

Lingwistyka komputerowa – czy po prostu lingwistyka?

Gdyśmy publikowali (w z. 1–2/2015 „Języka Polskiego”) zbiór artykułów podsumowujących naukową refleksję nad polszczyzną w latach 1989–2014, byliśmy świadomi, że nie obejmuje on pełnego spektrum zainteresowań naszych językoznawców okresu najnowszego. Jedną z najpoważniejszych luk stała się nieobecność materiału dotyczącego lingwistyki komputerowej – dyscypliny, która, choć dostrzegalna już w okresie wcześniejszym, rozwinęła się w pełni dopiero po roku 1989, a przyczyna tego jest całkiem prosta. Dopiero wtedy technika komputerowa wydoskonaliła się i upowszechniła (zwłaszcza w Polsce) w stopniu odpowiednim, aby badania z tej dziedziny nabrały rozmachu i przestały być tylko domeną wąskiej elity mającej dostęp do niebywale kosztownego sprzętu. Brak ów uzupełniamy w niniejszym zeszycie „Języka Polskiego”, który otwiera przeglądowy artykuł Magdaleny Derwojedowej, poświęcony lingwistyce komputerowej lat minionych, a po nim zamieszczamy kilka prac opisujących badania z tego zakresu obecnie prowadzone lub niedawno ukończone. Oczywiście, nie pokazują one całości dokonań polskiej lingwistyki komputerowej i inżynierii językowej. Prace takie są dziś prowadzone w tylu ośrodkach i przez tylu uczonych, że dopiero w przyszłości może się uda je syntetycznie ująć i opisać.

Gdy myślę dziś o historii polskiej lingwistyki komputerowej, przypomina mi się jedno znamienne zdarzenie. W 2005 roku odbył się w Warszawie zjazd Polskiego Towarzystwa Językoznawczego na temat *Media elektroniczne w języku i lingwistyce*. Zamiarem jego głównego organizatora, prof. Janusza Bienia, notabene jednego z pionierów lingwistyki komputerowej w Polsce, było m.in. upowszechnienie wśród większej grupy naszych językoznawców wiedzy o możliwościach i osiągnięciach komputerowej analizy języka. Zjazd ten okazał się wszak chyba jednym z najbardziej kameralnych w historii Towarzystwa; spotkali się na nim właściwie tylko przedstawiciele środowisk lingwistów komputerowych i ich nieliczni sympatycy. Szersze kręgi badaczy najwyraźniej uznały wydarzenie za nieinteresujące dla siebie. Zarejestrowanym gościom zjazdu rozdawano płytę CD ze zdigitalizowaną wersją słownika Knapiusza w formacie djvu (który wkrótce potem stał się standardem dla rozwijających się bibliotek cyfrowych). Traf chciał, że bezpośrednio z tego zjazdu udałem się na inną konferencję, poświęconą w głównej mierze historii języka polskiego, i tam pochwaliłem się ową płytą. Zaskoczenie niektórych z obecnych diachronistów było trudne do opisania. Nie wiedzieli bowiem ani o fakcie, że taka wersja najważniejszego siedemnastowiecznego słownika z językiem polskim istnieje, ani że można ją było otrzymać gratis na owym zjeździe, ani tym bardziej, że już kilka lat temu została opublikowana na CD w innej formie i była rozprowadzana płatnie. Doskonale to odzwierciedlało ówczesną sytuację, w której lingwiści i informatycy zajmujący się komputerową analizą języka oraz pozostali językoznawcy stanowili w zasadzie dwa całkiem odrębne środowiska, słabo się komunikujące ze sobą.

Minęło od tamtego czasu dwanaście lat. Upowszechniły się biblioteki cyfrowe, elektroniczne wersje słowników, powstał Narodowy Korpus Języka Polskiego, którego roli w rozwoju

warsztatu badań nad polszczyzną nie sposób przecenić i z którego korzystają niemal wszyscy językoznawcy, a wciąż trudno się oprzeć wrażeniu, że relacje między przedstawicielami lingwistyki komputerowej czy inżynierii językowej a pozostałą częścią środowiska dalekie są od ideału. Dominujące wydają się dwie postawy. Niektórzy językoznawcy, nazwijmy ich, tradycyjni, skłonni są uważać językoznawców, nazwijmy ich, informatycznych, za kogoś w rodzaju „panów od komputera”, których zwykli wzywać na pomoc w wypadku awarii sprzętu, którym ewentualnie można zlecić pewne czynności pomocnicze w procesie badania naukowego. Najwyraźniej nie mają świadomości, że w środowisku tym powstają również prace o charakterze teoretycznym, a nawet nowe modele opisu polszczyzny, adaptujące najnowsze osiągnięcia lingwistyki światowej. Inni zaś przeciwnie, uważają ich za posiadaczy „wiedzy tajemnej”, czarnoksiężników, którzy za pomocą jakichś zaklęć wydobywają z czeluści Internetu informacje, w rzeczywistości możliwe do uzyskania przez każdego, kto wie, gdzie szukać i jaką komendę do programu wpisać. Z kolei przedstawiciele lingwistyki komputerowej wydają się czasem zapominać, że znaczna część ich pracy opiera się (ciągle) na wiedzy wypracowanej przez lingwistów, i to nieraz w czasach, kiedy o komputerze mało komu się śniło. O ileż mniej moglibyśmy znaleźć w NKJP, gdyby w zamierzchłej dla wielu z dziś żyjących lingwistów przeszłości Jan Tokarski nie zaczął prac nad systematyzacją polskiej fleksji, gdyby tych prac potem nie kontynuował Zygmunt Saloni, sam i z zespołem, przy współpracy wielu osób, co w efekcie doprowadziło do uzyskania wiedzy morfologicznej i syntaktycznej zawartej w tzw. tagsecie korpusu. Czym byłaby *Słowosieć*, i czy byłaby w ogóle, gdyby nie tezaurus Rogeta, wydany w połowie XIX wieku, a przede wszystkim *Dobór wyrazów* Romana Zawilińskiego, jeden z najbardziej niedocenionych (zresztą przez samych językoznawców i leksykografów) polskich słowników ubiegłego stulecia? W czasach mody na „łączenie otwartych danych”, na „crowdsourcing” i inne podobne działania informatycy językowi wydają się czasami nie rozumieć, jaka jest różnica między słownikiem elektronicznym skompilowanym z innych a tym opracowanym merytorycznie od początku przez językoznawców czy leksykografów, opartym na analizie aktualnego materiału językowego, że nie każdy musi się zgodzić na włączenie wypracowanej przez siebie wiedzy jako składnika, cegiełki do produktu, który firmować będzie kto inny, a właściwy autor zostanie tylko wspomniany w atrybucji bibliograficznej. Szkodzi to wszystko wzajemnemu zaufaniu.

Ponieważ jednak wszystko wskazuje na to, że jakiegokolwiek badanie języka w XXI wieku bez odwoływania się do metod czy technik związanych z lingwistyką komputerową, z drugiej strony jakiegokolwiek rozwój lingwistyki komputerowej i inżynierii językowej bez korzystania z wiedzy uzyskanej w drodze analiz czysto językoznawczych, nie będą, poza nielicznymi wyjątkami, możliwe, a w każdym razie nie będą wystarczająco skuteczne i płodne poznawczo, oba środowiska – lingwistów komputerowych i pozostałych – powinny się nastawić na współpracę i takie ułożenie wzajemnych stosunków, które by u żadnej ze stron nie wywoływało uczucia dyskomfortu. W dalszej przyszłości zaś środowiska te pewnie powinny się zintegrować w jedno. Środowisko badaczy języka – czyli językoznawców.