Struktura ilościowa pól leksykalnych a procesy poznawcze człowieka

Adam Pawłowski

WPROWADZENIE

Pojęcie kodowania optymalnego tradycyjnie kojarzone jest w lingwistyce z binarnym zapisem jednostek poziomu fonetycznego lub innych równoważnych symboli, na przykład liter (Jassem 1973: 293-297; Herdan 1966: 259-303; Hammerl, Sambor 1990: 418-420). Jego celem jest ukazanie, iż zmienne częstości występowania badanych symboli mogą zredukować długość zakodowanej wiadomości, a więc pośrednio czas jej przetwarzania (zrozumienia, odtworzenia). Podobną technikę można jednak zastosować do innych podsystemów lub podzbiorów języka. W szczególności można tym sposobem zakodować pola leksykalne, traktowane tu jako podzbiory podsystemu leksykalnego. Łatwo jest zaobserwować, iż dla większości pól leksykalnych struktura częstości leksemów, uporządkowanych malejąco, przedstawia sobą krzywą malejącą niemonotonicznie o profilu przypominającym swym kształtem funkcję potęgową o ujemnym wykładniku (Rys. 1). Jest to zjawisko występujące w różnych językach i dotyczy struktury całego słownictwa, a nie tylko konkretnych pól leksykalnych. Z pozoru jego wyjaśnienie jest proste: można mianowicie odwołać się do kategorii r e a l i z m u p o z n a w c z e g o i przyjąć, zróżnicowanie częstości leksemów odzwierciedla ilościową strukturę rzeczywistych desygnatów odnośnych pojęć. Oznaczałoby to na przykład, że leksemy *czarny*, *biały*, *czerwony* itd. mają najwyższe częstości, ponieważ wskazują na najczęściej spotykane w środowisku człowieka barwy. Wyjaśnienie takie jest jednak niewystarczające. Chociaż pewien związek pomiędzy środowiskiem człowieka a ilościowym ukształtowaniem pól leksykalnych istnieje, wyjaśnienie nierównomiernych układów częstości w polach leksykalnych wymaga odwołania się do elementarnych kategorii epistemologicznych (aprioryzm, aposterioryzm), a także do wiedzy o procesach neurolingwistycznych, zachodzących podczas zapamiętywania i odtwarzania przez człowieka informacji językowej.

METODY MODELOWANIA STRUKTURY PÓL LEKSYKALNYCH

Z matematycznego punktu widzenia ten typ rozkładu częstości leksemów może być opisany wieloma sposobami. Jedną z często stosowanych metod jest estymacja jakiejś funkcji, stanowiącej w założeniu teoretyczny model opisywanego zjawiska (ten typ modelowania rozpowszechniony jest w niemieckiej szkole badań ilościowych, por. Altmann 2000, Köhler et al. 2005). Modele funkcyjne tego rodzaju mają wiele zalet, ukazują współzależność pomiędzy zmiennymi, pozwalają także na predykcję cech tekstów określonego typu. Ich wadą jest jednak niewielka moc eksplanacyjna, a więc brak możliwości wyjaśnienia istoty zjawiska, jego źródeł i konsekwencji. Podchodząc do tej kwestii minimalistycznie można oczywiście przyjąć, że przyczyną zmiany parametru *f(x)* jest zmiana wartości *x*, można także ukazać dynamikę tej zmiany. Ale badacz niewyjaśnionego zjawiska powinien pójść o krok dalej i postawić pytanie, d l a c z e g o taki a nie inny związek zachodzi.

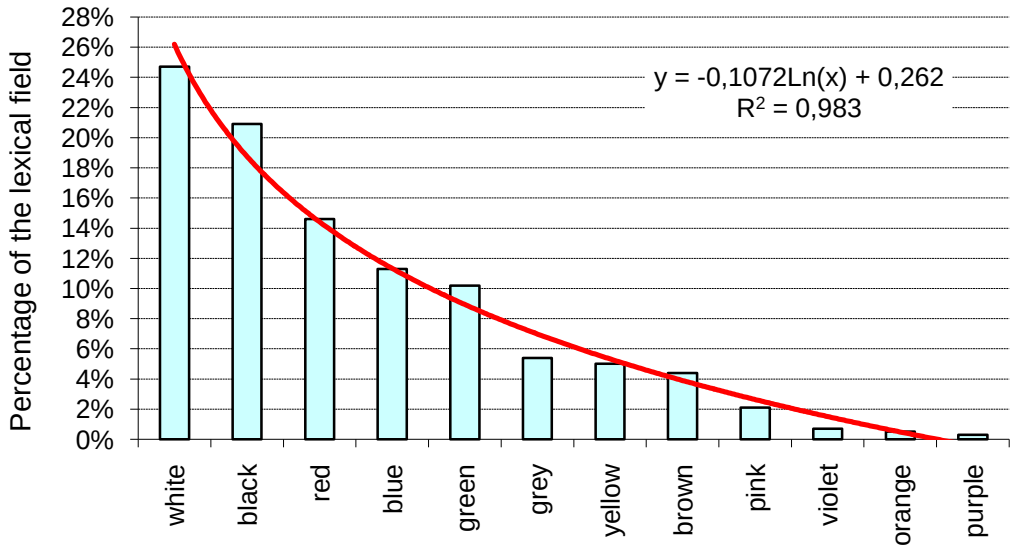
W niniejszej pracy zastosowano więc inne podejście, którego ideą przewodnią jest poszukiwanie przyczyn zjawiska, a nie jego wewnętrznej dynamiki. Przyjęto, że podzbiór lek-

sykonu tworzący pole leksykalne może być reprezentowany w pamięci człowieka jako binarna sekwencja, a co za tym idzie, także model powinien opierać się na skali dwuwartościowej. Podejście to jest zgodne z aktualną wiedzą na temat procesów neurologicznych, ponieważ zero i jedynka w modelu odpowiadają stanom nieaktywności i pobudzenia neuronu. W celu zakodowania sekwencji binarnych, odpowiadających pojedynczym leksemom, zastosowano dwie metody:

– kodowania prostego, opierającą się na zasadzie, zgodnie z którą sekwencje binarne odpowiadające poszczególnym leksemom są jednakowej długości;
– kodowania optymalnego, opierającą się na zasadzie, zgodnie z którą długość sekwencji binarnej zależna jest od częstości występowania leksemu, przy czym leksemy częste kodowane są sekwencjami krótszymi, a rzadkie dłuższymi (cf. Meyer-Eppler 1959, Herdan 1966: 277-278).

Po przeprowadzeniu kodowania porównano średnie długości sekwencji binarnych, odpowiadających leksemom należącym do pola leksykalnego kolorów, otrzymanych obiema metodami. Uzyskane wyniki poddano analizie i interpretacji, odwołując się do zasad najmniejszego wysiłku w procesach i systemach komunikacji oraz samoregulacji systemu językowego. Zwrócono także uwagę na ewolucyjne aspekty wzrostu efektywności procesów poznawczych mózgu ludzkiego.

Rys. 1 Udziały średnich częstości nazw kolorów w polu leksykalnym, na podstawie wielojęzycznego korpusu tekstów (Pawłowski 1999)[1]



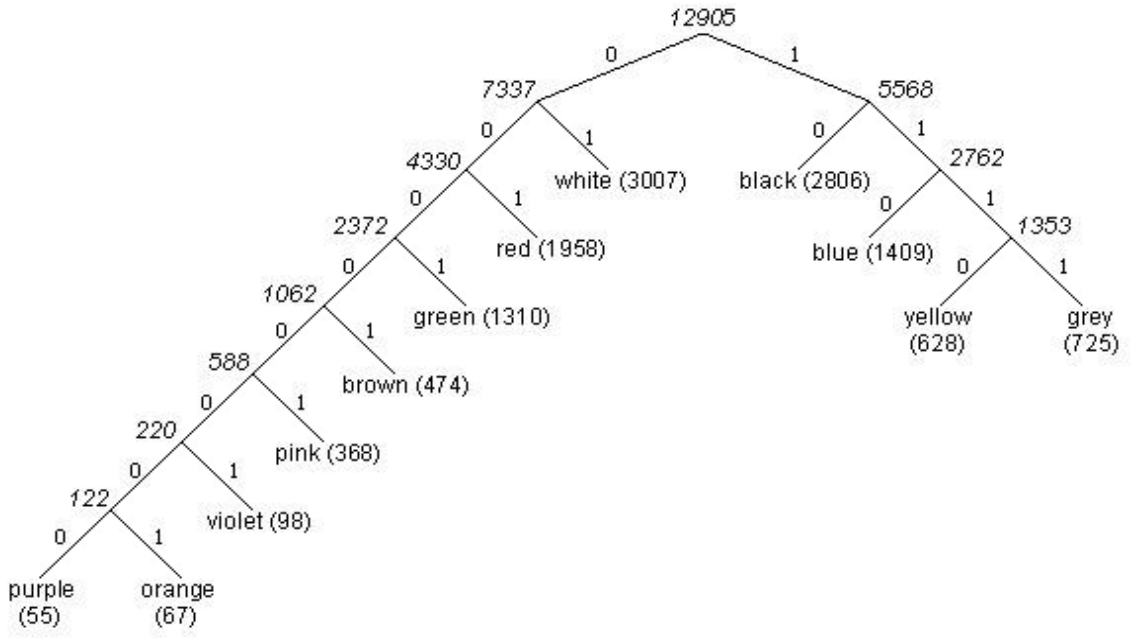
WYNIKI BADAŃ

Zgodnie z oczekiwaniami średnia długość sekwencji bitów kodowanych metodą Huffmana okazała się mniejsza od średniej długości sekwencji otrzymanej w toku kodowania równomiernego. Dla pola leksykalnego kolorów długość sekwencji przy kodowaniu równomiernym była stała i wynosiła 4 bity informacji, natomiast przy kodowaniu optymalnym spadła do 2,97

[1] Wykorzystano korpus o objętości 5,5 milina słów tekstowych, reprezentujący 10 języków indoeuropejskich.

bita (Tab. 1, Rys. 2). Oznacza to, że wzrost efektywności przetwarzania informacji zakodowanej metodą Huffmana wynosi około 25%, ponieważ o tyle właśnie skraca się średnia długość „słowa binarnego", a co za tym idzie średni czas jej odczytania lub zapisania. Pojęcie informacji jest oczywiście szerokie (w ogólnym sensie oznacza każdy bodziec zwiększający wiedzę organizmu na temat jego otoczenia). W tym jednak kontekście informację należy kojarzyć z przetworzeniem (zakodowaniem lub rozkodowaniem) bodźca odpowiadającego jednemu pojęciu lub leksemowi (nazwie koloru).

Rys. 2 Wartości kodów Huffmana przypisane elementom pola leksykalnego kolorów



Tab. 1 Kodowanie binarne elementów pola leksykalnego kolorów

C	*F*	*p*	*p.c.*	*H.c.*	*L*
white	3007	0,233	1111	10	2
black	2806	0,217	1110	01	2
red	1958	0,152	1101	100	3
blue	1409	0,109	1100	011	3
green	1310	0,102	1011	1000	4
grey	725	0,056	1010	1111	4
yellow	628	0,049	1001	0111	4
brown	474	0,037	1000	10000	5
pink	368	0,029	0111	100000	6
violet	98	0,008	0110	1000000	7
orange	67	0,005	0101	10000000	8
purple	55	0,004	0100	00000000	8

Oznaczenia:

C – nazwa barwy

F – częstość użycia terminów odpowiadających *C* w korpusie

p – empiryczne prawdopodobieństwo pojawienia się C w korpusie
$p.c.$ – kodowanie proporcjonalne (sekwencje równej długości)
$H.c.$ – kodowanie optymalne (sekwencje o zmiennej długości)
L – długość sekwencji kodowanej optymalnie

WNIOSKI

Rozumowanie przyczynowo-skutkowe, którego celem jest wyjaśnienie powszechnego zjawiska nierównomiernych rozkładów częstości jednostek leksykalnych, opiera się na zasadzie najmniejszego wysiłku. W kontekście systemu komunikacji oznacza ona, iż mechanizm samoregulacji prowadzi do ustanowienia stanu równowagi pomiędzy dwiema przeciwstawnymi siłami, jakimi są z jednej strony efektywność komunikacyjna człowieka i jego orientacja w środowisku, a z drugiej ograniczone możliwości rejestrowania i przetwarzania informacji przez mózg ludzki. Maksymalizacja pierwszego parametru, a więc najlepsza z możliwych orientacja człowieka w środowisku, wymagałaby przetwarzania w czasie rzeczywistym praktycznie nieograniczonej liczby bodźców odbieranych nieustannie przez perceptory. Zadanie takie przekracza jednak możliwości mózgu ludzkiego. Prawdopodobnie dlatego w toku procesów adaptacyjnych wytworzyły się wewnętrzne mechanizmy poznawcze, które niejako wtłaczają rejestrowany strumień bodźców w gotowe, upraszczające schematy (model takiego schematu pokazano na Rys. 2). Przedstawiona tu i wstępnie zweryfikowana „hipoteza kompromisu" pomiędzy tendencją do maksymalizacji ilości analizowanej informacji a ograniczonymi możliwościami przetwarzania informacji przez umysł ludzki jest dobrym punktem wyjścia prowadzącym do sformułowania uogólnień.

Pierwszym wnioskiem, jaki nasuwa się po analizie danych, jest stwierdzenie, iż procesy poznawcze człowieka mają charakter u m i a r k o w a n i e a p r i o r y c z n y. Oznacza to, że reprezentacja wiedzy zawarta w ludzkim mózgu nie jest determinowana bodźcami zewnętrznymi, lecz strukturą samej pamięci. Wymusza ona kategoryzację danych opartą na rozkładach nierównomiernych, składających się z około siedmiu lub ośmiu jednostek o malejących częstościach oraz dużej liczby jednostek o niskich częstościach („ogon" krzywej). Można powiedzieć, że ludzie postrzegają rzeczywistość w taki a nie inny sposób, ponieważ mózg nie wytrzymałby intensywności procesu poznawczego, podczas którego każdy postrzegany element świata byłby kategoryzowany zgodnie z jego cechami fizykalnymi, a jednostki należące do pól leksykalnych miałyby w dyskursach podobne częstości. Rozróżniając na przykład n odrębnych percepcyjnie barw (n może wynosić od kilkuset do kilkudziesięciu tysięcy, zależnie od cech osobniczych), człowiek redukuje tę wielość w swojej reprezentacji wiedzy do kilku jednostek dominujących. Wprawdzie następuje wtedy utrata informacji, ale jest ona rekompensowana zwiększoną szybkością przetwarzania mniejszej liczby kategorii, co w ostatecznym rozrachunku zwiększa orientację człowieka w jego środowisku. Zjawisko to jest oczywiście zwielokrotnione poprzez odniesienie do wszystkich kategoryzowanych językowo elementów doświadczenia.

Wniosek taki nie oznacza jednak, że zwerbalizowana reprezentacja wiedzy jest całkowicie oderwana od doświadczenia. Percepcja, a więc pośrednio środowisko człowieka, decyduje o tym, jakie kategorie znajdą się na poszczególnych pozycjach schematu analogicznego do tego, który przedstawiono na Rys. 1. Na przykład pierwotne doświadczenie światła, ciem-

ności, krwi i ognia sprawia, że odpowiadające tym prototypowym zjawiskom lub desygnatom barwy znajdują się we wszystkich językach, na których prowadzono badania lingwistyczne, na trzech pierwszych pozycjach schematu. Jednak sam układ malejących częstości kolejnych leksemów, prowadzący do subiektywnego przekonania użytkowników języka o różnej „ważności" lub „prototypowości" poszczególnych barw, jest już tylko wynikiem ograniczeń narzuconych przez ludzki mózg. Ponieważ analogiczne rozumowanie można przeprowadzić w odniesieniu do leksemów tworzących inne pola leksykalne, przedstawione tu wnioski można uznać za relewantne dla całokształtu procesów poznawczych człowieka.

Druga konkluzja ma charakter samoreferencyjny, a w skutkach może okazać się autodestrukcyjna. Skoro bowiem reprezentacja wiedzy w ludzkim mózgu posiada tak dużą autonomię w stosunku do rzeczywistości postrzeganej zmysłowo, być może również czynności poznawcze umysłu – a niniejszy abstrakt stanowi właśnie wynik rzeczonej aktywności – uznać należy jedynie za ćwiczenie intelektualne luźno powiązane z rzeczywistością.

REFERENCES

Altmann Gabriel (2000), *Einführung in die quantitative Lexikologie*. Trier: Wissenschaftlicher Verlag Trier.

Hammerl Rolf, Sambor Jadwiga (1990), *Statystyka dla językoznawców*. Warszawa: PWN.

Herdan Gustav (1966), *The Advanced Theory of Language Choice and Chance*. Berlin etc.: Springer.

Jassem W. (1974), *Mowa a nauka o łączności*. Warszawa: PWN.

Köhler Reinhard, Altmann Gabriel, Piotrowski R. (2005) (red.), Quantitative Linguistik / Quantitative Linguistics. Ein Internationales Handbuch / An International Handbook. Berlin, New York: Walter de Gruyter.

Meyer-Eppler W. (1959), *Grundlagen und Anwendungen der Informationstheorie*. Berlin etc.: Springer.

Pawłowski Adam (1999), The Quantitative Approach in Cultural Anthropology: Application of Linguistic Corpora in the Analysis of Basic Color Terms. In: *Journal of Quantitative Linguistics* 6/3, 1999, 222–234.