodně **doporučuji k obhajobě** a navrhuji hodnocení (dle výkonu u obhajoby) klasifikačním stupněm

„velmi dobře“ nebo i **„výborně“**.

Ing. Kamil Ekštein, Ph.D.
KIV FAV ZČU

V Plzni dne 26. července 2023

Doplňující otázky:

1. Pokoušel jste se nalézt/využít nějaký generátor syntaktického analyzátoru rekurzivním sestupem nebo jste od zahájení práce na řešení počítal s tím, že budete parser psát ručně? Co Vás případě vedlo k zavržení myšlenky využít generátor, který by možná dovolil zpracovávat rozsáhlejší gramatiku a umožnil tak zpracovat i komplexnější blasony?