Przedmowa

Cele, zakres i adresat pracy

Powstaniu niniejszej pracy przyświecały dwa cele.

Pierwszy cel to przybliżenie Czytelnikowi polskiemu tematyki przetwarzania tekstów oraz omówienie różnorodnych wykorzystywanych w praktyce technik składniowego przetwarzania powierzchniowego. O ile mi wiadomo, nie istnieją obecnie żadne prace w języku polskim, w których tematyka ta byłaby szeroko omawiana, a i w podręcznikach anglojęzycznych niewiele miejsca poświęca się parsowaniu powierzchniowemu, mimo jego szerokich praktycznych zastosowań. W publikacji niniejszej usiłuję tę lukę choćby częściowo wypełnić.

Drugi cel pracy to zaproponowanie nowego formalizmu przetwarzania powierzchniowego, o nazwie Spejd (skracanej do ♠). Formalizm ten ma kilka nowatorskich cech w porównaniu do wcześniej zaproponowanych rozwiązań, a ponadto powstała — i została publicznie udostępniona na zasadach oprogramowania swobodnego (ang. *Open Source*) — jego pełna implementacja. Formalizm Spejd i jego implementacja zostały sprawdzone w praktyce: stworzona została średniej wielkości powierzchniowa gramatyka języka polskiego, a jej implementacja została z powodzeniem wykorzystana w projekcie automatycznego wydobywania informacji lingwistycznych z dużych zbiorów tekstów.

Oprócz powierzchniowego przetwarzania składniowego, praca omawia także techniki tzw. dezambiguacji morfosyntaktycznej, a więc zadania choćby częściowego ujednoznacznienia interpretacji słów: czy na przykład słowo *piec* to w danym kontekście forma czasownikowa, czy rzeczownikowa, a jeżeli rzeczownikowa, to czy jest to mianownik, czy biernik? Jedną z nowatorskich cech formalizmu Spejd jest „bezszwowe" połączenie funkcjonalności parsera składniowego i regułowego systemu ujednoznaczniania morfosyntaktycznego.

Niniejsza publikacja adresowana jest do osób zainteresowanych automatycznym przetwarzaniem tekstów: zarówno do informatyków, jak i do lingwistów. Z jednej strony starałem się więc unikać w niej żargonu informatycznego i bardzo oszczędnie wprowadzać symbole i wzory matematyczne, z drugiej zaś — nie zakładam wcześniejszej wiedzy lingwistycznej, a terminy

lingwistyczne ilustruję przykładami. Obu grupom Czytelników pomóc w lekturze mają także skorowidze.

Część pierwsza książki pisana była z myślą o możliwym wykorzystaniu w wykładzie z przetwarzania języka naturalnego, na przykład jako uzupełnienie niedawno wydanego polskiego podręcznika Mykowiecka 2007 lub podręczników anglojęzycznych Manning i Schütze 1999, Jurafsky i Martin 2000, 2008 czy Nugues 2006. Formalizm przedstawiony w części drugiej, którego implementacja dostępna jest pod adresem `http://nlp.ipipan.waw.pl/Spejd/`, może natomiast posłużyć jako podstawa praktycznych prac w ramach zajęć laboratoryjnych lub projektów zaliczeniowych[1].

Mam na koniec nadzieję, że niniejsza publikacja przyczyni się do zwiększenia zainteresowania zagadnieniem przetwarzania tekstów polskich. Służyć ma temu nie tylko rozdział opisujący zasoby przydatne w przetwarzaniu tekstów polskich, ale także rozdział prezentujący konkretne propozycje reguł dla języka polskiego zapisanych w formalizmie Spejd. Ponieważ dostępne są zarówno teksty w formacie zakładanym przez obecną implementację tego formalizmu, jak i narzędzia do automatycznego znakowania tekstów taki format produkujące, nic więc nie stoi na przeszkodzie, by Czytelnik sam stworzył system powierzchniowego przetwarzania tekstów polskich!

Układ pracy

Niniejsza praca podzielona jest na dwie główne części, po których następują dodatki, bibliografia i skorowidze: tematyczny, terminów angielskich, skrótów i skrótowców, oraz nazwisk.

Pierwsza część pracy ma charakter przeglądowy. Wstęp zawiera ogólne wprowadzenie do dziedziny przetwarzania języka naturalnego oraz bardziej szczegółowe — do tematyki morfosyntaktycznego i syntaktycznego przetwarzania powierzchniowego, a także zawiera eksplikacje podstawowych pojęć, w tym pojęć użytych w niniejszym i następnym zdaniu. Techniki stosowane w przetwarzaniu powierzchniowym zostały szczegółowo przedstawione w dwóch kolejnych rozdziałach pracy: zadanie i metodologia ujednoznaczniania morfosyntaktycznego opisane zostały w rozdziale 2, zaś techniki składniowego przetwarzania powierzchniowego — w rozdziale 3. Pierwszą część pracy kończą: prezentacja polskich zasobów przydatnych w przetwarzaniu powierzchniowym (rozdz. 4) oraz opis wybranych narzędzi do przetwarzania tekstów (rozdz. 5).

Część druga niniejszej pracy zawiera natomiast materiał oryginalny. W pierwszym rozdziale tej części (rozdz. 6) opisuję główne cele przyświecają-

[1] Spejd jest wykorzystywany w ramach wykładu i laboratorium *Inżynieria lingwistyczna — konstrukcje*, prowadzonych przeze mnie w Instytucie Informatyki Uniwersytetu Warszawskiego.

ce stworzeniu systemu do powierzchniowego przetwarzania języka polskiego. Rozdział 7 poświęcony jest prezentacji formalizmu służącego do opisu związków składniowych, pozwalającego m.in. na znajdowanie w zdaniach polskich prostych konstrukcji składniowych. Co więcej, reguły zapisane w tym formalizmie mogą także posłużyć do eliminacji pewnych interpretacji morfosyntaktycznych w wypadku wyrazów o wielu interpretacjach, podobnie jak ma to miejsce w wypadku typowych regułowych dezambiguatorów morfosyntaktycznych. Kolejny rozdział, 8, prezentuje gramatykę powierzchniową dla języka polskiego zapisaną w formalizmie z rozdziału 7, wykorzystaną w zadaniu automatycznego wydobywania z dużych zbiorów odpowiednio znakowanych tekstów, przede wszystkim z Korpusu IPI PAN, informacji walencyjnych, a więc informacji o składniowych wymaganiach czasowników. Samo zadanie wydobywania informacji walencyjnych z korpusów i wyniki prac walencyjnych wykorzystujących niniejszą gramatykę zostaną bardziej szczegółowo omówione w rozdziale 10, wcześniej zaś, w rozdziale 9, pokażę natomiast, jak korzystać z informacji syntaktycznych przy przeszukiwaniu Korpusu IPI PAN, oraz jak obrazować częściowe reprezentacje syntaktyczne. Ostatni — nie licząc czterech dodatków — rozdział niniejszej monografii zawiera podsumowanie i przedstawia dalsze możliwe kierunki prac.

Istotną częścią publikacji jest obszerna bibliografia, która ma ułatwić Czytelnikowi polskiemu dotarcie do najważniejszych prac dotyczących różnorodnych aspektów przetwarzania powierzchniowego tekstów. W licznych odnośnikach bibliograficznych staram się trzymać następującej konwencji: w odwołaniach do prac rok podany jest bez nawiasów, zaś gdy mowa jest o autorach prac, rok jest pomijany lub podawany w nawiasach. Ponadto często pojawiają się w tekście apozycje typu „w *pracy Kowalska 2007*" czy „na podstawie *artykułu Kowalski 2007*" — niejako wbrew tradycji, ale zgodnie z chyba coraz częstszym uzusem[2].

Pracę kończą skorowidze: tematyczny, terminów angielskich, skrótów i skrótowców, oraz nazwisk. Ten ostatni obejmuje wszystkie cytowania prac współautorstwa danej osoby, nawet jeżeli w odnośniku bibliograficznym na stronie wskazywanej w skorowidzu nazwisko nie jest *explicite* wymienione (jak to ma miejsce w wypadku nieinicjalnych autorów w odnośnikach do prac wielu autorów).

Podziękowania

Powstanie niniejszej pracy było możliwe dzięki finansowaniu Komitetu Badań Naukowych (później Ministerstwa Nauki i Informatyzacji, następnie

[2] Takie formy odnośników bibliograficznych pojawiają się nawet w polskich artykułach językoznawczych, np.: „Konstrukcja... opisana jest... w pracy Saloni i Świdziński (1998: 128)" w artykule Miechowicz-Mathiasen i Witkoś 2007, str. 109.

Ministerstwa Edukacji i Nauki, a obecnie Ministerstwa Nauki i Szkolnictwa Wyższego) w ramach projektu numer 3 T11C 003 28 *Automatyczna ekstrakcja wiedzy lingwistycznej z dużego korpusu języka polskiego*, kierowanego przeze mnie od marca 2005 do marca 2008 w Instytucie Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN), a także dzięki projektom współpracy dwustronnej pomiędzy IPI PAN i Uniwersytetem w Ratyzbonie, Niemcy (projekt współfinansowany przez DAAD, realizowany w latach 2006–2007) oraz pomiędzy IPI PAN i CNRS w Nancy, Francja (projekt Polonium, realizowany w latach 2006–2007).

Chciałbym także gorąco podziękować Claire Grover za udostępnienie wersji beta narzędzi LT XML2 i LT TTT2, Kirilowi Simowovi za cierpliwe odpowiedzi na pytania dotyczące działania systemu CLaRK, Markowi Świdzińskiemu za udostępnienie słownika walencyjnego (Świdziński 1998), zaś Rolandowi Meyerowi między innymi za pomoc w zorganizowaniu dwóch miesięcznych pobytów w Ratyzbonie, dzięki którym mogłem się oderwać od codziennych obowiązków administracyjno-projektowych i skupić na pracy naukowej. Za pomoc wykraczającą poza zwykłe bibliotekarskie obowiązki dziękuję serdecznie Paniom Alicji Aloksie, Annie Bittner, Lidii Miernickiej, Annie Monkiewicz i Hannie Małgorzacie Werner z Biblioteki IPI PAN, a także Bibliotece Instytutu Języka Polskiego Uniwersytetu Warszawskiego i Bibliotece Uniwersytetu Warszawskiego (BUW).

Za szczegółowe uwagi do wcześniejszych wersji całej pracy pięknie dziękuję Elżbiecie Hajnicz i Marcinowi Wolińskiemu. Za komentarze do fragmentów niniejszej monografii winien jestem wdzięczność także Aleksandrowi Buczyńskiemu, Łukaszowi Dębowskiemu, Danielowi Janusowi, Agnieszce Mykowieckiej, Maciejowi Piaseckiemu, Jakubowi Piskorskiemu i Stanisławowi Szpakowiczowi. Praca zyskała także dzięki wnikliwym uwagom recenzenta wydawniczego, Krzysztofa Jassema. Oczywiście tylko ja ponoszę odpowiedzialność za wszelkie pozostałe niedoskonałości pracy.

Strona WWW

Z monografią niniejszą związana jest strona `http://nlp.ipipan.waw.pl/PPJP/`, na której będą sukcesywnie udostępniane kolejne wersje zasobów i narzędzi powstałych w ramach wspomnianego projektu MNiSW *Automatyczna ekstrakcja wiedzy lingwistycznej z dużego korpusu języka polskiego.*