# Oprogramowanie do analizy tekstów

Elżbieta Gajek

## Wstęp

Technika komputerowa zapewnia językoznawcom narzędzia do korzystania z baz danych, które nazywane są korpusami językowymi. Korpus jest to wybór tekstów zebranych według określonego przez twórców bazy kryterium. Kryteria bywają różne, np. korpusem jest zbiór tekstów jednego autora, jednego czasopisma, transkrybowane rozmowy grupy osób w określonym wieku, mieszkańców jednego miasta lub regionu, a także zbiór tekstów wydanych przez jedno wydawnictwo lub napisanych przez kandydatów zdających jeden egzamin.

Warto zająć się dwiema sprawami. Po pierwsze, powodem tworzenia baz językowych, zarówno przez instytucje, jak i indywidualnych entuzjastów stosowania techniki cyfrowej w językoznawstwie, po drugie, sposobem korzystania z nich za pomocą specjalistycznego oprogramowania.

## Potrzeby użytkowników językowych baz danych

**Potrzeby instytucjonalne.** Wydawnictwa produkujące słowniki jedno- i wielojęzyczne oraz materiały edukacyjne tworzą korpusy w celu zapewnienia sobie aktualnych danych językowych, aby w ich książkach prezentowany był język rzeczywiście pisany i mówiony przez jego rodzimych użytkowników.

Popularne na świecie egzaminy językowe są organizowane przez instytucje takie jak Uniwersytet Cambridge (z języka angielskiego) czy Instytut Goethego (z niemieckiego), które dysponują korpusami tekstów napisanych przez osoby zdające te egzaminy. Korpusy są wykorzystywane do badań nad językiem uczniów oraz do przygotowania materiałów edukacyjnych.

Instytucje akademickie tworzą korpusy dla celów badawczych, w celu rozwoju wiedzy o języku i zmianach w nim zachodzących. Teorie językoznawcze znajdują oparcie w danych pochodzących z korpusów. Dział językoznawstwa, zwany językoznawstwem korpusowym lub lingwistyką korpusową, zajmuje się metodami tworzenia korpusów, narzędziami elektronicznymi do pracy z nimi, metodami statystycznymi pozwalającymi na efektywne wykorzystanie danych językowych zgromadzonych w bazie.

W ostatnim czasie korpusy wielojęzyczne są przedmiotem zainteresowania instytucji ponadnarodowych, np. działających w strukturach Unii Europejskiej. Prowadzone są intensywne prace nad automatycznymi systemami tłumaczącymi dokumenty Komisji Europejskiej czy Parlamentu Europejskiego na języki państw członkowskich UE. W instytucjach UE rocznie tłumaczy się dziesiątki milionów stron. Celem automatyzacji jest skrócenie czasu tłumaczenia dokumentów oraz ograniczenie

do minimum czasochłonnej i kosztownej pracy ludzkiej.

Największe zainteresowanie językoznawców budzą ogromne zbiory tekstów, zawierające często setki milionów słów. Bazy danych językowych zawierających zarówno teksty pisane, jak i mówione są tworzone w celu uzyskania reprezentatywnego obrazu języka używanego przez społeczeństwo.

Przykłady reprezentatywnych korpusów narodowych dla języka:
- polskiego – korpus PELCRA,
- angielskiego – korpusy BNC (British National Corpus) i Bank of English, Collins-Cobuild Corpus,
- niemieckiego – korpus Manhaimer.

Wiele korpusów jest częściowo dostępnych na stronach internetowych instytucji, które je stworzyły.

Istnieją też liczne zbiory tekstów elektronicznych, które korpusami nie są, ale mogą służyć jako źródło danych dla korpusu nauczycielskiego lub uczniowskiego. Są to strony programu Gutenberg, zawierające teksty literatury narodowej konkretnego kraju. Teksty z Internetu mogą służyć do stworzenia bazy tekstów według własnego kryterium, bazy, która może być wykorzystywana do własnych celów edukacyjnych, natomiast nie może być publikowana, ponieważ twórca korpusu nie ma praw autorskich do zebranych tekstów.

**Potrzeby indywidualnych użytkowników języka.** Indywidualni twórcy baz językowych najczęściej opracowują je z powodu:
- zainteresowania językiem jako systemem,
- zainteresowania oprogramowaniem do przetwarzania języków naturalnych,
- zainteresowania językiem specjalistycznym lub używanym przez subkultury,
- samokształcenia w zakresie języka ojczystego i obcych,
- potrzeby wyjaśniania wątpliwości językowych.

Chcą oni poznać język używany w interesujących ich, wąskich specjalnościach nauki, techniki, biznesu lub sztuki. Zainteresowania specjalistyczne mogą znacznie podnieść motywację do nauki i rozwijania sprawności językowych uczniów zarówno młodych, jak i dorosłych.

**Potrzeby zawodowe nauczyciela języków.** Korpusy językowe są bardzo mało znane wśród nauczycieli języków obcych, mogą jednak być bardzo przydatne w wielu obszarach ich pracy. Nauczyciel może zachęcić młodzież do tworzenia własnego korpusu językowego w wybranej dziedzinie wiedzy lub kultury i samodzielnych studiów nad językiem wybranej specjalności. Integracja formalnej nauki szkolnej i pozaszkolnej pracy nad językiem może mieć wpływ na odpowiedzialność ucznia za własną naukę i pozwolić mu lepiej przygotować się do nauki przez całe życie, a także do samodzielności w rozwiązywaniu problemów językowych.

Samokształcenie językowe jest niezbędnym elementem pracy nauczyciela, który przez całe życie zawodowe jest jednocześnie zaawansowanym uczniem języka, którego uczy. Praca nauczyciela wymaga bezustannego doskonalenia osiągniętych podczas studiów sprawności językowych z dwóch powodów. Jednym z nich jest naturalne zapominanie języka, np. podczas ciągłej pracy z początkującymi uczniami, a drugim – zachodzące w nim zmiany.

Opisane poniżej sposoby korzystania z gotowych baz danych do rozwiązywania problemów językowych pozwalają nauczycielom na pracę samokształceniową i podtrzymywanie własnych sprawności językowych oraz uaktualnianie wiedzy o języku.

Praca dydaktyczna nauczyciela języka obcego wymaga znajomości sposobów przygotowania materiałów dydaktycznych dla uczniów o specjalnych potrzebach językowych, np. w nauczaniu języka zawodowego, tj. lekarzy, inżynierów, ekonomistów, wojskowych, często specjalizujących się w bardzo wąskich dziedzinach i będących na różnych poziomach zaawansowania językowego. Wiedza językowa zdobyta na studiach nie wyczerpuje wszystkich możliwych dziedzin życia. Na rynku brakuje materiałów dydaktycznych dla bardzo wąskich specjalności zawodowych, nauczyciel musi więc samodzielnie rozwiązywać problemy w konkretnej sytuacji edukacyjnej i często sam tworzyć odpowiednie materiały. Znajomość korpusów i oprogramowania oraz metod lingwistyki korpusowej okazuje się wówczas pomocna. Dotyczy to również nauczycieli języka polskiego, którzy coraz częściej uczą cudzoziemców języka specjalistycznego i zmuszeni są rozwiązywać problemy wynikające z międzyjęzykowych skojarzeń.

Nauczyciel zarówno języka obcego, jak i ojczystego jest często proszony o wyjaśnienie np. różnicy w użyciu synonimów lub zasad występowania obok siebie dwóch lub kilku wyrazów. Okazuje się, że słownik nie zawsze jest przydatny, ponieważ podane przykłady nie wystarczają. Wtedy korpus staje się niezastąpiony. Dzięki niemu można porównać wiele przykładów i dostrzec różnice w kontekście użycia słów. Niejednokrotnie ważna jest częstość używania słowa. Ponadto język jest żywy, ciągle tworzymy neologizmy. Uczniowie, którzy mają dostęp do różnych tekstów, oczekują od nauczyciela języka obcego pomocy i wyjaśnień dotyczących takich słów – wówczas korpusy pomagają rozwikłać problem. Wielu słów rzadkich lub specjalistycznych nie można znaleźć w słownikach, ponieważ słowniki, mimo wysiłków leksykografów, często nie nadążają za rejestracją zmian, wówczas wykorzystanie korpusu i sieci jako korpusu okazuje się niezastąpione.

**Potrzeby zawodowe językoznawców.** Dane językowe zgromadzone w korpusach służą językoznawcom, powiększając ich wiedzę o języku jako systemie, w tym w szczególności:

- leksykografom przygotowującym słowniki,
- semantykom badającym znaczenia słów,
- syntaktykom badającym struktury gramatyczne,
- fonologom badającym brzmienie języka,
- socjolingwistom zajmującym się odmianami języka używanymi przez różne społeczności,
- psycholingwistom zajmującym się indywidualnymi różnicami użycia języka,
- lingwistom wspierającym lekarzy w diagnozowaniu uszkodzeń mózgu,
- logopedom i terapeutom do identyfikacji schorzeń i zaburzeń mowy.

**Oprogramowanie.** Zbudowanie korpusu to dopiero początek pracy z danymi językowymi. W celu uzyskania dostępu do danych powstaje wiele rodzajów specjalistycznego oprogramowania:

- oprogramowanie zliczające liczbę słów w korpusie oraz liczbę słów różnych,
- oprogramowanie do tworzenia list frekwencyjnych, czyli sortujące słowa w porządku częstotliwości występowania, porządku alfabetycznym lub *a tergo*[1],
- oprogramowanie konkordancyjne, czyli wykrywające słowo w kontekście. Programy te mają zwykle interfejs pokazujący listę znalezionych słów z kilkoma słowami poprzedzającymi je i kilkoma następującymi po nim,
- oprogramowanie, które dodaje do korpusu specjalne znaki – *tagi* – które ułatwiają wyszukiwanie, np. części mowy (rzeczowników, czasowników itd.) lub części zdania (podmiot, orzeczenie itd.),
- oprogramowanie do porównywania dwóch korpusów w tym samym języku,
- oprogramowanie do porównywania korpusów w różnych językach, z których jeden powstał w wyniku tłumaczenia tekstów zgromadzonych w drugim.

W sieci dostępne są liczne programy przydatne w analizie korpusów. Poniżej zostaną podane przykłady oprogramowania komercyjnego i wolnego dostępu – do pobrania z sieci.

Programy **WordSmith Tools i MonoConc** to programy komercyjne, dobrze działające z danymi w wielu językach.

Zestaw aplikacji WordSmith Tools jest jednym z najpopularniejszych programów stosowanych w lingwistyce korpusowej. Jego wersja demonstracyjna, np. *WordSmith Tools Demo* 4.0, pozwala na działanie wszystkich jego funkcji. Ograniczona jest tylko liczba wyświetlanych wyników.

Program można pobrać ze strony http://www.lexically.net/wordsmith/version4. Wersję demonstracyjną programu MonoConc 2.2. można pobrać ze strony http://www.camsoftpartners.co.uk/monoconc.htm. Liczba wyświetlonych wyników jest jednak ograniczona do 20. W porównaniu z aplikacją WordSmith Tools program jest znacznie prostszy w stosowaniu, jednak jego możliwości są mniejsze.

**Simple Concordance Program**, dostępny na stronie http://www.textworld.com/scp, i **AntConc,** dostępny na stronie http://www.antlab.sci.waseda.ac.jp, są przykładami oprogramowania, które można pobrać ze stron internetowych ich autorów. Zarówno jeden, jak i drugi program działają bardzo dobrze z danymi anglojęzycznymi. Podają podstawowe wartości liczbowe opisujące wielkość korpusu, listy częstotli-

[1] A *tergo*, czyli o kolejności decyduje ostatnia litera wyrazu, potem przedostania, np. la**ba**, ba**ca,** ra**fa,** kla**sa**.

wości, wyświetlają konkordancje wybranych słów. Oba programy są bardzo przyjazne i zawierają proste i kompletne instrukcje w języku angielskim, nie ma więc potrzeby opisywać ich tutaj szczegółowo.

W celu porównania częstotliwości słów w dwóch lub więcej korpusach można wykorzystać program **Range** ze strony jego twórcy Paula Notiona http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx. Przykładem programu *freeware* umożliwiającego uzyskanie podstawowych danych numerycznych jest **Frequency** tego samego autora do pobrania z jego strony.

Wersję demonstracyjną programu **ParaConc** służącego do porównywania danych w różnych językach można pobrać ze strony http://www.athel.com/para.html.

W wyniku działania programów uzyskuje się np. podstawowe numeryczne parametry korpusu określające jego wielkość, czyli liczbę słów oraz częstotliwość występowania słowa w korpusie, czyli dane z listy frekwencyjnej. Dane te są podstawą obliczania wszelkich testów statystycznych oraz wstępnej analizy jakościowej użytego języka. Okazuje się, że w wyniku zastosowania różnych programów w tym samym korpusie uzyskuje się różne wyniki. Pierwszym problemem, który się pojawia, jest definicja słowa. Dla informatyka słowo to ciąg znaków pomiędzy spacjami. Takiej definicji nie może zaakceptować językoznawca, ponieważ według niej słowo obejmuje również znak interpunkcyjny, który może po nim wystąpić w zdaniu. Niedogodność ta jest zwykle w prosty sposób usuwana w stosowanych w językoznawstwie programach. Kolejne wątpliwości dotyczą na przykład słów pisanych z łącznikiem czy adresów internetowych. Z powodu różnych definicji słowa, w programach uzyskuje się często różne liczby określające wielkość korpusu. Dalsze rozważania w tym artykule będą jednak prowadzone z punktu widzenia początkującego użytkownika programów do analizy językoznawczej, a nie programisty – ich twórcy. Zachęcam Czytelników do samodzielnego eksperymentowania z użyciem wielu programów do ustalania wielkości korpusu.

Najprostszym sposobem określenia rozmiaru korpusu jest zastosowanie edytora tekstu Microsoft Word, który podaje liczbę słów w tekście w menu „Narzędzia ⇒ Statystyka wyrazów".

W celu szybkiego ustalenia wielkości korpusu i jego listy frekwencyjnej warto skorzystać z programu **Web Frequency Indexer** Uniwersytetu w Georgetown – http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html#doit. Wyniki zarówno w języku angielskim, jak i polskim uzyskuje się w ciągu ułamków sekund.

**Internetowe wyszukiwarki konkordancji.** Zastosowanie programów opisanych powyżej nie jest jedyną drogą uzyskania wiedzy o tym, w jakim sąsiedztwie występuje słowo. W celu znalezienia konkordancji słów można skorzystać z jednej z następujących wyszukiwarek:

**WebCorp** – http://www.webcorp.org.uk/index.html, **WebConc** – http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?sprache=en&art=google, **Lexware Culler** – http://82.182.103.45/lexware/concord/culler.html. Czasami wyszukiwanie wyników, w zależności od słowa i parametrów technicznych sprzętu i łącza internetowego, może trwać nawet kilka minut. **KwicFinder** – http://www.kwicfinder.com/KWiCFinder.html wymaga rejestracji i pobrania plików. **Glossanet** – http://glossa.fltr.ucl.ac.be – po rejestracji przysyła wyniki poszukiwań pocztą elektroniczną. Do porównania listy częstotliwości małego korpusu z listą częstotliwości BNC (*British National Corpus*) można skorzystać ze stron zawierających listy częstotliwości brytyjskiego korpusu referencyjnego, np. Companion Website for: Word Frequencies in Written and Spoken English: based on the British National Corpus – http://www.comp.lancs.ac.uk/ucrel/bncfreq. Warto korzystać również z wygodnych interfejsów ułatwiających korzystanie z danych korpusu anglojęzycznego BNC np. **VIEW** – http://view.byu.edu – Marka Davisa lub **PIE** (*Phrases in English*) – http://pie.usna.edu.

**Problemy.** Korpus językowy zapewnia dużą liczbę przykładów użycia języka, jednak ich analiza bywa czasochłonna. Wymaga myślenia indukcyjnego, postrzegania podobieństw i uogólniania przedstawionych danych. Często wnioski z samodzielnej analizy danych pozostają w sprzeczności z opisanymi w literaturze językoznawczej i w podręcznikach do nauki języka regułami i zasadami. Dla wielu uczniów i nauczycieli, przyzwyczajonych do definicji słownikowych opracowanych przez leksykografów oraz do definicji gramatycznych, samodzielna analiza danych językowych

może być bardzo trudna, a wnioski z niej płynące także trudne do zaakceptowania[2]. Jednak użytkownik o wysokim poziomie znajomości języka często musi ogarnąć bogactwo jego użycia w różnych dziedzinach i zmiany szybko w nim zachodzące, a wówczas korzystanie z przygotowanych przez kogoś wcześniej opracowań leksykograficznych i gramatycznych nie wystarczy – samodzielna analiza korpusu nawet na najprostszym poziomie staje się niezbędna.

## Podsumowanie

Łatwość tworzenia korpusów językowych z tekstów elektronicznych oraz dostępność oprogramowania zachęca do korzystania z korpusów w nauce i nauczaniu języków. Lingwistyka korpusowa jest fascynującym działem wiedzy, łączącym językoznawstwo i technikę komputerową. Dla czytelników, który zainteresowali się tematem, podaję obszerną bibliografię.

## Bibliografia

1. Duszak A. Gajek E. Okulska U. Red. *Korpusy w angielsko-polskim językoznawstwie kontrastywnym*. Universitas, Kraków 2006.
2. Gajek E. *Standardy przygotowania nauczycieli w zakresie technologii informacyjnej w kontekście kształcenia nauczycieli języków obcych.* Języki obce w szkole, nr 6, 2003.
3. Gajek E. *Edukacja językowa w Unii Europejskiej. Informator i przewodnik internetowy dla nauczycieli.* Fraszka Edukacyjna, Warszawa 2004.
4. Leech G. *Teaching and Language Corpora: a Convergence* [w:] Wichman A. Fligelstone S. McEnery T. Knowles G. Eds. *Teaching and Language Corpora*. Longman, Londyn 1997.
5. Lewandowska-Tomaszczyk B. Red. *Podstawy Językoznawstwa Korpusowego*. Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2005.
6. McEnery T. Wilson A. *Corpus Linguistics An Introduction*. Edinburgh University Press, Edinburgh 2001.
7. Przepiórkowski A. *Korpus IPI PAN*. Instytut Podstaw Informatyki PAN, Warszawa 2004.
8. Sinclair J. *Corpus Concordance Collocation*. Oxford University Press, Oxford 1991.
9. Tribble C. Jones G. *Concordances in the classroom: a resource book for teachers*. Longman, Harlow 1990.
10. Wichman A. Fligelstone S. Mc Enery T. Knowles G. Eds. *Teaching and Language Corpora*. Longman, London 1997.

## Netografia

1. AntConc [dostęp 5 lipca 2007: http://www.antlab.sci.waseda.ac.jp].
2. Books online [dostęp 10 września 2005: http://digital.library.upenn.edu/books].
3. British National Corpus [dostęp 5 lipca 2007: http://sara.natcorp.ox.ac.uk/lookup.html].
4. EVA English Vocabulary Assistant [dostęp 5 lipca 2007: http://poets.notredame.ac.jp/cgi-bin/wn].
5. EVA English Vocabulary Helper [dostęp 10 września 2005: http://poets.notredame.ac.jp/cgi-bin/wn].
6. Frequency [dostęp 5 lipca 2007: http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx].
7. Glossanet [dostęp 10 września 2005: http://glossa.fltr.ucl.ac.be].
8. Korpus IPI PAN [dostęp 10 września 2005: www.korpus.pl].
9. Korpus PELCRA [dostęp 10 września 2005: http://korpus.ia.uni.lodz.pl].
10. Korpus PWN [dostęp 10 września 2005: http://www.pwn.com.pl].
11. KwicFinder [dostęp 10 września 2005: http://www.kwicfinder.com/KWiCFinder.html].
12. Lexware Culler [dostęp 10 września 2005: http://82.182.103.45/lexware/concord/culler.html].
13. MonoConc 2.2. Demo [dostęp 5 lipca 2007: http://www.camsoftpartners.co.uk/monoconc.htm].
14. Online newspapers [dostęp 10 września 2005: http://www.onlinenewspapers.com].
15. Online newspapers [dostęp 5 lipca 2007: http://www.onlinenewspapers.com].
16. Phrases in English [dostęp 10 września 2005: http://pie.usna.edu].
17. Range [dostęp 5 lipca 2007, http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx].

[2] Leech G. *Teaching and Language Corpora: a Convergence* [w:] Wichman A. Fligelstone S. McEnery T. Knowles G. Eds. *Teaching and Language Corpora*. Longman, London 1997.

18. Rusiecki J. *Context, Concordance, And What Next? Suggestions For Computer-Assisted Teaching Of Reading in a Foreign Language. Teaching English with Technology* [dostęp 10 września 2005: http://www.iatefl.org.pl/call/callnl8.htm].
19. SARA [dostęp 5 lipca 2007: http://sara.natcorp.ox.ac.uk/lookup.html].
20. Simple Concordance Program [dostęp 5 lipca 2007: http://www.textworld.com/scp].
21. The online books page [dostęp 5 lipca 2007: http://digital.library.upenn.edu/books].
22. Think Map Visualthesaurus [dostęp 10 września 2005: http://www.visualthesaurus.com].
23. Thinkmap Visual Thesaurus [dostęp 5 lipca 2007: http://www.visualthesaurus.com].
24. University of Michigan Corpus of Academic Spoken English [dostęp 10 września 2005: http://www.hti.umich.edu/m/micase].
25. VIEW [dostęp 5 lipca 2007: http://corpus.byu.edu/bnc].
26. Web Concordancer [dostęp 10 września 2005: http://www.edict.com.hk/concordance].
27. WebConc [dostęp 10 września 2005: http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?sprache=en&art=google].
28. WebCorp [dostęp 10 września 2005: http://www.webcorp.org.uk/index.html].
29. WordNet 2.1 Vocabulary Helper [dostęp 10 września 2005: http://poets.notredame.ac.jp/cgi-bin/wn].
30. WordNet online 3.0 [dostęp 5 lipca 2007: http://wordnet.princeton.edu/perl/webwn].
31. WordSmith Tools Demo 4.0 [dostęp 5 lipca 2007: http://www.lexically.net/wordsmith/version4].

**Autorka jest adiunktem w Instytucie Lingwistyki Stosowanej Uniwersytetu Warszawskiego i nauczycielem konsultantem w Ośrodku Edukacji Informatycznej i Zastosowań Komputerów w Warszawie**

*Ważne jest by nigdy nie przestać pytać.*
*Ciekawość nie istnieje bez przyczyny.*

*Wystarczy więc,*

*jeśli spróbujemy zrozumieć*
*choć trochę tej tajemnicy każdego dnia.*
*Nigdy nie trać świętej ciekawości.*
*Kto nie potrafi pytać nie potrafi żyć.*

Albert Einstein