PRZEGLĄD PSYCHOLOGICZNY, 2012, TOM 55, Nr 2, 195-209

Automatyczna analiza tekstu: polska adaptacja programu LIWC Jamesa Pennebakera

Bartosz Szymczyk*
Wyższa Szkoła Zarządzania i Prawa im. Heleny Chodkowskiej w Warszawie
Wojciech Żakowicz
Uniwersytet Adama Mickiewicza w Poznaniu
Katarzyna Stemplewska-Żakowicz
Wyższa Szkoła Zarządzania i Prawa im. Heleny Chodkowskiej w Warszawie

COMPUTERIZED TEXT ANALYSIS:
POLISH ADAPTATION OF JAMES PENNEBAKER'S LIWC DICTIONARY

Abstract. The article describes work of creating Polish language adaptation of the Linguistic Inquiry and Word Count software, launched originally by James Pennebaker. It gives a brief history of the development of the original dictionary and word categories and a synopsis of the software mechanisms in order to show the conceptual and linguistic work of adapting the dictionary to a different language. Finally, the

* Adres do korespondencji: Bartosz Szymczyk, Instytut Psychologii, Wyższa Szkoła Zarządzania i Prawa im. Heleny Chodkowskiej, Al. Jerozolimskie 200, 02-486 Warszawa; e-mail: bartosz.szymczyk@gmail.com

Wojciech Żakowicz, Uniwersytet Adama Mickiewicza w Poznaniu; e-mail: wojciech.zakowicz@gmail.com

Katarzyna Stemplewska-Żakowicz, Instytut Psychologii, Wyższa Szkoła Zarządzania i Prawa im. Heleny Chodkowskiej, Al. Jerozolimskie 200, 02-486 Warszawa; e-mail: kastem@gmail.com

Granty wewnętrzne w ramach BST Szkoły Wyższej Psychologii Społecznej (2008/2009) oraz Wyższej Szkoły Zarządzania i Prawa im. Heleny Chodkowskiej (2010/2011).

Serdecznie dziękujemy członkom i sympatykom zespołu DIP, w szczególności Kamilowi Jaworskiemu, Małgorzacie Jędrasik-Styła, Annie Gabińskiej i Adrianowi Wójcikowi, którzy na różnych etapach prac nad polską adaptacją słownika przyczynili się do ich postępu.

article shows results of the first two trial studies in which the Polish dictionary was used in its test version. The studies correspond with the research of the LIWC's Spanish language adaptation team and were aimed at verifying the equivalence and (to some initial extend) the validity of the Polish adaptation. The results of the studies are discussed as guidelines for the Polish dictionary fine tuning before launching its official version.

Termin „analiza tekstu" (*text analysis* – por. Silverman, 2007; Stemplewska-Żakowicz, 2009) oznacza analizowanie próbek ustnych lub pisemnych wypowiedzi, które mogą mieć charakter prywatny (listy, dzienniki, pamiętniki), publiczny (artykuły prasowe, dokumenty urzędowe, przemówienia) lub pośredni (blogi, czaty). Analizie można poddawać różne aspekty wypowiedzi: treść, strukturę, zasoby leksykalne, formy gramatyczne i inne.

Ze względu na sposób analizowania można wyróżnić dwa podejścia: ilościowe i jakościowe. Podejście jakościowe, jeśli ma nie być czysto impresyjne i pozbawione jakichkolwiek standardów, wymaga zaangażowania sędziów kompetentnych i zainwestowania dużej ilości czasu i pracy w ich trening – bez tego wskaźniki rzetelności metody są niezadowalające.

Z kolei podejście ilościowe, oferujące analizę całkowicie zobiektywizowaną, wymaga drobiazgowego zliczania wszystkich przypadków wystąpienia w analizowanym tekście określonej kategorii leksykalnej (np. słowa „kobieta" we wszystkich przypadkach), gramatycznej (np. czasowników w czasie przeszłym bądź przyszłym itp.) lub specyficznych struktur (np. współwystępowania określonych par słów). Wykonywanie tego typu analiz przez człowieka jest niezwykle żmudne, jednak dla komputera nie stanowią one żadnej trudności, toteż popularność ilościowej analizy tekstu wzrasta obecnie wraz z rozwojem technologii informatycznych. Ilość przechodzi w jakość: poprzez enumeratywne wyliczenie wszystkich desygnatów możliwe staje się obecnie definiowanie nawet złożonych kategorii semantycznych czy psychologicznych. Powstają coraz bardziej wszechstronne i przyjazne programy komputerowe do przeprowadzania takich analiz, a jednym z nich jest program Linguistic Inquiry and Word Count (LIWC), stworzony przez Jamesa Pennebakera i współpracowników (Pennebaker, Francis, Booth, 2001; Pennebaker, Booth, Francis, 2007).

LIWC – PROGRAM DO AUTOMATYCZNEJ FREKWENCYJNEJ ANALIZY TEKSTU

LIWC to program służący do analizy tekstu metodą zliczeń. Podczas analizy każde słowo w tekście porównywane jest z zawartością wbudowanego słownika. Jeżeli słownik zawiera sczytywany wyraz, jego wystąpienie zostaje zarejestrowane i powoduje automatyczne zaliczenie użycia w tekście określonych kategorii, do których w słowniku LIWC przypisane jest dane słowo (Pennebaker i in., 2001, 2007).

Słowa zawarte w oryginalnym angielskim słowniku LIWC 2007 przypisane są do 76 kategorii językowych i psychologicznych. Dwadzieścia dwie kategorie językowe grupują wyrazy według kryteriów gramatycznych (np. zaimki, słowa w pierwszej osobie liczby pojedynczej, czasowniki w czasie przeszłym), ale także funkcjonalnych (np. negacja lub przekleństwa). Na trzydzieści dwie

kategorie psychologiczne składa się sześć kategorii głównych, zawierających podkategorie. Te kategorie to: (a) procesy społeczne z podkategoriam, takimi jak rodzina, przyjaciele, komunikacja; (b) afekt z podkategoriami pozytywnego i negatywnego afektu; (c) procesy poznawcze, włączając w to słowa wskazujące na wgląd czy przyczynowość; (d) percepcja z podkategoriami widzenie, słyszenie itd.; (e) procesy biologiczne obejmujące stany i funkcje ciała czy seksualność; (f) przestrzeń i czas. Siedem kolejnych kategorii dotyczy dziedzin, takich jak praca, pieniądze czy dom. Do jednego wyrazu w słowniku może być przypisanych kilka z wyżej wymienionych kategorii, np. słowo płakał (*cried*) to (1) czasownik; (2) czas przeszły; (3) afekt; (4) negatywne emocje oraz (5) smutek (Pennebaker i in., 2007).

Pozostałe kategorie dotyczą komunikacyjnych właściwości wypowiedzi (np. słowa-wypełniacze czy chrząknięcia, co szczególnie przydatne w analizie transkrypcji wywiadów czy sesji terapeutycznych) lub sposobu pisania (kategorie związane z interpunkcją). Poza 76 kategoriami, LIWC podaje także statystyki opisujące analizowany tekst, takie jak liczba słów, odsetek słów zawartych w słowniku (stosunek słów wychwyconych przez słownik do wszystkich słów w tekście) czy średnia długość zdań.

Słownik angielski LIWC 2007 liczy 4500 słów lub samych członów, które zakończone są symbolem funkcyjnym asterysku (*). Na przykład człon *hungr** w angielskim słowniku obejmuje słowa takie, jak *hungry*, *hungrier*, *hungries* (Pennebaker i in., 2007).

Wybór słów obecnych w ostatniej wersji angielskiego słownika jest rezultatem procesu jego rozwijania, który obejmował następujące etapy: (1) wybór słów ze skal dotyczących uczuć, np. skali PANAS (Watson, Clark, Tellegen, 1988), wybór ich synonimów z tezaurusa, burze mózgów czy wyszukiwanie słów w słowniku; (2) dwie fazy ocen sędziów kompetentnych, którzy oceniali zgodność słów z kategoriami i wybierali synonimy, a także usuwali słowa, co do których nie było zgodności, po czym oceniali hierarchiczność kategorii; (3) empiryczna ocena psychometrycznych właściwości słownika; (4) odrzucenie słów nie występujących na angielskiej liście frekwencyjnej; (5) końcowa analiza dużej liczby obszernych próbek tekstów, dzięki której dokonano dalszych podziałów kategorialnych i zmian względem poprzednich wersji (zob. Pennebaker i in., 2007).

PRZYKŁADY BADAŃ Z ZASTOSOWANIEM PROGRAMU LIWC

LIWC powstał jako narzędzie, które miało w prosty i efektywny sposób analizować treść wypowiedzi, zebranych w ramach badań nad związkiem między ujawnianiem uczuć a zdrowiem (Pennebaker, Francis, 1996). Wkrótce jednak okazało się, że słowa związane z treścią, choć stanowią 95% słownika języka angielskiego, odpowiadają jedynie za 55% wypowiedzi. Pozostałe 45% wypowiedzi to słowa funkcyjne (zaimki, przedimki, spójniki itp. – Tausczik, Pennebaker, 2010). Współczesna wersja programu LIWC umożliwia identyfikację zarówno słów niosących znaczenie treściowe, jak i tych, które są spoiwem języka.

Znaczenie słów funkcyjnych

W badaniach nad użyciem języka w depresji (Mehl, Pennebaker, 2003; Rude, Gortner, Pennebaker, 2004) osoby z objawami tego zaburzenia różniły się od osób zdrowych częstszym stosowaniem zaimków pierwszoosobowych liczby pojedynczej (ja, moje, mnie), co wskazywało na ich większą koncentrację na własnym ja. Częstość tych zaimków okazała się lepszym wskaźnikiem depresji niż używanie słów opisujących negatywne emocje (Chung, Pennebaker, 2007). Podobne wyniki przyniosła analiza poezji (Stirman, Pennebaker, 2001): twórcy przypuszczalnie w depresji, którzy ostatecznie popełnili samobójstwo, częściej w swych wierszach używali zaimka „ja", niż czynili to poeci bez tendencji samobójczych.

O ile użycie zaimków pierwszoosobowych wskazuje na koncentrację na ja, o tyle używanie innych zaimków, a szczególnie zaimka w pierwszej osobie liczby mnogiej (my, nasze, nam), może wskazywać na zróżnicowane zachowania (np. dystansowanie się vs szukanie związku z innymi; por. Pennebaker, Lay, 2002). Odwoływanie się do innych poprzez zaimki (ona, jego, oni itp.) korelowało w badaniach z adaptacyjnymi sposobami radzenia sobie ze stresem i w rezultacie ze zdrowiem somatycznym (Chung, Pennebaker, 2007).

Sposób użycia zaimków jest też przejawem reakcji na wydarzenia traumatyczne. Analizy dyskusji internetowych zarówno po śmierci księżnej Diany (Stone, Pennebaker, 2002), jak i po jednej z tragedii na uniwersytecie w Teksasie, w której zginęło 12 studentów, ujawniły częstsze używanie zaimków w pierwszej osobie liczby mnogiej. Najwięcej danych na ten temat dostarcza analiza porównawcza wpisów na około 1000 blogach z okresu przed i po 11 września 2011 roku. Okazało się, że około 10 dni po ataku terrorystycznym nastąpił spadek użycia zaimków pierwszoosobowych w liczbie pojedynczej (ja, moje) i jednoczesny wzrost użycia zaimków w pierwszej osobie liczby mnogiej (my, nasze), co wskazuje na nasilone przeżywanie siebie jako części pewnej wspólnoty. Po kilku kolejnych tygodniach sposób użycia zaimków był podobny do tego sprzed ataku (Cohn, Mehl, Pennebaker, 2004).

Jak się okazało sposób użycia zaimków był także dobrym wskaźnikiem kłamstwa i prawdomówności (Newman i in., 2003). Osoby mówiące prawdę częściej personalizowały opowieść, również poprzez użycie zaimków w pierwszej osobie liczby pojedynczej. Używały one także procentowo więcej słów wyłączających (oprócz, ale, pomimo), co wskazuje na większą złożoność poznawczą ich narracji (Pennebaker, King, 1999).

Znaczenie słów niosących treść

Program LIWC pozwala na trafną identyfikację słów dotyczących emocji. Używanie negatywnych emocjonalnie słów okazało się częstsze u osób, które kłamały niż u mówiących prawdę (Newman i in., 2003). Z kolei użycie pozytywnych emocjonalnie słów jest wskaźnikiem zgodności w interakcji między dwoma osobami (Chung, Pennebaker, 2007). Częstsze używanie takich słów – ale tylko przez mężczyzn! – okazało się wskaźnikiem większej satysfakcji ze związku odczuwanej przez oboje partnerów.

Inną zmienną analizowaną w badaniach (Pasupathi, 2007) było użycie czasu. Okazało się, że osoby piszące o przeszłym zdarzeniu, którego wcześniej nikomu nie ujawniały, stosują czas teraźniejszy, inaczej niż w przypadku pi-

sania o czymś, co już zostało ujawnione, kiedy to używany jest czas przeszły. Zdaniem badaczy wskazuje to na niedomknięty charakter wydarzeń nieujawnionych, które zdają się wciąż trwać.

W użyciu czasu przejawia się nasze zainteresowanie rzeczywistością. W przypadku kampanii politycznych te wypowiedzi i reklamy partii, które oceniane są pozytywnie, charakteryzują się większą liczbą czasowników w czasie teraźniejszym i przyszłym, a te, których ocena jest negatywna, częściej odwołują się do przeszłości (Gunsch i in., 2000).

Zbadano też rolę słów oznaczających przyczynowość (ponieważ, więc itp.) oraz wgląd (myślę, wiedzą, rozważ). Z badań nad tzw. emocjonalnym pisaniem (*emotional writing*) wynika, że pisanie o emocjach pomaga wtedy, kiedy prowadzi do ponownej oceny poznawczej opisywanych wydarzeń, a słowa należące do kategorii przyczynowości oraz wglądu są tego wskaźnikiem i prognostykiem poprawy zdrowia (Pennebaker, Mayne, Francis, 1997). Również badania nad wydarzeniami traumatycznymi (Kross, Ayduk, 2008) czy rozstaniem (Boals, Klein, 2005) pokazują, że słowa opisujące przyczynowość i wgląd pełnią istotną rolę w nadawaniu struktury doświadczeniom. Natomiast wskaźnikami narracji nieustrukturyzowanej okazały się (Pasupathi, 2007) słowa oznaczające wahanie (możliwe, wydaje się, może) i tzw. wypełniacze (ich klasycznym przykładem jest fredrowskie „mocium panie", a przykłady współczesne to „rozumiesz" lub „jakby").

POLSKA ADAPTACJA PROGRAMU LIWC

Program LIWC doczekał się do tej pory kilku adaptacji językowych. Słownik dostępny jest w wersjach hiszpańskiej, holenderskiej, niemieckiej, norweskiej, portugalskiej i włoskiej (Pennebaker i in., 2007). Kilka innych tłumaczeń jest w fazie przygotowania, w tym na ukończeniu jest już słownik polski, nad którym aktualnie pracują autorzy niniejszego artykułu. Poniżej opisujemy drogę, jaką doszliśmy do obecnej wersji polskiego słownika, oraz wstępne badania nad jego trafnością i innymi parametrami psychometrycznymi.

Tworzenie polskiego słownika LIWC

Prace nad polską adaptacją LIWC rozpoczęliśmy od tłumaczenia słownika angielskiego na polski słowo po słowie. Krótko po rozpoczęciu tego procesu napotkaliśmy trzy znaczące problemy (dwa natury językowej i jeden językowo-merytoryczny), zasadniczo wpływające na docelowy kształt słownika.

Pierwszy z nich, dotyczący wszystkich tłumaczeń, polega na braku ścisłej odpowiedniości między dowolnymi dwoma językami – w każdym z nich rozkład znaczeń wśród wyrazów rysuje się trochę inaczej, dlatego też nie jest możliwe idealne tłumaczenie tekstu z jednego języka na drugi, tym bardziej gdy chodzi o tekst słownikowy, który jest nie tylko wytworem języka, lecz również narzędziem jego opisu (więcej na temat tłumaczenia tekstu słownikowego – por. Żakowicz, 2009). Przykład: jednym z możliwych angielskich odpowiedników polskiego wyrazu *stanąć* jest *to stop*, które na polski można również przetłumaczyć jako *zatrzymać*; to z kolei łączy się z angielskim czasownikiem *to keep*, tłumaczonym czasem jako *kontynuować* – oczywiście ten wyraz ma swój bliższy budową odpowiednik angielski: *to continue*. Z powodu

tego mechanizmu naszą pracę można tyleż nazwać tłumaczeniem słownika, co tworzeniem nowego.

Drugi problem wiąże się z gramatycznymi różnicami między językiem polskim a angielskim. Język angielski w zasadzie nie ma systemu fleksyjnego (prawie nie występuje w nim odmiana), zatem dany wyraz występuje jedynie w kilku formach (czasem jednej), a dodatkowe znaczenia, np. osoba czy rodzaj, są przenoszone przez osobne wyrazy. W języku polskim natomiast wyrazy mogą przenosić o wiele bogatsze znaczenia dzięki dobraniu odpowiedniej końcówki, np. czasownik *poszedłem* określa zarówno podjętą czynność, jak też osobę (pierwsza), liczbę (pojedyncza), rodzaj (męski), czas (przeszły), stronę (czynna), tryb (oznajmujący) czy nawet iteratywność (jednokrotny). Oznacza to jednak, że polskie wyrazy dysponują znaczną mnogością form wyrazowych, które słownik musi zliczać i rozróżniać, jeśli chce różnicować np. kategorię czasu. Przedrostek *po-* w formie *poszedłem* informuje o aspekcie dokonanym tego czasownika, w opozycji do niedokonanego *szedłem* lub innych dokonanych *przyszedłem*, *doszedłem*. Należy zauważyć, że informacje przekazywane przez takie doklejane człony często są w języku angielskim przekazywane za pomocą osobnych wyrazów lub poprzez zastosowanie zupełnie innego czasownika. W efekcie tego polskie formy są jeszcze liczniejsze w porównaniu z angielskimi. Jako przykład podamy czasownik *to read*, który w pisanym angielskim występuje jedynie w trzech formach: *read*, *reads* oraz *reading*. Po polsku będą to 63 formy (tabela 1).

Tabela 1.

Formy czasownika „czytać”

czytać	czytałeś	czytali	czytałabyś	czytałyby	czytającą	czytane
czytam	czytałaś	czytały	czytałby	czytając	czytających	czytanego
czytasz	czytał	czytano	czytałaby	czytający	czytającymi	czytanemu
czyta	czytała	czytaj	czytałoby	czytająca	czytanie	czytanej
czytamy	czytało	czytajmy	czytalibyśmy	czytające	czytania	czytaną
czytacie	czytaliśmy	czytajcie	czytałybyśmy	czytającego	czytaniu	czytanym
czytają	czytałyśmy	czytałbym	czytalibyście	czytającemu	czytaniem	czytanych
czytałem	czytaliście	czytałabym	czytałybyście	czytającej	czytany	czytanymi
czytałam	czytałyście	czytałbyś	czytaliby	czytającym	czytana	

Jeśli jednak dodamy do nich wszystkie formy czasowników *czytywać*, *doczytać*, *doczytywać*, *przeczytać*, *odczytać*, *odczytywać*, *podczytać*, *podczytywać*, a nawet *sczytać* oraz *sczytywać* (każdy niedokonany czasownik liczy 63 formy, dokonany 53 z powodu odmiennych imiesłowów), uzyskamy liczbę 643 polskich form wyrazowych, odpowiadających trzem formom angielskim. Ta dysproporcja przekłada się oczywiście na różnice we frekwencji – o ile można wnioskować o pokaźnej liczbie użyć wszystkich trzech angielskich form, o tyle większość polskich nie pojawia się prawie nigdy (o użytych narzędziach do pomiaru frekwencji piszemy niżej).

Trzeci problem związany był z naszymi zastrzeżeniami co do kategorii w oryginalnym słowniku. Niektóre wyrazy funkcjonują nieco inaczej w języku polskim niż ich angielskie odpowiedniki. Na przykład wyraz *rich* jest zaliczony zarówno do kategorii „pieniądze", jak i kategorii „emocje pozytywne" – i jest to trafne, gdy myślimy o kulturze anglosaskiej. Jednak, jak wiemy z badań Bogdana Wojciszke (2010), dla Polaków fakt, że ktoś jest *bogaty* wcale nie musi być jednoznacznie pozytywny. Jeśli do jego oceny zostanie zastosowane kryterium moralności, a nie kompetencji, to *bogaty* może wręcz nabrać znaczenia negatywnego. Jak widzimy, nie jest możliwe literalne przeniesienie powiązania słów z kategoriami wprost z angielskiego do polskiego słownika. Główną przyczyną tego są różnice językowo-kulturowe, ale czasem można też podejrzewać niedoskonałości oryginalnego słownika (jak np. przy zaliczonym do kategorii „religia" wyrazie *jew*, który może także występować w czysto etnicznym, niebiblijnym znaczeniu).

Widząc, że nie możemy zbudować słownika identycznego z angielskim, postanowiliśmy stworzyć obszerniejszy, który obejmowałby wszystkie użyteczne formy językowe, zliczając dodatkowo specjalnie utworzone kategorie gramatyczne, takie jak część mowy lub rodzaj (co nie byłoby możliwe w języku angielskim z uwagi na wspomniany brak systemu fleksyjnego). W tym celu musieliśmy zrezygnować z ułatwienia, jakie stanowi znak specjalny asterysk (tzw. gwiazdka), interpretowany przez programy komputerowe jako dowolny ciąg znaków, i zliczać każdą formę osobno. Zamierzenie stworzenia tak ekstensywnego słownika okazało się zbyt ambitne. Wstępna wersja tego słownika liczyła ponad 100 tys. form wyrazowych, podczas gdy LIWC obsługuje słowniki nie większe niż 15 000 form (dla porównania: słownik angielski LIWCa z 2001 roku liczy ich 2319).

W tej sytuacji konieczne było wznowienie prac od samego początku. Tym razem skupiliśmy się na naśladowaniu oryginalnego słownika wraz z jego strukturą i filozofią przewodnią. Zrezygnowaliśmy z dodatkowych kategorii, dzięki czemu mogliśmy używać gwiazdki do skracania wszystkich części mowy oprócz czasownika (aby LIWC zliczał kategorię czasu, obecną również w angielskim słowniku, czasy gramatyczne musiały być rozróżnione, mimo to wszystkie formy czasu przeszłego z reguły można zredukować do 2-3 form zakończonych gwiazdką).

Postanowiliśmy też dodawać do słownika tylko wyrazy często używane – aby je wyłonić, posłużyliśmy się dwoma listami frekwencyjnymi, utworzonymi specjalnie na potrzeby naszych prac. Pierwsza z nich, uzyskana dzięki uprzejmej współpracy dr hab. Adama Przepiórkowskiego, została wygenerowana na podstawie Korpusu Języka Polskiego, stworzonego w Instytucie Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN). Drugą listę utworzyliśmy sami na podstawie zebranych samodzielnie próbek wypowiedzi potocznych, zawierających różnego rodzaju dialogi i dyskusje (fora internetowe, czaty, spisane wywiady psychologiczne). W końcowej części prac posłużyliśmy się też samym korpusem IPI PAN oraz korpusem *Słownika frekwencyjnego polszczyzny współczesnej* (Kurcz i in., 1990), dostępnymi nieodpłatnie na stronie poświęconej Korpusowi Języka Polskiego IPI PAN (Zespół Inżynierii Lingwistycznej IPI PAN, 2006) za pośrednictwem dedykowanego oprogramowania *Poliqarp*. Korpusy te były również bardzo przydatne przy określaniu dominującego znaczenia wyrazów wieloznacznych oraz frekwencji poszczegól-

nych form wyrazów (przy znaczących różnicach we frekwencji poszczególnych form w obrębie jednego leksemu pomijaliśmy te nieużywane). Dzięki tym zmianom nowy słownik miał być niewielki i możliwie najlepiej dopasowany do języka w powszechnym użyciu, zaś osiągane wyniki – podobne do wyników oryginalnego słownika angielskiego i dzięki temu umożliwiające badania międzykulturowe. W tym celu nie wystarczyło jednak ograniczyć się do słów częstych, ponieważ ich znaczenia bywają liczne, szerokie lub metaforyczne. Dlatego też oprócz wyrazów popularnych staraliśmy się też zawrzeć trochę rzadszych, ale kategorialnie jednoznacznych (dobrym przykładem są tu nazwy chorób, również obecne w angielskim słowniku, mimo niewielkiej frekwencji). Jak dotychczas najpóźniejsza, wciąż testowa wersja tego słownika liczy prawie 5000 form wyrazowych.

Badania własności polskiego słownika LIWC

Planując badania sprawdzające trafność polskiego słownika i jego równoważność ze słownikiem angielskim, wzorowaliśmy się na podobnych badaniach, dotyczących słownika LIWC w wersji hiszpańskiej (Ramirez-Esparza i in., 2007).

Równoważność. Aby sprawdzić równoważność słowników w wersji polskiej i angielskiej, wykorzystaliśmy siedem pozycji literatury pięknej, każda w dwóch wersjach językowych (w sumie było to 14 utworów). Utwory reprezentowały różne gatunki literackie i pochodziły z różnych epok. Użyto m.in. Pana Tadeusza Adama Mickiewicza (Księga I), powieści *Harry Potter i kamień filozoficzny* J. K. Rowling czy dramatu *Czekając na Godota* Samuela Becketta. Wszystkie teksty w polskiej i w angielskiej wersji językowej poddano analizie programem LIWC, ze słownikiem odpowiednio polskim i angielskim, a następnie wyniki dla wszystkich kategorii porównano testem *t* Studenta.

Na 84 analizowane pozycje (kategorie LIWC) wersja polska nie różniła się od wersji angielskiej w 36 przypadkach. Dla 48 pozycji wynikowych zaobserwowano istotną różnicę. Przyjrzymy się niektórym kategoriom z tej drugiej grupy.

Pierwsza różnica dotyczy stopnia pokrycia tekstu przez słownik, czyli zasięgu naszej wersji. Wersja polska na obecnym etapie jej rozwijania wychwyciła średnio około 48% słów, zawartych ogółem w analizowanych utworach literackich, dla wersji angielskiej było to 72%. Różnica ta jest istotna – $p<0{,}001$. Oznacza to, że wciąż istnieje duże pole do zwiększenia zasięgu polskiego słownika. Daje to także dwie informacje o charakterystyce języka polskiego: a) większa liczba form wyrazowych w słowniku nie przekłada się na większy zasięg, gdyż w większości są to różne formy jednego wyrazu, a nie nowe słowa; b) być może inaczej wygląda profil użytkowania słów dla języka polskiego, np. silniej zarysowuje się efekt długiego ogona – duża liczba rzadziej używanych słów ma większy niż w języku angielskim udział w ogóle słów będących w powszechnym użyciu.

Kolejne różnice dotyczą użycia zaimków. W wersji angielskiej zaimki stanowiły istotnie ($p<0{,}01$) częściej występującą kategorię (wersja angielska: M = = 13,2%; SD = 3,32; wersja polska: M = 8,4%; SD = 0,9). Podobne wyniki uzyskano także dla poszczególnych podkategorii, grupujących odrębnie zaimki

w pierwszej, drugiej i trzeciej osobie liczby pojedynczej i mnogiej. Są to jednak różnice zrozumiałe i oczekiwane, jeśli weźmie się pod uwagę różnice fleksji obu języków. Pisaliśmy o tym powyżej, omawiając trzy ogólne problemy z tłumaczeniem słownika na inny język. Znaczenie, jakie w języku angielskim przekazuje zaimek osobowy, w języku polskim często wyrażają same końcówki czasowników (por. omawiany wcześniej przykład formy *poszedłem*).

Innym przykładem otrzymanej w naszym badaniu różnicy, którą w całości tłumaczą różnice gramatyczne między językiem angielskim i polskim, są rodzajniki (*articles*). W analizowanych utworach literackich w języku angielskim rodzajniki stanowiły średnio 6,7% (*SD* = 1,67), a w ich polskich odpowiednikach oczywiście odsetek rodzajników był zawsze zerowy, ponieważ taka część mowy nie występuje w języku polskim. Tak więc, choć różnica była statystycznie istotna ($t(12) = -10{,}79$; $p<0{,}001$), nie jest ona dowodem nierównoważności słowników i nie wymaga niwelowania.

W analizach pojawiły się jednak i takie istotne różnice, które trudno wytłumaczyć w powyższy sposób. Dotyczą one na przykład większej w języku polskim niż angielskim procentowej zawartości słów oznaczających zaprzeczanie ($M_{pol} = 2{,}9$, $M_{ang} = 2{,}2$; $t(12) = 2{,}17$; $p<0{,}05$) i potwierdzanie ($M_{pol} = 0{,}94$, $M_{ang} = 0{,}23$; $t(12) = 6{,}53$; $p<0{,}001$). Różnice w odwrotnym kierunku dotyczą z kolei większości kategorii opisujących emocje (zob. tab. 2).

Tabela 2.
Średnie odsetki słów (w nawiasach odchylenia standardowe) opisujących emocje w różnych utworach literackich w języku polskim i angielskim

Kategoria	Średni % słów w tekście (*SD*)		*t*	*df*
	Polskie	Angielskie		
Afekt	2,59 (0,25)	3,65 (0,48)	-5,27***	12
Pozytywne emocje	1,48 (0,17)	2,01 (0,31)	-4,01**	12
Pozytywne uczucia	0,19 (0,09)	0,43 (0,15)	-3,82**	12
Optymizm	0,42 (0,07)	0,39 (0,12)	0,73	9,67
Negatywne emocje	1,03 (0,13)	1,63 (0,29)	-5,06***	12
Napięcie i lęk	0,13 (0,04)	0,26 (0,05)	-5,98***	12
Gniew	0,26 (0,10)	0,43 (0,16)	-2,51*	12
Smutek	0,18 (0,05)	0,50 (0,26)	-3,10**	12

* $p<0{,}05$; ** $p<0{,}01$; *** $p<0{,}001$ (dwustronne testy *t* Studenta)

Tego typu różnic między słownikiem polskim a angielskim nie można wytłumaczyć odmiennościami obu języków. Różnic podobnych do opisanych powyżej odnotowaliśmy więcej. W większości przypadków słownik angielski okazał się bardziej kategorialnie czuły od słownika polskiego – wychwytywał więcej słów z kategorii, dla których występują różnice.

Wyniki te mają pierwszorzędne znaczenie dla dalszej pracy nad polskim słownikiem. Jak pisaliśmy wyżej, polski słownik na obecnym etapie rozpoznaje średnio tylko niespełna połowę słów w analizowanym tekście, co oczywiście oznacza konieczność kontynuowania starań o podwyższenie tego parametru. Znajomość konkretnych kategorii, których dotyczą różnice między dwiema wersjami słownika, pozwala na wyznaczenie kierunków tej pracy. Na przykład opisane różnice dotyczące kategorii afektywnych skłaniają do wzbogacenia słownika o kolejne polskie słowa określające i wyrażające emocje. Zapewne trzeba będzie przy tym nieco rozluźnić kryterium częstości występowania i dodać słowa także spoza ścisłej czołówki listy frekwencyjnej. Z pewnością też w dalszych badaniach trzeba będzie z wielką uwagą dobrać materiał, ponieważ – na co zwrócił nam uwagę Recenzent – specyfika treściowa niektórych analizowanych utworów literackich mogła wpłynąć na uzyskane wyniki.

Badanie trafności kategorii polskiego słownika w analizach wypowiedzi osób cierpiących z powodu objawów depresji. Przygotowując hiszpański słownik LIWC (Ramirez-Esparza i in., 2007), jego autorzy sprawdzili w odrębnym badaniu, czy będzie on równie przydatny do analizy wypowiedzi osób cierpiących na depresję, co słownik angielski (Rude i in., 2004). Okazało się, że bez względu na to, czy osoby depresyjne mówią po angielsku, czy też po hiszpańsku, ich język ma wiele wspólnych właściwości, które można uznać za „markery" depresji. Są to zwiększone częstości użycia słów z trzech kategorii: ja, emocje negatywne i stany ciała. Natomiast zaobserwowane rozbieżności między badaniami hiszpańskimi i angielskimi dotyczyły sposobów używania słów o nacechowaniu negatywnym – w myśl badań Rude i współpracowników (2004) osoby depresyjne używają ich mniej niż osoby niedepresyjne, co znalazło jedynie częściowe potwierdzenie w języku hiszpańskim. Inna rozbieżność dotyczyła używania słów odwołujących się do procesów społecznych. W badaniach Rude i współpracowników (2004) kategoria ta nie różnicowała osób depresyjnych i niedepresyjnych. W badaniach Ramireza-Esparzy i współpracowników (2007) osoby depresyjne istotnie częściej używały słów związanych z relacjami, rodziną, znajomymi itp., co autorzy tłumaczą jako wynik dzielenia się trudnościami z najbliższymi.

Pracując nad polskim słownikiem, postanowiliśmy powtórzyć badania zespołu hiszpańskiego. W tym celu losowo wybraliśmy dziesięć fragmentów blogów i wpisów na forach, w których ich autorzy opisywali po polsku swoje doświadczanie objawów depresji, oraz dziesięć fragmentów innych polskojęzycznych blogów i wypowiedzi internetowych. Użyte fragmenty liczyły średnio 2738 słów. Następnie porównaliśmy zawartość kategorialną tych dwóch rodzajów wypowiedzi za pomocą polskiej wersji słownika. Tabela 3 przedstawia wyniki tych porównań dotyczące kategorii, które wypadły tak samo w badaniach angielskich i hiszpańskich.

Tabela 3.
Średnie odsetki (w nawiasach odchylenia standardowe) zaimków osobowych oraz słów opisujących emocje w polskojęzycznych wypowiedziach zamieszczonych na portalach internetowych poświęconych depresji oraz innym tematom

Kategoria ogólna	Podkategoria	Średni % słów w tekście (*SD*)		*t*	*df*
		depresja	inne tematy		
Zaimki osobowe	zaimki osobowe ogółem	10,44 (1,22)	5,89 (1,78)	6,68***	18
	ja	3,36 (0,78)	1,11 (0,73)	6,66***	18
	my	0,51 (0,32)	0,39 (0,34)	0,80	18
	ty + wy	0,86 (0,46)	0,67 (0,29)	1,13	18
	on, ona, ono + oni	2,12 (0,70)	1,98 (0,41)	0,56	18
Procesy afektywne	procesy afektywne ogółem	4,33 (0,66)	3,24 (1,68)	1,91	18
	pozytywne emocje	2,15 (0,44)	2,49 (1,82)	-0,58	10,06
	pozytywne uczucia	0,32 (0,15)	0,45 (0,90)	-0,46	18
	optymizm	0,49 (0,14)	0,75 (0,69)	-1,14	9,77
	negatywne emocje	2,11 (0,41)	0,72 (0,35)	8,18***	18
	napięcie i lęk	0,29 (0,13)	0,06 (0,06)	5,00***	12,67
	gniew	0,41 (0,19)	0,19 (0,11)	3,27**	18
	smutek	0,52 (0,24)	0,15 (0,13)	4,25***	18
Procesy fizyczne	procesy fizyczne ogółem	1,06 (0,42)	0,48 (0,42)	3,12**	17,99
	ciało	0,76 (0,36)	0,32 (0,33)	2,86**	18
	seks	0,11 (0,10)	0,15 (0,10)	-0,89	18
	jedzenie	0,09 (0,10)	0,07 (0,08)	0,39	18
	sen	0,14 (0,10)	0,02 (0,03)	3,69**	10,18

* $p<0,05$; ** $p<0,01$; *** $p<0,001$ (dwustronne testy *t* Studenta)

Jak wynika z tabeli 3, polskojęzyczne wypowiedzi na stronach www poświęconych depresji różnią się od wypowiedzi zamieszczanych na stronach o innej tematyce pod względem szeregu kategorii słów. Co więcej, są to różnice podobne do tych, jakie pojawiały się we wcześniejszych badaniach angielskich

i hiszpańskich. Rzeczywiście, w tekstach osób cierpiących na depresję zaimki, zwłaszcza te pokazujące odwołanie do *ja*, miały istotnie większy udział. Nie zaznaczyły się natomiast różnice dotyczące zaimków w drugiej i trzeciej osobie. Jest to kolejne po badaniach hiszpańskich potwierdzenie wniosków z badań amerykańskich (Stirman, Pennebaker, 2001; Rude i in., 2004), które mówiły o większej koncentracji osób cierpiących na depresję na sobie niż na innych.

Wyniki amerykańskie i hiszpańskie zyskały także potwierdzenie w różnicach dotyczących udziału słów z kategorii opisujących emocje. Wypowiedzi osób depresyjnych zawierały istotnie więcej odniesień do emocji negatywnych, lęku, smutku i gniewu, a nie różniły się od blogów niedepresyjnych, jeśli chodzi o kategorie opisujące emocje pozytywne. Oznacza to, że osoby depresyjne nie pomijają pozytywnych emocji, ale znacznie więcej miejsca poświęcają opisywaniu stanów negatywnych.

Jak wiadomo, w depresji nierzadko następuje zwiększenie zainteresowania własnym ciałem jako przedmiotem cierpienia. Objawy często obejmują trudności somatyczne (ociężałość, zmiana wagi) i problemy ze snem. Znalazło to swoje odzwierciedlenie w naszym porównaniu. Blogi osób cierpiących na depresję charakteryzowały się istotnie większym udziałem słów z kategorii Procesy fizyczne, Ciało i Sen. To kolejny wynik naszego badania, który jest spójny z wynikami badań angielskich i hiszpańskich. Wystąpiły też inne interesujące różnice, jednak ze względu na ograniczone ramy artykułu, nie omawiamy ich szerzej w tym miejscu (badanie powtórzymy, gdy polski słownik LIWCa uzyska ostateczną postać).

Podsumowując, można uznać, że nasze porównanie potwierdziło wnioski z wcześniejszych badań. Blogi osób depresyjnych ujawniały ich większą koncentrację na sobie, na swoich negatywnych emocjach i swoim ciele. Zarazem wnioski te wskazują na zasadniczą trafność słownika polskiego pod względem rozpoznawania specyficznych właściwości wypowiedzi osób cierpiących z powodu objawów depresji. Można stwierdzić, że już na obecnym etapie polski słownik jest użytecznym narzędziem do badań, choć oczywiście wymaga dalszego doskonalenia.

WNIOSKI

W niniejszym artykule zaprezentowaliśmy program LIWC, służący do komputerowej analizy tekstu, i opisaliśmy amerykańskie badania, przeprowadzone z jego zastosowaniem. Przedstawiliśmy też własne prace nad polską adaptacją słownika LIWC oraz wykonane przez nas badania, inspirowane badaniami nad hiszpańską wersją słownika LIWC, mające na celu ocenę równoważności i niektórych aspektów trafności słownika polskiego.

Obecny etap prac nie jest końcowy. Jednym z celów dalszych prac jest zwiększenie odsetka rozpoznawanych przez słownik słów w analizowanych tekstach. Kolejnym celem, nierozłącznym z poprzednim, jest zwiększenie czułości kategorialnej słownika polskiego poprzez celowe wzbogacanie słownika w tych miejscach, w których występują znaczne różnice między wersją polską a angielską. Należy przy tym przeanalizować, czy wersja angielska jest właściwym punktem odniesienia – czy nie jest przesadnie czuła dla kilku kategorii.

Wyzwaniem pozostaje kwestia adekwatności kategorii gramatycznych. Można uznać, że kategoriami odpowiadającymi funkcjonalnie angielskim zaimkom byłyby specyficzne dla polskiej adaptacji kategorie ukazujące osobowe formy czasowników. Oznaczałoby to jednak dużą zmianę i ograniczało możliwości porównań międzykulturowych. Być może należy zbudować osobny słownik, rejestrujący kategorie gramatyczne zamiast tradycyjnych kategorii LIWC-a, i przyszłe teksty analizować obydwoma.

Uzyskanie udoskonalonego polskiego słownika otworzy duże pole do badań. Możliwe stanie się powtórzenie w Polsce wielu interesujących badań zespołu Pennebakera i przeprowadzenie porównań międzykulturowych. Będzie można też tworzyć kolejne słowniki, wyspecjalizowane w pewnych konkretnych typach analizy (np. w wykrywaniu kłamstwa lub ocenie poziomu stresu u autora wypowiedzi). Nietrudno wyobrazić sobie zastosowanie słownika LIWC (lub szerzej – automatycznej frekwencyjnej analizy tekstu) w wielu obszarach psychologii stosowanej – od wsparcia diagnozy klinicznej poprzez dostarczenie nowych narzędzi dla selekcji i rekrutacji pracowników po analizę przemówień i doniesień medialnych, pomocną w marketingu politycznym. Mamy nadzieję, że opracowanie kompletnego słownika podstawowego LIWC dla języka polskiego okaże się tylko wstępnym etapem, który otworzy szeroki i wieloraki nurt nowych badań.

BIBLIOGRAFIA

Boals, A., Klein, K. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, 24, 252-268.

Chung, C. K., Pennebaker, J. W. (2007). The psychological functions of function words. [W:] K. Fiedler (red.), *Social communication* (s. 343-359). New York: Psychology Press.

Cohn, M. A., Mehl, M. R., Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15, 687-693.

Gunsch, M. A., Brownlow, S., Haynes, S. E., Mabe, Z. (2000). Differential linguistic content of various forms of political advertising. *Journal of Broadcasting Electronic Media*, 44, 27-42.

Kross, E., Ayduk, O. (2008). Facilitating adaptive emotional analysis: Distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction. *Personality and Social Psychology Bulletin*, 34, 924-938.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., Woronczak, J. (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Instytut Języka Polskiego PAN.

Mehl, M. R., Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and conversations. *Journal of Personality and Social Psycholology*, 84, 857-870.

Newman, M. L., Pennebaker, J. W., Berry, D. S., Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29, 665-675.

Pasupathi, M. (2007). Telling and the remembered self: Linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory*, 15, 258-270.

Pennebaker, J. W., Booth, R. E., Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC2007 – Operator's manual. Austin, TX: LIWC.net.; www.psy.utexas.edu/Pennebaker (wrzesień 2011).

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., Booth, R. J. (2007). The development and psychometric properties of LIWC2007. [Software manual]. Austin, TX; www.liwc.net (wrzesień 2011).

Pennebaker, J. W., Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion*, 10, 601-626.

Pennebaker, J. W., Francis, M. E., Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC 2001*. Mahwah, NJ: Erlbaum; www.psy.utexas.edu/Pennebaker (wrzesień 2011).

Pennebaker, J. W., King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality Social Psychology*, 77, 1296-1312.

Pennebaker, J. W., Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality*, 36, 271-282.

Pennebaker, J. W., Mayne, T., Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, 72, 863-871.

Ramirez-Esparza, N., Pennebaker, J. W., Garcia, F. A., Suria, R. (2007). La psychología del uso de las palabras: Un programa de comutadora que analiza textos en Español [Psychologia używania słów: Program komputerowy do analizowania tekstu w języku hiszpańskim]. *Revista Mexicana de Psicología*, 24, 85-99.

Rude, S. S., Gortner, E. M., Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition Emotion*, 18, 1121-1133.

Silverman, D. (2007). *Interpretacja danych jakościowych. Metody analizy rozmowy, tekstu i interakcji*. Tł. M. Głowacka-Grajper, J. Ostrowska. Warszawa: Wydawnictwo Naukowe PWN.

Stemplewska-Żakowicz, K. (2009). *Diagnoza psychologiczna. Diagnozowanie jako kompetencja profesjonalna*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Stirman, S. W., Pennebaker, J. W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine*, 63, 517-522.

Stone, L. D., Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology*, 24, 172-182.

Tausczik, Y., Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24-54.

Watson, D., Clark, L. A., Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063-1070.

Wojciszke, B. (2010). *Sprawczość i wspólnotowość. Podstawowe wymiary spostrzegania społecznego*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., Kordy, H. (2008). Computergestuetzte quantitative Textanalyse: Aequivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count [Computerized quantitative text analysis: Equivalence and robustness of the German adaptation of Linguistic Inquiry and Word Count]. *Diagnostica*, 54, 85-98.

Zespół Inżynierii Lingwistycznej IPI PAN (2005). Korpus IPI PAN; http://korpus.pl (wrzesień 2011).

Żakowicz, W. (2009). Język, słownik i komputer. [W:] Interdyscyplinarne Koło Badań nad Językiem Wydziału Psychologii UW (red.), *II Studenckie Forum Badań nad Językiem: Teksty Pokonferencyjne* (s. 156-163). Warszawa: Oficyna Wydawnicza WBK; http://www.psychologia.pl/forum2010/wydawnictwa/ (wrzesień 2011).