Recenzja rozprawy doktorskiej Pawła Skórzewskiego "Wydajne algorytmy parsowania dla języków o szyku swobodnym"

Stanisław Szpakowicz

Analiza składniowa tekstów w językach naturalnych o szyku swobodnym to temat fascynujący, więc trochę szkoda, że rozprawa pana Skórzewskiego poświęca szykowi swobodnemu *per se* stosunkowo mniej miejsca niż można by oczekiwać po jej tytule. Rozprawa wpisuje się w program naukowy jednostki macierzystej Autora. Powstaje pytanie, na ile ta praca jest znaczącym wkładem w ów program, a na ile przyczynkiem. Moim zdaniem, niepodważalne osiągnięcie *naukowe* to definicja PTgBG,[1] z twierdzeniami i dowodami, chociaż jest to w istocie naturalne rozszerzenie istniejącego już formalizmu TgBG.[2]

Inne elementy kontekstu naukowego tej rozprawy to zestaw narzędzi informatycznych PSI-Toolkit, przeznaczony do przetwarzania tekstów w językach naturalnych, i system tłumaczenia maszynowego Translatica. Autor opisuje starannie re-implementację i usprawnienie parsera wyjętego z systemu Translatica tak, aby działał on jako moduł w zestawie PSI-Toolkit. Usprawnienia dokonano dzięki zastosowaniu poprawnych technik inżynierii oprogramowania, w rozprawie doktorskiej poświęconej przetwarzaniu języków naturalnych jest to jednak osiągnięcie bardziej konstrukcyjne niż naukowe.

Rozdział 4 i rozdział 6 rozprawy przedstawiają zatem dokonania *autorskie*, to pierwsze o wyraźnie większej wadze. Rozdział 1 i rozdział 7 sprawiają zawód: są bardzo krótkie, zwłaszcza podsumowanie, w którym ogromnie mi brakuje choć kilku uwag o konsekwencjach tych prac i o kontynuacji opisanych tu poczynań.

Zamysłem pozostałych rozdziałów było, jak rozumiem, zaprezentowanie tła prac Autora nad gramatykami i parserami. Prezentacja ta jest nierówna i nie wszędzie widać, jakim celom naukowym ma ona służyć w tej pracy. Doceniam konieczność wprowadzenia niezbędnych pojęć, ale wiele z tych pojęć powinno być dobrze znane przeciętnemu absolwentowi studiów informatycznych, które uczą podstawowych faktów o językach formalnych i podstaw sztucznej inteligencji.

Rozdział 2 opiera się przede wszystkim na podręczniku Hopcrofta i Ullmana, ale nie podaje wszystkich niezbędnych faktów, w tym dowodów, po które Autor odsyła czytelnika do owego podręcznika.

Przegląd formalizmów opisu języków o szyku swobodnym w rozdziale 3 wprowadza kilka systemów notacyjnych, z których tylko TgBG odgrywa rolę w dalszych rozważaniach. Otwierająca rozdział lista typów nieciągłości w języku polskim (s. 22) jest długa, ale nie jest bynajmniej jasne, czy obejmuje *wszystkie* typy.[3] Gramatyki ID/LP to chyba zawężenie, a nie *rozszerzenie* nieuporządkowanych gramatyk bezkontekstowych ☺. Niezależnie od zalet teoretycznych obu tych klas gramatyk, ich sens praktyczny jest wątpliwy, skoro posługiwanie się nimi to problem NP-zupełny. Zastanawia brak cytowań literatury językoznawczej w punkcie 3.2.1; pozycje 31, 3 i 21 są techniczne, nie teoretyczne.[4]

Podrozdział 3.3 opiera się na pracy z pisma *Journal of Functional Programming*, która nie ma wiele wspólnego z przetwarzaniem języków naturalnych.[5] Nie jest to jednak ważne, bo formalizm GF do niczego Autorowi w jego pracy badawczej nie służy. To samo można powiedzieć o formalizmie FROG pokazanym w podrozdziale 3.4, chociaż te gramatyki są akurat opisane starannie, a nie – jak poprzednie gramatyki – każda na jednym małym przykładzie. Definicja 3.5 jest częściowo niepoprawna: w trzech ostatnich wzorach na wartość funkcji δ drugi argument to nie 0, tylko $\rightarrow$.

[1] probabilistyczne gramatyki binarne generujące drzewa

[2] gramatyki binarne generujące drzewa

[3] Nie bardzo widać, po co zacytowano pozycję 29, bo nieciągłości w języku ukraińskim i w języku nowogreckim nie są w żaden sposób istotne dla tej rozprawy.

[4] *Nota bene*, praca Sandry Kübler w ogóle nie wspomina *raising*.

[5] Z drugiej strony, przydałyby się cytowania o gramatykach kategorialnych i o kompilatorach kompilatorów.

Istotny dla niniejszej rozprawy jest dopiero opis formalizmu TgBG.[6] Notacja jest nieco myląca: operacji rozszerzenia odpowiada symbol *ext* (równanie 3.5), ale też symbol *xt* (s. 41); ta niejednoznaczność przewija się przez resztę rozprawy. Przy okazji: ciekaw jestem, dlaczego nieokreślony wynik operacji na drzewach oznacza się (s. 36) symbolem nieskończoności ∞.

Nie bardzo rozumiem, dlaczego Autor w ogóle nie wspomina o systemie Świgra. Warto odwiedzić strony internetowe `http://nlp.ipipan.waw.pl/~wolinski/swigra/` i w szczególności stronę "Przykłady wyników analizy", gdzie można między innymi znaleźć zdania o szyku swobodnym. Świgra, system oparty na bardzo szczegółowej składnikowej gramatyce formalnej zbudowanej przez Marka Świdzińskiego, jest parserem polszczyzny *par excellence* głębokim, a szyk składników rozpoznaje dowolny. Co więcej, jest on w pełni dostępny i wydaje się, że podlega stałej konserwacji.[7]

Przegląd algorytmów parsowania w rozdziale 5 przedstawia w miarę szczegółowo klasyczny algorytm CYK,[8] bardzo szczegółowo adaptację do analizy składniowej klasycznego algorytmu wyszukiwania heurystycznego A* i pobieżnie kilka innych metod. Algorytm A* i przykład jego zastosowania zajmują ponad 7 stron, czyli mniej więcej 1/14 pracy (nie licząc dodatków i bibliografii).[9] Taka objętość dziwi, bo A* w ogóle się nie przydaje w pracach opisanych w następnych rozdziałach.

Powróćmy do rozdziału 4, który przedstawia najistotniejsze, moim zdaniem, dokonanie tej rozprawy. Gramatyki PTgBG to naturalne (i dlatego przewidywalne ☺) rozszerzenie formalizmu TgBG. Sporo miejsca w tym rozdziale zajmują dowody. Są one dokładne i technicznie poprawne, ale z jednej strony raczej nieuchronne (standardowa indukcja matematyczna albo strukturalna), a drugiej strony niekiedy nużąco długie i słabo czytelne, bez komentarzy przy wielu nie całkiem ewidentnych przejściach. Nie bardzo wiadomo, kto jest adresatem tych dowodów, bo mało który czytelnik zechce poświęcić im tyle uwagi, ile ich pełne zrozumienie wymaga.

Relacja wyprowadzalności $\vdash_G$ została zdefiniowana prawidłowo, ale dziwi mnie fakt, że Autor nie zdefiniował przedtem w punkcie 3.4.2 jej odpowiednika w gramatyce TgBG. Definicja 4.7 powiada, że funkcja prawdopodobieństwa $\mathbf{P}_G$ ma wartości w zbiorze liczb rzeczywistych: dlaczego nie po prostu w przedziale [0, 1]? W tabeli 4.2 reguła w postaci ogólnej ma sens tylko dla $n > 2$. Ta sama tabela poucza, że reguły "pośrednie" dla x_1 mają prawdopodobieństwo 1. Takie właśnie jedynki sumuje się w drugim wierszu formuły 4.14 z innymi prawdopodobieństwami, jak zatem całość może mieć wartość 1?

Lemat 4.2 lepiej by chyba było umieścić w środku dowodu twierdzenia 4.1. Poza tym pożytecznie byłoby wskazać, w którym miejscu w wywodzie na s. 49 się ten lemat przydaje. Widać je dobrze, ale odrobina "łopatologii" nie zawadzi. Lemat – formuła 4.22 – opisuje pewien fakt. Bardzo brakuje zwięzłego, intuicyjnego wyjaśnienia tego faktu.

Przykład 4.3 pokazuje regułę, która "może zostać zastosowana nieskończoną liczbę razy". Co to ma wspólnego z nieograniczoną liczbą poddrzew? Zazwyczaj każde zastosowanie reguły schodzi na niższy poziom w drzewie. Należało pewnie wyjaśnić, że chodzi o specjalną operację dołączania na tym samym poziomie. Poza tym w zwykłej gramatyce bezkontekstowej takie reguły też wydają się zupełnie możliwe.

Twierdzenia 4.3 i 4.4 zostały podane bez dowodów. Nawet jeżeli są to dowody oczywiste, warto to właśnie powiedzieć. Ograniczenie sformułowane w twierdzeniu 4.5 należało uzasadnić, bo nie jest ono oczywiste. Nie od rzeczy byłoby też podać interpretację lematu 4.6.

[6] Żal, że nie podano tu adresów internetowych kluczowych pozycji 14 i 15, aby czytelnik mógł sobie te prace bez trudu ściągnąć.

[7] Instytut Podstaw Informatyki PAN prowadzi też prace nad zależnościową głęboką analizą składniową języka polskiego.

[8] Cocke-Younger-Kasami

[9] Przykład analizy syntaktycznej, jak wszystkie inne takie przykłady w tej pracy, opiera się na *bardzo* prostym zdaniu polskim. Jest to zapewne dydaktycznie uzasadnione, ale przykłady są nie całkiem pouczające.

Przejdę teraz do omówienia rozdziału 6. Chociaż jest on napisany klarownie i przekonująco, nie zmienia to faktu, że – jak już to zauważyłem – Autor przedstawia tu operacje wymagające umiejętności programistycznych znacznie bardziej niż talentu badawczego. Pomysł trenowania gramatyk PTgBG na danych z anglojęzycznego korpusu BNC wydaje się nieco nie na miejscu w pracy nad językami o szyku swobodnym, zważywszy, że angielski takim językiem nie jest.[10] Rozumiem, że Autorowi chodziło o sprawdzenie pewnej hipotezy, ale zastanawiam się, dlaczego nie można było sięgnąć na przykład do Narodowego Korpusu Języka Polskiego. Swoją drogą, cel wyrażony na s. 78 jest zastanawiająco mało ambitny: "znalezienie takiej gramatyki probabilistycznej, która da wyniki parsowania *możliwie zbliżone* [moja kursywa] do wyników parsowania uzyskanych przy użyciu gramatyki wyjściowej".

Konstrukcja gramatyki, intuicyjnie prosta, opiera się na "typowych" zdaniach, ale typowość ta nie została zdefiniowana. Prawdopodobieństwa obliczane są też w prosty sposób, według wzoru 6.1, bez jakiejkolwiek próby wygładzania (*smoothing*). Eksperyment porównawczy przeprowadzono bardzo okrężną drogą, poprzez tłumaczenie maszynowe, proces znany z nienajlepszej dokładności; jakość wyników "oceniana była przez człowieka".[11]

Wyniki eksperymentu podane w tabeli 6.2 rozczarowują. Mniejsza już o błąd w gramatyce, który nie pozwala poprawnie tłumaczyć zdań angielskich z przyimkiem *of*.[12] Przypisywanie wag losowo daje w praktyce ten sam wynik co metoda, którą proponuje Autor, wbrew stwierdzeniu na s. 81: "Zgodnie z oczekiwaniami, zdania zostały lepiej przetłumaczone przy użyciu gramatyki probabilistycznej niż przy użyciu losowych wag." Różnica jest statystycznie najzupełniej nieznacząca. Zastanawia też zdanie: "Gdyby jako korpusu treningowego użyć banku drzew poprawnie oznaczonego, przygotowanego lub choćby sprawdzonego ręcznie, to wyniki byłyby zapewne lepsze." Co stało na przeszkodzie takim właśnie czynnościom?

Podrozdział 6.4 przedstawia operację przeniesienia parsera z systemu Translatica do zestawu narzędzi PSI-Toolkit. Opis jest rzetelny i w miarę pouczający, jeżeli ktoś chciałby się zapoznać z trudnościami, jakich nastręcza "wycięcie" pojedynczego modułu ze zintegrowanego systemu. Podrozdział 6.5 omawia usprawnienie tak uzyskanego modułu, a czyni to znowu w sposób rzetelny i ciekawy dla tych, którzy takich usprawnień w swoich programach potrzebują, chociaż na pewno przydałaby się im głębsza znajomość kodu parsera w języku C++ i w szczególności struktur danych.

Pozostały mi do omówienia dodatki. Pierwszy z nich pokazuje drzewa składniowe trzech krótkich zdań polskich. Opisy nie są przesadnie pomocne, bo brakuje wyjaśnienia symboli Z, FC, FR, C, R, P i ZRo. Drugi dodatek opisuje najpierw sposób instalowania zestawu PSI-Toolkit w kilku wersjach systemu operacyjnego Linux.[13] Instrukcja jest skomplikowana, a instalacja czasochłonna, toteż mała jest szansa, że potencjalny użytkownik bez solidnego przygotowania technicznego potrafiłby z niej skorzystać. Ponownie zatem nie całkiem wiadomo, kto jest adresatem tej pracy. Następnie dowiadujemy się, jak korzystać z zestawu PSI-Toolkit. W teorii dodatek B pozwala cierpliwemu czytelnikowi zapoznać się z działaniem systemu. Szukałem w całej rozprawie informacji, która by mi pozwoliła ocenić konkretne wyniki w tej rozprawie przedstawiane — bez sukcesu. Nie ma na przykład adresu internetowego, pod którym mógłbym znaleźć dane testowe albo pełne wyniki eksperymentów. Krótko mówiąc, mogę tylko przyjąć na wiarę wszystko, co Autor postanowił zawrzeć w swojej rozprawie.

*

[10] Odnotuję przy okazji, że potrzeba cytowań o BNC na s. 77 i systemie CLAWS na s. 78.

[11] Nie będę się rozwodził nad zasadami przeprowadzania i oceniania doświadczeń z udziałem adnotatorów. Są one daleko bardziej złożone niż to, o czym pisze Autor.

[12] Wyraz *of* pojawia się na znanych mi angielskich listach frekwencyjnych na miejscu drugim, trzecim albo czwartym.

[13] Nie liczyłbym na wersję działającą w systemie Windows, a zresztą nie potrzebowałbym takiej wersji ☺, ale szkoda, że Autor nie rozważył instalacji w systemie Unix na komputerach Macintosh.

Rozprawa została napisana językiem jasnym, choć – co zapewne nieuniknione ☺ – miejscami lekko żargonowym. Trochę za dużo jest anglicyzmów (na przykład ewaluować → oceniać; interwał → przedział; wolumen → tom; framework → ? [s. 29, s. 92]) i nieprzetłumaczonych terminów angielskich (na przykład *raising* → *podnoszenie*; kilka terminów na s. 58; if → jeżeli [s. 43, s. 44]). Redakcja jest staranna, a w całej pracy znalazłem dosłownie kilka potknięć.[14]

Bibliografia jest adekwatna, ale nie do końca zadowalająca, niekompletna (wyliczam w recenzji szereg miejsc, gdzie potrzebne są cytowania) i niekiedy lekko przestarzała. Szkoda, że prawie nie ma w niej adresów internetowych wersji cytowanych prac w PDF albo w innym popularnym formacie; takie adresy są nadzwyczaj przydatne. Znalazłem też kilka usterek.[15]

Strona formalna rozprawy stosuje się, jak się domyślam, do lokalnego wzorca. Osobiście wolałbym zobaczyć krótkie streszczenie na początku, a zaraz potem listy rysunków i tabel. Brakuje mi listy oznaczeń i krótkiego słownika terminów.

*

Ustawa o stopniach naukowych i tytule naukowym mówi: "Rozprawa doktorska [...] powinna stanowić oryginalne rozwiązanie problemu naukowego [...] oraz wykazywać ogólną wiedzę teoretyczną kandydata w danej dyscyplinie naukowej [...] oraz umiejętność samodzielnego prowadzenia pracy naukowej [...]". Niniejsza rozprawa rozwiązuje jeden, w miarę oryginalny, problem naukowy; ukazuje dobre zrozumienie szeregu aspektów przetwarzania języków naturalnych; i niczym nie wskazuje, aby Autor nie mógł prowadzić samodzielnej pracy naukowej. Moja recenzja wylicza pewne zastrzeżenia, ale uważam, że w ostatecznym rachunku rozprawa spełnia jednak ustawowe wymogi. Wnioskuję zatem o dopuszczenie do publicznej obrony.

*

W pozostałej części recenzji znajdują się uwagi o pomniejszych usterkach. Wskazuję też drobne błędy i niekonsekwencje.

Dominacja języków analitycznych (s. 7) – dominacja w jakim sensie?

Oznaczenie V = { S, N, V } (s. 12) jest lekko mylące.

We wszystkich formułach i wywodach dłuższych niż jeden wiersz ostatni symbol – zwykle operator – powtarza się w następnym wierszu (pierwszy taki wypadek pojawia się na s. 13). Nie bardzo rozumiem sens tej konwencji.

Definicje 2.20 i 2.21 są nieścisłe. Język kontekstowy generuje pewna gramatyka kontekstowa, ale żadna gramatyka bezkontekstowa. Język bezkontekstowy generuje pewna gramatyka bezkontekstowa, ale żadna gramatyka regularna.

Nieścisłe są też definicje 2.23 i 2.24. Notacja (c, (e_1, t_1), ..., (e_n, t_n))) jest oczywista, ale powinno się ją jednak formalnie wprowadzić. Symbol $\hat{\mathcal{T}}$ jest niezdefiniowany, a indeks k bierze się nie wiadomo skąd.

Co to jest "nieistotna rola składniowa"? Nieistotna?

W definicji 2.27 występuje termin *łuk* zamiast terminu *krawędź*.

Skąd wiadomo, że *prawdopodobieństwo łańcucha* (definicja 2.33) jest w istocie prawdopodobieństwem?

Czy można powiedzieć (s. 27), że formalizmy nie są *w pełni* równoważne? Równoważność albo zachodzi, albo nie zachodzi.

Na rysunku 3.3 pojawiają się gwiazdki. Dobrze by było tę notację wyjaśnić.

[14] języka Format → Format [s. 26]; sugerują → sugeruje [s. 35]; stwierdzenia → twierdzenia [s. 46]; literówka (k **jeżeli**) i niedopasowane czcionki w wyrazie "wskaźnikami" [s. 65]; został przez → został wprowadzony (?) przez [s. 73]; to kraty → do kraty [s. 83]; mięć → mieć [s. 86]

[15] Pozycja 4: brak nazwiska redaktora. Pozycja 5: błąd (polecam `http://publications.uvt.nl/repository/harry.bunt/publications.html`). Pozycja 8: "i in."? Pozycja 10: bez wielkich liter? Pozycja 11: brakuje nazwy szkoły. Pozycja 23: dwie literówki. Pozycja 37 jest niekompletna. Pozycja 66: literówka.

Czy aspekty mogą przyświecać (s. 29)?

Wyraz *w* jest przyimkiem, nie przedimkiem (s. 33)!

Definicja 3.7 przywołuje oznaczenia $\mathcal{T}$(T, V, $\mathcal{R}$) i σ. Odesłałbym po ich definicje kilkanaście stron wstecz, do rozdziału 2.

Praca Mielczuka i Polguerre'a [34] wspomina tłumaczenie maszynowe raz i w przelocie, czyli cytowanie na s. 35 jest niewłaściwe. *À propos*, praca ukazała się w roku 1987, nie 1984.

Co to jest *płaska* reprezentacja (s. 35)?

Operacja *li* dołącza drzewo t_1 do *jedynego* poddrzewa t_2, ale ten warunek nie wynika z równań 3.10 i 3.11.

Cytowanie książki (s. 40) bez podania numerów stron nie ułatwia czytelnikowi życia.

Co oznacza symbol =>= na s. 46?

Jeżeli gramatyka "generuje identyczne języki bądź daje takie same prawdopodobieństwa" (s. 46), to czy jest to alternatywa wyłączna?

W dowodach przez indukcję lepiej nie pisać "przypuśćmy", bo przecież nie chodzi o *przypuszczenie* indukcyjne.

Przywoływanie na s. 60 przykładu sprzed 40 stron nie jest wygodne. Powtórzyłbym go, tym bardziej, że nie jest duży. Konwencje w przykładzie 5.1 warto by wyjaśnić: pogrubianie, kursywa, ukośnik?

Pojęcie hipergrafu z wagami (s. 62) nieźle by było pokrótce zdefiniować i przy okazji wyjaśnić, dlaczego wagi krawędzi to odwrotności logarytmów prawdopodobieństw. Na s. 64 dobrze by było zdefiniować, też zwięźle, pojęcie multigrafu.

Pojęcie agendy pojawia się na s. 63 bez uprzedniej definicji.

Metoda *beam search* obniża średni koszt wyszukiwania, ale czy naprawdę zmniejsza złożoność?

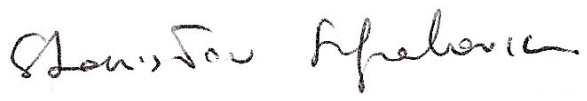
Lepiej mówić o maksymalizacji wartości funkcji niż o maksymalizacji funkcji (s. 64).

Tłumaczenie *weighted constraint dependency grammar* jako "ograniczone gramatyki zależnościowe z wagami" jest niedokładne.[16]

"Konwersja X do Y" brzmi gorzej niż "przekształcenie X w Y" (s. 76). "Uczenie gramatyk" brzmi gorzej niż "trenowanie gramatyk" (s. 77). Przy okazji: każdy wie, co to są dane treningowe, ale dla kompletności powinno się chociaż wspomnieć, że w pracy będzie się teraz wykorzystywać metody uczenia maszynowego i że stąd bierze się potrzeba takich właśnie danych.

W krótkim akapicie na środku s. 82 (System wspiera ...) brakuje kilku cytowań.

Licencja LGPL (s. 82) to wersja, do której gnu.org silnie zniechęca.



7 maja 2014

[16] System WCDG traktuje analizę języka jako optymalizację sterowaną wymuszanymi warunkami (*constraint optimization*). Nie znaczy to, że gramatyka i język przez nią generowany są ograniczone.