http://dx.doi.org/10.17651/POLON.38.2

POLONICA XXXVIII
PL ISSN 0137-9712

JAROSŁAW FOLTMAN

Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków

# Jak mówią postaci „Lalki" Bolesława Prusa. Analiza stylometryczna

## Wstęp

Tekst „Lalki" w znacznej mierze złożony jest z wypowiedzi postaci, dialogów, myśli bohaterów i „pamiętników starego subiekta". Bolesław Prus, zmniejszając udział narratora, bardzo często oddawał głos bohaterom swojej powieści. Z językowego punktu widzenia mogą nasunąć się zatem pytania: Jak bardzo zróżnicowane stylistycznie są te poszczególne wypowiedzi? Czy głosy różnych postaci zmieniają się w zakresie stosowanych w ich kwestiach form gramatycznych? Czy autor, budując odmienne sylwetki, prócz rysowania oczywistych różnic w charakterze czy światopoglądzie, wyposażał je również w odmienne sposoby mówienia? W końcu: czy częstość używania poszczególnych słów lub części mowy zmienia się w zależności od tego, kto mówi?

W artykule podjęto próbę odpowiedzi na powyższe pytania. W tym celu zostaną przedstawione wnioski z analizy wyników kilku eksperymentów stylometrycznych przeprowadzonych na tekście powieści, a dokładniej — na wypowiedziach jej bohaterów.

W szerokim rozumieniu stylometria jest metodą analizy tekstu badającą jego ilościowe i jakościowe parametry; stosuje się ją głównie z zamiarem tworzenia statystyk występowania określonych elementów języka, wzorów i powtórzeń słów, odszukiwania pewnych regularności i zależności między tekstami. Najczęściej przyjmowanym celem badań stylometrycznych są dążenia do przypisania autorstwa anonimowym utworom, ustalanie chronologii ich powstawania, wykrywanie plagiatów czy wreszcie eksploracja różnic stylistycznych w tekstach odmiennych autorów lub epok.

Zakres użycia różnorakich metod stylometrycznych zwiększa się z każdym rokiem, na co ma wpływ nie tylko szybki rozwój możliwości współczesnych komputerów i oprogramowania, ale również stale zwiększająca się baza tekstów dostępnych w formie elektronicznej. Ogromna liczba utworów literackich i publicystycznych, prace naukowe, całe roczniki gazet i wiele innych publikacji zbierane są w licznych cyfrowych bazach danych, dając sposobność takiego spojrzenia na wiele dziedzin humani-

stycznych, w tym językoznawczych, o jakim nie było mowy jeszcze w latach 70. czy 80. XX wieku. Oczywiście jak na razie żaden program komputerowy nie jest w stanie przeprowadzić analizy obejmującej wszystkie aspekty języka. Wciąż bardzo trudno bowiem za pomocą nawet najbardziej wyrafinowanych algorytmów zinterpretować wieloznaczność struktur gramatycznych czy warunkowane kontekstem odcienie semantyczne wypowiedzi. Ale nawet jeśli komputery nie przeniknęły jeszcze do głębokich struktur mowy ludzkiej, to z pewnością dzięki nim udaje się przynajmniej przeprowadzić wiele eksperymentów językowych, które trudno byłoby zrealizować metodami klasycznymi.

Niniejszy artykuł dotyczy podstawowych elementów języka, bez dociekania czynników definiujących styl i bez rozważania zależności między częstościami wyrazów a stylem właśnie[1]. Omówione będą sposoby stosowania poszczególnych części mowy i frekwencja pojedynczych jednostek leksykalnych. Być może w dalszej perspektywie badawczej powinna się również pojawić pogłębiona refleksja na temat tego, jakimi innymi sposobami Prus mógł urozmaicać postaci swoich utworów (można bowiem osiągnąć ten efekt choćby za pomocą modyfikacji składniowych), ale nie będzie to tematem tego opracowania.

## Co literaturoznawcy mówią o języku „Lalki”?

W „Lalce” występują bohaterowie należący do różnych klas społecznych. Są tu przedstawiciele klasy niższej, robotniczej, usługowej — służący, woźnice, szwaczki itd. Znaczną część postaci stanowi klasa średnia: mieszczanie, subiekci, kupcy, adwokaci. Szlachtę i arystokrację reprezentują prawie tak samo liczne grupy.

W tekstach poświęconych językowej analizie powieści niemal od samego początku pojawiały się stwierdzenia, że pisarz używał prostego, przejrzystego języka, czasem nawet potocznego, dbając jednocześnie, aby wszystkie postaci mówiły językiem właściwym swojemu pochodzeniu, środowisku, wykształceniu i grupie społecznej. Odrobinę zatem archaiczna mowa Rzeckiego (przykładowe zwroty: *kamraci*, *facecje*, *acan*) różnić się ma od języka arystokracji, pełnego zapożyczeń z francuskiego i angielskiego (przykładowe zwroty: *dystyngowany*, *dżentelmen*, *maniery*), od mowy kupców, operujących terminologią handlową (np. *ajent*, *obstalunek*, *pryncypał*, *rabat*), czy od mowy kelnerów i fryzjerów, używających zdrobniałych form (np. *numerek kaloszy*, *przybić literki*).

Jeszcze w pierwszej połowie XX w. Halina Kurkowska dowodziła, że narracja „Lalki” podlega rygorowi konstrukcji języka pisanego, natomiast przytoczenia przynoszą luźny szyk i wykolejenia charakterystyczne dla żywej mowy (Kurkowska 1948: 2–8).

Józef Bachórz pisał:

[1] Już samo ustalenie, co rozumiane jest przez styl, wykracza znacznie poza ramy tego artykułu. Przegląd definicji oraz sposobów rozumienia pojęcia, również w ujęciu stylometrycznym, znaleźć można na przykład w opracowaniu „Revisiting style, a key concept in literary studies” (Herrmann et al. 2015).

> Silnie indywidualizujące i zarazem osadzające w środowisku mówienie każdej postaci „po swojemu" jest w „Lalce" stylizowane niezawodnie […]. W żydowskich odkształceniach polszczyzny starego Szlangbauma, niemieckich starego Mincla, w nalotach angielszczyzny hrabiego Licińskiego i wtrętach francuskich w języku panny Izabeli są wspomnienia lekcji stylizacyjnych z komedii i powieści dawniejszych; w sposobie mówienia Suzina posłyszeć można echo kapitana Rykowa z „Pana Tadeusza". Język Węgielka przybarwia się gwarowo, lokaj Wokulskiego po warszawsku „szadzi" (Bachórz 1998: LVIII).

Tadeusz Budrewicz podsumowywał, że

> można […] w „Lalce" wyróżnić styl liryczny, gawędowy, reporterski, satyryczny itd. Wymieniają się wraz z przeplataniem się tematów i stanowiskiem autora wobec poruszanych kwestii. W tym sensie wolno mówić o stylistycznej różnorodności. Różnorodność zapewnia też indywidualizacja i charakterystyka językowa postaci, widoczna w ich wypowiedziach ustnych i monologach wewnętrznych (Budrewicz 1990: 186)[2].

## Metoda

Wspólną cechą przytoczonych opisów, jak i większości innych opracowań, jest skoncentrowanie uwagi badaczy na najbardziej charakterystycznych, znaczących i unikatowych słowach czy też dystynktywnych zwrotach dostrzegalnych nawet przy pobieżnym czytaniu, używanych bowiem wyłącznie przez jedną postać. Takie podejście spełnia oczywiście swoją funkcję, jednak ma dość poważny brak: nie bierze zwykle pod uwagę najczęściej występujących, synsemantycznych elementów tekstu, jak przyimki, zaimki, spójniki, przysłówki itp. A te stanowią przecież ogromną część każdej wypowiedzi, frekwencyjnie przyćmiewając udział wyrazów autosemantycznych i stanowiąc zwykle około 40–50% tekstu. Wyrażenia funkcyjne są przez autorów stosowane automatycznie, tylko jako elementy konieczne do poprawnego używania określonych zwrotów, osnowa myśli wyrażanej konkretnymi jednostkami znaczącymi.

Wpływ wyrażeń funkcyjnych na styl jest znany językoznawcom od dawna, a sposoby badania częstości ich używania są wypracowywane na całym świecie (kilka znaczących publikacji: Burrows 2002, Eder 2011, Hoover 2002, Rybicki 1997/1998). John Frederick Burrows, jeden z najwybitniejszych teoretyków i praktyków stylometrii w historii, w wydanej w 1987 r. książce „Computation into Criticism. A Study of Jane Austen's Novels and an Experiment in Method" tak pisał o bardzo wysokiej frekwencji wyrazów synsemantycznych i ich znaczeniu:

> Rozkłady częstości bardzo częstych słów dlatego są tak przydatne, że słowa nie funkcjonują w tekście jako indywidualne jednostki. Ponieważ swe pełne znaczenie uzyskują dopiero wchodząc w najróżniejsze związki między sobą, można widzieć w nich wyznaczniki tych związków i wszystkiego,

[2] Przytoczono tutaj z oczywistych względów tylko niewielką część opracowań „Lalki". Jednak koniecznie należy jeszcze wspomnieć o pracy Teresy Smółkowej „Słownictwo i fleksja «Lalki» Bolesława Prusa. Badania statystyczne", która w szczegółowy i wyczerpujący sposób zbiera wiele językowych danych statystycznych „Lalki", jak choćby częstości występowania poszczególnych części mowy czy form przypadkowych; praca nie porusza jednak istotnej dla niniejszego artykułu kwestii różnic w sposobach wypowiadania się poszczególnych bohaterów powieści.

> co owe związki oznaczają w sensie semantycznym. I tak: tam, gdzie spotykamy więcej najczęstszych przyimków, mamy zazwyczaj do czynienia ze stylem bardziej opisowym czy refleksyjnym; jeżeli z kolei jest ich mało, oznacza to, że akcja rozgrywa się na znacznie skromniej udekorowanej scenie (Burrows 1987: 12, fragment w tłumaczeniu J. Rybickiego).

Owe pozornie nieznaczące elementy języka mogą być istotnym wskaźnikiem zróżnicowania stylistycznego tekstów, równie doniosłym jak stosowanie takich czy innych wyrażeń charakterystycznych o silnym zabarwieniu stylistycznym. O ile dla badaczy posługujących się tradycyjnymi metodami ważne może być kilka wyjątkowych, powtarzanych słów, to dla rozważań stylometrycznych istotniejsze jest, z jaką częstością w poszczególnych tekstach są stosowane wyrazy funkcyjne.

W przypadku materiału omawianego w tym opracowaniu interesujące jest więc, jakie są względne frekwencje użycia słów funkcyjnych przez indywidualnych bohaterów „Lalki". Czy używają ich z różnym natężeniem? Czy wprowadzając do ich słownictwa elementy mocno znaczące, przypisujące przez frazeologię do grup społecznych, autor stosował jednocześnie rozróżnienie w tej niższej, „mechanicznej" warstwie języka?

Ale jednocześnie szerzej można zapytać, czy i ewentualnie w jaki sposób Prus odróżnił bohaterów swojej powieści w innych warstwach ich wypowiedzi. Czy zróżnicował język postaci poprzez zróżnicowanie częstości używania przez nie określonych form gramatycznych lub części mowy? Które czynniki można zmierzyć, by móc powiedzieć coś więcej o stylu wypowiedzi bohaterów?

Poniższe rozważania są zapisem poszukiwań odpowiedzi na część tych pytań.

Jakiekolwiek doświadczenie stylometryczne, niczym każde inne badanie statystyczne, wymaga podjęcia szeregu decyzji odnośnie do używanych narzędzi i ich odpowiedniej konfiguracji. Do najważniejszych czynników należą w przypadku omawianego eksperymentu: metoda mierzenia stopnia podobieństwa między tekstami, minimalna liczba wypowiadanych przez postać słów oraz badana liczba najczęściej wypowiadanych słów; ważne są również programy, za pomocą których przeprowadza się pomiary, oraz sposoby wizualizacji danych.

Jeśli chodzi o zastosowaną w niniejszym badaniu statystyczną metodę mierzenia podobieństwa między tekstami, to wybrano Deltę Burrowsa (2002), jako metodę uznawaną za jedną z najlepiej sprawdzonych i najczęściej stosowanych w tego typu doświadczeniach. Mimo że podejmowano wielokrotne próby udoskonalania tej miary odległości (m.in. Hoover 2004, Argamon 2008, Eder et al. 2013), to jednak panuje przekonanie, że daje ona jednolite wyniki, które są zbliżone do innych proponowanych metod.

Do prezentowania wyników zastosowano skalowanie wielowymiarowe oraz drzewo zgodności. Skalowanie wielowymiarowe bierze pod uwagę różnice między frekwencjami pojawiania się słów u każdej z postaci, po czym prezentuje dane tak, by większe rozbieżności sygnalizowane były większymi odległościami w dwuwymiarowej przestrzeni wykresu. Drzewo zgodności w analogiczny sposób rozpoznaje dane wejściowe, wyniki przedstawiając jednak w formie drzewa, na którego gałęziach po-

łożone są wypowiedzi postaci; zbliżone do siebie idiolekty powinny znajdować się na stosunkowo bliskich gałęziach, te o znacznym stopniu zróżnicowania pojawią się na zupełnie odległych odnogach (więcej o tych metodach m.in. w: Baayen 2008: 136, Levshina 2015: 351–366).

Kolejnym istotnym czynnikiem wymagającym wstępnego rozważenia jest liczba słów wypowiadanych przez poszczególnych bohaterów, umożliwiająca włączenie ich do eksperymentów. Dla przeprowadzenia jakiegokolwiek badania konieczny jest określony zasób danych, nie da się bowiem budować zestawień statystycznych na podstawie zaledwie kilku dostępnych elementów. Jeśli zatem któraś z postaci pojawia się w utworze zbyt rzadko i wypowiada za mało kwestii, nie może być obiektem analizy. Toczą się dyskusje, jak obszerna powinna być próbka, by można było mówić o wiarygodnych wynikach doświadczeń stylometrycznych (Eder 2014, 2017), ale w przypadku badania pojedynczej powieści należy chyba mieć na względzie dodatkowy czynnik, jakim jest mocno ograniczony zasób słów dostępnych w analizowanym tekście. Jeśli więc w przypadku prób atrybucji autorskiej różnych utworów, z których każdy liczy setki tysięcy słów, rozsądne wydaje się rozpatrywanie próbek obejmujących 2000 czy 5000 jednostek i więcej, to dla pojedynczej powieści jest to wartość trudna do utrzymania. Dlatego w przypadku niniejszego badania „Lalki" zdecydowano o wyborze postaci wypowiadających tylko około 800 słów, co i tak w praktyce spowodowało wykluczenie sporej grupy bohaterów.

Po ustaleniu preferowanej minimalnej objętości próbki należało określić, ile najczęstszych słów będzie brane pod uwagę w doświadczeniach. Idealnym rozwiązaniem byłoby przeprowadzenie cyklu testów ze zmieniającą się wielkością parametru, na przykład dla wartości 50, 100, 200, 500, 1000 najczęściej występujących słów, ale tak samo jak w przypadku długości próbki trzeba mieć na względzie dostępne dane i ograniczenia narzucone przez objętość powieści. W związku z tym, że większość bohaterów „Lalki" wypowiada od 1000 do 2250 słów (Tabela 1), trudno o tworzenie zestawienia setek czy tysięcy najczęściej występujących słów. Postanowiono więc przeprowadzić niezależne badania na 25, 50, 75, 100 i 150 najczęściej występujących słowach i porównać wyniki. Ostatecznie poniżej zostaną zaprezentowane dane dla 75 najczęściej występujących słów, ponieważ dały one najwyraźniejsze rezultaty, chociaż i w przypadku pozostałych rozpatrywanych wartości wyniki były zbliżone.

## Zagadnienia techniczne

Z „Lalki" zostały wyodrębnione wypowiedzi poszczególnych postaci i zapisane w formie osobnych plików tak, że wszystkie wypowiedzi danego bohatera stawały się jednym ciągłym tekstem. Następnie te zestawy wypowiedzi zostały w trzech wariantach zapisane jako:

1. Tekst w słowoformach takich jak w powieści, z tym zastrzeżeniem, że wszystkie wyrazy zapisano małymi literami w celu uniknięcia sytuacji, w której to samo sło-

wo, pisane małą lub dużą literą, byłoby traktowany jako dwie odrębne słowoformy. Z wypowiedzi została usunięta interpunkcja. Ze względu na nieciągły, fragmentaryczny kształt tekstu (wynikający z usunięcia komentarzy narratora, opisów fabularnych itp.) każde słowo zostało przeniesione do osobnego akapitu, tak by całość tworzyła ciąg wypowiedzianych przez danego bohatera słów; np.: *mówili*, *że*, *wielmożny*, *pan*, *już*, *na*, *ostatnich*, *nogach*, *ale*, *widzę*.

2. Wyrazy zlematyzowane, czyli sprowadzone do podstawowej formy gramatycznej (formy słownikowej); podobnie jak w poprzednim przypadku ułożone w formie listy; np.: *mówić*, *że*, *wielmożny*, *pan*, *już*, *na*, *ostatni*, *noga*, *ale*, *widzieć*. Do uzyskania form zlematyzowanych użyto tagera języka polskiego TaKIPI wraz z analizatorem morfologicznym Morfeusz.
3. Formy słownikowe zamienione na etykiety odpowiadających im części mowy, również składające się na listę; np.: prate, conj, adj, subst, ub, prep, adj, subst, conj, fin.

Analizowane były wypowiedzi adwokata, arystokratów, barona Dalskiego, Florentyny, Geista, Heleny Stawskiej, Ignacego Rzeckiego (w całości, dialogi, pamiętniki), Izabeli Łęckiej, Jadwigi Misiewiczowej, hrabiny Joanny Karolowej, Juliana Ochockiego, Kazimiery Wąsowskiej, Kazimierza Starskiego, baronowej Krzeszowskiej, barona Krzeszowskiego, księcia, Maruszewicza, subiekta Mraczewskiego, służących, Stanisława Wokulskiego (w całości, dialogi, myśli), Szlangbauma, doktora Szumana, stolarza Węgiełka, rządcy Wirskiego, prezesowej Zasławskiej.

Do zestawu włączony został także tekst całej książki („całość"), plik obejmował cały tekst powieści, łącznie z wypowiedziami bohaterów, narracją autorską, opisami fabularnymi itd.

Pozostałe występujące w powieści postaci zostały pominięte ze względu na niewielką liczbę wypowiadanych kwestii, udaremniającą budowanie odpowiednich zestawień frekwencyjnych. Najmniejsza rozpatrywana próbka składała się z 783 słów.

Wypowiedzi postaci jednoznacznie przypisywanych do grupy społecznej służących, ale nieprzywołane imiennie włączono do jednego pliku „służący". Taka sama sytuacja miała miejsce w przypadku nienazwanych arystokratów — plik „arystokraci".

Ze względu na specjalne miejsce zajmowane w powieści przez Rzeckiego w jego przypadku stworzono trzy odrębne pliki: jeden zawierający tylko te wypowiedzi, które pojawiają w rozdziałach „pamiętnika starego subiekta", drugi z jego wypowiedziami dialogowymi, trzeci zbierający wszystkie kwestie (pisane i mówione). Rozgraniczenie to miało na celu sprawdzenie, czy Prus zróżnicował styl jego mówienia i pisania.

Podobnie w przypadku Wokulskiego przygotowano trzy pliki: jeden całościowy, z wszystkimi wypowiedziami, drugi z jego kwestiami dialogowymi oraz trzeci zawierający tylko przemyślenia i dialogi wewnętrzne; te ostatnie stanowią bowiem znaczną część jego kwestii (około 45%).

Tabela 1. Zestawienie analizowanych postaci wraz z liczbą wypowiadanych przez nie słów

| **Postać** | **Liczba wypowiadanych słów** |
|---|---|
| książę | 783 |
| służący | 808 |
| Florentyna | 936 |
| Helena Stawska | 1043 |
| Kazimierz Starski | 1043 |
| Maruszewicz | 1087 |
| adwokat | 1273 |
| subiekt Mraczewski | 1505 |
| Szlangbaum | 1581 |
| prezesowa Zasławska | 1752 |
| hrabina Joanna Karolowa | 1861 |
| stolarz Węgiełek | 1935 |
| arystokraci | 1940 |
| baron Dalski | 2028 |
| Jadwiga Misiewicz | 2040 |
| Geist | 2046 |
| rządca Wirski | 2047 |
| baronowa Krzeszowska | 2064 |
| baron Krzeszowski | 2247 |
| Ignacy Rzecki dialogi | 4617 |
| Kazimiera Wąsowska | 4914 |
| Julian Ochocki | 4939 |
| doktor Szuman | 6498 |
| Izabela Łęcka | 7058 |
| Stanisław Wokulski myśli | 13044 |
| Stanisław Wokulski dialogi | 15594 |
| Stanisław Wokulski | 28448 |
| Ignacy Rzecki pamiętniki | 39811 |
| Ignacy Rzecki | 75555 |
| **całość** | **245234** |

Do przeprowadzenia analiz statystycznych i wizualizacji wyników użyto pakietu stylo() (Eder, Rybicki 2011).

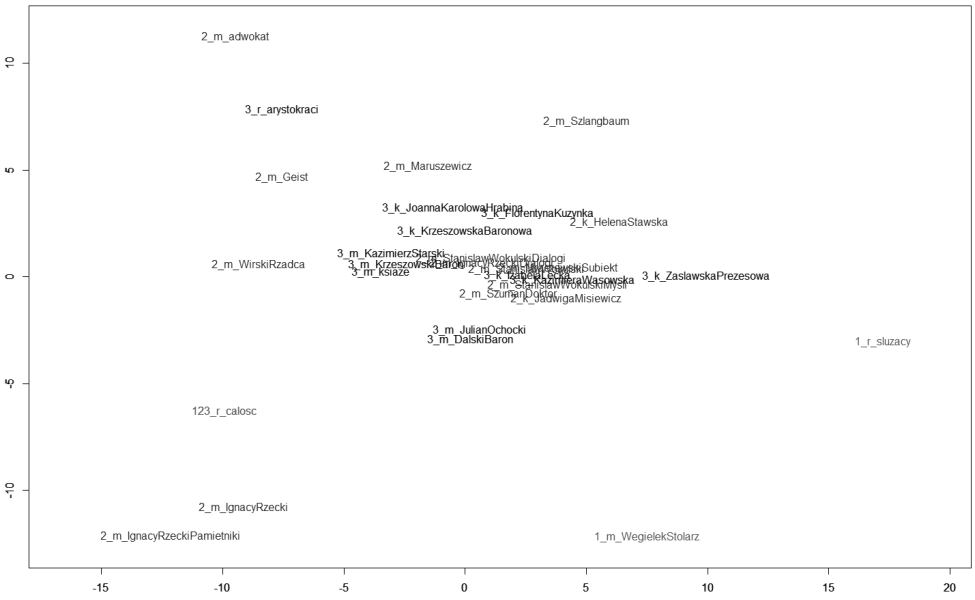
## Wyniki

Rozpatrując wyniki doświadczeń przeprowadzanych na omawianych tekstach, należy wziąć pod uwagę fakt, że jakkolwiek stylizowana by była mowa postaci „Lalki", to cały czas badanie dotyczy jednego i tego samego autora piszącego powieść w ciągu niecałych dwóch lat — a więc w niezbyt długim odcinku czasu. Stąd zastrzeżenie, że nawet jeśli statystyki dotyczące poszczególnych bohaterów będą się od siebie odróżniać, to nie należy spodziewać się aż takich zmian, jak gdyby pod rozwagę byli brani zupełnie różni autorzy.

### Części mowy

Analizując, jakich części mowy używają bohaterowie powieści, można się dopatrzyć kilku dość charakterystycznych tendencji. Wykres 1 pokazuje w formie graficznej wynik skalowania wielowymiarowego dokonanego na tekście.

Większość postaci zlokalizowana jest wyraźnie w środkowej części wykresu, co oznacza zbliżoną frekwencję używania poszczególnych części mowy; czyli na przykład kuzynka Florentyna używa rzeczowników średnio w 25,73% przypadków (statystycznie na 100 wypowiedzianych przez nią słów około 25 będzie rzeczownikami), podobnie jak Kazimiera Wąsowska, dla której średnia wynosi 25,37% itd. Podobnie rzecz wygląda w przypadku znacznej części bohaterów, jednak kilkoro w widoczny




Wykres 1. Różnice w użyciu części mowy (cyfra 1 poprzedzająca wpisy sygnalizuje arbitralne przypisanie postaci do najniższej klasy społecznej, 2 — do klasy średniej, 3 — do szlachecko-arystokratycznej. Litery następujące po cyfrach porządkują płcie: k — kobiety, m — mężczyźni, r — grupa mieszana).

sposób odróżnia się pod tym względem od pozostałych. Analiza liczb potwierdza tę obserwację.

Głównie uwagę zwraca zgrupowanie pamiętników Rzeckiego, jego wszystkich wypowiedzi oraz całego tekstu „Lalki". „Pamiętnik starego subiekta" stanowi znaczną część powieści (około 16%), wszystkie wypowiedzi Rzeckiego jeszcze mocniej wpływają na jej kształt, stanowiąc około 19% tekstu. Skoro zatem wypowiedzi tej postaci stanowią tak dużą część tekstu, to naturalny wydaje się fakt, że językowy charakter powieści jest przez nią mocno konstytuowany. Ciekawy jest przypadek wypowiedzi dialogowych tego bohatera, które na wykresie położone są razem z większością postaci. Wypływa stąd wniosek, że Rzecki-pamiętnikarz i Rzecki-rozmówca posługują się innymi idiolektami.

Drugą znacząco odstającą grupę tworzą służący (potraktowani zbiorczo) i stolarz Węgiełek, będący jedynymi przedstawicielami najniższej klasy społecznej w zestawieniu. W obydwu przypadkach położenie tych postaci wskazuje na znaczne różnice w sposobie budowania wypowiedzi. Analiza kilku używanych form daje znaczące odchylenia w stosunku do wartości przeciętnych. I tak na przykład średnia użycia przymiotników wynosi 11,25%, a dla tych dwu grup odpowiednio: służący — 7,26%, Węgiełek — 9,27%, co może oznaczać próbę budowania przez Prusa prostych, mało opisowych wypowiedzi, pozbawionych elementów wartościujących, skoncentrowanych jedynie na konkretnych faktach, przedstawiających rzeczowo wydarzenia. Podobnie dzieje się w przypadku zaimków, gdzie przy średniej 8,23% służący mają 9,13%, ale Węgiełek już 12,82%; takie rozłożenie wartości może znów wskazywać stylizację na język osób prostych, często używających zaimków w funkcji uzupełniającej, przecinkowej. Pozostałe wartości dla tych dwóch bohaterów, choć mniej charakterystyczne, w każdym przypadku zauważalnie odchylają się od średnich wartości pozostałych postaci.

Znamienne wydaje się także usytuowanie na wykresie „arystokracji" (potraktowanej zbiorczo) po drugiej stronie zestawienia, niczym na przeciwnym biegunie „służących". Analiza danych liczbowych przynosi potwierdzenie takiego stanu rzeczy. Wartości liczbowe wskazują, że nienazwane osoby zaliczane do kategorii klasy wyższej charakteryzują się wynikami niemal dokładnie odwrotnymi niż klasy niższej. Tam, gdzie służący mają wartość poniżej średniej — arystokraci zwykle wykazują użycie z częstością wyższą niż średnia. Tam znów, gdzie służący używają jakiejś części mowy częściej niż średnia wszystkich postaci — mowa arystokratów charakteryzuje się niższymi wartościami. Na przykład rzeczowniki mają średnią użycia równą 25,5%, służący — 24,12%, arystokraci — 28,93%; przymiotniki odpowiednio 11,25%, 7,26%, 14,29%; spójniki — 9,43%, 11,94%, 8,08%.

Prócz powyższych na wzmiankę zasługuje postać adwokata, która być może ze względu na charakterystyczny żargon prawniczy została obdarowana przez Prusa dodatkowymi właściwościami językowymi, co znalazło odzwierciedlenie na wykresie i w danych liczbowych. Najczęściej ze wszystkich używa rzeczowników oraz liczebników (na 2. miejscu), w kilku innych kategoriach znajduje się blisko skrajnych pozycji listy.

Stosunkowo niedaleko od adwokata usytuowani są Szlangbaum, Maruszewicz i Geist, tak samo jak odrobinę bardziej po stronie Rzeckiego znajdują się rządca Wirski, baron Dalski i Julian Ochocki. Wskazuje to oczywiście na jakieś różnice, jednak skoro odległości nie są tak znaczące jak w przypadku wcześniej omawianych grup lub postaci, to i liczby nie przynoszą większych rozbieżności.

Na podstawie danych i wykresu trudno dopatrzeć się jakiejkolwiek zależności między sposobem mówienia a płcią bohatera. Chociaż wszystkie kobiece postaci zgrupowane są w centrum, otoczone kilkoma męskimi, to nie wydaje się to w jakikolwiek sposób działaniem celowym lub uporządkowanym.

**Najczęściej używane słowa**

Drugie z doświadczeń polegało na stworzeniu tabeli 75 najczęściej występujących w powieści słów, a następnie sprawdzeniu, z jaką częstością pojawiają się u indywidualnych bohaterów. Tak zbudowany profil każdej figury został zestawiony z pozostałymi postaciami. Wyniki prezentuje Wykres 2 (skalowanie wielowymiarowe) oraz Wykres 3 (drzewo zgodności). Obydwa są oparte na tych samych danych, ale w nieco różny sposób prezentują wyniki.

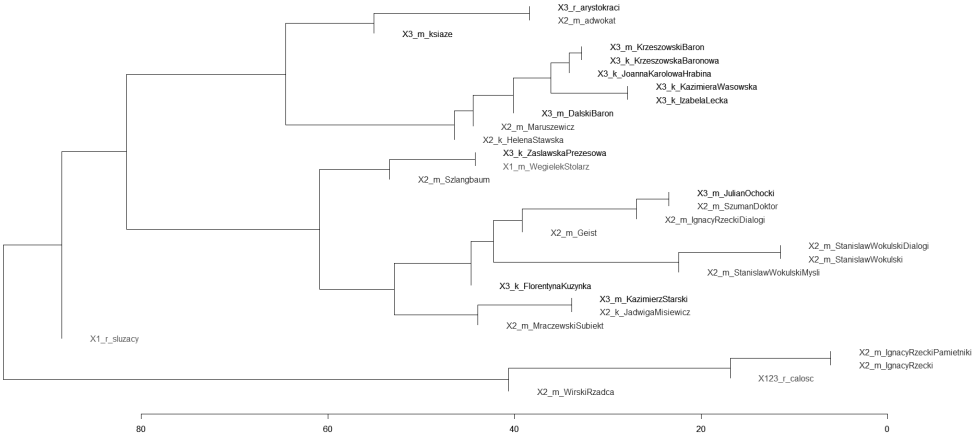
Lista 75 najczęściej używanych słów zawiera następujące wyrazy: *i*, *się*, *nie*, *w*, *na*, *z*, *że*, *do*, *a*, *to*, *pan*, *o*, *ale*, *co*, *za*, *jest*, *jak*, *mnie*, *pani*, *mi*, *tak*, *ja*, *wokulski*, *po*, *już*, *od*, *nawet*, *tylko*, *czy*, *go*, *mu*, *może*, *jej*, *więc*, *on*, *tym*, *jeszcze*, *panie*, *ma*, *ten*, *bo*, *ze*, *dla*, *sobie*, *który*, *jeżeli*, *bardzo*, *był*, *ażeby*, *tej*, *albo*, *jego*, *kiedy*, *cóż*, *przez*, *u*, *tego*, *no*, *tu*, *dziś*, *panna*, *rubli*, *pana*, *ją*, *nic*, *tam*, *jestem*, *wszystko*, *nas*, *było*, *nim*, *przed*, *przy*, *będzie*, *być*.

Ograniczenie obserwacji do 75 najczęściej występujących słów pozwoliło na wykluczenie z zestawienia większości nazw własnych (Izabela, Rzecki itd.), poza Wokulskim, który jako centralna postać powieści pojawia się bardzo często w wypowiedziach pozostałych postaci.

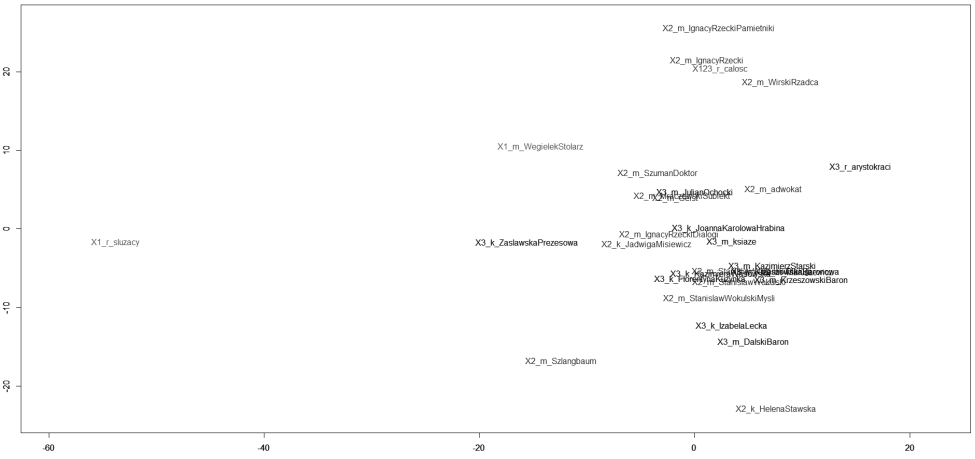
Jak widać, poza kilkoma przypadkami (*panna*, *rubli*, *być*) zestawienie obejmuje niemal same przyimki, spójniki i zaimki, a więc niesamodzielne znaczeniowo części języka, co potwierdza wcześniej przedstawione założenia teoretyczne.

Analizując graficzne reprezentacje frekwencji słów, tak w jednym, jak i w drugim przypadku należy zwrócić uwagę na zdecydowane odłączenie od wszystkich postaci samego tekstu powieści oraz pamiętników i całościowych wypowiedzi Rzeckiego, podobnie jak to miało miejsce w przypadku użycia części mowy. Zatem i w tym ujęciu Rzecki pisze w sposób zbliżony do tego jak narrator powieści, jednocześnie mocno odstając od niemal wszystkich pozostałych bohaterów, włączając w to nawet Rzeckiego jako postać mówiącą.

To, co nie było aż tak widoczne przy częściach mowy, a tutaj jest wyraźnie sygnalizowane, to fakt pojawienia się w tej samej grupie jeszcze wypowiedzi rządcy Wirskiego. Z formalnego punktu widzenia postać ta mówi więc językiem bardzo zbliżonym do Rzeckiego i narratora.




Wykres 2. Najczęściej występujące słowa, skalowanie wielowymiarowe




Wykres 3. Najczęściej występujące słowa, drzewo zgodności

O wiele większą różnicę prezentują służący. Nie sposób nie zauważyć, jak mocno miejsca reprezentowane przez tę kategorię postaci są graficznie oddzielone od pozostałych osób mówiących w tekście. Sprawdzenie danych przynosi potwierdzenie wyraźnego wykresu, a analiza tabeli frekwencji w wielu przypadkach pokazuje, że służący używają pewnych słów z częstotliwością najwyższą lub najniższą dla wszystkich postaci, rzadziej mieszcząc się w wynikach średnich. Na liście pokazującej frekwencję 10 najczęstszych słów służący zajmują następujące miejsca: *i* — 3 miejsce, *się* — 1, *nie* — 1, *w* — 30 (ostatnie), *na* — 6, *z* — 24, *że* — 3, *do* — 27, *a* — 26, *to* — 22. Wynika z tego, że służący albo używają danych słów o wiele częściej niż wynosi przeciętna, albo o wiele rzadziej.

Podobnie jak w wynikach analizy części mowy tutaj również dość dobrze widoczna jest rozbieżność między mową służby i mową arystokracji. I w tym przypadku zarówno na wykresach, jak i w liczbach można dostrzec biegunowo różne stosowanie określonych słów. Zwykle te grupy są oddzielone na liście rankingowej dla danego wyrażenia przynajmniej kilkunastoma miejscami. Na drzewie pokazującym drzewo zgodności znajdują się natomiast na dwóch skrajnie od siebie oddalonych gałęziach.

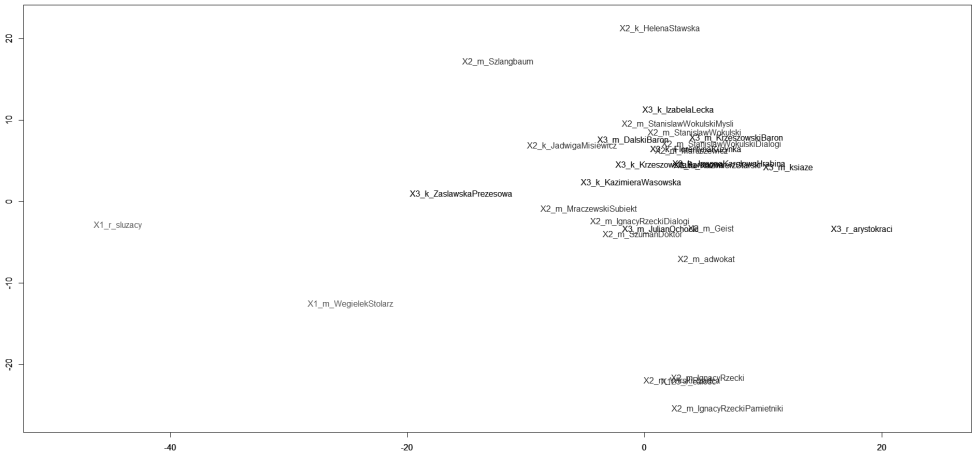
Drzewo zgodności ponownie uwypukla pokrewność stylu arystokracji i adwokata oraz w niedalekiej odległości również postaci księcia, a nieco dalej większości osób zaliczonych do klasy wyższej. Zdawałoby się to potwierdzać tezę o próbach oddawania przez Prusa odmiany języka charakterystycznej dla arystokracji, różnej od sposobu wysławiania się pozostałych, niżej urodzonych bohaterów.

**Najczęściej używane słowa, warianty zlematyzowane**

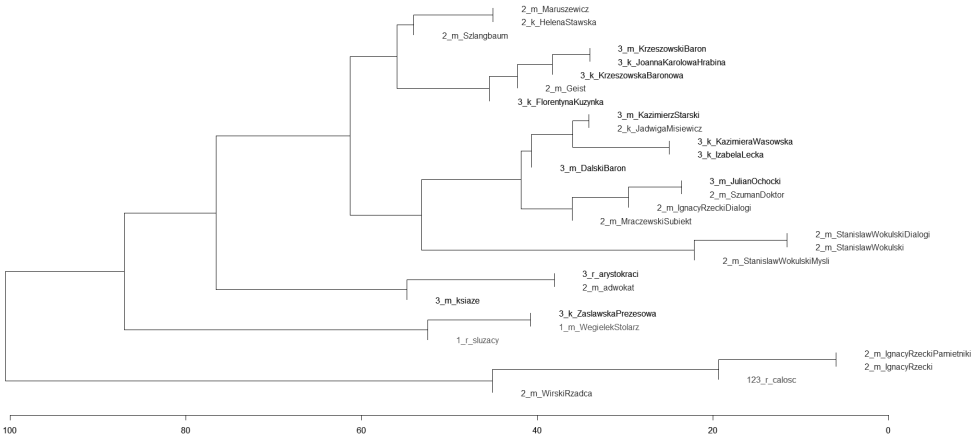
Trzecie i ostatnie z badań zostało przeprowadzone w sposób podobny do tego dotyczącego najczęściej używanych słów. Jedyna różnica polegała na tym, że analizie poddane zostały warianty zlematyzowane, wyciągnięte z wypowiedzi postaci. Wykres 4 oraz Wykres 5 ponownie prezentują wyniki graficznie.

Lista 75 najczęściej używanych słów zlematyzowanych zawiera następujące wyrazy: *on*, *być*, *się*, *i*, *nie*, *w*, *z*, *na*, *ja*, *że*, *pan*, *to*, *do*, *a*, *ten*, *co*, *mieć*, *który*, *o*, *ale*, *Wokulski*, *za*, *pani*, *jak*, *by*, *tak*, *po*, *od*, *móc*, *już*, *mówić*, *mój*, *nawet*, *człowiek*, *tylko*, *czy*, *ty*, *taki*, *panna*, *my*, *więc*, *rzec*, *wiedzieć*, *swój*, *jeszcze*, *jeden*, *sam*, *chcieć*, *raz*, *bo*, *dla*, *bardzo*, *myśleć*, *ażeby*, *kiedy*, *jaki*, *jeżeli*, *sklep*, *kobieta*, *rok*, *ręka*, *albo*, *cóż*, *dzień*, *zrobić*, *jakiś*, *wszystko*, *przez*, *u*, *widzieć*, *no*, *gdyby*, *rubel*, *tu*, *chwila*.

Zgodnie z oczekiwaniami doświadczenie to nie pokazało zbyt wielkich różnic w liczbach w stosunku do tych, które dały się zaobserwować w poprzednim ekspery-




Wykres 4. Najczęściej występujące słowa zlematyzowane, skalowanie wielowymiarowe




Wykres 5. Najczęściej występujące słowa zlematyzowane, drzewo zgodności

mencie, większość bowiem spośród 75 najczęstszych słów to takie, które występują w tylko jednej, nieodmiennej formie (np. spójniki lub przyimki), więc i ich zlematyzowane warianty przyjmują dokładnie taki sam wygląd.

Tym niemniej dostrzegalne są zmiany w częstotliwościach pojawiania się określonych słów. Poniżej porównanie pierwszych 10 najczęściej występujących słów w obydwu wariantach:

Tabela 2. Zestawienie 10 najczęstszych słowoform tekstowych i 10 najczęstszych wariantów zlematyzowanych

| Lp. | Słowoformy | Zlematyzowane |
|---|---|---|
| 1 | i | on |
| 2 | się | być |
| 3 | nie | się |
| 4 | w | i |
| 5 | na | nie |
| 6 | z | w |
| 7 | że | z |
| 8 | do | na |
| 9 | a | ja |
| 10 | to | że |

Różnice frekwencji wynikają z prostej przyczyny — w przypadku wariantów zlematyzowanych do jednej postaci zostają sprowadzone rożne formy gramatyczne, np. *panie*, *panu*, *panem* dadzą postać mianownika liczby pojedynczej *pan*, zwiększając frekwencję użycia właśnie tego jedynego słowa.

Jeśli chodzi o analizę danych, to po raz kolejny zauważalna staje się odmienność idiolektyczna pamiętników Rzeckiego i mowy autorskiej oraz wypowiedzi rządcy Wirskiego.

Dodatkowo odrobinę bardziej jest wyeksponowana zbieżność wypowiedzi służących i należącego do niższej klasy stolarza Węgiełka. To ostatnie zaburza jednak położeniem w tej samej grupie prezesowej Zasławskiej, zaklasyfikowanej do warstwy szlacheckiej.

Pozostałe wyniki zdają się mocniej rozmyte niż w przypadku poprzednich eksperymentów, grupy klasowe nie są ułożone w łatwe do rozpoznania struktury.

**Dyskusja**

Analizując kolejno wyniki doświadczeń, można wysnuć pewne wnioski o zróżnicowaniu stylistycznym mowy bohaterów „Lalki" Bolesława Prusa.

Przede wszystkim należy więc stwierdzić, że owa odmienność stylistyczna jest przeprowadzona zupełnie wyraźnie i konsekwentnie w kilku aspektach języka. Niektóre postaci używają z różnymi częstościami nie tylko pojedynczych słów, ale też odmiennie posługują się częściami mowy. Zróżnicowanie to widoczne jest zarówno w zestawieniu ich wypowiedzi z językiem narratora, jak i pomiędzy indywidualnymi bohaterami.

„Pamiętniki starego subiekta" budowane są podobnie do fragmentów narracyjnych powieści, a jednocześnie zupełnie różnie od wypowiedzi dialogowych samego Rzeckiego. Te ostatnie wyraźnie łączą się stylistycznie z pozostałymi bohaterami. Rzecki mówi jak postać powieści, pisze jak narrator.

Podobnego rozróżnienia nie otrzymał główny bohater utworu Wokulski, którego kwestie dialogowe noszą te same cechy co te wypowiadane w myśli.

Jednocześnie można dopatrzeć się pewnej granicy oddzielającej mowę klas społecznych. Wydaje się, że choć w wynikach eksperymentów dla większości postaci pojawiły się zbieżne rezultaty, to w kilku przypadkach da się wskazać znaczne różnice między klasą niższą a arystokracją i ich poszczególnymi członkami. Język służby zawsze znamionuje mocne odstępstwo od języka pozostałych klas, jeszcze wyraźniej dostrzegalne w porównaniu z językiem arystokracji. Obydwie te grupy układają się zwykle na przeciwnych krańcach zestawień.

Poza rozróżnieniem klasowym nie jest dostrzegalne żadne inne. Ani wiek, ani płeć zdają się nie mieć wpływu na to, w jaki sposób budowana jest mowa postaci.

**Bibliografia**

Argamon S., 2008, Interpreting Burrows's Delta: Geometric and Probabilistic Foundations, Literary and Linguistic Computing 23, s. 131–147.
Bachórz J., 1998, Wstęp, [w:] B. Prus, Lalka, Wrocław–Warszawa–Kraków.
Baayen R.H., 2008, Analyzing linguistic data. A practical introduction to statistics, Cambridge.
Budrewicz T., 1990, „Lalka". Konteksty stylu, Kraków.
Burrows J.F., 1987, Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method, Oxford.
— 2002, "Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing 17, s. 267–287.
Eder M., 2011, Style-markers in authorship attribution: A cross-language study of the authorial fingerprint, Studies in Polish Linguistics 6, s. 99–114.
— 2014, Does size matter? Authorship attribution, small samples, big problem, Literary and Linguistic Computing, Volume 30, Issue 2, s. 167–182.
— 2017, Short samples in authorship attribution: a new approach, Digital Humanities 2017: Conference Abstracts, s. 221–224.
Eder M., Rybicki J., 2011, Stylometry with R, Digital Humanities 2011: Conference Abstracts, s. 308–311.
— 2013, Do birds of a feather really flock together, or how to choose test samples for authorship attribution, Literary and Linguistic Computing 28, s. 221–228.
Herrmann J.B., Van Dalen-Oskam K.H., Schöch Ch., 2015, Revisiting style, a key concept inliterary studies, Journal of Literary Theory 9, s. 25–52.
Hoover D.L., 2002, Frequent Word Sequences and Statistical Stylistics, Literary and Linguistic Computing 17, s. 157–180.
— 2004, Delta Prime?, Literary and Linguistic Computing 19, s. 477–495.
Jannidis F., Pielström S., Schöch Ch., Vitt T., 2015, Improving Burrows' Delta — An empirical evaluation of text distance measures, Digital Humanities 2015: Conference Abstracts.
Kurkowska H., 1948, O języku „Lalki" Prusa, Poradnik Językowy, z. 1, s. 2–8.
Levshina N., 2015, How to do Linguistics with R. Data exploration and statistical analysis, Amsterdam/Philadelfia.
Rybicki J., 1997/1998, Stylometria komputerowa w służbie tłumacza, Rocznik Przekładoznawczy, nr 3/4, s. 171–181.
— 2007, Twelve Hamlets: a stylometric analysis of major character's idiolects in three English versions and nine translations, Digital Humanities 2007: Conference Abstracts, s. 191–192.
Smółkowska T., 1974, Słownictwo i fleksja „Lalki" Bolesława Prusa. Badania statystyczne, Wrocław.

**SUMMARY**

**How the characters of "The Doll" by Bolesław Prus speak. A stylometric analysis**

**Keywords:** quantitative linguistics, idiolect, Delta method, most frequent words, parts of speech.

**Słowa kluczowe:** językoznawstwo kwantytatywne, idiolekt, metoda Delta, najczęściej używane słowa, części mowy.

This article investigates if there is some distinctive way in which characters of Bolesław Prus novel "The Doll" are speaking. It is well known that author endowed his characters with different social backgrounds, views, ethics. Literary critics often emphasized differences in vocabulary and numerous language stylizations among social classes, but in most of the cases only the most characteristic words were taken into con-

sideration and function words (prepositions, conjunctions, personal pronouns and other) were omitted in the research. The aim of this study, in contrast to the previous, is to examine individual characters' way of speaking by measuring frequencies of using most frequent words and given parts of speech. Tests have been performed using the Delta method with a couple of data visualization techniques. The results show some significant variations in individual idiolects.