Andrzej Bogusławski
Uniwersytet Warszawski

# Glosa do sprawy „300 tysięcy polskich słów"

Tytanopodobna robota edytorów rejestru pt. „300 tysięcy polskich słów" mogłaby zostać rozszerzona o biliony, jeżeli nie decyliardy (lepiej: np. $10^{50}$) obiektów, które *można* nazwać polskimi słowami. M.in. można dodać do tej listy wyraz *tytanopodobny*, który jest niezaprzeczalnym faktem tekstowym od chwili, kiedy pojawił się jako pierwszy wyraz graficzny w tekście głównym niniejszej wypowiedzi.

Wydrukowanie takiego rozszerzonego rejestru jest zupełnym niepodobieństwem. Byłaby to rzecz przypominająca ewentualny rejestr wszystkich form fleksyjnych odnotowanych w takich czy innych drukach polskich wyrazów, por. np. celownik *perpendykułowi* w zdaniu *Jaś przyglądał się temu perpendykułowi.* wydrukowanym tu właśnie (wiadomo, że celownik to kategoria o szczególnie niskiej frekwencji *występowania* w realnych tekstach w roli formy rzeczowników nieożywionych, ale przecież nie taka, by miała właściwość *nieistnienia* w tej klasie rzeczowników).

Produktywnym operacjom fleksyjnym i słowotwórczym przysługuje moc niezwykła. Co nie znaczy, że przechodzimy tu na liczby nieskończone. Nie, będziemy mieli zawsze do czynienia z listą skończoną: bo właściwe *operandy* stanowią zbiór skończony, a liczba stosownych operacji jest również skończona, przy czym nieograniczona rekurencja jest wykluczona. Chodzi jedynie o to, że dotykamy tu liczb *naprawdę baaardzo wielkich.*

To samo dotyczy indywidualnych, nieregularnych produktów słowotwórczych, czyli absolutnych hapaksów. Jeżeli mamy dwuwiersz: *Kosz z poezją się otworzy, gdy twe serce się rozsroży.*, to jest jasne, że występuje tu nie analogon wyrazu *rozweseli się* (choć i w wypadku tego ostatniego wyrazu nie wchodzi w grę żadna operacja tworząca klasę otwartą), lecz analogon ros. *razljubit'* 'przestać kochać', por. też pol. *rozkrochmalić się.* A zatem jest to kompletnie nowy neologizm, który może stanąć obok mnóstwa neologizmów odnotowanych w omawianym dziele i obok którego mogą stanąć niezliczone inne. W całej wieczności będą one jednak tak czy inaczej należeć do listy skończonej.

Jeżeli wziąć pod uwagę umowy międzynarodowe, to nie można nie uznać równouprawnienia zwrotu *umowa rurytańsko-taukitiańska* (przypominam piosenkę Wysockiego o „Taukitianach" w pewnej galaktyce i dość dobrze znany z lingwistyki przykład nazwy fikcjonalnego kraju *Rurytania*) ze zwrotem *umowa polsko-węgierska*, *umowa chińsko-fińska* czy *umowa fińsko-chińsko-estońska*. I analogicznie będzie z wyrażeniami, z jednej strony, *mecz Niemcy-Holandia*, a z drugiej, *mecz Patagonia-Rurytania*. Znaczy to, że nawet w tych ograniczonych kategoriach zetkniemy się z wielkościami wielomilionowymi.

W zgodzie z tym, co powiedziałem, muszę wyrazić szczere uznanie dla następujących bardzo ważnych słów Profesora Jana Wawrzyńczyka w jego broszurze 1000 słów zadośćuczynienia (Warszawa 2016) poświęconej „300 tysiącom":

> „Zdajemy sobie sprawę, że dla części osób wertujących nasz opasły tom obecność w nim jednostek w rodzaju **polsko-angielski** [...] będzie czymś przykrym. [...] Tworów tych nie możemy pomijać, ignorować, choć nie spełniają one kryterium elementarności nakładanego na tzw. leksykalne jednostki języka w nowszych opisach polszczyzny. Rygoryzmu tych opisów nie da się pogodzić z parowiekową tradycją leksykograficzną."

Nasuwa się tu myśl o możliwości zastosowania jakichś „kwalifikatorów" leksykograficznych wskazujących na oczywistość statusu produktu dobrze ustalonej operacji przysługującego danej wokabule, brak takiego statusu lub dyskusyjność każdego z członów tej alternatywy.

Doradzałbym rozważenie wskazanej możliwości w dalszych przedsięwzięciach.

Ale zdaję sobie sprawę z tego, że odpowiednie decyzje w ogromnej liczbie wypadków byłyby nawet *skrajnie trudne* – jeżeli miałyby być przynajmniej w miarę konkluzywne. I wymagałyby one wprowadzenia do gry choćby szkicu *teorii* zajmującej się zjawiskami, z jakimi mamy tu do czynienia.

Łukasz Borchmann
Uniwersytet im. A. Mickiewicza w Poznaniu

# Słowa, których nie ma w książkach

## 1. Motywacja

Zbiór *300 tysięcy polskich słów* Jana Wawrzyńczyka i Piotra Wierzchonia (2016) jest niebagatelnym krokiem w stronę odpowiedzi na pytanie sformułowane przez pierwszego z autorów, który zapytywał w swej broszurze i wcześniejszych publikacjach, ile słów obejmuje współczesna polszczyzna – *300 tysięcy czy milion(y)*? Charakteryzując fazę trzecią rozwoju polskiej leksykografii jednojęzycznej, Wawrzyńczyk rysuje wizję ekscerpcji totalnej, nieselektywnej i zupełnej – całkowicie pozbawionej przypadkowości i niekonsekwencji w konstruowaniu hasłowników (Wawrzyńczyk, 2015).

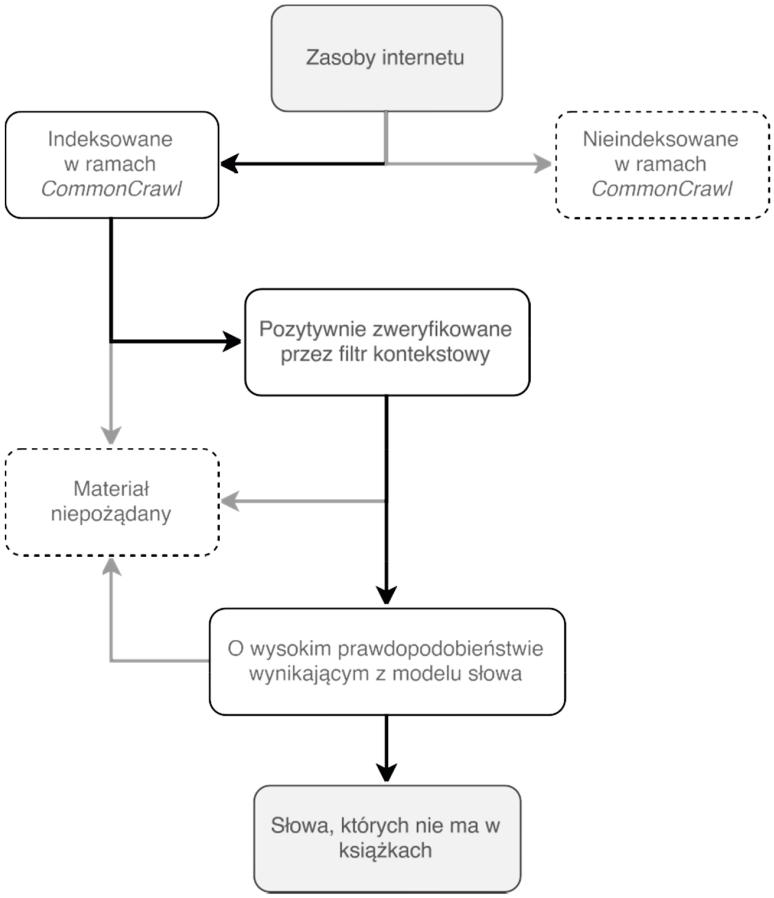
Ową ekscerpcję totalną prowadzić można z materiałów drukowanych, takich jak prasa czy książki, podążając w kierunku, w którego forpoczcie lokują się niewątpliwie autorzy *300 tysięcy*… W tym samym kierunku, ale drogą – powiedzieć by można – komplementarną, kroczyć można skupiając się na tych jednostkach, których przywoływany indeks nie obejmuje i objąć nie mógł, ponieważ istnieją w rzeczywistości tekstowej od materiałów drukowanych niezależnej.

Jako że największym niedrukowanym i dostępnym cyfrowo zasobem leksyki polskiej jest internet, zasadnym jest podjęcie analogicznego przedsięwzięcia w odniesieniu do tekstów dostępnych w sieci. Niniejszy artykuł prezentuje metodę poszukiwania na stronach internetowych jednostek języka nienotowanych przez zbiór *300 tysięcy* – jednostek, które niebezzasadnie, w uznaniu dla rozmiarów indeksu i pod wrażeniem ogromu materiałów, które poddawano ekscerpcji by go stworzyć, nazywać można *słowami, których nie ma w książkach*.

## 2. Metoda

Zmarły niedawno, wybitny polski językoznawca – Witold Mańczak – zwykł rozpoczynać swoje prace cytatem z francuskiego poety Nicolasa Boileau-Despréaux: *Ce que l'on conçoit bien s'énonce clairement*. Tej myśli czyniły zadość jego teksty, w których złożone koncepcje wyrażane były słowami prostymi i jasnymi.

Z wyżej zarysowanej motywacji, w opisie zastosowanej metody oraz jej założeń starano się wyłożyć problem tak, by był zrozumiały przez każdego, kto sięgnie po niniejszy tom, bez względu na to jakie przygotowanie filologiczne lub techniczne posiada.




**Rysunek 1.** Schemat obrazujący proces gromadzenia i filtrowania jednostek, które weszły w skład finalnej listy. Ciemniejsze strzałki prezentują ścieżkę, którą przeszły pozytywnie zweryfikowane jednostki.

## 2.1. Analizowany korpus

Od zakrojonego na szeroką skalę projektu należałoby oczekiwać jak najszerszej indeksacji zasobów polskojęzycznego internetu. Na jego początkowym etapie zasadne jest jednak ustalenie dotychczas dokonanych przedsięwzięć tego typu oraz wykorzystanie ich efektów. Jednym z takich, ponadnarodowych projektów jest *CommonCrawl* – otwarte repozytorium pozyskane na drodze indeksowania zawartości internetu.

Do niewątpliwych wad tego zbioru należy fakt, że w ogromnym (setki terabajtów danych tekstowych po skompresowaniu) archiwum *CommonCrawl*, dane polskie stanowią ledwie wycinek, który choć pokaźny może nie zaspokajać ambicji i wszystkich potrzeb badawczych. Wspomniany zasób wymaga także pewnego przetwarzania wstępnego, przed jego dalszym wykorzystaniem, na które składa się przede wszystkim wyfiltrowanie polskich stron internetowych oraz eliminacja spamu (zob. Graliński et al. (2016)).

Powyższe trudności zdaje się rekompensować łatwość dostępu do materiału, a w przekonaniu o słuszności rozważanego kroku utwierdzać mogą inne, bazujące na *CommonCrawl*, przedsięwzięcia związane z badaniem języka.

Przezwyciężywszy zarysowane problemy stanąć musimy przed kolejnym – jak odróżnić poprawne słowa w języku polskim od błędów i przypadkowych zbitek, które stanowią niebagatelny odsetek w materiale.

## 2.2. Sygnał i szum

W toku prac nad Narodowym Fotokorpusem Języka Polskiego wypracowano różne metody odróżniania poprawnych słów w języku polskim od szumu. Szczególnie obiecująca ze względu na efektywność i skuteczność jest metoda porównywania nieznanej jednostki z bazą potencjalizmów (*verba possibilia*), tj. słów, które z punktu widzenia systemu leksykalnego i derywacyjnego polszczyzny są możliwe do skonstruowania (mimo, że ich istnienie jest nieznane leksykografii).

Jedną z koncepcji wykorzystanych przy filtrowaniu zbioru pod kątem tych drugich jest automatyczne generowanie potencjalizmów. Należą do nich m.in. *composita* typu *administracyjno-biurowy* i słowa utworzone przez konkatenację słów z przedrostkami (tj. derywaty prefiksalne, zob. Wiśnicki (2010)).

Co istotne, model potencjalizmu pozwala na automatyzację tworzenia opisu gramatycznego nieznanego słownictwa, stąd znajdzie zastosowanie nie tylko w projektach leksykograficznych, ale we wszystkich przedsięwzięciach z zakresu językoznawstwa komputerowego i przetwarzania języka naturalnego, w których występuje problem słów spoza leksykonu (tzw. OOV, od ang. *out-of-vocabulary words*).