Biuletyn
Polonistyczny

20.06.2024

7 pytań, które warto sobie zadać tworząc korpus do badań

Korpusy, czyli zbiory danych tekstowych, dobranych zgodnie z określonymi kryteriami i odpowiednio opisane, stają się coraz bardziej popularnym narzędziem pracy naukowców i naukowczyń reprezentujących różne dyscypliny. Korzystają z nich przedstawiciele i przedstawicielki dyscyplin takich jak językoznawstwo, ale i socjologia, historia, psychologia, a od pewnego czasu także literaturoznawstwo. Dzięki wykorzystaniu korpusów możliwe jest przetwarzanie zbiorów danych wielokrotnie przekraczających możliwości analityczne człowieka-badacza, co pozwala na stawianie zupełnie nowych pytań badawczych oraz poszukiwanie nowych odpowiedzi na pytania już znane.

Jednak aby badanie z wykorzystaniem korpusu było wiarygodne i wniosło pozytywną jakość w naszą pracę naukową, powinien on być odpowiednio przygotowany i opracowany. Oto siedem pytań, które warto sobie zadać, przystępując do tworzenia korpusu.

1. Co chcę zbadać?

To pytanie może wydawać się banalne, ale jeśli nie zaczniemy od udzielenia sobie na nie bardzo precyzyjnej odpowiedzi może się okazać, że korpus, który stworzymy, tak naprawdę nie będzie dostosowany do naszych potrzeb.

Staranne sformułowanie pytania badawczego będzie też kluczowe dla odpowiedzi na wszystkie następne pytania — pozwoli na sprawne wyznaczenie kryteriów wyboru tekstów wchodzących w skład korpusu, zdeterminuje dobór narzędzi i źródła tekstów itp.

Na przykład badanie języka mówionego na podstawie korpusu tekstów pisanych prawdopodobnie będzie chybione. Jeśli zaś chcemy badać język doniesień prasowych, to oparcie się tylko na jednym zbiorze — np. jednym tytule czy jednym roczniku — może okazać się niewystarczające.

2. Jak ustalić skład korpusu?

Skład korpusu powinien być mocno związany z naszym pytaniem badawczym. Dlatego na podstawie tegoż pytania musimy ustalić, jakie kryteria są ważne, aby dobrane teksty faktycznie odpowiadały naszym potrzebom. Zależnie od przedmiotu

badania, niektóre kryteria mogą być kluczowe lub nieistotne w konkretnym przypadku. Oto przykłady:

- **autorstwo** – w przypadku części badań informacja o autorstwie tekstu może być bardzo istotna. Dla niektórych analiz nie ma znaczenia nazwisko autora, ale liczyć może się jego płeć, pozycja społeczna, rok lub miejsce urodzenia, wykształcenie itp. Jednocześnie dla niektórych badań istotne może być ustalenie autorstwa na bardziej ogólnym poziomie – np. w przypadku analizy języka prasowego ważniejsze może być, w jakim czasopiśmie dany tekst został opublikowany. Należy zastanowić się, które kryteria mają znaczenie w naszym przypadku.
- **publikacja** – jeśli badamy teksty, które miały realny wpływ na język swojej epoki, powinniśmy skupić się na tych, które były wydane w danym czasie. Jednak jeśli interesują nas teksty stanowiące odbicie jakiejś tendencji, może okazać się, że kluczowa będzie data powstania – nawet jeśli tekst nie doczekał się publikacji w swojej epoce. Teksty szerzej czytane lub popularne zwykle miały więcej wydań czy wznowień, choć z dzisiejszej perspektywy możemy mieć pokusę, aby uwzględniać przede wszystkim te utwory, które weszły do kanonu. Uwzględnienie czasu i miejsca publikacji jako kategorii pozwala też na ograniczenie czasowe populacji tekstów, którą badamy. To zwykle istotne kryterium doboru źródeł do korpusu.
- **kwestie społeczne i polityczne** – zwłaszcza w przypadku tekstów prasowych czy publicystyki, ale często też w przypadku diarystyki czy korespondencji, wpływ poglądów i programów politycznych i społecznych na używane słownictwo jest nie do przecenienia. Warto zadać sobie pytanie, jakie ma to znaczenie dla naszego badania. Jeśli chcemy zbadać przekrojowo język prasy danego okresu, powinniśmy postarać się o reprezentację jak najszerszy przegląd stanowisk. Jeśli jednak próbujemy zbadać różnice między językiem używanym przez przedstawicieli różnych obozów, przydadzą nam się korpusy porównawcze zbierające teksty reprezentujące odmienne postawy. Podobnie pochodzenie społeczne czy poziom wykształcenia mają istotny wpływ na język, którym posługują się autorzy tekstów. Zależnie od problemu badawczego, możemy ograniczyć wybór źródeł korpusu do tekstów spełniających jedno kryterium, lub zdecydować się na teksty możliwie różnorodne.
- **kwestie gatunkowe i stylistyczne** – teksty literackie mają inną rangę od tekstów użytkowych, a w ich obrębie zróżnicowanie gatunkowe mogą sprawiać, że słownictwo będzie miało odmienną charakterystykę. Nawet w obrębie jednego gatunku różnice stylistyczne czy tematyczne mogą być źródłem istotnych różnic językowych, których pominięcie może prowadzić do przekłamania wyników badania.
- **objętość** – korpusy co do zasady mogą być próbkowane bądź nie. W przypadku korpusów próbkowanych mamy do czynienia ze sztucznym ograniczeniem objętości wybranych tekstów, np. po to, aby uniknąć dysproporcji w ich rozmiarach. Jednak takie podejście ma swoje minusy — analiza języka całego tekstu różni się od analizy jego fragmentu. Zadajmy

sobie więc pytanie, na ile objętość poszczególnych tekstów ma znaczenie dla naszego badania.

To jedynie przykładowe kryteria doboru tekstów do korpusu. Szczegółowa ich lista powinna być opracowana na podstawie pytania badawczego i do niego dostosowana.

3. Jak skomponować korpus?

Teoretycznie, dobrze skomponowany korpus powinien być jednocześnie reprezentatywny i zrównoważony. Reprezentatywność korpusu oznacza, że zawiera on wszystkie elementy badanej odmiany języka; czyli jeśli badamy XIX-wieczną poezję, powinniśmy uwzględnić różne nurty i różnych autorów. Zrównoważenie korpusu zakłada, że żaden z nurtów nie powinien dominować nad innymi (czyli korpus XIX-wiecznej poezji nie powinien składać się jedynie z dzieł Mickiewicza i Słowackiego).

W praktyce spełnienie obu tych kryteriów jednocześnie jest bardzo trudne, o ile w ogóle możliwe. Powinniśmy więc zastanowić się, która z tych wartości jest ważniejsza z punktu widzenia naszego badania. Czy zależy nam na odwzorowaniu populacji tekstów, co prawdopodobnie będzie oznaczało dominację tekstów reprezentujących wybrane cechy, czy jednak istotniejsze jest uniknięcie przewagi jakiegoś typu tekstów, kosztem utraty pełnej reprezentatywności korpusu? Jeszcze innym przypadkiem są korpusy tekstów jednego autora, których ambicją jest zgromadzić wszystkie dzieła danego twórcy.

Powinniśmy więc zastanowić się, jakie znaczenie opisane uprzednio kryteria będą miały dla kompozycji korpusu. Przykładowo, jeśli uznamy, że ważnym kryterium jest dla nas pochodzenie autora, musimy zdecydować, czy chcemy, aby korpus składał się w równych proporcjach z dzieł autorów pochodzących z określonych miejsc (np. urodzonych we wszystkich trzech zaborach), czy raczej będziemy starali się, żeby udział procentowy dzieł autorów urodzonych w poszczególnych zaborach odzwierciedlał ich udział w całej populacji XIX-wiecznej produkcji poetyckiej.

Jak widać, bardzo ważnym aspektem wpływającym na ukształtowanie korpusu jest jak najbardziej precyzyjne określenie całej populacji tekstów, do których się odnosimy.

4. Skąd pozyskać teksty do korpusu?

Teksty do korpusu możemy pozyskać na wiele różnych sposobów i podobnie jak w przypadku wcześniejszych pytań, odpowiedź będzie mocno związana z tym, co konkretnie chcemy zbadać. Jeśli zależy nam na uzyskaniu dużego zbioru danych szybko, a ich precyzja ma dla nas nieco mniejsze znaczenie (korpus *big and dirty*), warto rozważyć zdobycie tekstów dostępnych np. w portalach typu Wolne Lektury czy POLONA. Dane możemy pozyskiwać automatycznie dzięki narzędziom do

netscrappingu, które czytają kod strony internetowej i są w stanie dokonać ekstrakcji tekstu (co zwykle wymaga kontroli użytkownika).

W przypadku badań literaturoznawczych może się jednak okazać, że pozyskane w ten sposób dane będą zbyt niskiej jakości i nie będą się nadawały do przeprowadzenia badania, na którym nam zależy. Możliwe też, że dostępne zasoby nie będą zawierały materiałów, które chcemy zbadać. W takiej sytuacji konieczne będzie przygotowanie korpusu samodzielnie, z wykorzystaniem wysokiej jakości danych, kwalifikowanych wydań i ich starannej korekty. Należy pamiętać, że znacząco wydłuża to czas konieczny na stworzenie korpusu, jednak efektem będzie wysokiej jakości zbiór danych, który może posłużyć nie tylko nam, ale i kolejnym badaczom.

5. Jak opisać korpus?

Korpus składa się nie tylko z danych (czyli samych tekstów), ale i z metadanych, czyli różnego rodzaju informacji o tekstach, które mogą same w sobie być interesującym źródłem danych badawczych (więcej na ten temat pisał Cezary Rosiński w swoim poście). Zbierając dane do badania zwykle zbieramy też wiele metadanych , choćby pochodzących z analizy tekstów pod kątem kryteriów, które wyznaczyliśmy w punkcie drugim. Wybór sposobu zapisu tych metadanych może nam pozwolić na znacznie szersze korzystanie ze stworzonego zbioru, a zastosowanieistniejących standardów zapisu metadanych ułatwi potencjalne połączenie naszego korpusu z innymi korpusami – w tym zagranicznymi — w przyszłości.

6. Jakich narzędzi użyć do przechowywania i przetwarzania korpusu?

To kolejne pytania, które są bezpośrednio związane z problemem badawczym. Wśród wielu dostępnych dla naukowców narzędzi część jest bezpłatna, inne wymagają wykupienia licencji. Nie można wykluczyć, że bardziej zaawansowane badania będą wymagały wsparcia specjalistów i przygotowania narzędzia precyzyjnie odpowiadającego potrzebom konkretnego projektu.

Decydując się na wybór narzędzia należy wziąć pod uwagę przede wszystkim, czy udostępnia ono funkcje, które będą nam niezbędne do przeprowadzenia badania. Ważne aspekty to także intuicyjność i wygoda użytkowania czy dostępność w otwartej licencji. Przykładowe narzędzia służące do przetwarzania korpusów i prowadzenia podstawowych analiz zostały omówione w tym tekście.

Od tego, z jakich narzędzi chcemy skorzystać, będzie zależał także wybór odpowiedniego formatu, w którym zapisujemy pliki. Warto rozważyć to już na początku, aby uniknąć konieczności zmian i poprawek. Tworząc korpus, można skorzystać z narzędzi czyszczących wstępnie pliki (np. z elementów kodu strony) czy zapisujących je w pasującym formacie.

Wybór narzędzi ma też związek z wyborem określonego formatu zapisu metadanych. Wiele dostępnych serwisów przynajmniej część metadanych może odczytać z nazwy pliku; inne wymagają zapisania ich osobno.

7. Udostępnianie korpusu

Zgodnie z dobrymi praktykami otwartej nauki, przygotowany korpus warto udostępnić innym badaczom, by mogli oni wykorzystać go do kolejnych badań. Rozważając formę udostępnienia, powinniśmy jednak wziąć pod uwagę, czy teksty, z których skorzystaliśmy, nie są chronione prawami autorskimi i ich udostępnienie w całości nie będzie stanowiło naruszenia tych praw. Jeśli kwestie prawne są uporządkowane, możemy przygotować korpus w jednym z popularnych formatów i umieścić go w repozytorium lub np. na GitHubie. W opisie korpusu powinniśmy podać wszystkie informacje, które go dotyczą — nazwiska twórców, możliwie dokładny opis składu korpusu, a także informacje dotyczące praw autorskich i możliwości dalszego przetwarzania. Serdecznie zachęcamy do publikacji korpusów literaturoznawczych, jako że jest to najlepszy sposób na rozwijanie cyfrowego literaturoznawstwa. Tworzenie wysokiej jakości korpusów jest zajęciem żmudnym i pracochłonnym – warto udostępnić efekty swojej pracy innym, by mogli zweryfikować nasze wnioski, ale także przeprowadzić własne badania na naszym zbiorze lub poszerzyć go o własne dane. Istnieją też projekty w całości oparte o dostarczanie tekstów korpusów w różnych językach, np. DraCor dla tekstów dramatycznych czy PoeTree dla poezji. Możliwość skorzystania z takich zbiorów danych otwiera przed badaczami nowe pola działalności, w tym komparatystykę różnych literatur narodowych.

Pierwodruk: *7 pytań, które warto sobie zadać tworząc korpus do badań*, *OtwartaHumanistyka* (20 czerwca 2024), https://otwartanauka.hypotheses.org/1099

Linki:

- https://otwartanauka.hypotheses.org/1099

Słowa kluczowe: korpus, literaturoznawstwo cyfrowe, analiza akustyczna, korpusomat

Autor: Anna Mędrzecka-Stefańska