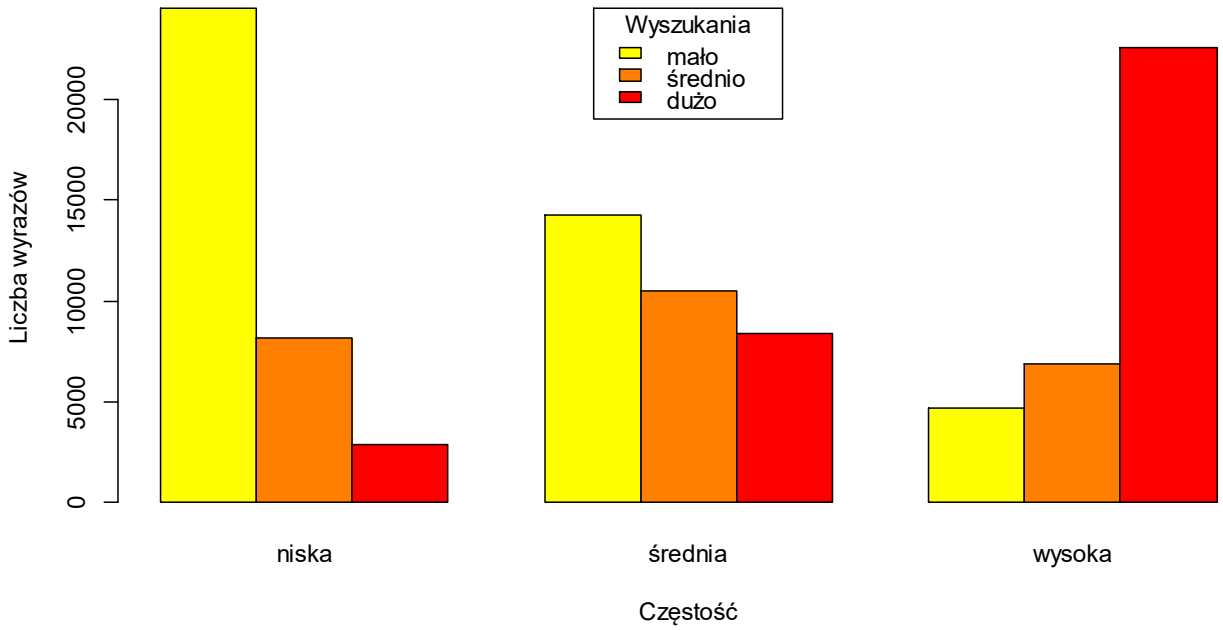
Nr rejestracyjny: 2020/39/B/HS2/00923; Kierownik projektu: prof. dr hab. Robert Adam Lew

Czynniki leksykalne wyjaśniające popularność wyrazów w angielskim słowniku Wiktionary

Stworzenie dobrego słownika jest ogromnym przedsięwzięciem, wymagającym od zespołu ekspertów (zwanych *leksykografami*) lat intensywnej pracy (a i dziesięcioleci przy dużych projektach). Wbrew potocznej opinii, liczba słów w języku nie jest zbiorem skończonym, zatem w każdym projekcie leksykografowie muszą podejmować decyzje, które słowa uwzględnić, a które pominąć.

Najbardziej użyteczny słownik zawiera słowa, których użytkownicy najczęściej poszukują. Które to są słowa? Gdybyśmy mieli zbiór danych o aktywności użytkowników słownika, to powinniśmy móc policzyć, ile razy każde słowo było sprawdzane. Następne ciekawe pytanie, to jakie właściwości słów sprawiają, że są popularne (lub nie) wśród użytkowników słownika. Co sprawia, że niektóre słowa są szczególnie ważne? Najnowsze badania sugerują, że *częstość leksykalna*, czyli to, jak często słowo pojawia się w tekstach pisanych i mówionych, jest istotnym czynnikiem; przy czym wyrazy częstsze są poszukiwane częściej. Efekt ten widoczny jest na poniższym wykresie (za De Schryver, Wolfer i Lew 2019, http://dx.doi.org/10.17576/gema-2019-1904-01):



W tym projekcie planujemy zbadać wpływ co najmniej trzech dalszych czynników:

(1) wiek, w którym dane słowo opanowuje typowe dziecko mówiące tym językiem (w żargonie naukowym nazywa się to *wiek akwizycji*);

(2) stopień, w jakim słowo jest znane dorosłym native speakerom (*rozpowszechnienie leksykalne*); oraz

(3) ile różnych znaczeń ma wyraz (czyli *polisemiczność*).

W tym celu musimy pozyskać, a następnie zestawić kilka dużych zbiorów danych leksykalnych. Po pierwsze, planujemy pobrać kompletne zapisy z serwerów angielskiego Wikisłownika i wyłowić z nich informacje o częstości wyszukiwania wszystkich haseł Wikisłownika: dzięki temu będziemy wiedzieć, które słowa użytkownicy wyszukują częściej, a które rzadziej. Przechodząc do informacji o czynnikach, które potencjalnie o tym mogą decydować, częstość leksykalną ustalimy, zliczając wystąpienia słów w bardzo dużych zbiorach tekstów angielskich (zwanych korpusami). Dalej, dobrej jakości dane dotyczące *wieku akwizycji* i *rozpowszechnienia* wyrazów angielskich są od niedawna dostępne dzięki badaniom innych specjalistów. Jeśli chodzi zaś o liczbę znaczeń, możemy policzyć je w istniejących hasłach Wikisłownika (nie ręcznie, naturalnie, jako że są ich miliony).

Wszystkie te dane należy następnie powiązać. Następnym krokiem będzie opracowanie modeli matematycznych, które będą, z największą możliwą precyzją, „zgadywać" popularność wyrazu na podstawie jego cech leksykalnych. W tym celu planujemy zastosować zaawansowane metody modelowania. Nasze wyniki podpowiedzą twórcom słowników, na których wyrazach powinni skupić się w pracach leksykograficznych, dzięki czemu szybciej powstanie bardziej pomocny słownik. Wiedza o tym, co sprawia, że użytkownicy słowników wyszukują te słowa a nie inne, jest także interesująca teoretycznie, ponieważ mówi nam coś o tym, jak funkcjonuje język w naszych umysłach.