„LingVaria"
Rok V (2010), nr 1 (9)

Agnieszka Kułacka
King's College London
University of London
Londyn

PRAWO SHERMANA

Wstęp

Prawa Shermana są to modele opisujące rozkład długości zdań w tekście. Nadano im tę nazwę, ponieważ L. A. Sherman był pierwszym językoznawcą, który uznał, że długość zdania złożonego może stanowić zmienną lingwistyczną.

W 1888 r. Sherman opublikował artykuł, w którym opisał analizę zmian długości zdania złożonego w tekstach angielskich, począwszy od Geoffreya Chaucera, a skończywszy na współczesnej mu literaturze (Sherman 1888). Długość zdań mierzył liczbą występujących w nich wyrazów. Autor zauważył, że w danym tekście nie ma istotnych wahań długości zdań. Na przykład w opowiadaniu Chaucera *The Tale of Melibeus* pierwsze 50 zdań liczy 2572 wyrazów, następne 50 – 2536, kolejne – 2199, 2099, 2640, 2338, a w pozostałych 40 zdaniach autor wykorzystuje 2345 wyrazów. Uzyskane średnie dla tych odcinków tekstu zamyka przedział od 42 do 59 wyrazów, a średnia obliczona dla całego tekstu wynosi około 49 wyrazów. Dla tekstów bardziej współczesnych średnia ta jest niższa, np. w artykule *Father Taylor: A Man of Genius* napisanym przez C. A. Bartol na początku XIX w. na zdanie przypada średnio 16 wyrazów. Sherman nie wyciągnął z tego wniosków o zmianach średniej długości zdania w tekście następujących w czasie (od najwcześniejszych tekstów pisanych do bardziej współczesnych), zauważył jednak, że średnie zależą od rodzaju tekstu. W tekstach symbolicznych są one wyższe. Autor podał również procent zdań pojedynczych przypadających na całkowitą liczbę zdań w danym tekście.

Gabriel Altmann (1992) podsumował dotychczasowe badania nad prawami Shermana. Długość zdania jest własnością tekstu, której nie można uznać za chaotyczną, deterministyczną, czy też rządzoną pewnymi zasadami. Można jednak powiedzieć, że poddaje się ona prawu współpracy i rywalizacji pewnych sił działających podczas powstawania tekstu. Zaliczyć do nich można: (1) usiłowania autora, by być oryginalnym, (2) ograniczenia gatunkowe tekstu, (3) potencjalnych odbiorców tekstu, (4) zakłócenia powstałe wskutek zastosowania dużej liczby zdań podrzędnych w zdaniach złożonych. Altmann zauważył również, że modele opisujące rozkład długości tekstu

mogą znacząco się różnić w zależności od stosowania różnych jednostek miary, tj. tego, czy długość zdania mierzy się liczbą występujących w nim wyrazów, czy też liczbą zdań podrzędnych. Innym istotnym czynnikiem wpływającym na uzyskane wyniki jest sposób próbkowania. Altmann uznał, że badanie należy przeprowadzać na zamkniętych odcinkach tekstów, np. pełnych rozdziałach książki. Ten sposób badania wykorzystałam podczas badań nad prawem Menzeratha-Altmanna (Kułacka 2008).

Celem niniejszego artykułu jest ilustracja prawa Shermana na przykładzie wybranych tekstów. Przyjrzymy się rozkładom zdań złożonych mierzonych liczbą zdań składowych w kilku tekstach w dwóch wersjach językowych − polskiej i angielskiej oraz w dwóch gatunkach − literackim i naukowym. Postaram się ustosunkować do wniosków Shermana: (1) nie istnieją istotne wahania długości zdania na danym tekście, (2) średnie długości zdań zależą od rodzaju tekstu. Celem porównania tych samych tekstów w dwóch wersjach językowych oraz analizy częstości względnych zdań złożonych we wszystkich badanych tekstach jest dostrzeżenie potencjalnych prawidłowości.

Procedura analizy danych

Zgodnie z wnioskami Altmanna, przytoczonymi w pierwszej części artykułu, analizowałam wszystkie zdania pochodzące z zamkniętych całości, czyli rozdziałów. Zdania złożone oraz składające się na nie zdania podrzędne lub współrzędne liczyłam „ręcznie". Założyłam, że znakiem oddzielającym jedno zdanie złożone od drugiego jest końcowy znak interpunkcyjny. Liczbę zdań składowych ustaliłam natomiast na podstawie liczby finitywnych form oraz imiesłowów uprzednich i współczesnych.

Zauważmy najpierw, że w celu analizy średnich długości zdania w poszczególnych tekstach możemy brać pod uwagę albo medianę, zwaną też drugim kwartylem lub środkową wartością w danym ciągu liczb, albo ich średnią arytmetyczną. Średnią arytmetyczną oblicza się, dzieląc sumę wartości w danym ciągu liczb przez liczbę tych wartości. Mediana jest natomiast wartością środkową uporządkowanego ciągu liczb, np. w ciągu liczb 2, 3, 4, 5, 7, 7, 8 mediana to czwarta wartość. W dużej próbie, a z takimi będziemy mieć tutaj do czynienia, mediana to wartość stojąca na pozycji n/2 uporządkowanego ciągu liczb, gdzie *n* jest liczbą wartości w danym ciągu. Inną wartością średnią jest moda, czyli wartość o najwyższej częstości. W naszym przykładzie jest to liczba 7.

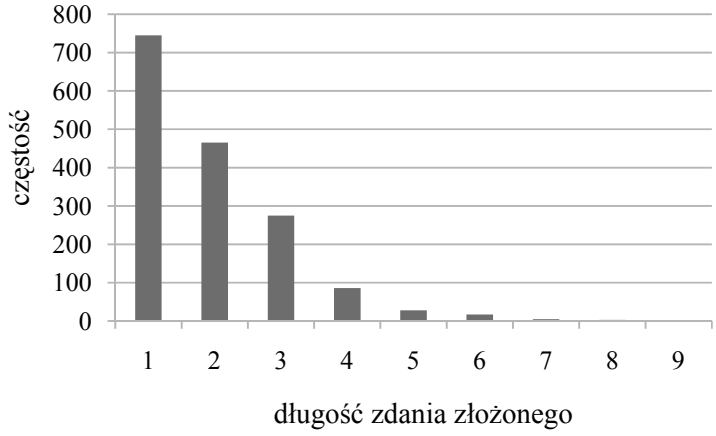
Spójrzmy na dane pochodzące z fragmentu *Kodu Leonarda da Vinci* autorstwa Dana Browna. W tabeli 1. x oznacza liczbę zdań składowych zdania złożonego, a f – częstość tych zdań.

Tabela 1. Dane na podstawie *Kodu Leonarda da Vinci*

x	1	2	3	4	5	6	7	8	9	Suma
f	745	465	275	86	28	17	5	3	2	1626

Dane z tabeli 1. umieściłam na wykresie 1. Można zauważyć, że dane te charakteryzuje skośność dodatnia, tj. większość danych to długości najkrótsze. Formalnie skośność można potwierdzić, obliczając kwartyle. Są to wartości środkowe z pierwszej i drugiej połowy uporządkowanego ciągu wartości. Kwartyl dolny to wartość na pozycji $n/4$, a kwartyl górny to wartość na pozycji $3n/4$ uporządkowanego ciągu wartości. Rozpatrzmy ponownie przykład przytoczony wyżej, czyli ciąg wartości 2, 3, 4, 5, 7, 7, 8. Kwartyl dolny to wartość druga (7/4 = 1,75, po zaokrągleniu w górę to 2), czyli 3, a kwartyl górny to wartość 21/4 = 5,25, po zaokrągleniu w górę to 6), czyli 7.

Wykres 1. Częstości długości zdania złożonego



Pozycja kwartyla dolnego to 1626/4 = 406,5, a zatem kwartyl dolny to długość zdania na pozycji 407: $Q_1 = 1$. Podobnie, pozycja kwartyla górnego to (3 x 1626)/4 = 1219,5, a zatem kwartyl górny oznacza długość zdania na pozycji 1220: $Q_3 = 3$. Pozycja mediany, drugiego kwartyla, to 1626/2 = 813, a więc mediana to długość zdania na pozycji: $Q_2 = 2$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 1$. Jak widać, odstępy te są równe, co oznacza, że w analizie możemy zastosować albo średnią arytmetyczną albo medianę. Ponieważ nie jest to wykres symetryczny, nie można zastosować mody, która w przypadku powyższych danych wynosi 1. Średnia arytmetyczna obliczona dla tych danych wynosi 1,95 (z dokładnością do dwóch miejsc po przecinku), tym samym nie różni się w sposób znaczący od mediany, która ma wartość 2.

Przed analizą danych każdorazowo będę obliczać kwartyle oraz średnią i wybierać do analizy tę wartość, która będzie lepiej reprezentować dane.

Analizowane teksty

Do analizy włączyłam trzy teksty w dwu wersjach językowych (polskiej i angielskiej). Dane są zawarte w tabelach 2–7, gdzie *x* oznacza długość zdania złożonego

mierzonego liczbą zdań składowych, a *f1* i *f2* są częstościami tych zdań: *f1* – częstością dla pierwszej części tekstu, a *f2* – dla pierwszej części tekstu powiększonej o 40% zdań.

Tabela 2. Dane na podstawie *Hobbita* Tolkiena w wersji angielskiej

x	1	2	3	4	5	6	7	8	9	10	13	Suma
f1	282	311	251	137	88	50	35	15	9	2	2	1182
f2	438	427	350	202	119	67	43	25	11	4	2	1688

Dla pierwszej części tekstu uzyskujemy $Q_1 = 2$, $Q_2 = 2$ i $Q_3 = 3$. $Q_2 - Q_1 = 0$ oraz $Q_3 - Q_2 = 1$, co oznacza formalnie stwierdzoną skośność dodatnią. Średnia arytmetyczna dla tych danych wynosi 2,91 (z dokładnością do dwóch miejsc po przecinku). Dla całości badanego odcinka tekstu otrzymujemy $Q_1 = 1$, $Q_2 = 2$ i $Q_3 = 4$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 2$, ponownie jest to więc formalnie stwierdzona skośność dodatnia. Średnia arytmetyczna dla tych danych wynosi 2,85 (z dokładnością do dwóch miejsc po przecinku).

Tabela 3. Dane na podstawie *Hobbita* Tolkiena w wersji polskiej

x	1	2	3	4	5	6	7	8	9	10	11	12	Suma
f1	302	351	244	142	81	43	29	12	5	1	2	1	1213
f2	471	482	337	189	108	54	41	14	10	2	3	1	1712

Dla pierwszej części tekstu uzyskujemy $Q_1 = 2$, $Q_2 = 2$ i $Q_3 = 4$. $Q_2 - Q_1 = 0$ oraz $Q_3 - Q_2 = 2$, to zaś oznacza formalnie stwierdzoną skośność dodatnią. Średnia arytmetyczna dla tych danych wynosi 2,77 (z dokładnością do dwóch miejsc po przecinku). Dla całości badanego odcinka tekstu mamy $Q_1 = 1$, $Q_2 = 2$ i $Q_3 = 3$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 1$, a zatem brak formalnego potwierdzenia skośności. Średnia arytmetyczna dla tych danych wynosi 2,70 (z dokładnością do dwóch miejsc po przecinku).

Tabela 4. Dane na podstawie *Semantyki 2* Lyonsa w wersji angielskiej

x	1	2	3	4	5	6	7	8	9	10	20	Suma
f1	67	66	67	46	27	16	14	7	5	1	1	317
f2	104	116	102	73	45	23	17	7	6	2	1	496

Dla pierwszej części tekstu mamy $Q_1 = 2$, $Q_2 = 3$ i $Q_3 = 4$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 1$, zatem brak formalnego potwierdzenia skośności. Średnia arytmetyczna dla tych danych wynosi 3,29 (z dokładnością do dwóch miejsc po przecinku). Dla całości badanego odcinka tekstu uzyskujemy $Q_1 = 2$, $Q_2 = 3$ i $Q_3 = 4$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 1$,

nie mamy więc formalnego potwierdzenia skośności. Średnia arytmetyczna dla tych danych wynosi 3,15 (z dokładnością do dwóch miejsc po przecinku).

Tabela 5. Dane na podstawie *Semantyki 2* Lyonsa w wersji polskiej

x	1	2	3	4	5	Suma
f1	172	100	42	15	7	336
f2	276	163	69	20	9	537

Dla pierwszej części tekstu otrzymujemy $Q_1 = 1$, $Q_2 = 1$ i $Q_3 = 2$. $Q_2 - Q_1 = 0$ oraz $Q_3 - Q_2 = 1$, czyli mamy formalnie stwierdzoną skośność dodatnią. Średnia arytmetyczna dla tych danych wynosi 1,76 (z dokładnością do dwóch miejsc po przecinku). Dla całości badanego odcinka tekstu uzyskaliśmy $Q_1 = 1$, $Q_2 = 1$ i $Q_3 = 2$. $Q_2 - Q_1 = 0$ oraz $Q_3 - Q_2 = 1$, mamy zatem formalnie stwierdzoną skośność dodatnią. Średnia arytmetyczna dla tych danych wynosi 1,74 (z dokładnością do dwóch miejsc po przecinku). Tłumacz tekstu angielskiego nie zachował struktury składniowej tekstu, jego zdania charakteryzują się mniejszą złożonością, czego konsekwencją jest większa liczba zdań na tym samym odcinku tekstu.

Tabela 6. Dane na podstawie *Zarysu logiki matematycznej* Grzegorczyka w wersji angielskiej

x	1	2	3	4	5	6	7	8	9	Suma
f1	160	114	64	32	20	4	3	2	0	399
f2	209	155	99	57	36	4	4	3	1	568

Dla pierwszej części tekstu uzyskaliśmy $Q_1 = 1$, $Q_2 = 2$, $Q_3 = 3$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 1$, brak więc formalnego potwierdzenia skośności. Średnia arytmetyczna dla tych danych wynosi 2,18 (z dokładnością do dwóch miejsc po przecinku). Dla całości badanego odcinka tekstu otrzymaliśmy $Q_1 = 1$, $Q_2 = 2$ i $Q_3 = 3$. $Q_2 - Q_1 = 1$ oraz $Q_3 - Q_2 = 1$, nie mamy formalnego potwierdzenia skośności. Średnia arytmetyczna dla tych danych wynosi 2,30 (z dokładnością do dwóch miejsc po przecinku).

Tabela 7. Dane na podstawie *Zarysu logiki matematycznej* Grzegorczyka w wersji polskiej

x	1	2	3	4	5	6	7	8	Suma
f1	194	101	41	28	12	6	1	0	383
f2	275	140	66	33	16	9	1	1	541

Dla pierwszej części tekstu uzyskujemy $Q_1 = 1$, $Q_2 = 1$ i $Q_3 = 2$. $Q_2 - Q_1 = 0$ oraz $Q_3 - Q_2 = 1$, czyli mamy formalnie stwierdzoną skośność dodatnią. Średnia arytmetyczna dla tych danych wynosi 1,92 (z dokładnością do dwóch miejsc po przecinku).

Dla całości badanego odcinka tekstu otrzymujemy $Q_1 = 1$, $Q_2 = 1$ i $Q_3 = 2$. $Q_2 - Q_1 = 0$ oraz $Q_3 - Q_2 = 1$, czyli mamy formalnie stwierdzoną skośność dodatnią. Średnia arytmetyczna dla tych danych wynosi 1,91 (z dokładnością do dwóch miejsc po przecinku).

Wahania długości zdania

Przyjrzyjmy się zestawieniu danych w tabeli 8. Niezależnie od tego, czy do analizy weźmiemy medianę czy średnią arytmetyczną, uzyskamy ten sam wniosek: średnia długość zdania nie zmienia się na danym tekście, nie zależy więc od liczby analizowanych zdań. Należy jednak zauważyć, że w wypadku każdego z badanych tekstów analizowane były zamknięte całości, w tym wypadku rozdziały książek. Potwierdzony został wniosek Shermana dotyczący braku wahania średniej długości zdania w odcinkach tego samego tekstu. Sherman doszedł do tego wniosku, mierząc zdania liczbą wyrazów. W badaniu opisanym w niniejszym artykule zdania złożone mierzone były liczbą zdań składowych. Wniosek ten potwierdzono zatem niezależnie od wybranej jednostki miary.

Tabela 8. Zestawienie danych uzyskanych na badanych tekstach

Tekst	Wersja	Mediana	Średnia arytmetyczna
Hobbit	angielska	2	2,91
		2	2,85
	polska	2	2,77
		2	2,70
Semantyka 2	angielska	3	3,29
		3	3,15
	polska	1	1,76
		1	1,74
Zarys logiki matematycznej	angielska	2	2,18
		2	2,30
	polska	1	1,92
		1	1,91

Gatunek literacki tekstu

Sherman uznał, że gatunek tekstu może determinować średnią długość zdania. Przyjrzyjmy się jeszcze raz danym zawartym w tabeli 8. Mamy dwa gatunki tek-

stu: tekst literacki oraz dwa teksty naukowe. Nie można dostrzec żadnej zależności między średnią długością zdania mierzonego liczbą zdań składowych a gatunkiem tekstu. Średnie długości zdania w tekście *Zarys logiki matematycznej* w obu wersjach językowych są krótsze niż w tekście *Hobbita*. W porównaniu z *Semantyką 2* w wersji angielskiej *Zarysu* średnia długość zdania jest dłuższa niż w angielskiej wersji *Hobbita*, a w wersji polskiej krótsza. Być może zastosowanie innej jednostki do badania tekstów, tj. liczby zdań składowych, a nie liczby wyrazów, spowodowało niesystematyczne zachowanie się tekstów. Tekst *Semantyki 2* może być potwierdzeniem tego przypuszczenia. W wersji angielskiej pojawiają się zdania bogato złożone, natomiast w wersji polskiej zdania są najwyżej pięciokrotnie złożone. Mimo to w obu wersjach całkowita liczba badanych zdań jest zbliżona.

Wersje językowe

Tabela 8 pokazuje również, że średnia długość zdania złożonego mierzonego liczbą zdań składowych może nie zależeć od języka tekstu. Mediana wersji angielskich wynosi 1, 2 oraz 3, wersji polskich − 1 i 2. Średnie arytmetyczne dla wersji angielskich przyjmują wartości od 2,18 do 3,29, a dla wersji polskich – w przedziale między 1,74 a 2,77. Oczywiście badaliśmy znikomą liczba tekstów, można jednak zauważyć, że przedział wartości średnich arytmetycznych dla wersji polskich jest przesunięty w lewo w stosunku do przedziału wartości średnich arytmetycznych dla wersji angielskich. Większa liczba badanych tekstów mogłaby zweryfikować ten wniosek w sposób ostateczny.

Rozkłady częstości względnych

W tej części artykułu przyjrzymy się częstościom względnym badanych tekstów. Dane zgromadziłam w tabelach 9 i 10, biorąc pod uwagę jedynie całe odcinki badanych tekstów oraz około 95% ich najdłuższych zdań w celu porównania tych częstości względnych.

Tabela 9. Częstości względne badanych tekstów w wersjach angielskich

x	1	2	3	4	5	6	7
Hobbit	0,2595	0,2530	0,2073	0,1197	0,0705	0,0397	–
Semantyka 2	0,2097	0,2339	0,2056	0,1472	0,0907	0,0464	0,0343
Zarys logiki matematycznej	0,3680	0,2729	0,1743	0,1004	0,0634	–	–

Tabela 10. Częstości względne badanych tekstów w wersjach polskich

x	1	2	3	4	5	6
Hobbit	0,2751	0,2815	0,1968	0,1104	0,0631	0,0315
Semantyka 2	0,5140	0,3035	0,1285	–	–	–
Zarys logiki matematycznej	0,5083	0,2588	0,1220	0,0610	–	–

Możemy zauważyć, że 95% zdań w polskiej wersji badanych tekstów jest najwyżej sześciokrotnie złożonych, podczas gdy 95% zdań w wersji angielskich jest najwyżej siedmiokrotnie złożonych. Zauważmy też, że oryginalny *Zarys logiki matematycznej* powstał w języku polskim, a wersja angielska jest jego tłumaczeniem. Tłumacz *Semantyki 2*, jak już wcześniej zaobserwowaliśmy, nie zachował oryginalnej struktury tekstu oryginału angielskiego. Obie wersje *Hobbita* strukturalnie są bardzo zbliżone, co możemy stwierdzić porównując rozkład częstości względnych długości zdań złożonych mierzonych liczbą zdań składowych. Przyjrzyjmy się jeszcze rozkładom częstości względnych długości zdań złożonych w czterech innych tekstach w obu wersjach językowych, podanych w tabelach 11 i 12.

Tabela 11. Częstości względne badanych tekstów w wersjach angielskich

x	1	2	3	4	5	6
Kod Leonarda da Vinci	0,4582	0,2675	0,1691	0,0529	–	–
Zagadnienia teorii składni	0,2119	0,3436	0,2284	0,1317	0,0494	–
Zabić drozda	0,2656	0,2949	0,2028	0,1088	0,0678	0,0273
Demony dobrego Dextera	0,5165	0,2989	0,1108	0,0441	–	–

Tabela 12. Częstości względne badanych tekstów w wersjach polskich

x	1	2	3	4	5
Kod Leonarda da Vinci	0,4338	0,3045	0,1586	0,0671	–
Zagadnienia teorii składni	0,3143	0,3295	0,1905	0,1010	0,0343
Zabić drozda	0,3302	0,3231	0,1807	0,0828	0,0435
Demony dobrego Dextera	0,5910	0,2936	0,0927	0,0326	–

Podobnie jak w tekstach rozważanych wcześniej także tutaj możemy zauważyć, że rozkłady częstości względnych długości zdań złożonych są odmienne dla różnych tekstów oraz dla ich tłumaczeń. Potwierdza to trudność, jaką napotykają inni badacze poszukujący modeli tych rozkładów. Jedyną wspólną cechą tekstów jest to, że udział zdań najwyżej czterokrotnie złożonych stanowi co najmniej 75% wszystkich zdań złożonych w danym tekście mierzonych liczbą zdań składowych. Dotychczas

nie znaleziono zadowalającego modelu, który po uwzględnieniu pewnych parametrów dobrze przybliżałby rozkłady długości zdań złożonych danego tekstu.

Perspektywa dalszych badań

Dalsze badania nad prawem Shermana powinny uwzględnić teksty tego samego autora, co pozwoliłoby stwierdzić, czy obliczone średnie długości zdań złożonych są charakterystyczne dla danego pisarza. Wydaje się, że adekwatną jednostką służącą do badania tych długości jest liczba zdań składowych. Ze względu na typologię języki tę samą treść mogą oddawać różną liczbą wyrazów, a tą samą liczbą zdań składowych. Po uzyskaniu wyników takich badań można będzie więcej powiedzieć na temat tego, jakie parametry należy uwzględnić w zaproponowanych modelach statystycznych. Na tym etapie badań można jedynie stwierdzić, że rozkłady te może charakteryzować skośność dodatnia, jednak kształt potencjalnej funkcji gęstości prawdopodobieństwa będzie różny dla tekstów różnych autorów.

Literatura

ALTMANN, G., 1992, *Sherman's Laws of Sentence Length Distribution*, *What is language synergetics?*, „ Acta Universitatis Ouluensis" 16, s. 38–39.

KUŁACKA, A., 2008, *Badania nad prawem Menzeratha-Altmanna*, „LingVaria" nr 2 (6), s. 167–174.

SHERMAN, L. A., 1888, *Some Observations upon the Sentence-Length in English Prose,* „University Studies" 1, nr 2, s. 119–130.

Sherman's Law
Summary

The aim of this article is to familiarise the audience with Sherman's laws, which is a set of laws describing the average sentence-length of texts and possible models describing distributions of sentence-lengths in texts with respect to either number of clauses or words. We also confirm Sherman's conclusion regarding the invariance of an average sentence-length in a given text. However, after the analysis of our data we concluded that this length does not depend on a sort of text, neither on language used. We also analysed the distributions of sentence-length with respect to the number of clauses and concluded that the characteristics of the possible statistical models are positive skewness of the data and general indeterminacy of the shape of a density function or its dependence on a given text.