Mirosław Bańko
(Uniwersytet Warszawski)

Obrazy Google jako źródło informacji lingwistycznej[1]

1. Wprowadzenie

Funkcja o nazwie Grafika wyszukiwarki internetowej Google udostępnia zdjęcia i inne pliki graficzne zindeksowane na stronach WWW i przyporządkowane słowom, np. podpisom pod zdjęciami. Dzięki niej można, wychodząc od słów, znajdować powiązane z nimi obrazy. Na przykład kwerenda *koty syjamskie* wyświetla galerię syjamskich kotów.

Wyszukiwarka obrazów Google może mieć różne zastosowania. Miliony internautów korzystają z niej, aby znaleźć potrzebne zdjęcie. Osoby uczące się języków obcych mogą nieraz skuteczniej sprawdzić znaczenie słowa za pomocą obrazów Google niż słownika. Dwaj brytyjscy artyści – jak donosi witryna amerykańskiego stowarzyszenia leksykografów Dictionary Society of North America – ułożyli słownik obrazkowy, zastępując każde kolejne hasło standardowego słownika pierwszym z obrazów na liście wyników wyszukiwarki[2]. Obrazy Google posłużyły też do badań naukowych, np. nad automatyczną kategoryzacją pojęć (Fergus i in. 2005).

W niniejszym artykule rozpatrujemy obrazy Google jako źródło informacji do badań nad językiem. Wyszukiwanie w Internecie obrazów za pomocą przypisanych im słów można traktować jako rozszerzenie wyszukiwania informacji tekstowych, a ponieważ Internet bywa wykorzystywany jako korpus językowy (Andrzejczuk, Czupryniak 2008, Grafenstette 2002), uwzględnienie w badaniach nad językiem informacji ikonograficznych wydaje się krokiem naturalnym i potencjalnie korzystnym. Dzięki obrazom przypisanym do danego słowa jego charakterystyka funkcjonalna staje się pełniejsza. Można

[1]Artykuł powstał w ramach projektu „Recepcja i adaptacja wyrazów obcych w języku polskim i czeskim. Uwarunkowania językowe, psychologiczne i historyczno-kulturowe", finansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji nr DEC-2011/03/B/HS2/02279. Zob. witrynę projektu pod adresem: http://www.approval.uw.edu.pl.

[2] http://www.dictionarysociety.com/2012/05/dictionary-from-google-image-search.html, dostęp: 24.12.2012.

powiedzieć wręcz, że w pełnym opisie funkcjonalnym wyrazu – obejmującym jego cechy desygnacyjne, konotacyjne, kolokacyjne, emotywne, stylistyczne, etymologiczne, strukturalne, gramatyczne i in. – nie powinno zabraknąć miejsca na charakterystykę ikonograficzną, czyli informację o tym, co o danym wyrazie mówią obrazy z nim skojarzone.

Możliwości, jakie otwiera wyszukiwarka obrazów w badaniach nad językiem, są szerokie: od sprawdzenia referentów nazwy i ustalenia jej zakresu przez ocenę względnej częstości znaczeń słowa polisemicznego aż po badanie konotacji semantycznych słów. Badania porównawcze, np. ocena stopnia semantycznej bliskości synonimów albo stopnia adekwatności odpowiedników różnojęzycznych, to jeszcze jedno przykładowe zastosowanie. Nigdy dotąd językoznawcy nie dysponowali tak łatwym w użyciu narzędziem kojarzącym znak z jego referentem. W tej sytuacji zaskakujące jest, że dotąd z niego nie korzystają[3].

W niniejszym artykule omawiamy zastosowanie obrazów Google w opisie szeroko rozumianego znaczenia leksykalnego. W szczególności interesuje nas możliwość wykorzystania informacji ikonograficznych w porównawczej analizie wyrazów bliskoznacznych. Wybór ten ma związek z realizowanym obecnie projektem badawczym (zob. przyp. 1), w którym jedno z zadań do wykonania polega na sprawdzeniu, jak zapożyczenia na pozór zbyteczne – niemotywowane potrzebami nominatywnymi – adaptują się w języku-biorcy, tzn. jak ich znaczenia stabilizują się w opozycji do znaczeń ich rodzimych synonimów. Za wykorzystaniem obrazów Google do tego celu przemawia to, że analiza obrazów zindeksowanych na stronach WWW pomaga uchwycić różnice między synonimami, które trudno zidentyfikować w inny sposób.

Wyjdziemy od wyjaśnienia, jak Google przypisuje obrazy do słów i jak je szereguje w wynikach wyszukiwań. Kwestia ta ma praktyczne znaczenie dla wszystkich, którzy chcą, aby ilustracje zamieszczone przez nich w sieci znalazły się wysoko na liście wyników – czy to z powodów komercyjnych, czy ambicjonalnych. Dla badaczy języka szczegóły dotyczące powiązania obrazów z tekstem są też istotne, gdyż pomagają zrozumieć, co właściwie znajduje wyszukiwarka obrazów, a także trafniej formułować kwerendy i lepiej interpretować wyniki.

Następnie odniesiemy się do pytania, w jaki sposób informacje ikonograficzne mogą pomóc w semantycznym opisie wyrazów. Jest rzeczą ważną, aby zrozumieć, że obrazy Google nie tylko pomagają ustalić zakres danej nazwy, czyli jej znaczenie desygnacyjne, ale też zwykle wnoszą coś do opisu innych komponentów jej szeroko rozumianego znaczenia.

W centralnej części tekstu omówimy wyniki kilku przykładowych kwerend. Pierwsza dotyczy wyrazu wieloznacznego i nasuwa pytanie o przydatność obrazów Google w ocenie względnej częstości znaczeń. Następne kwerendy mają za zadanie pokazać, że informacje ikonograficzne pozwalają czasem szybko ujawnić takie składniki znaczenia wyrazu, które w inny sposób są trudne do zauważenia, bądź też dostarczają dodatkowych argumentów za czymś, co wynika z analizy innych źródeł, np. danych korpusowych.

[3] Nie udało nam się natrafić na żaden artykuł, polski ani obcy, dotyczący zastosowania wyszukiwarek obrazów w badaniach nad językiem.

Artykuł ma na celu wstępne rozpoznanie problematyki i zwrócenie uwagi na łatwo dostępne, a dotychczas mało wykorzystywane źródło informacji potencjalnie interesującej dla lingwistów. Wnioski z tego rekonesansu znajdują się w końcowej części tekstu.

2. Jak Google przypisuje obrazy do słów?

Podstawowe informacje o tym, jak wyszukiwarka Google przypisuje obrazy do słów, można znaleźć na stronie internetowej Google[4]. Są to wskazówki dla webmasterów i innych osób zainteresowanych tym, by obrazy umieszczane przez nie w sieci były pozycjonowane wysoko w wynikach wyszukiwań.

Z podanych informacji wynika, że przypisanie odbywa się na podstawie co najmniej kilku kryteriów. Po pierwsze, algorytm kojarzący obrazy ze słowami bierze pod uwagę nazwy plików graficznych i stara się z nich wydobyć zawartość werbalną. Po drugie, wykorzystywany jest tzw. tekst alternatywny, opisany atrybutem *alt* w kodzie strony, a używany do opisu zawartości plików graficznych. Po trzecie, uwzględniany jest tzw. tekst zakotwiczenia, czyli tekst występujący w linkach do danej strony znajdujących się na innych stronach WWW. Po czwarte, algorytm uwzględnia elementy tekstowe bliskie obrazowi i prawdopodobnie związane z nim treścią, zwłaszcza tytuły i podpisy.

Obrazy zindeksowane przez Google są układane w porządku domniemanej trafności: na samej górze te, które prawdopodobnie najlepiej odpowiadają oczekiwaniom osoby wykonującej kwerendę. Mówiąc ściślej, pozycja danej strony na liście wyników jest tym wyższa, im więcej linków prowadzi do niej z innych stron WWW i im wyższa jest pozycja owych stron w rankingu[5]. Algorytm pozycjonujący strony (tzw. *PageRank*) działa więc według podobnych zasad, jakimi kierują się ludzie w ocenie wartości źródeł, np. materiałów bibliograficznych: im więcej cytowań jakiejś pracy naukowej i im ważniejsze są prace, które ją cytują, tym ważniejsza wydaje się praca w nich cytowana.

Dodatkowo Google stosuje wiele innych kryteriów pozycjonowania stron – mówi się, że ponad 250 – ale większość z nich trzyma w tajemnicy, aby utrudnić działania spamerów, którzy mogliby wpływać na wyniki kwerend i w ten sposób dać przewagę konkurencyjnym wyszukiwarkom[6]. Brak pełnej jawności kryteriów pozycjonowania obrazów – w połączeniu z groźbą zakłócania wyników przez sprytnych programistów – może budzić nieufność do korzystania z funkcji Grafika w badaniach nad językiem. Zarazem jednak, nawet gdyby firma ujawniła wszystkie kryteria, większość z nas i tak nie potrafiłaby z wiedzy tej uczynić żadnego użytku. Nie należy więc z góry przekreślać wartości diagnostycznej obrazów Google tylko dlatego, że niektóre drugorzędne czynniki rządzące ich wyborem są nieznane.

Nie warto także przeceniać efektownych doniesień medialnych, które podrywają zaufanie do wyszukiwarki. Na przykład portal Slate (po polsku ‘zmieszać z błotem’) pod datą 10.10.2012 informuje, że w odpowiedzi na kwerendę *completely wrong* Google

[4] http://support.google.com/webmasters/bin/answer.py?hl=pl&answer=114016, dostęp: 24.12.2011.

[5] http://en.wikipedia.org/wiki/PageRank, dostęp: 24.12.2012.

[6] http://en.wikipedia.org/wiki/Google_search, dostęp: 24.12.2012.

wyświetla galerię fotografii Mitta Romneya, kandydata na prezydenta USA[7]. Nie ma w tym jednak żadnego oszustwa ani złośliwego działania: Romney sam określił swoją wcześniejszą wypowiedź, po której stracił w sondażach, jako *completely wrong* 'całkowicie błędną', a internauci, cytując ją wielokrotnie, mimowolnie spowodowali jej wypromowanie na listach wyników.

Slate zauważa trafnie, że algorytm pozycjonowania wyników stosowany przez Google działa jak samonapędzający się mechanizm: im więcej ludzi cytuje jakąś stronę i daje link do niej, tym bardziej rośnie jej pozycja w rankingach, co czyni ją jeszcze bardziej popularną, i tak dalej. Mechanizm ten powoduje okresowe wzrosty popularności niektórych stron i może na pewien czas utrudniać korzystanie z obrazów Google do badań nad znaczeniem słów (np. gdyby jakaś publicznie znana osoba wypowiedziała się na temat kotów syjamskich, to wyszukiwarka Google mogłaby przez pewien czas pokazywać jej twarz na samej górze listy wyników kwerendy *koty syjamskie*). Takie okresowe fluktuacje można jednak zneutralizować przez filtrowanie wyników. W szczególności Google pozwala ograniczyć wyniki do stron zindeksowanych w określonym czasie i w ten sposób wykluczyć wydarzenie medialne, które mogło spowodować tymczasowe zaburzenie długookresowych tendencji.

W praktyce o użyteczności obrazów Google językoznawca najlepiej przekona się sam za pomocą przykładowych kwerend. Wśród wyników znajdzie zazwyczaj odpowiedzi nietrafne, ale te może zignorować albo wyeliminować przez odpowiednie filtrowanie wyników. Zanim przejdziemy do omówienia konkretnych przykładów, które będą związane z opisem znaczenia wyrazów, należy się odnieść do pytania o status informacji ikonograficznej w opisie semantycznym.

3. Miejsce informacji ikonograficznej w opisie znaczenia

Na gruncie językoznawstwa, filozofii i psychologii istnieje wiele koncepcji znaczenia, które mogą być podstawą konkretnych metod jego opisu (Grzegorczykowa 2002: 13–28). Istnieje też długa tradycja opisu znaczeń w słownikach (częściowo autonomiczna, czyli rozwijająca się niezależnie od wyżej wymienionych dyscyplin badawczych). Literatura na temat znaczenia jest ogromna i odniesienie się tu do niej w całości w związku z pytaniem o rolę informacji ikonograficznej w opisie znaczeń byłoby niemożliwe. Ograniczymy się zatem do kilku ogólnych uwag.

Przede wszystkim należy przypomnieć, że znaczenie jest czymś innym niż referencja. Znaczenie istnieje w języku i jest względnie trwałe, podczas gdy referencja istnieje w tekście i zmienia się zależnie od wypowiedzi. Znaczenie przysługuje znakom ujętym *in abstracto* i może być różnie koncypowane, zależnie od orientacji metodologicznej badacza, np. jako zbiór desygnatów znaku, zbiór ich cech wspólnych, a jednocześnie istotnych, wyobrażenie typowego desygnatu, warunki prawdziwości zdania, klasa typowych kontekstów znaku, klasa znaków synonimicznych, a nawet określony stan komórek nerwowych w mózgu nadawcy. Referencja natomiast wiąże konkretne użycie znaku

[7] http://www.slate.com/blogs/future_tense/2012/10/10/google_image_search_for_completely_wrong_returns_mitt_romney_photos.html, dostęp: 24.12.2012.

w określonej wypowiedzi z jego referentem, czyli obiektem, do którego się on odnosi. Ów obiekt może być przedmiotem materialnym, osobą, sytuacją, miejscem, relacją itp.

W typowej kwerendzie – takiej jak *koty syjamskie* – obrazy Google to wyobrażenia referentów użytej nazwy, a nie reprezentacje jej znaczenia. Przez obserwację referentów można jednak wyciągać wnioski o zakresie i o znaczeniu wyrazu. Sytuacja badacza zbliża się tu do sytuacji dziecka uczącego się języka ojczystego: zarówno badacz, jak i dziecko polegają do pewnego stopnia na wnioskowaniu indukcyjnym, czyli dokonują kategoryzacji na podstawie przykładów, z którymi się zetknęli. Nie umniejszając wagi innego rodzaju informacji – zwłaszcza kontekstu wyrazowego i sytuacyjnego, eksplicytnych instrukcji ze strony innych osób, a także definicji słownikowych – nie można referentom nazwy odmówić wartości diagnostycznej, gdy chodzi o ustalenie zakresu i znaczenia desygnacyjnego wyrazu.

Wartość obrazów Google na tym jednak się nie kończy. Informują one także o pozadesygnacyjnych elementach znaczenia wyrazu, np. o cechach kojarzonych z jego typowym referentem, ale niekoniecznych, a czasem nawet nieprzysługujących mu obiektywnie, określanych zwykle jako konotacja (Bartmiński, red., 1988). W związku z tym obrazy Google mogłyby być źródłem do badań nad tzw. językowym obrazem świata, utrwalonym w języku i udokumentowanym w różnych przekazach kulturowych (Bartmiński, red., 1990, krytyczna analiza: Kiklewicz, Wilczewski 2011*).

O wartości materiałów ikonograficznych w analizie semantycznej wspomina Anna Wierzbicka (1993), analizując znaczenie słowa *mysz* w artykule poświęconym definiowaniu nazw zwierząt. Choć nie wpisuje się otwarcie w program badań nad językowym obrazem świata, metodologicznie jest mu bliska, gdy znaczenie wyrazu ustala na podstawie takich przesłanek, jak jego znaczenia przenośne, związki frazeologiczne przezeń fundowane, wyrazy pochodne i wyrazy powiązane tematycznie. Dodatkowo postuluje wykorzystanie „pozajęzykowych dowodów etnograficznych", np. schematycznych rysunków i popularnych zabaw dziecięcych. Część badaczy z pewnością uzna, że tylko niektóre elementy zaproponowanego przez nią opisu semantycznego *myszy* należą do znaczenia desygnacyjnego tej nazwy, pozostałe zaś mają inny status. To właśnie chcemy podkreślić: informacje ikonograficzne mogą wnieść coś cennego do różnych komponentów szeroko rozumianego znaczenia wyrazu.

Aby lepiej docenić potencjalną wartość obrazów Google, można dokonać tu przeglądu różnych składników charakterystyki funkcjonalnej wyrazu. Geoffrey Leech (1974: 26) wyodrębnia ich siedem, nazywając je wszystkie „znaczeniami". Na pierwszym miejscu wskazuje znaczenie konceptualne, określane też przezeń jako logiczne, kognitywne lub denotacyjne. Po nim wylicza pięć innych komponentów szeroko rozumianego znaczenia, ujmowanych łącznie jako asocjacyjne. Są to kolejno: znaczenie konotacyjne, stylistyczne, afektywne, odbite[8] i kolokacyjne. Na koniec autor wymienia

---

* Krytyczna analiza A. Kiklewicza i M. Wilczewskiego doczekała się z kolei krytycznego komentarza Adama Głaza *Prostowanie zwierciadła. Przyczynek do (jeszcze?) niezaistniałej ogólnokrajowej dyskusji nt. kondycji lubelskiej etnolingwistyki*, przyjętego do druku w 68. tomie „Biuletynu Polskiego Towarzystwa Językoznawczego" [przyp. red. – W.Ch.].

[8] W oryginale *reflected meaning*, co dotyczy skojarzeń niesionych przez inne znaczenia tego samego wyrazu. Wyrazistym przykładem może tu być *nowotwór językowy* – w terminologii niektórych lingwistów neu-

znaczenie tematyczne, mające związek z tym, jak dane słowo wpisuje się w strukturę tematyczno-rematyczną wypowiedzi.

Koncepcja ta znalazła naśladowców, np. Renata Przybylska (2003: 182) – nie powołując się na Leecha, ale wyraźnie w duchu jego propozycji – wymienia następujące składniki znaczenia wyrazu: znaczenie przedmiotowe, ekspresywne, asocjacyjne, stylistyczne, gramatyczne i strukturalne. Niektóre z nich, jak się wydaje, pokrywają się ze znaczeniami Leecha i mają tylko inne nazwy, inne jednak są nowe, u Leecha nieobecne. Do wyżej wymienionych można by jeszcze dorzucić znaczenie etymologiczne, jeśli uznać, że nie mieści się ono w znaczeniu strukturalnym.

Nie ma tu miejsca na szczegółowe omówienie tych wszystkich „znaczeń" bądź też – jak byłoby bezpieczniej mówić – składników charakterystyki funkcjonalnej wyrazu. Nie będziemy też zastanawiać się nad pytaniem, czy w pełnej charakterystyce funkcjonalnej wyrazu powinno zawierać się „znaczenie ikonograficzne", obejmujące to, co wynika z analizy obrazów Google. Poprzestaniemy na ostrożniejszym stwierdzeniu: wiele z cech funkcjonalnych wyrazu – czyli „znaczeń", mówiąc językiem Leecha – może znaleźć odzwierciedlenie w obrazach Google. Co za tym idzie, wiele z nich można dzięki obrazom Google opisać trafniej, wnikliwiej lub przynajmniej szybciej. Przykłady omówione niżej powinny starczyć za egzemplifikację.

Na użytek dalszej części artykułu konfigurujemy wyszukiwarkę obrazów tak, aby wynajdywała tylko strony w języku polskim i tylko strony zindeksowane na serwerach w Polsce. Filtr ten nie jest w pełni skuteczny, ponieważ polega na metadanych, o których webmasterzy czasem zapominają. Ponadto jego działanie jest często niewidoczne, tzn. nie wpływa w istotny sposób na wyniki. Aby jednak wykluczyć z wyników wyszukiwania strony obcojęzyczne, na których wystąpiło obce słowo przypadkowo równokształtne z polskim słowem użytym w kwerendzie, warto filtrować wyniki co najmniej ze względu na język.

Pozostałe parametry działania wyszukiwarki w omawianych niżej kwerendach są zgodne z konfiguracją domyślną. Wszystkie kwerendy wykonano bez ujmowania słowa szukanego w cudzysłów, co oznacza, że wyszukiwarka uwzględniała formy odmiany wyrazu, a także niektóre wyrazy pochodne. Wyrywkowo sprawdzano wyniki za pomocą analogicznych kwerend z cudzysłowem, wyłączających odmianę, czyli ograniczających wyniki do szukanego słowa. W szczegółach dały się zauważyć różnice, ale zaobserwowane tendencje pozostały te same, co przemawia na korzyść uzyskanych danych.

4. Mysz i myszka

Zaczniemy od *myszy* – przykładu rozpatrywanego w cytowanym artykule Wierzbickiej. Słowo to ma kilka znaczeń, a ponieważ wyszukiwarka nie przeprowadza analizy semantycznej, na liście wyników obrazy odpowiadające różnym znaczeniom są wymieszane. Zjawisko to można ograniczyć, budując bardziej precyzyjne kwerendy,

tralny synonim *neologizmu*. Trudno jednak o neutralną percepcję tej nazwy, skoro ewokuje ona skojarzenie z *nowotworem* w znaczeniu medycznym. Nazwą neutralną jest *neologizm*, a nazwą odbieraną pozytywnie – *innowacja językowa*.

z uwzględnieniem wyrazów, które powinny lub które nie powinny występować na tej samej stronie. Można też, korzystając z funkcji wyszukiwania przyrostowego Google[9], wybrać jeden z typowych kontekstów, w których pojawia się słowo kluczowe. Nas jednak interesować teraz będzie ocena względnej frekwencji znaczeń za pomocą wyszukiwarki obrazów, toteż nie będziemy eliminować wieloznaczności z wyników wyszukiwania.

Dziesięć pierwszych obrazów zindeksowanych przez Google dla kwerendy *mysz* przedstawia zwierzę (czasem jako obiekt drugorzędny, np. z większym od niej kotem), a dopiero na dalszych pozycjach pojawia się mysz komputerowa na zmianę z myszą – zwierzęciem. Wśród stu pierwszych obrazów mysz komputerowa przeważa w stosunku 60 do 40, a w dalszej części listy przewaga jej jeszcze rośnie[10]. Podobnie jest, gdy analizuje się zawartość tekstową stron WWW – na początku listy wyników przeważa pierwsze słownikowe znaczenie *myszy*, ale dalej przewagę zdobywa znaczenie drugie, przenośne[11]. Natomiast w Narodowym Korpusie Języka Polskiego (dalej: NKJP) jest na odwrót, przewagę ma *mysz* – zwierzę nad *myszą* – częścią komputera[12]. Różnica ta nie dziwi, jeśli zważyć, że internauci są jako społeczność bardziej zainteresowani urządzeniami komputerowymi niż przeciętny użytkownik polszczyzny i że naprzeciw tym zainteresowaniom wychodzą właściciele sklepów internetowych oraz innych witryn dotyczących sprzętu komputerowego. Interesujące natomiast jest to, że *mysz* – zwierzę mimo wszystko przeważa w górnej części listy wyników wyszukiwarki, zarówno w obrazach Google, jak i w ogóle na stronach WWW.

Dla porównania dodajmy, że lista obrazów Google dla słowa *myszka* zaczyna się od zdjęć myszy komputerowej, dalej jednak ich przewaga spada i coraz liczniej widoczne są żartobliwe rysunki myszy – zwierzęcia (przeważnie w stylu disnejowskim), przeplatane zdjęciami prawdziwych myszy. Rozkład znaczeń słowa *myszka* w obrazach Google jest więc zasadniczo odwrotny niż rozkład znaczeń słowa *mysz*. Natomiast w NKJP żywych *myszek* jest więcej niż komputerowych, co znaczy, że w korpusie hierarchia dwóch głównych znaczeń *myszy* i *myszki* jest identyczna.

Różnica między korpusem a Internetem w zakresie względnej frekwencji znaczeń *myszy* i *myszki* może wynikać z odmiennego charakteru tych zasobów lub ze sposobu, w jaki tworzone są wyniki kwerend. Przypomnijmy, że Google szereguje wyniki według domniemanej trafności czy też istotności – o miejscu danej strony na liście decyduje liczba linków do niej i ranga stron, na których te linki się znajdują. Lista wyników kwerendy w typowym korpusie językowym (także w NKJP) obejmuje natomiast wszystkie konteksty zgodne z zapytaniem, nie uhierarchizowane według istotności. Wprawdzie można skonstruować narzędzia do automatycznego wyboru cytatów najtraf-

[9] Na stronach wyszukiwarki jest ono nazywane wyszukiwaniem dynamicznym.

[10] Dane liczbowe z Internetu, tu i dalej w artykule, pochodzą z końca grudnia 2012 roku. Ponieważ Google indeksuje strony nieprzerwanie, a Internet jest zasobem dynamicznym, wyniki te w każdej chwili mogą się zmienić. Zakładamy jednak, że zaobserwowane tendencje utrzymują się w dłuższym czasie.

[11] Ściśle biorąc, *mysz (komputerowa)* jest w polszczyźnie wynikiem zapożyczenia semantycznego z angielskiego *mouse*, natomiast w angielskim komputerowe znaczenie tego słowa powstało na drodze metaforyzacji.

[12] Dane z NKJP, tu i dalej, pochodzą z podkorpusu zrównoważonego.

niejszych z jakiegoś punktu widzenia (np. Kilgarriff i in. 2008 przedstawiają narzędzie selekcjonujące z korpusu cytaty użyteczne dla leksykografa), ale kryteria takiej selekcji będą jakościowo inne niż stosowane przez Google przy pozycjonowaniu stron WWW.

Nasuwa się pytanie: czy w wypadku słów wieloznacznych, do jakich należy *mysz*, proporcje na liście obrazów Google są wiarygodnym świadectwem frekwencji ich znaczeń w tekstach? Innym sposobem oceny frekwencji znaczeń może być kwerenda w korpusie językowym, powstaje więc drugie pytanie: do jakich celów lepiej używać jednej metody, a do jakich drugiej? Kwestie te musimy tutaj zostawić bez odpowiedzi.

5. Traktor i ciągnik

Słowniki polskie ujmują *traktor* i *ciągnik* jako wyrazy synonimiczne. Nie wiadomo, czy wynika to z ograniczenia opisu do języka ogólnego, w którym zakresy wymienionych wyrazów istotnie się pokrywają, czy też z nieuwagi. Dość, że tradycja ujmowania ich jako synonimów jest dość długa, sięga co najmniej SIJP Arcta (wyd. 3, 1929). Ciekawe, że *ciągnik* odesłano tam do nieużywanego już dziś wyrazu *ciągówka*, co dowodzi, że w użyciu były wówczas trzy nazwy synonimiczne. Nieco wcześniej, bo na początku lat 20., założoną jeszcze pod koniec XIX wieku fabrykę w Ursusie przemianowano na Fabrykę Silników i Traktorów Ursus, wtedy też wyjechały z niej pierwsze „ciągówki". Słowo *ciągnik* pojawiło się w nazwie zakładów w Ursusie dopiero pół wieku później (1972 – Zrzeszenie Przemysłu Ciągnikowego Ursus, 2001 – Fabryka Ciągników Ursus)[13].

Fachową wiedzę o *traktorze* i *ciągniku* przekazują encyklopedie. Można się z nich dowiedzieć, że znaczenia wymienionych wyrazów pozostają w stosunku inkluzji: każdy traktor jest ciągnikiem, ale tylko ciągniki rolnicze są nazywane traktorami. Ciągniki o innym przeznaczeniu, mianowicie drogowe, leśne czy artyleryjskie, nie są traktorami ani ich nie przypominają.

Ograniczając dalszą analizę do języka ogólnego, a więc wyłączając fachowe użycia słowa *ciągnik*, postanowiliśmy sprawdzić, jakie różnice zachodzą między *traktorem* a *ciągnikiem* (*rolniczym*), jeśli wyjść poza ich znaczenie desygnacyjne. Pierwsza próba polegała na zbadaniu ich kolokacji w NKJP. Kolokator wbudowany w wyszukiwarkę PELCRA nie znalazł między nimi żadnej istotnej różnicy – z wyjątkiem hasła „Kobiety na traktory", które przypomina o polityce rolnej i społecznej we wczesnych latach PRL i nie ma odpowiednika ze słowem *ciągnik*. Bardziej przydatna okazała się funkcja Profil PELCR-y, która podaje częstość występowania danego słowa w określonej kategorii tekstów. Z porównania profili wynika, że w przeliczeniu na milion słów *ciągnik* jest dwukrotnie częstszy od *traktora* w prasie, natomiast dwukrotnie rzadszy w książkach, m.in. w beletrystyce. Dane te wydają się zgodne z intuicją, która podpowiada, że *ciągnik* to słowo oficjalne, a *traktor* – neutralne, wspólne różnym odmianom polszczyzny.

Tyle korpus, a co nowego wnosi analiza obrazów Google? Na liście wyników kwerendy *traktor* dużo jest programów komputerowych do miksowania muzyki, nazywanych „traktorami" od nazwy własnej, pod którą są sprzedawane. Większość z nich moż-

[13] http://www.ursus.com.pl/Historia, dostęp: 28.01.2013.

na wyeliminować, jeśli kwerendę uzupełni się o słowo *muzyka* poprzedzone minusem – powoduje to wykluczenie stron zawierających słowo *muzyka* z listy wyników. Nawet jednak i bez tego uściślenia uderzające jest, że wśród obrazów przypisanych słowu *traktor* dość licznie, i to już od samego początku listy, występują traktory zabawki. Zresztą dostępna w wyszukiwarce Google funkcja wyszukiwania przyrostowego podpowiada, gdy wpisuje się słowo kluczowe, frazę *traktory zabawki*, co równie dobitnie pokazuje rangę tej kategorii obiektów na stronach WWW.

Dla słowa *ciągnik* funkcja wyszukiwania przyrostowego nie podpowiada frazy *ciągniki zabawki*, a na liście obrazów kwerendy *ciągnik* zabawki są wyraźnie mniej liczne niż poprzednio. I ten wynik zdaje się zgodny z intuicją: studenci poloniści, którzy zapoznali się z nim na seminarium, uznali go za oczywisty, por. symptomatyczną wypowiedź studentki: „Traktorem mogłabym się bawić, ale ciągnikiem?!" Zgodność danych ikonograficznych z intuicją można było jednak ocenić dopiero *ex post*, gdyż żaden z uczestników zajęć nie przewidział różnicy, którą ujawniły obrazy Google.

Z powyższego nie należy oczywiście wyciągać wniosku, że możliwość bycia zabawką jest jakąś szczególnie istotną cechą semantyczną słowa *traktor*. Polisemia typu „przedmiot – jego imitacja służąca do zabawy" jest regularna i seryjna, ale fakt, że z dwóch nazw synonimicznych jedna zdecydowanie częściej służy jako nazwa zabawki, sugeruje istnienie ważniejszej różnicy między *traktorem* a *ciągnikiem*, dotyczącej dystrybucji tych wyrazów w języku i ich charakterystyki stylistycznej. Przykład ten pokazuje, że jeden rzut oka na obrazy Google może wystarczyć, aby nadać analizie potencjalnie obiecujący kierunek.

6. Kartofel i ziemniak

W kilku starszych słownikach – SWil, SW, SJPDor – *ziemniak* został odesłany do *kartofla* bądź otrzymał definicję synonimiczną, która jest też rodzajem odesłania. W dwóch nowszych – ISJP, USJP – jest na odwrót: *kartofel* podano z definicją synonimiczną. Zmiana ta świadczy o większym obecnie upowszechnieniu wyrazu rodzimego, czego dodatkowym przejawem jest frekwencja w NKJP: *ziemniaki* pojawiają się tu dwa i pół raza częściej od *kartofli*. Można też zauważyć, że *ziemniaki* to nazwa handlowa, przeważająca w sklepach i w restauracjach, w kartach dań. *Kartofle* natomiast stały się słowem raczej potocznym.

Jako częstszy ma *ziemniak* w NKJP więcej kolokatów. W szczególności występuje w kontekście większej liczby nazw innych potraw, co sprzyja postrzeganiu *ziemniaków* jako mniej pospolitych. Na przykład zarówno *kartofle*, jak i *ziemniaki* kolokator PELCR-y pokazuje w kontekście *buraków*, *cebuli*, *kapusty* i *marchwi*, ale tylko *ziemniaki* w sąsiedztwie *brokułów*, *fasoli*, *papryki*, *pomidorów* i *selera*. Zarówno *ziemniaki*, jak i *kartofle* można *obierać*, *skrobać*, *gotować*, *smażyć*, *piec*, *tłuc* i oczywiście *jeść*, ale tylko *ziemniaki kroi się* w kostkę lub plasterki albo *trze*, tylko *ziemniaki* można *polać* lub *posypać* czymś i tylko z przymiotnikiem *ziemniaczany* występują słowa *sałatka* i *purée*. Obraz kolokacyjny *ziemniaków* jest wytworniejszy: to samo warzywo okazuje się składnikiem kuchni bardziej wykwintnej, gdy oznaczane jest słowem *ziemniak*, niż gdy występuje jako *kartofel*.

Wrażenie pospolitości *kartofla* pogłębiają derywaty: potocznej *kartoflówce* i *kartoflance* nie odpowiada derywat prosty od słowa *ziemniak*, lecz zestawienie *zupa ziemniaczana*, które samą swoją budową sugeruje, że chodzi o kuchnię mniej pospolitą (*zupa kartoflana* też występuje w NKJP, ale dwukrotnie rzadziej i głównie w beletrystyce, a nie – jak *zupa ziemniaczana* – głównie w prasie).

Jest ponadto *kartoflowaty nos*, czyli *nos jak kartofel* (niekiedy po prostu: *kartofel*) – duży, nieforemny, niełady, „plebejski"; por. cytat z NKJP: „[Czeczot] Preferował zdecydowanie plebejuszy, swojskich, jurnych, takich trochę kartoflowatych w kształtach, przyrośniętych do ziemi..." (*Polityka*). I jest inwektywa *kartofel*, odnoszona – podobnie jak *burak* – do kogoś, kogo uważa się za osobę nieobytą, zacofaną, przynoszącą wstyd innym.

Ten szkicowy z konieczności obraz *kartofla* i *ziemniaka*, oparty na danych słownikowych i korpusowych, zestawimy teraz z tym, co pokazują obrazy Google. Lista wyników dla słowa *ziemniak* obejmuje przede wszystkim nieobrane warzywa, takie, jakie widać w sklepie. Dość liczne są także wymyślne potrawy z ziemniaków. Rzadkie natomiast są wizerunki twarzy i żartobliwe rysunki, w szczególności karykatury.

Całkiem inaczej przedstawia się lista obrazów dla słowa *kartofel*: tu przeważają dziwaczne okazy pojedynczych kartofli, żartobliwe rysunki lub przetworzone komputerowo zdjęcia, w tym portrety (m.in. braci Kaczyńskich, których niemiecka gazeta „Tageszeitung" przyrównała w 2006 roku do kartofli). Lista wyników jest bardziej zróżnicowana, a poszczególne obrazy rzadziej pełnią funkcję czysto ilustracyjną, częściej – ekspresywną. Większość z nich należy odbierać inaczej niż w poprzedniej kwerendzie: nie dosłownie, lecz jako nawiązania, aluzje.

W sumie obrazy Google zindeksowane dla słów *kartofel* i *ziemniak* – podobnie jak w wypadku *traktora* i *ciągnika*, lecz w sposób bardziej złożony i zawikłany – przekazują różnicę w szeroko rozumianym znaczeniu porównywanych wyrazów, obejmującym ich cechy stylistyczne, konotację i aspekty emotywne. Jeśli nawet nie od razu wiadomo, jak tę różnicę zinterpretować, jeden rzut oka wystarczy, aby spostrzec, że ona istnieje, a skoro tak, to jest powód, aby poszukać jej potwierdzenia w innych źródłach, np. korpusowych. Ważne, że informacje ikonograficzne z Internetu i informacje wynikające z innych źródeł wzajemnie się uzupełniają.

7. Podsumowanie

Zamiarem naszym było zwrócenie uwagi na łatwo dostępne źródło informacji o znaczeniu wyrazów, dotychczas mało wykorzystywane przez lingwistów. Zaprezentowane przykłady pokazują, że do obrazów Google warto sięgać dla uzupełnienia lub potwierdzenia wniosków wynikających z analizy innych źródeł, np. słowników, korpusów lub ankiet. Informacje ikonograficzne mogą wzbogacić opis różnych komponentów szeroko rozumianego znaczenia wyrazu – nie tylko znaczenia desygnacyjnego.

Wystarczającym powodem, aby korzystać z obrazów Google w opisie znaczenia wyrazów, jest to, że znaczenie wiąże znak z jego referentem (zgodnie z klasycznym trójkątem oznaczania Ogdena i Richardsa, zob. np. Grzegorczykowa 2002: 15). Mając dostęp do zdjęć lub innych wyobrażeń referentów danej nazwy, a także do obrazów

z nią kojarzonych, można trafniej i pełniej opisać jej znaczenie. Niekiedy jedno spojrzenie na listę wyników wystarczy, aby zauważyć elementy szeroko rozumianego znaczenia, które w analizie słownikowej lub korpusowej są niewidoczne lub nie od razu widoczne.

Nie zawsze funkcja Grafika Google daje wyniki łatwe do zinterpretowania. Obrazy przypisane do rzeczowników konkretnych, a także niektórych innych wyrazów mających wyrazisty desygnat są zwykle bezpośrednią ilustracją ich znaczeń, np. dla *kot* Google pokazuje zdjęcia kotów, dla *pływać* – zdjęcia pływaków i innych osób w wodzie, dla *czarny* – czarne prostokąty i inne ciemne motywy graficzne. W wypadku pozostałych wyrazów interpretacja wyników jest mniej lub bardziej pośrednia, np. *pustka* to głównie puste krajobrazy, puste pomieszczenia i portrety (znaczące jest, że przeważają tu zdjęcia czarno-białe), *przyjaźń* to postaci ludzkie i motywy symboliczne, np. splecione dłonie, *wiedza* to okładki książek oraz pochodzące z nich rysunki i diagramy, a *szybki* to motocykle, pociągi i potrawy, które można szybko przygotować.

Wyniki kwerend są zapośredniczone przez różne elementy szeroko rozumianego znaczenia wyrazu (w tym jego typowe konteksty), a także konwencje znakowe przyjęte w danej kulturze. Dla słowa *czekoladka* wśród zdjęć czekoladowych przysmaków wyszukiwarka pokazuje stosunkowo liczne portrety ciemnoskórych lub opalonych, skąpo ubranych kobiet, dla słowa *łasiczka* – prócz zwierzęcia wyświetla obraz Leonarda da Vinci i jego malarskie lub graficzne trawestacje, dla *Google* – przede wszystkim logotyp wyszukiwarki, dla *Polska* – godło, flagi i mapy, a w dalszej części listy m.in. migawki z meczów piłkarskich z udziałem polskiej drużyny (część tych obrazów jest przyporządkowana w rzeczywistości nie do słowa *Polska*, lecz *polski*, tego efektu nie można jednak uniknąć, gdyż wyszukiwarki internetowe nie ujednoznaczniają słów).

Przyporządkowanie obrazów do słów na stronach WWW nie jest arbitralne – jak w wypadku internetowych archiwów fotograficznych – lecz wynika z ich naturalnego sąsiedztwa. Śledzenie związku obrazów i słów w tym naturalnym, dynamicznym i na pozór chaotycznym – jak samo życie – środowisku pozwala opisać wiele elementów naszego kodu kulturowego, nie zawsze łatwo widocznych. Ponieważ na liście wyników wyszukiwarki Google pierwszeństwo mają strony popularne – ściślej: takie, do których prowadzą liczne linki z innych stron, zwłaszcza też popularnych – obrazy Google dają wgląd w funkcjonowanie słów w kulturze masowej.

W wypadku wyrazów wieloznacznych interpretacja wyników może być utrudniona. Warto rozważyć ujednoznacznienie kwerendy przez uwzględnienie wyrazów, które powinny lub które nie powinny wystąpić na tej samej stronie, co słowo szukane. Można wpisać je ręcznie lub wybrać jedną z kwerend sugerowanych przez mechanizm wyszukiwania przyrostowego. Proporcja obrazów odpowiadających różnym znaczeniom szukanego słowa może być jednak sama w sobie interesująca jako miara rozpowszechnienia znaczeń.

Należy pamiętać o wyborze właściwego języka: wyszukiwarka automatycznie wybiera język zgodny ze swoją wersją językową, więc jeśli chcemy szukać stron w innym języku, trzeba zmienić opcje wyszukiwania lub skorzystać z obcojęzycznej wersji wyszukiwarki. Dodatkowo można ograniczyć wyszukiwanie do stron zindeksowanych na serwerach w określonym kraju.

Aby zniwelować wpływ chwilowych wzrostów popularności niektórych stron, można ograniczyć wyszukiwanie do stron zindeksowanych w określonym przedziale czasu. Wskazówki dotyczące zapytań można znaleźć na stronie Google[14]. Warto wypróbować różne opcje i różne warianty tych samych kwerend.

Bibliografia

Andrzejczuk A., Czupryniak M., 2008, *O wykorzystaniu zasobów internetowych w pracy językoznawcy*, „Polonica", t. 29, s. 189–204.

Bartmiński J., red., 1988, *Konotacja*, Lublin.

Bartmiński J., red., 1990, *Językowy obraz świata*, Lublin.

Fergus R., Fei-Fei L., Perona P., Zisserman A., 2005, *Learning Object Categories from Google's Image Search*, [w:] *Computer Vision in Human-Computer Interaction. International Conference on Computer Vision*, vol. 2, eds. N. Sebe, M.S. Lew, T.S. Huang, Berlin–New York, s. 1816–1823, http://eprints.pascal-network.org/archive/00001128/01/fergus05a.pdf.

Grafenstette G., 2002, *The WWW as a Resource for Lexicography*, [w:] *Lexicography and Natural Language Processing. A Festschrift in Honour of B.S.T. Atkins*, ed. M.-H. Corréard, Grenoble, s. 199–215.

Grzegorczykowa R., 2002, *Wprowadzenie do semantyki językoznawczej*, Warszawa.

ISJP: *Inny słownik języka polskiego*, red. M. Bańko, Warszawa 2000.

Kiklewicz A., Wilczewski M., 2011, *Współczesna lingwistyka kulturowa: zagadnienia dyskusyjne (na marginesie monografii Jerzego Bartmińskiego „Aspects of Cognitive Linguistics")*, „Biuletyn Polskiego Towarzystwa Językoznawczego", t. 67, s. 165–178.

Kilgarriff A., Husak M., McAdam K., Rundell M., Rychlý P., 2008, *GDEX: Automatically Finding Good Dictionary Examples in a Corpus*, [w:] *Proceedings of the XIII EURALEX International Congress*, eds. E. Bernal, J. DeCesaris, Barcelona, s. 425–432.

Leech G., 1974, *Semantics*, Harmondsworth.

NKJP: Narodowy Korpus Języka Polskiego, http://nkjp.pl.

Przybylska R., 2003, *Wstęp do nauki o języku. Podręcznik dla szkół wyższych*, Kraków.

SIJP: *M. Arcta Słownik ilustrowany języka polskiego*, t. 1–2, wyd. 3, Warszawa 1929.

SJPDor: *Słownik języka polskiego PAN*, red. W. Doroszewski, t. 1–11, Warszawa 1958–1969.

SW: *Słownik języka polskiego*, red. J. Karłowicz, A. Kryński, W. Niedźwiedzki, t. 1–8, Warszawa 1900–1927.

SWil: *Słownik języka polskiego*, wydany staraniem i kosztem Maurycego Orgelbranda, t. 1–2, Wilno 1861.

USJP: *Uniwersalny słownik języka polskiego*, red. S. Dubisz, t. 1–6, Warszawa 2003.

Wierzbicka A., 1993, *Nazwy zwierząt*, [w:] *O definicjach i definiowaniu*, red. J. Bartmiński, R. Tokarski, Lublin, s. 251–267.

[14] https://support.google.com/websearch/?hl=pl, dostęp: 24.12.2012.