Biuletyn
Polonistyczny

03.08.2018

Rozmowy „Biuletynu Polonistycznego” - KorBa

Przygotowany i zrealizowany przez Pracownię Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego PAN we współpracy z Zespołem Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN projekt „Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)” miał swoją uroczystą premierę 18 czerwca 2018 r. Najważniejszym rezultatem projektu jest udostępniony w Internecie Elektroniczny Korpus Tekstów Polskich XVII i XVIII w. (do 1772 r.), w skrócie nazywany Korpusem Barokowym. Od tej ostatniej nazwy urobiony został akronim KorBa.

Projekt był finansowany ze środków Narodowego Programu Rozwoju Humanistyki na lata 2013-2018.

Z panem profesorem Włodzimierzem Gruszczyńskim, kierownikiem projektu, oraz panią doktor Renatą Bronikowską, koordynatorką projektu, rozmawiamy o „wczytywaczu”, „tagerach”, i „transkryberach” – a także innych narzędziach humanistyki cyfrowej.

Biuletyn Polonistyczny: Czym jest korpus KorBa, po co powstał i dlaczego?

Włodzimierz Gruszczyński: Trzeba wyjaśnić dwie rzeczy. Po pierwsze, czym jest korpus językowy w ogóle, a po drugie, czym jest korpus tekstów historycznych XVII–XVIII-wiecznych. Najprościej mówiąc, korpus tekstów, nieprecyzyjnie nazywany korpusem języka, to jest zbiór tekstów, dziś najczęściej w postaci elektronicznej. Ale zwyczajowo przyjęło się, przede wszystkim wśród językoznawców, zajmujących się korpusami, że to jest taki zbiór tekstów, które są dobrze opisane w sensie metadanych i bardzo dobrze oznakowane pod jakimiś względami. Najczęściej dla takiego języka jak język polski poszczególne słowa czy segmenty są oznakowane cechami gramatycznymi. Dzieje się tak dlatego, że jeśli ktoś w korpusie tekstów angielskich chce znaleźć słowo „dog”, to wpisuje sobie „dog” i co najwyżej musi jeszcze dodać „dogs”, wtedy ma wszystkie „dogi” i „dogsy” i nie ma żadnej wątpliwości, że wpisał właściwe frazy, żeby znaleźć wszystkie rzeczowniki o znaczeniu „pies”. W naszym korpusie trzeba było do form wszystkich wyrazów dopisywać tzw. tagi, zawierajace informację o cechach gramatycznych danej formy i – przede wszystkim – o jej przynależności do leksemu. Jeśli wszystkie słowa w tekstach są w ten sposób oznakowane, czyli – mówiąc slangowo – otagowane, to wtedy można zadać – za pomocą odpowiedniej komendy – pytanie o leksem, a w

odpowiedzi dostaje się wszystkie wystąpienia wszystkich form na przykład rzeczownika „pies”. To jest w skrócie odpowiedź na pierwszą część pytania – taki jest w zasadzie każdy korpus tekstów polskich.

Oczywiście w każdym cywilizowanym kraju, gdzie istnieje językoznawstwo, takie korpusy powstają. Najpierw – rzecz jasna – zaczęły powstawać korpusy tekstów współczesnych. W Polsce powstał Narodowy Korpus Języka Polskiego (NKJP), po wielu latach prac, po różnych dyskusjach środowiskowych. NKJP ma oczywiście swoje wady i zalety, jak wszystkie tego typu zasoby. W pewnym momencie pojawiła się myśl, że może warto stworzyć korpus tekstów dawnych, który mógłby być wykorzystywany do badań historycznojęzykowych. Nasz zespół stanął przed takim zadaniem, ponieważ na co dzień tworzymy Słownik języka polskiego XVII i XVII wieku na podstawie kartoteki, która – choć jest tworzona od lat 50. XX wieku i ma ok. 2 800 000 fiszek, to – wbrew pozorom – jest za mała (siostrzany słownik opracowywany w Instytucie Badań Literackich – Słownik polszczyzny XVI wieku powstaje na podstawie kartoteki niemal ośmiomilionowej). Nasze prawie 3 mln to – jak powiedziałem – za mało. Robienie w dzisiejszych czasach kolejnych kartotek papierowych, byłoby wyrzucaniem pieniędzy i marnowaniem czasu. W związku z tym postanowiliśmy zrobić korpus tekstów XVII- i XVIII-wiecznych. Oczywiście współczesne korpusy muszą być wyposażone w narzędzia, które pozwolą wyszukiwać to, czego potrzebujemy. Przy tworzeniu korpusu tekstów współczesnych jest dużo pracy, dużo problemów teoretycznych, ale przy korpusie tekstów historycznych jest ich dużo, dużo więcej. Na szczęście, jest o tyle łatwiej, że przetarte są szlaki, wiadomo, do czego można się odwołać – mam tu na myśli przede wszystkim osiągnięcia zespołu, który zbudował NKJP, z którego członkami często współpracowaliśmy. Trzeba przyznać, że mieliśmy trochę mniej pod górkę niż ci, którzy robili tamten korpus.

BP: Można zatem powiedzieć, że szeroko pojmowane narzędzie cyfrowe, które Państwo wykorzystujecie, z jednej strony służy tworzeniu korpusu, z drugiej – służy użytkownikom tego narzędzia?

Renata Bronikowska: Przy tworzeniu korpusu wykorzystywanych było wiele narzędzi. To, co teraz widzą użytkownicy, to jest wyszukiwarka. Natomiast „w środku” był cały proces tagowania, wspomagany bardzo mocno przez różnorodne narzędzia informatyczne.

WG: Może warto przedstawić, krok po kroku, w jaki sposób powstawał korpus.

BP: Bardzo proszę, to bardzo ciekawe.

RB: Najpierw były teksty. Teksty, które dla nas były szczególnie ciekawe, to przede wszystkim starodruki. One są dostępne w postaci skanów w bibliotekach cyfrowych. Ale postać skanów to nie jest to, czego oczekujemy, to nie jest przydatna dla nas postać docelowa. Zatem te wszystkie teksty musiały zostać przepisane. Oczywiście nie wszystkie teksty w korpusie, bo ten zawiera nie tylko starodruki, ale też wydania współczesne tekstów XVII-wiecznych. Te ostatnie były

dostępne w postaci elektronicznej, wystarczyło je tylko „zoceerować” i zrobić korektę.

WG: Tylko! (śmiech)

RB: Jednak większość tekstów (ok. dwie trzecie) to są starodruki, jest też trochę rękopisów i one musiały być przepisane ręcznie przez naszą grupę „skryptorów”.

BP: Czy Państwo w jakiś sposób ujednolicali teksty, które miały funkcjonować w brzmieniu docelowym?

WG: Teksty były transliterowane, z dużą dokładnością, jednak znaczenie mniejszą niż na potrzeby Słownika polszczyzny XVI wieku, bo pomijaliśmy np. różne warianty liter „s” czy „r”, od razu rozwiązywaliśmy ligatury. Nie jest to zatem tak, jak w wydawnictwach typu „A” według znanej starej książki Woronczaka et consortes (mowa o Zasadach wydawania tekstów staropolskich. Projekt – przyp. red.).

RB: Choć oczywiście przepisywaliśmy bardzo dokładnie, i na wspomnianej książce także się opieraliśmy.

WG: Oczywiście. Ale jednak, jeżeli jakaś różnica graficzna systematycznie pojawiająca się nie odzwierciedlała żadnej cechy fonetycznej albo gramatycznej, to wydawało nam się, że możemy sobie pozwolić na jej pominięcie. I pozwoliliśmy sobie. Natomiast wszystkie cechy istotne lingwistycznie zostały zachowane.

RB: Pytał Pan zapewne o transkrypcję – ta warstwa transkrypcyjna również się w korpusie pojawia, to jest kolejny etap. Ale zanim do tego doszło, trzeba powiedzieć, że równolegle z przepisywaniem teksty były wstępnie znakowane. Znakowana była struktura tekstu i fragmenty obcojęzyczne. Dzięki temu możemy teraz zadać w wyszukiwarce pytanie, w którym odrzucimy obce segmenty i nie będziemy mieli tego „śmiecia” w postaci zwrotów obcojęzycznych.

WG: A latynista może sobie wyszukać XVII-wieczne wystąpienia jakiejś formy łacińskiej, i z kolei on nie dostanie „śmiecia”, którym dla niego będą zwroty polskojęzyczne.

RB: Na etapie przepisywania były też znakowane numery stron, co jest istotne, bo teraz, gdy wyszukiwarka podaje nam wyniki, wiemy, na której stronie starodruku ten tekst się znajduje. To wszystko było robione w odpowiednio dostosowanym szablonie dokumentu „Word”, zatem żadne „wielkie” narzędzie nie zostało zaangażowane (śmiech). Ale kolejny etap to było przekształcenie formatu „doc” w wersję „xml”. To jest to, co potem wszystkie narzędzia informatyczne czytają i wiedzą, co z tym robić. Do tego etapu zostało stworzone narzędzie zwane „wczytywaczem”.

WG: „Wczytywacz” był ciekawy z jeszcze jednego powodu – był pomocny z punktu widzenia organizacji pracy. Redaktor, czyli ktoś z naszego zespołu, przydzielał skryptorowi tekst, tekst otrzymywał we „wczytywaczu” swój identyfikator i podstawowe metadane, następnie skryptor tekst przepisywał. Kiedy skryptor

skończył przepisywanie, próbował wczytać automatycznie do bazy danych, a „wczytywacz” za pierwszym razem bardzo często mówił „nie”. Bo np. skryptor wprowadził znacznik początku strony, ale zabrakło znacznika końca strony, jest początek przypisu, a nie ma końca przypisu itd. „Wczytywacz” sprawdzał zatem materiał pod kątem poprawności oznakowania.

RB: Na tym etapie wpisywaliśmy również dokładniejsze metadane. Było bardzo dużo informacji, które staraliśmy się zawrzeć. Nie tylko takie, jak autor, tytuł, data wydania, data powstania, ale też różne informacje dotyczące stylu, gatunku, tematu, tego, czy jest to mowa wierszowana czy niewierszowana, czy jest to tekst serio czy buffo, tego, jaki region dany tekst reprezentuje. Słowem – wszystko, co uznaliśmy, że może być później przydatne przy wyszukiwaniu.

WG: Dzięki temu można w tej chwili zadać wyszukiwarce pytanie o jakiś typ form, np. rzeczowniki kończące się na „anie” w mianowniku, występujące tylko w tekstach powstałych w latach 1601–1610, z obszaru Mazowsza, napisanych wyłącznie w konwencji serio.

RB: To może służyć na przykład badaniu języka danego autora, języka danego stylu czy gatunku.

WG: Dzięki znacznikowi „serio”/ „nie serio” można w zasadzie wyszukać wszystkie testy rybałtowskie, ponieważ w tamtym okresie to głównie one mają prześmiewczy charakter i często zawierają naprawdę bardzo dziwne formy językowe. Wróćmy jednak do etapu prac, w którym mamy tekst w „xmlu”…

RB: Tak, a to jest dopiero początek! Kolejnym dużym etapem było oznakowanie półmilionowego korpusu. Ta liczba odnosi się do segmentów, a segment to jest mniej więcej słowo tekstowe, od spacji do spacji (przyjęliśmy rozumienie segmentu za NKJP). Ręczne oznakowanie części korpusu miało służyć temu, by można było wytrenować tagery, czyli narzędzia do automatycznego znakowania. Ręczne znakowanie wymagało zaangażowania wielu ludzi. Najpierw szef (profesor Włodzimierz Gruszczyński – przyp. red.) stworzył tagset czyli zestaw znaczników…

WG: Znaczników gramatycznych, przede wszystkim fleksyjnych, uwzględniający swoisty podział na części mowy, bo nie mógł być to podział szkolny. Staraliśmy się być jak najbliżej systemu NKJP, żeby kiedyś można wyszukiwać za pomocą tej samej wyszukiwarki, żeby to nie były dwa zupełnie różne światy. Oczywiście XVII-wieczny język polski i język współczesny różnią się dość istotnie. Trzeba było na przykład wprowadzić liczbę podwójną, nieoczekiwane dla inżynierów lingwistycznych (którzy naprawdę świetnie znają gramatykę) stopniowanie imiesłowów przymiotnikowych…

RB: Na przykład „bolątszy”, „pachniątszy”…

WG: Czyli taki, który bardziej pachnie… I tak powstał tagset, czyli system znaczników gramatycznych, którymi trzeba było oznakować wspomniany półmilionowy korpus. Trudno sobie wyobrazić, żeby tak zwani anotatorzy, czyli ci,

którzy przypisują charakterystykę fleksyjną czy morfosyntaktyczną poszczególnym słowom, mieli przed sobą tylko „goły tekst”. Liczba błędów byłaby wówczas niewiarygodna. Zwykle robi się tak, że w pierwszej kolejności wykorzystywany jest program komputerowy, który nazywany jest automatycznym analizatorem morfologicznym. My wykorzystywaliśmy analizator, który nazywa się pięknie MORFEUSZ. Trzeba jednak pamiętać, że on produkuje wszystkie możliwe interpretacje, działa bezkontekstowo. Jeśli zatem mamy segment „miał” to może to być rzeczownik rodzaju męskiego („miał węglowy”), jak również forma czasownika „mieć”. Chcieliśmy, aby powstała „postarzona” wersja MORFEUSZA, czyli taka, która uwzględni nieistniejące dzisiaj słowa i formy, trzeba było je wprowadzić do analizatora i uwzględnić kategorie gramatyczne, których dzisiaj nie ma. To zajmowało dużo czasu. Tutaj przydał się nasz niedokończony Słownik języka polskiego XVII i XVIII wieku. Zespół na gwałt wprowadzał do niego nowe hasła, zawierające na razie wyłącznie formy fleksyjne. Chodziło o to, by tych form było jak najwięcej, szczególnie tych wyrazów, co do których intuicja podpowiadała, że nie ma ich we współczesnym języku polskim. Trochę to pomogło, ale nadal MORFEUSZ nie rozpoznawał wielu form, więc nasi koledzy z Instytutu Podstaw Informatyki PAN próbowali na podstawie tego, co mamy w Słowniku, wygenerować pewne formy automatycznie. Na przykład słowo „komputer” – jest we współczesnym języku polskim, jest zatem w MORFEUSZU. Jeśli mielibyśmy się zdecydować, że „komputer” ma zostać w analizatorze „postarzonym”, co nie jest takie do końca bezsensowne…

RB: Albo słowo „kursor”…

WG: Tak! „Kursor” istniał w XVII wieku. Ale trzeba mu było dotworzyć formy liczby podwójnej – „dwa kursora” itd. Odbywało się to taką metodą, że częściowo było pobierane z naszego słownika, a częściowo odtwarzane. I co potem? Potem było losowanie. Z pełnego korpusu, który liczy około 13 milionów segmentów, trzeba było wylosować próbki o łącznej długości pół miliona segmentów, ale takie, które byłyby reprezentatywne dla całego korpusu. Te próbki zostały przeanalizowane przez analizator morfologiczny „postarzony”, który nazwaliśmy KORBEUSZEM.

RB: Jeszcze jedna ważna rzecz – wcześniej był etap transkrypcji, ponieważ KORBEUSZ nie bardzo poradziłby sobie z transliteracją (był tworzony – jak wspomnieliśmy – na bazie MORFEUSZA, który jest analizatorem współczesnym), nie rozpoznałby np. „a” kreskowanego. Dlatego słowa musiały być doprowadzone do w miarę współczesnej postaci. To działo się półautomatycznie. Skorzystaliśmy z narzędzia stworzonego na potrzeby innego projektu, które udoskonaliliśmy. Zostały dopisane formuły, za pomocą których określone sekwencje liter zostały zamienione na inne. Najprostszy przypadek – każde „a” kreskowane zostało zamienione na „a” zwykłe.

WG: Największe problemy były oczywiście z „y” i „i”…

RB: No tak, bo czasami powinno być zamienione na „j”, czasem na „i”, czasem na „yj”. Chodziło o określenie kontekstu, w którym te zamiany miały następować. Tych

reguł jest tam ponad 1000. Dopiero próbki tekstów przekonwertowane na postać transkrypcji były anotowane i tak powstał korpus treningowy. Na tej podstawie zostały wytrenowane dwa tagery i całość korpusu została oznakowana automatycznie.

WG: Za pomocą dwóch tagerów.

RB: Właśnie, dlatego mamy dwie interpretacje gramatyczne, które czasem się od siebie różnią. Czasem jeden tager się myli, czasem drugi. Czasem można przyjąć, że są to dwie możliwe interpretacje – to może być dla bardziej zaawansowanych użytkowników korpusu ważna informacja.

WG: Oczywiście my tutaj mówimy o interpretacji gramatycznej, ale najważniejsze dla zwykłego użytkownika naszej KorBy jest to, że tager lematyzuje, czyli daną formę traktuje jako formę konkretnego leksemu, daje hasło. Dzięki temu możemy wyszukiwać wszystkie formy danego leksemu za jednym zamachem, nie wypisując literalnie każdej formy osobno. To jest bardzo wygodne, ważne, ale zaawansowany użytkownik musi mieć świadomość tego, o czym koleżanka powiedziała, że lematyzacja została dokonana automatycznie, a więc są tam błędy, jest ich naprawdę dużo.

RB: Błędy wynikają z niedoróbek na różnych etapach powstawania korpusu. To może być błąd w przepisywaniu, błąd w transkrypcji…

WG: A nawet wcześniej – błąd mógł być też w starodruku…

RB: Oczywiście, ale tu już nie ponosimy odpowiedzialności (śmiech).

WG: Poniekąd! Bo skryptor miał prawo poprawić, oczywiście w nawiasach ostrych, ale nie zawsze poprawiał…

RB: Przyjmijmy, że to błąd na etapie przepisywania. Potem – błąd w transkrypcji, bo mogła być reguła, która „nie złapała” jakiegoś elementu, jakiejś zamiany, która powinna nastąpić, albo złapała za dużo, nie ma wyjątku, który by ją ograniczył. Potem znowu – przy ręcznej anotacji. Anotator też człowiek i mógł się pomylić, a przez to tager źle się nauczył.

WG: A tager nie jest bytem doskonałym, zwłaszcza jeżeli ma sobie radzić z rzeczami dawnymi. Sprawność dobrego tagera w odniesieniu do tekstów polskich współczesnych dochodzi do 90% poprawnych rozstrzygnięć, oczywiście przy tekście XVII- czy XVIII-wiecznym jest dużo mniej. Zakładaliśmy na początku projektu, że to będzie tylko i wyłącznie eksperyment i że będziemy zadowoleni, jeśli poprawnie rozpoznanych form będzie powyżej 50%, a w tej chwili to jest około 80%. Eksperyment na tyle się powiódł, że to, co jest wystawione w Internecie, jest otagowane. Wiemy, że jeśli szukamy wystąpień określonego leksemu, to nie możemy wierzyć w 100%, że dostaniemy po naszym zapytaniu tylko to, o co pytaliśmy, i jednocześnie wszystko, o co pytaliśmy – to jest absolutnie niemożliwe. Jest to jednak pierwsze przybliżenie, i jeśli dostaniemy w 70–80% to, czego

szukamy, to wydaje się, że gra jest warta świeczki. Narzędzia cały czas się ulepszają – jeden z tych tagerów, które były zaprzęgnięte do pracy w naszym projekcie jest zupełnie nowy, to jest tager najnowszej generacji pracujący na tak zwanych sieciach neuronowych i jest cały czas mocno ulepszany przez jego twórczynię, Katarzynę Krasnowską-Kieraś. W najbliższych dniach korpus zostanie ponownie przez ten tager „przepuszczony" – będzie nowa wersja KorBy – ulepszona. Z jednej strony twórczyni tagera dojrzała jakiś błąd, z drugiej strony pytano nas, czy nie widzimy problemów, które można przy okazji rozwiązać. Dosłownie dziś wysłaliśmy kilka sugestii. Skróty, które były ogromnym wyzwaniem, będą dużo lepiej obsługiwane przez tager i użytkownik będzie dostawał mniej głupich rozstrzygnięć.

BP: Jak długo trwało powstawanie koncepcji i wdrażanie przyjętych rozwiązań?

WG: Koncepcja powstawała tak długo, jak długo powstawał wniosek (o przyznanie finansowania – przyp. red.). Trwało to kilka miesięcy, trzeba powiedzieć, że to nie była rzecz robiona zupełnie od zera, ponieważ – jak powiedziałem – korpusy współczesnego języka są, narzędzia są, wiedza na temat tego, jak ludzie na świecie robią korpusy, również jest duża. Niemniej od samego początku było wiadomo, że dużym problemem będzie na przykład ustalenie tagsetu, że nie da się tak łatwo dostosować tagsetu NKJP do naszych potrzeb. Na konferencji, którą mieliśmy w Pałacu Staszica, byli nasi koledzy ze Słownika polszczyzny XVI wieku – oni są w tej chwili krok za nami. Dla nich najistotniejsze było pytanie o to, jak radzimy sobie z tagsetem, a przede wszystkim z rodzajem rzeczowników – to jest klucz, bo to jest kategoria, która najbardziej się zmieniała w ciągu ostatnich pięciuset lat w systemie gramatycznym języka polskiego i nie za bardzo wiadomo, jak sobie z nią poradzić.

RB: Realizacja to pięć lat projektu (tyle trwał grant Narodowego Programu Rozwoju Humanistyki).

WG: Nie przedłużyliśmy nawet o miesiąc, co było dla nas zaskoczeniem. Inna sprawa, że granty NPRH są raczej trudne do przedłużenia…

BP: Komu korpus będzie służył przede wszystkim?

RB: Przede wszystkim będzie służył nam (śmiech).

WG: Będą z niego korzystać leksykografowie i historycy dawnej polszczyzny. Ale myślę, że nasi koledzy z Instytutu Badań Literackich również, bo dopóki nie ma kompletnego słownika polszczyzny XVII wieku, to właśnie KorBa może dać jakąś podpowiedź tym, którzy badają lub wydają XVII-wieczny tekst literacki i wpadają na „minę" – słowo, które pierwszy raz widzą na oczy, a którego nie ma ani w słowniku XVI wieku, ani w słowniku „warszawskim", ani u Lindego. Co prawda w KorBie nie będzie ono objaśnione, ale jeżeli trafią się ze cztery konteksty i to od dwóch autorów, to na ich podstawie wydawca, edytor, specjalista, który pracuje z tekstami średniopolskimi, na pewno sobie poradzi – te nowe konteksty mu pomogą.

Na pewno KorBa służyć będzie również historykom języka, którzy zajmują się rekonstrukcją systemu gramatycznego…

RB: Rozwojem słownictwa, zmianami znaczeniowymi. Korpus daje możliwości prześledzenia tych zmian, to jest coś bezcennego. Także badania nad językiem pisarzy. Możemy ograniczyć wyszukiwanie do jednego autora, możemy zatem badać konteksty, w jakich używał danego słowa czy konstrukcji.

BP: Czy są plany rozwoju korpusu, kontynuacji prac?

RB: Jeśli tylko uda nam się zdobyć kolejny grant.

BP: Z jakimi kłopotami trzeba będzie się zmierzyć, jakich narzędzi będą Państwo potrzebować, co trzeba zmienić, ulepszyć?

RB: Kontynuacja będzie polegać na różnych rzeczach. Chcemy powiększyć nasz zasób do końca XVIII wieku, wejdą zatem teksty zupełnie nowe, z nowej epoki literackiej.

WG: I to jest ogromne wyzwanie. Nie powiem, że wybór tekstów XVII- i XVIII wiecznych był łatwy, ale jednak nasza wiedza i wiedza naszych poprzedników, dokumentacja Słownika sprawiały, że było znacznie łatwiej. Jeśli dojdzie do tego, że będziemy poszerzać zasób o ostatnie ćwierćwiecze XVIII wieku, to z pewnością będziemy potrzebowali konsultacji z ludźmi, którzy znają się na tekstach oświeceniowych. Nie chodzi tu tylko o teksty literackie. Chcemy mieć pełen przekrój, reprezentację całej produkcji wydawniczej czy nawet rękopiśmiennej – podręczniki medycyny, rolnictwa, hippiki, fizyki, muzyki itd., pamiętniki, przewodniki po obcych miastach, publicystykę.

RB: Będą to teksty z nowej epoki. Być może trzeba będzie dostosować tagset, być może zmienić transkryber.

WG: Komisja Edukacji Narodowej przyczyniła się do tego, że z jednej strony było ujednolicenie pisowni, z drugiej z końcem XVIII wieku zaczęło się pojawiać na przykład „e” kreskowane. Będzie kolejny kłopot i wyzwanie. Ale największym wyzwaniem będzie konieczność zrobienia korekty dotychczasowego materiału. Wiemy, które teksty nie do końca zostały skorygowane, jest sporo błędów literowych. Nie wystarczyło nam czasu i pieniędzy, ale to jest ból każdego wielkiego korpusu, zwłaszcza takiego, który jest przepisywany. Największą zmianą będzie próba połączenia dwóch zasobów: KorBy ze słownikiem. Naszą ambicją jest to, by redaktor hasła w słowniku, który zaczyna robić nowe hasło, nie musiał wszystkiego przepisywać ręcznie. Będziemy chcieli automatycznie pozyskiwać z korpusu wszystkie formy gramatyczne wyrazów wraz z ich charakterystyką. Skoro forma jest rozpoznana, wystarczy to sprawdzić. Druga rzecz to pozyskiwanie automatyczne cytatów. W tej chwili jesteśmy na etapie „kopiuj-wklej” – to już jest genialne, ale jeszcze lepiej byłoby, gdyby automat przerzucał przykłady, które są w KorBie, a redaktor tylko dzielił je na znaczenia, dokładał definicję i – hasło prawie gotowe. Na świecie tak się już robi, ale oczywiście diabeł tkwi w szczegółach –

najwięcej takich narzędzi jest do języka angielskiego, który nie jest fleksyjny. Niemniej wzory są, są też informatycy, z którymi współpracujemy od lat.

BP: Projekt jest realizowany w Państwa pracowni, natomiast przedstawicieli jakich innych dyscyplin doprosili Państwo do prac?

WG: Informatyków przede wszystkim. W projekcie bierze udział instytucja partnerska – to jest Instytut Podstaw Informatyki PAN (IPI PAN), a dokładnie Zespół Inżynierii Lingwistycznej. To są ludzie, z którymi współpracujemy od lat. Ja jestem polonistą-językoznawcą od lat przeszło czterdziestu i od początku współpracuję z informatykami, więc dobrze znam tych, którzy rozumieją te problemy i którzy są nimi zainteresowani. Ciekawostką jest to, że w czasie kiedy my robiliśmy KorBę, w IPI PAN jeden z kolegów, Marcin Woliński, dostał finansowanie grantu na stworzenie systemu CHRONOFLEKS, czyli systemu, który będzie pokazywał rozwój fleksyjny jakiegoś wyrazu, typu, klasy wyrazów na osi czasu (projekt „Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja” – przyp. red.). Często robiliśmy łączone spotkania i duża część narzędzi powstała w koprodukcji, np. anotatornia, czyli narzędzie, za pomocą którego były ręcznie wprowadzane oznakowania fleksyjne, powstała w znacznym stopniu w ramach tamtego projektu, a w naszym projekcie rozwiązania były testowane i dostosowywane do naszego materiału i potrzeb.

RB: Poza tym dokooptowujemy licznych współpracowników do przepisywania, do anotowania. To muszą być ludzie, którzy znają gramatykę historyczną języka polskiego, a przy tym nie boją się komputerów, nowego narzędzia.

WG: Zespół był ogromny, ale przede wszystkim składał się z lingwistów i informatyków.

RB: Mieliśmy sporo współpracowników-studentów, ludzi świeżo po studiach. To dla nich z pewnością ciekawe doświadczenie – konfrontacja wiedzy teoretycznej z praktyką.

WG: Wadą projektu z pewnością było to, że mieliśmy zbyt mało pieniędzy, ponad pół miliona mniej niż wnioskowaliśmy. Wielu ludzi wykonywało bardzo ciężką pracę za stosunkowo niskie stawki. Z tego wynika również to, że nie wszystko zostało skorygowane.

BP: Dziękuję za rozmowę.

Rozmawiał: Piotr Bordzoł

prof. dr hab. Włodzimierz Gruszczyński

Absolwent Wydziału Polonistyki Uniwersytetu Warszawskiego, językoznawca, uczeń prof. Zygmunta Saloniego. Kierownik Pracowni Historii Języka Polskiego XVII i XVIII wieku w Instytucie Języka Polskiego PAN w Warszawie, której zadaniem jest opracowanie „Elektronicznego słownika języka polskiego XVII i XVIII wieku”. Pracuje też w SWPS Uniwersytecie Humanistycznospołecznym w Warszawie.

dr Renata Bronikowska

Absolwentka Wydziału Prawa i Administracji oraz Wydziału Polonistyki Uniwersytetu Warszawskiego. Adiunkt w Pracowni Historii Języka Polskiego XVII i XVIII wieku w Instytucie Języka Polskiego PAN w Warszawie. Uczestniczy w tworzeniu „Elektronicznego słownika języka polskiego XVII i XVIII w.”. Koordynatorka projektu „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)”.

Słowa kluczowe: bazy danych, korpus

Rozmówca: Włodzimierz Gruszczyński
Rozmówca: Renata Bronikowska
Rozmówca: Piotr Bordzoł