Contents

1 Informacje o korpusach

1.1 Jakie korpusy będziemy omawiać?

Na naszych zajęciach będziemy wykorzystywać kilka korpusów dla języka polskiego oraz angielskiego. Jeżeli chodzi o język polski, to skupimy się na Narodowym Korpusie Języka Polskiego, British National Corpus, Corpus of Contemporary American English oraz kilku mniej znanych korpusach dostępnych przez portal SketchEngine. Zobaczymy również jak można stworzyć własny korpus tekstów oraz jakie problemy związane są z tworzeniem korpusów.

1.1.1 Narodowy Korpus Języka Polskiego

Narodowy Korpus Języka Polskiego jest wspólnym projektem Instytutu Podstaw Informatyki PAN, Instytutu Języka Polskiego PAN, Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego i Wydawnictwa Naukowego PWN. Jest to jak do tej pory największy (oprócz korpusów budowanych automatycznie z materiałów dostępnych w Internecie) korpus języka polskiego. Oprócz samego korpusu grupa odpowiedzialna za NKJP stworzyła również rozmaite narzędzia pozwalające pracować z korpusami. Część z nich omówimy w czasie naszego kursu. W ramach projektu NKJP powstały tak naprawdę trzy korpusu:

- automatycznie anotowany korpus zrównowazony o wielkości 300M segmentów

- automatycznie anotowany korpus niezrównoważny o wielkości 1800M segmentów
- ręcznie anotoany korpus referencyjny o wielkości 1.2M segmentów

Jeśli chodzi o przedział czasowy tekstów zawartych w korpusie, to znajdują się tam teksty (i transkrybowane wypowiedzi ustne) z lat XXXX-XXXX.

1.1.2 British National Corpus

British National Corpus został stworzony przez wydawnictwo Oxford University Press. Zawiera on próbki brytyjskiej odmiany języka angielskiego. Był jednym z pierwszych tak dużych i szeroko dostępnych korpusów angielszczyzny. BNC zawiera około 100M słów i jest korpusem zrównoważonym. Zawiera teksty z lat 1960-1993, czyli dość współczesne.

1.1.3 Corpus of Contemporary American English

Corpus of Contemporary American English został stworzony przez Marka Daviesa z Brigham Young University. W odróżnieniu od BNC koncentruje się na amerykańskiej odmianie języka angielskiego. COCA zwiera ponad 560M słów. Jako korpus zrównoważony zawiera próbki z rozmaitych gatunków i kanałów. Teksty (i transkrypcje) zawarte w tym korpusie pochodzą z lat 1990-2017, więc oddaje on bardzo współczesną angielszczyznę. Mark Davies skonstruował również między innymi korpus historycznej angielszczyzny (COHA), którym nie bedziemy się jednak zajmować bezpośrednio na kursie.

1.2 Z jakich narzędzi będziemy korzystać?

Jeśli chodzi o narzędzia do korzystania z korpusów, to będziemy korzystać z takich programów, które umozliwiają pracę z korpusami za pomocą aplikacji w przeglądarce internetowej. Będziemy więc głównie korzystać z:

- wyszukiwarki korpusowej PELCRA oraz kolokatora, która umożliwia pracę z NKJP;
- wyszukiwarki korpusowej Poliqarp, która umożliwia prace z NKJP;
- multiwyszukiwarki i kolokatora corpus.byu.edu opracowanej przez Marka Daviesa, umożliwiającej pracę z COCA, BNC i wieloma innymi dostepnymi korpusami
- programu do zarzadzania i przetwarzania korpusów SketchEngine

Każda z tych aplikacja ma swoje plusy i minusy oraz udostępnia nieco inne funkcje. Dodatkowo umiejętność korzystania z kilku narzędzi pozwala łatwo nauczyć się funkcji innych podobnych programów (które, być może, pojawią się w przyszłości).

1.3 Jak stworzone są korpusy?

Potencjalny twórca korpusu praktycznie na każdym kroku. Istnieją, na przykład, różne podejścia doboru tekstów do korpusów. Według jednego podejścia należy z każdego dostępnego tekstu wybrać próbkę pewnej długości (np. 2000 słów) i uwzględnić w korpusie jak najwięcej tekstów. Taki sposób postępowania ma tę zaletę, że zwiększa się zróżnicowanie materiałów w korpusie oraz łatwiej uzyskać później zgodę podmiotów dysponujących prawami autorskimi do tekstów na ich redystrybucje w formie korpusu. Inna filozofia głosi, że należy do korpusu włączać w miarę możliwości całe teksty (np. John Sinclair stoi na takim stanowisku) bo "szatkowanie" tekstów zabija ich integralność. Innym ważnym problemem jest kwestia zrównoważenia korpusu. Jeżeli tworzymy ogólny korpus dla jakiegoś języka, skierowany do szerokiego grona użytkowników, to możemy chcieć, aby struktura tekstowa korpusów (rodzaje tekstów i ich objętość) była jak najbardziej zróżnicowana. Inną strategią jest takie dobranie stuktury tekstów, aby jak najlepiej odzwierciedlała rzeczywistą praktykę językową. O obu tych podejściach więcej powiemy w dalszej części kursu. Czasami jednak do kwestii wyboru tekstów trzeba podejść bardziej oportunistycznie, szczególnie jeśli mówimy o językach wymierających bądź wymarłych lub takich, w których nie mamy zbyt wiele źródeł pisanych.

Zebranie (i ewentualna transkrypcja/digitalizacja) tekstów nie jest końcem pracy przy tworzeniu korpusów. Większość z korpusów jest anotowana. Kiedyś zadanie anotacji korpusu wykonywane było ręcznie i zajmowało ogromną ilość czasu dość dużym zespołom ludzi. Dzisiaj komputerowe systemy (np. tagery morfosyntaktyczne) automatycznie przypisujom słowom w korpusie właściwe części mowy. W zależności od korpusu i wykorzystanych narzędzi dostępne są różne poziomy anotacji, pewnym standardowym minimum jest jednak anotacja częściami mowy (*part of speech* - w skrócie POS). Oczywiście możliwa jest również bardziej zaawansowana anotacja składniowa czy sensami słów. My w ramach kursu nie będziemy zajmować się tymi narzędziami, warto jednak pamiętać, że wykorzystywane są do tworzenia korpusów.

1.4 Struktura tekstowa korpusów.

Każdy z korpusów, ktorymi będziemy się zajmować, przyjał inną filozofię doboru tekstów do korpusów. Ogólnie rzecz biorąc korpus można charakteryzować jako "zrównoważony" bądź "reprezentatywny". Często można spotkać się z synonimicznym rozumieniem tych dwóch terminów w odniesieniu do korpusów, Górski i Łaziński (2012) definiują te pojęcia inaczej, wprowadzając bardzo przydatne rozróżnienie:

- **reprezentatywność** - odnoszenie się do jakiejś rzeczywistości istniejącej poza korpusem
- **zrównoważenie** - dbałość o taką budowę korpusu, żeby żaden składnik na żadnym z poziomów nie dominował nad innym.

W przypadku **reprezentatywności**, która może wydawać się bardziej atrakcyjna dla twórcy korpusu, musimy - zdaniem Górskiego i Łazińskiego - zdecydować się na to jaką pozakorpusową rzeczywistość oddawać ma struktura tekstów w korpusie. Korpus może odzwierciedlać populację twórców tekstów, populację tekstów, produkcję tekstów i recepcję tekstów. Autorzy korpusu NKJP zdecydoali się na strategię, zgodnie z którą struktura tekstów odzwierciedlać ma recepcję polszczyzny pisanej. Szczegółowe informacje na temat tego w jaki sposób estymowali ilościowe proporcje recepcji polszczyzny pisanej znajdują się w artykule, należy jednak wspomnieć, że posłuiwali się np. pochodzącymi z badań socjologicznych danymi dotyczącymi czytelnictwa w Polsce.

Zazwyczaj jednak twórcy korpusu nie wprowadzają tego rozróżnienia i koncentrują się raczej na tym, co Górski i Łaziński nazywają **zrównoważeniem**. Żeby osiągnąć taki zwrównoazony korpus należy umiejścić w odpowiednich proporcjach teksty pochodzące z różnych źródeł. Poniżej umieszczone są informacje dotyczące struktury tekstowej korpusów, z którymi pracowć będziemy podczas naszego kursu:

1.4.1 Narodowy Korpus Języka Polskiego

Górski i Łaziński (2012) podają następujące proporcje tekstów:

- Publicystyka i krótkie wiadomości prasowe: 50%
- Literatura piękna: 16%
- Literatura faktu: 5,5%

- Typ informacyjno-poradnikowy: 5,5%
- Typ naukowo-dydaktyczny: 2,0%
- Inne teksty pisane: 3,0%
- Książka niebeletrystyczna nieklasyfikowana: 1,0%
- Teksty konwersacyjne, mówione medialne i quasi-mówione razem: 10,0%
- Teksty internetowe statyczne i dynamiczne razem: 7,0%

Powyższa struktura zachowana jest w pełni tylko dla ręcznie anotowanego korpusu refrencyjnego, liczącego 1,2M segmentów. W przypadku korpusu zrównoważonego, który liczy 300M segmentów, zachowana została wzorcowa proporcja następujących kategorii: prasa, beletrystyka, książki non fiction, inne teksty pisane, wszystkie trzy rodzaje tekstów mówionych. Nie została zachowana dokladnie wzorcowa proporcja podtypów książek non-fiction.

1.4.2 British National Corpus

Jeśli chodzi o część korpusu zawierającą teksty pisane (stanowiącą 90% obiętości). Podobnie jak w przypadku NKJP autorzy posługiwali się danymi z badań społecznych aby ustalić odpowiednią strukturę tekstową korpusu. Rozkład poszczególnych typów tekstów przedstawia się następująco (źródło: `http://www.natcorp.ox.ac.uk/docs/URG.xml?ID=BNCdes`):

- Imaginative: 27.10%
- Informative: natural & pure science: 3.67%
- Informative: applied science: 7.15%
- Informative: social science: 13.99%
- Informative: world affairs: 16.00%
- Informative: commerce & finance: 7.66%
- Informative: arts: 6.43%
- Informative: belief & thought:: 3.03%
- Informative: leisure: 14.92%

1.4.3 Corpus of Contemporary American English

W przypadku COCA struktura korpusu jest trochę prostsza - zawiera on po 20% tekstów z następujących kategorii (źródło: (`https://corpus.byu.edu/coca/`):

-Spoken: (118 million words [118,167,133]) Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer, etc). [See notes on the naturalness and authenticity of the language from these transcripts).

- Fiction: (113 million words [113,404,735]) Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts.

- Popular Magazines: (118 million words [118,450,563]) Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc). A few examples are Time, Men's Health, Good Housekeeping, Cosmopolitan, Fortune, Christian Century, Sports Illustrated, etc.

- Newspapers: (114 million words [114,341,164]) Ten newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. In most cases, there is a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.

- Academic Journals: (112 million words [111,537,393]) Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), both overall and by number of words per year

1.5 Jakie informacje są w korpusach? (metadane, tagset itp)

W każdym z korpusów, którymi będziemy się zajmować, zawarte są teksty i - co oczywiste - słowa, z których składają się te teksty. Nie są to jednak wszystkie informacje zawarte w korpusie. Oprócz tego mamy pewne metadane dotyczące fragmentów tekstu zawartych w korpusie oraz dodatkowe dane

dotyczące słów/zdań na poziomie anotacji korpusu. Dla przykładu w ręcznie anotowanym subkorpusie Narodowego Korpusu Języka Polskiego mamy dostęp do następujących danych dotyczących każdego fragmentu tekstu:

- **Autor**
- **Tytuł tekstu**
- **Źródło**
- **ISBN** (w przypadku publikacji)
- **Rok publikacji**
- **Wydawca**
- **Miejsce publikacji**
- **Typ tekstu**:
- literatura piękna
  - proza,
  - poezja,
  - dramat,
- literatura faktu,
- publicystyka i krótkie wiadomości prasowe,
- typ naukowo-dydaktyczny,
- typ informacyjno-poradnikowy,
- książka niebeletrystyczna niesklasyfikowana,
- inne teksty pisane
  - typ urzędowo-kancelaryjny,
  - teksty perswazyjne (ogłoszenia, reklamy, propaganda polityczna),
  - krótkie teksty instruktażowe ,
- listy,
- Internet

    - interaktywne strony WWW (fora, chaty, listy dyskusyjne itp.),
    - statyczne strony WWW,
- teksty mówione konwersacyjne,
- teksty mówione medialne,
- teksty quasi-mówione.
- **Kanał**:
    - prasa
        * prasa – dziennik,
        * prasa – tygodnik,
        * prasa – miesięcznik,
        * prasa – inne,
    - książka,
    - Internet,
    - mówiony,
    - ulotka,
    - rękopis.

Na przykład korzystając z wyszukiwarki Poliqarp możemy dowiedzieć się, że zawarte w korpusie zdanie "Aby zdjąć ze mnie ten straszny obowiązek, ten rozkaz piekielny, ksiądz zabije innego człowieka" pochodzi z następującej publikacji:

- **autor**: Jarosław Iwaszkiewicz
- **tytuł**: Brzezina i inne opowiadania Kościół w Skaryszewie
- **źródło**: Brzezina i inne opowiadania Kościół w Skaryszewie
- **ISBN**: 9788307030838
- **rok publikacji**: 2006
- **wydawca**: Czytelnik
- **miejsce publikacji**: Warszawa
- **typ**: literatura piękna

- **kanał**: książka

Jeśli chodzi o informacje na poziomie anotacji, przeanalizujmy słowo "informowania" w zdaniu "Kto miał wg Ciebie obowiązek informowania opinii publicznej o tej sprawie?" pochodzącym z usenetu. W NKJP znajdziemy następujące informacje o tym słowie:

[informować:ger:sg:gen:n:imperf:aff]

Oznacza to, że:

- forma bazowa tego słowa to "informować" (lemat)
- klasa gramatyczna tego słowa to rzeczownik odczasownikowy (*ger*)
- jest to rzeczywnik w liczbie pojedynczej (*sg*)
- przypadek tego rzeczownika to dopełniacz (*gen*)
- rodzaj tego rzecownika to rodzaj nijaki (*n*)
- czasownik, od którego derywowana jest ta forma jest czasownikiem niedokonanym (*imperf*)
- jest to forma niezanegowana (czyli nie jest to np. *niepoinformowanie*) (*aff*)

Informacje dotyczące tagsetu (czyli zestawu znaczników morfosyntaktycznych) używanego przez NKJP znajdują się na stronie: `http://nkjp.pl/poliqarp/help/plse2.html#x3-20002`

Inne korpusy mogą (i pewnie robią to) używać innego zestawu znaczników. Warto zajrzeć do informacji o korpusie, aby zobaczyć jaki tagset używany jest w danym korpusie.