

Patryk Gujda

Uniwersytet Jagielloński

patryk.gujda@gmail.com

ORCID: 0009-0008-4488-7308

Przyczynek do badań nad możliwością zastosowania algorytmu k-średnich do typologii postaci z bajek ludowych

An Introduction to the Possibility of Applying the k-means Algorithm to the Typology of Characters from Folk Tales

DOI: 10.12775/LL.1.2.2026.009 | CC BY-ND 4.0

ABSTRACT: The classification of folktale characters through the application of statistical methods – particularly those associated with unsupervised machine learning – remains a largely unexplored domain within folklore studies. Several intellectual traditions may be regarded as providing the conceptual foundation for such an inquiry, including formalism, structuralism, and cognitive approaches to the study of mythical forms, all of which share an emphasis on identifying recurrent patterns in folklore texts. In recent years, clustering algorithms have undergone significant development. Their current state of advancement appears sufficient to support the basic requirements of introductory analyses in folklore studies. The study presented in this article was designed to yield unequivocal and interpretable results, and this objective was successfully achieved. While the findings suggest that the application of clustering algorithms to folklore research holds considerable promise, the issue of continuity between the resulting groups remains problematic. Addressing this challenge calls for further theoretical reflection and the development of methodological frameworks capable of accommodating the fluid and overlapping nature of folktale character classifications.

KEYWORDS: computational folkloristics, folktale, character analysis, myth theory, clustering

Wstęp

Celem prezentowanego w niniejszym artykule badania było zastosowanie obliczeniowych metod analizy skupień (metody nienadzorowanego uczenia maszynowego wykorzystywanej do klasyfikowania obserwacji) do charakterystyki postaci występujących w polskiej bajce ludowej. Analizie został poddany niewielki zbiór bajek w celu sprawdzenia, czy proponowana procedura daje wyniki dające się interpretować.

Pytanie o możliwość formalnego opracowania katalogu postaci występujących w mitach, eposach czy bajkach ludowych (na wzór proppowskich funkcji z poziomu wydarzeniowego) wciąż pozostaje otwarte (Czeremski 2011: XXIV). Schemat charakterystyki był proponowany m.in. przez Jamesa Jakóba Liszkę (1989: 121–124), przy czym może być on rozszerzony o takie zmienne jak kategoria ontologiczna (pojęcie zaczerpnięte z kognitywnych analiz mitu; zob. Atran 2013: 117; Czeremski 2021: 261–272) czy specyficzne atrybuty, zachowania, postawy etyczne postaci. Niektóre wyżej wymienione kategorie mogą być, za Claudem Lévi-Straussem, zakodowane jako *bricolage* abstrakcyjnych idei (szczególnie inspirujące jest rozkodowanie mitu w idee dotyczące sytuacji społecznych; Lévi-Strauss 2001: 31–39, 2010: 61–63).

Jednocześnie w ostatnich latach dynamicznie rozwija się subdziedzina folklorystyki obliczeniowej (ang. *computational folkloristics*)¹, co wydaje się sugerować, iż wykorzystanie metod informatycznych, by znaleźć odpowiedzi na pytania tradycyjnie stawiane przez folklorystów, jest obiecującym polem badawczym.

Należy spodziewać się, że zastosowanie zautomatyzowanych metod klasyfikujących da podobne wyniki, co bazujące na tradycyjnych sposobach analizy specjalistów. Niemniej sądzę, że ich wykorzystanie może pozwolić na efektywniejsze badanie zarówno samych typów, jak i ich ewolucyjnego rozwoju oraz dyfuzyjnego przemieszczania się w przestrzeni geokulturowej.

Zbiór danych

Dane uzyskano ze zbioru *Jak chłop u diabła pieniądze pożyczał* pod redakcją Anny Michalec i Stanisławy Niebrzegowskiej-Bartmińskiej, przy czym dla zredukowania nakładu manualnej pracy związanej z kodowaniem w tym wstępnym badaniu ograniczono się do sekcji z bajkami, które zostały przez redaktorki tomu podzielone na 20 fabuł w 37 wariantach (Michalec, Niebrzegowska-Bartmińska 2019: 21–85); w ich obrębie zidentyfikowano 129 postaci możliwych do scharakteryzowania².

1 Wśród prac z tego zakresu uwagę zwracają np. te nad generatorami bajek opartymi na modelu Władimira Proppa (Finlayson 2017: 284–300). Innym obszarem, szczególnie interesującym z punktu widzenia niniejszego badania, są klasyfikacje tekstów (np. Grundkiewicz, Graliński 2011: 29–36).

2 Ze względu na brak środków finansowych za kodowanie danych odpowiada autor niniejszego artykułu. Nie przeprowadzono weryfikacji w postaci drugiego kodowania przez niezależną osobę. Autor wyraża nadzieję, że przyczynkarski charakter badania częściowo usprawiedliwia tę sytuację oraz że w przyszłości uda się pozyskać środki na kontynuację badań z możliwością sprawdzenia poprawności kodowania.

W dalszej części artykułu zachowano oznaczenia bajek z tekstu oryginalnego, a więc numeryczne dla tekstu głównego i literowe dla wariantów (np. 1A). Zbiór ten został wybrany z uwagi na łatwą dostępność transkrypcji rzeczywistych przekazów ustnych, co pozwala chwilowo pominąć problem „zniekształceń” dokonanych przez zbieracza (o możliwych źródłach tych zniekształceń zob. Libera 2021: 185–200). Jednocześnie fakt, że publikacja ta zawiera warianty poszczególnych fabuł, sprawia, iż możemy w tym przypadku uzyskać dane stosunkowo do siebie zbliżone, co jest warunkiem badania podobieństw i różnicowania w ogóle. Przy niewielkiej ilości tekstów bowiem, gdyby były one zbyt różne (tzn. nie reprezentowały fenomenu, który folklorysty nazywają wariantami typu), mogłoby dojść do sytuacji zafałszowania wyników przez zbyt dużą zmienność w danych poprzez nadreprezentację odmiennych obserwacji. Stąd uważa się, że w małym badaniu taki dobór tekstów, w których pojawia się zarówno podobieństwa (warianty tego samego typu), jak i różnice (różne typy), jest pożądany.

Kodowanie danych

Schemat kodowania informacji i wyzwania z nim związane zostały w znacznej mierze opracowane przez Liszkę w jego pracy *The Semiotic of Myth* (Liszka 1989: 121–125). Zgodnie z przedstawioną tam koncepcją postacie występujące w badanym tekście można scharakteryzować za pomocą makrokategorii: cech biofizycznych, relacji pokrewieństwa i społecznych, ról ekonomicznych, statusu politycznego i rangi. Wyzwaniem w opisywaniu konkretnych przypadków mogą być cechy implikowane przez kontekst kulturowy. Najbanalniejszym przykładem niech będzie przynależność stanowa bohatera bajki ludowej. Choć nie jest nigdy określony jednoznacznie mianem „chłopa”, to znajomość kontekstu kulturowego badanego tekstu w połączeniu z relacjami postaci w ramach fabuły (np. poddańcza rola wobec „dziedzica”) zdają się taki status sugerować. Te konteksty to np. paranie się uprawą roli albo zawód rzemieślniczy (tradycyjne zajęcia trzeciego stanu), albo występowanie innej postaci, *explicite* nazywanej „panem” lub „dziedzicem” (co sugeruje różnicę klasową między bohaterami).

Liszka stosuje kategorię osób naznaczonych (ang. *marked*) na określenie tych, które nie spełniają wymogów stawianych przez system normatywny kultury (np. kobieta przewodząca armii jest „naznaczona” swoją kobiecością). Opozycje aktorów naznaczonych i nienaznaczonych tworzą napięcie fabularne i „potencjał dla akcji” (Liszka 1989: 122). W tej analizie chciałbym podkreślić każdą różnicę, nie tylko taką na osi normatywny/prototypowy–przełamujący normę. Jednak trudno nie zgodzić się z Liszką, że człowiek trzeciego stanu zajmujący się wyrębem drewna będzie mniej różny (pod względem zajęcia) od osoby tego samego stanu wykonującej zawód kowala niż od chłopca dowodzącego armią. Z tego względu stosuję również hierarchę porównań: odległość między obserwacjami (różnica między postaciami) jest obliczana na podstawie zarówno przynależności stanowej, jak i wykonywanego zajęcia.

Domyślne cechy stanowią problem dla kodowania, jeśli bierzemy pod uwagę „dziedziczenie” cech z bardziej ogólnych kategorii lub porównań pomiędzy tekstami.

W pierwszej kolejności rozważmy przykład bajki o diabelskim pomocniku (Michalec, Niebrzegowska-Bartmińska 2019: 29–36, bajki 3A–C). Ogólny schemat takiej bajki prezentuje się następująco: (1) diabeł wyrządza szkodę człowiekowi, (2) Lucyfer nakazuje diabłu odpracować szkodę, (3) diabeł zatrudnia się u chłopca i prezentując nadludzką siłę, czyni nowego pracodawcę bogatym. Wyróżnikiem diabła jest niezwykła siła. Jednak czy cecha ta powinna zostać zakodowana również dla Lucyfera? Skoro diabły mają nadludzką siłę, a Lucyfer (jak się domyślamy) jest diabłem, to jego również należałoby scharakteryzować w ten sposób.

Drugi przykład jest podobny. Jeśli wiele bajek opisuje diabły jako „czarne”, to czy jeśli w jakimś przypadku ta cecha nie występuje, ale też nie ma oczywiście jej zaprzeczenia, powinno się domyślnie ją zakodować?

Uważam, że w obu przypadkach, mimo pierwotnej intuicji, należy zrezygnować z automatycznego kodowania tych cech. W przeciwnym razie: (1) klasyfikacja opierałaby się na apriorycznych założeniach, a więc wykazywała taką systematyzację kategorii aktorów, jaką badacz uprzednio założył; (2) powstałyby problemy z potencjalnym podziałem wewnątrz ogólnie znanych kategorii: jeśli zakodowalibyśmy wszystkie diabły tak samo, algorytm nie mógłby znaleźć ich „podgatunków”.

Z tego względu informacje niewynikające z kategorii ontologicznej zostały zakodowane z pominięciem nadmiernego ekstrapolowania cech wynikających z jednych obserwacji na inne. Pewien wyjątek stanowią przypadki, gdzie nazwa własna sugeruje pewne cechy. Naczelnym przykładem jest Śmierć. W tekstach 18 (A–D) i 19 (Michalec, Niebrzegowska-Bartmińska 2019: 75–81) jedną z bohaterek jest personifikacja śmierci, która nie zawsze jest opisywana jako odznaczająca się jednoznacznie trupimi cechami (jak choćby bladeść czy kośćistość). Czasem narrator ogranicza się wyłącznie do podania nazwy własnej. W tym przypadku traktuje się tę nazwę jako wskazującą na cechy trupy.

Logika związku z pewną ideą wydaje się często lepszym sposobem kodowania niż trzymanie się ściśle litery opowiadania. Na przykład w wariancie B bajki nr 18 (Michalec, Niebrzegowska-Bartmińska 2019: 75–80) bohater jest „siódmym lub ósmym” dzieckiem, w wariancie D zaś to ojciec „dwunastu lub trzynastu” staje się protagonistą. W obu przypadkach mamy do czynienia z wielodzietnością, która staje się specyficznym nadmiarem uruchamiającym ciąg zdarzeń. Bohaterowie ci, w mojej opinii, powinni być więc kodowani jako identyczni (na pewnym, koniecznym przecież, poziomie ogólności): będą więc określani cechą „wielodzietność”, zamiast: „jest 7/8 dzieckiem” (wariant B) i „ma 12/13 dzieci” (wariant D). Ten, lévi-straussowski w istocie, sposób kodowania wydaje się najbardziej odpowiedni do relacji społecznych.

Jedną z kodowanych makrokategorii powinny być również podstawowe kategorie ontologiczne (Atran 2013: 117–121). Są to w klasycznym modelu: czło-

wiek, zwierzę, roślina, przedmiot i artefakt. W przypadku analizowanego zbioru korzystna jest jednak drobna modyfikacja: zastąpienie przedmiotu i artefaktu kamieniem oraz dodanie kategorii trupa. Taka zmiana wynika z charakteru danych, nie mamy bowiem aktorów przejawiających cechy kategorii, z których zrezygnowano, z wyjątkiem motywu zmiany w kamień oraz osobników w oczywisty sposób przejawiających cechy osób zmarłych. W ramach modularnej teorii umysłu, na której opiera się ta część kodowania, kamień jest oczywiście specyficznym przypadkiem przedmiotu, a trup sam w sobie – wyobrażeniem kontrintuicyjnym: martwy (przedmiot), jednak z cechami człowieka.

Kodowaniem została również objęta ogólna rola fabularna, w jakiej znajduje się określony aktor. Ta zmienna nie bierze udziału w dalszej klasyfikacji, ale jej obecność może pomóc w odpowiedzi na pytanie, czy pewne typy z większym prawdopodobieństwem występują w określonych rolach.

Baza danych jest zbudowana hierarchicznie: wektory charakterystyk poszczególnych postaci składają się z kategorii (zmiennych)³ zgrupowanych w 11 nadrzędnych makrokategoriach⁴:

- kategorie ontologiczne⁵,
- cechy biofizyczne,
- relacje społeczne (w tym pokrewieństwo),
- klasa społeczno-ekonomiczna,
- status społeczno-ekonomiczny,
- pochodzenie i sposób istnienia,
- powiązania przestrzenne,
- atrybuty,
- magiczne zachowania i cechy,
- cechy charakteru i moralne,
- podstawowa funkcja fabularna.

Kodowanie odbywa się na poziomach poniżej makrokategorii (zob. aneks 1: zmienne oznaczone znakiem X).

Trzeba także wspomnieć, że zrezygnowano z objęcia kodowaniem aktorów będących swoistymi „rekwizytami”, tzn. nieposiadających własnej charakterystyki.

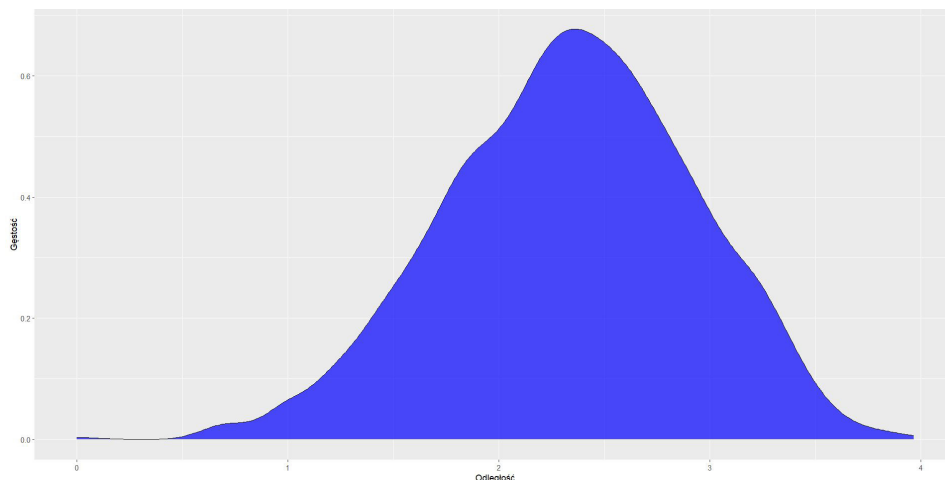
3 Makrokategorie w tym badaniu mają charakter porządkujący, jednak można by też testować bazujące na nich nadawanie wag odległościom – tzn. jeśli zaprojektować funkcję, która najpierw wylicza odległości między obserwacjami w ramach makrokategorii, a później wypracowuje na ich podstawie całkowitą miarę podobieństwa, jedne kategorie traktując jako istotniejsze, a inne jako poboczne. Pewne kategorie wydają się bowiem ważniejsze niż inne dla wskazania typu, do jakiego należy określona obserwacja (np. kategoria ontologiczna *versus* powiązania przestrzenne). Wagi jednak musiałyby być arbitralne.

4 Szczegółowe uwagi dotyczące użytych zmiennych oraz sposobu kodowania danych znajdują się w aneksie 1.

5 W analizowanym zbiorze przytłaczająca większość bohaterów ma postać ludzką i ewentualne cechy innych kategorii, więc pominięto problem rozróżniania ludzi z cechami zwierzęcymi i zwierząt z ludzkimi.

Analiza danych

Zakodowano tablicę danych, zawierającą 128 zmiennych i 129 obserwacji. Obliczono macierz odległości euklidesowych między obserwacjami, co dało następującą funkcję gęstości:



Wykres 1. Histogram odległości euklidesowych między obserwacjami. Źródło: opracowanie własne

Już analiza wizualna sugeruje rozkład normalny, co jednak zostało potwierdzone testem Shapiro-Wilka, który zwrócił znikomą wartość parametru p .

Normalny rozkład odległości z jednej strony sugeruje, że dane mają charakter naturalny, tj. zostały poprawnie zebrane i w wystarczającym stopniu reprezentują rzeczywistość. Z drugiej strony, gdyby dane dzieliły się na wyraźne odrębne grupy, należałoby spodziewać się rozkładu bimodalnego, czyli takiego, w którym odległości małe i duże byłyby liczniejsze od średnich. Nie jest to jednak dowód na niewystępowanie naturalnych grup, lecz sugestia, że mogą one być sobie bliższe bądź że różnice między nimi są niejednakowe, co po uśrednieniu daje wykres rozkładu normalnego, w którym najliczniejsze są wartości średnie.

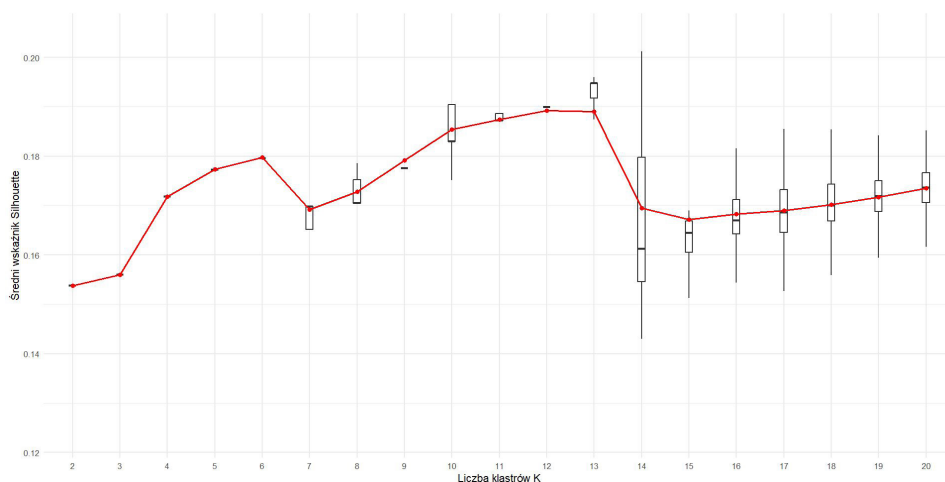
Badanie możliwości grupowania przeprowadzono za pomocą algorytmu k -średnich, do ustalenia optymalnej liczby klastrów użyto zaś współczynnika Silhouette oraz metody łokcia.

Algorytm k -średnich polega na wyborze centroidów, czyli centrów potencjalnych grup, następnie zaś na iteracyjnej optymalizacji odległości w ramach poszczególnych klastrów i pomiędzy nimi, tak aby zminimalizować pierwszą, a zmaksymalizować drugą. Położenie centroidów jest dostosowywane w każdej kolejnej iteracji tak, żeby funkcje odległości były jak najlepiej dopasowane względem tych dwóch parametrów (Wierzchoń, Kłopotek 2017: 88–118). Tutaj posłużono się standardowym algorytmem wbudowanym w pakiecie R (The R Foundation for Statistical Computing 2024: *kmeans*).

Wybór optymalnej liczby klastrów metodą współczynnika Silhouette polega na obliczeniu go dla każdej obserwacji, a następnie uśrednieniu dla całej badanej populacji oraz selekcji optymalnej średniej pomiędzy wynikami k-średnich dla różnej liczby klastrów. Współczynnik ten określa, w jakim stopniu obserwacja jest dopasowana do własnego klastra w stosunku do jej oddalenia od innych klastrów. Przyjmuje wartości od -1 do 1 , przy czym wartości bliżej 1 oznaczają lepsze dopasowanie (Rousseeuw 1987: 53–65; Wierzchoń, Kłopotek 2017: 172–173).

Zastosowanie metody łockia polega na obliczeniu modelu k-średnich dla różnych wartości k (liczby klastrów), podobnie jak w poprzedniej metodzie. Następnie oblicza się sumę kwadratów odległości obserwacji od centroidów klastrów, do których są one przypisane. Jako że w algorytmie k-średnich optymalizuje się funkcję odległości wewnątrz klastrów do minimum, należy się spodziewać, że również w tym przypadku wraz ze wzrostem jakości dopasowania współczynnik WCSS (*within cluster sum of squares*) będzie malał. Optymalną liczbą klastrów jest ta, od której WCSS zaczyna się stabilizować w okolicy minimum (Wierzchoń, Kłopotek 2017: 41–42).

W przypadku współczynnika Silhouette wykonano analizę jego wartości dla różnej liczby klastrów przy ustawieniu ziarna (ang. *seed*) na wartości od 1000 do 1 000 000, co 1000. Uzyskano następujący wykres:



Wykres 2. Optymalna liczba klastrów – metoda Silhouette. Pudełka wykresu oznaczają wartości wskaźnika w zależności od ziarna (zwraca uwagę, że dla klastrów 2–6, 9 i 12 pudełka są znikome). Czerwona linia określa średnie wartości między ziarnami. Źródło: opracowanie własne

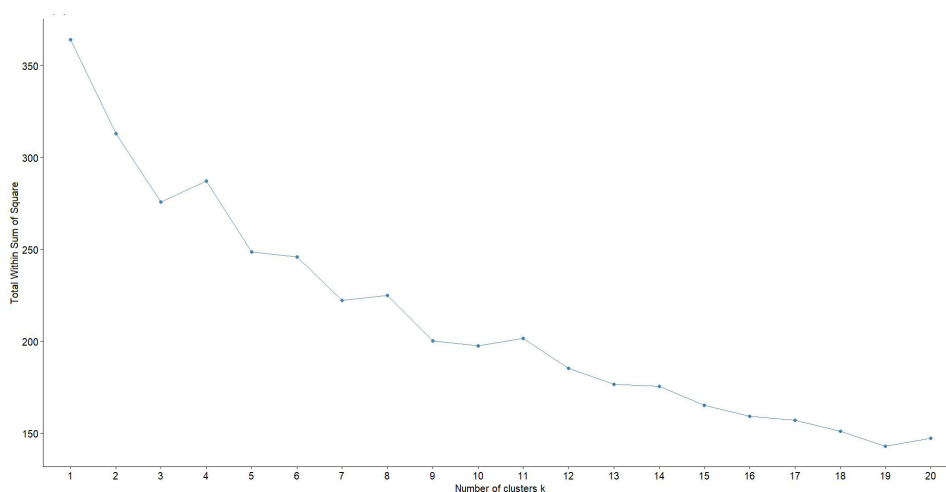
Uwagę zwraca fakt, że wskaźnik Silhouette przyjmuje stosunkowo niskie wartości. Może się tak dziać dlatego, że pracujemy na dużej liczbie zmiennych, a dla każdej obserwacji są one w większości puste, co oznacza, że w takim wariancie są do siebie bardziej podobne niż różne. Można sobie wyobrazić

pustą formę złożoną ze 129 okienek, z których zazwyczaj tylko 5–7 jest wypełnionych. Na pierwszy rzut oka różne warianty są do siebie podobne, bo forma jest prawie pusta. Nie oznacza to jednak, że nie istnieją schematy układania tych 5–7 wypełnień. Odległość euklidesowa jest odpowiednia dla algorytmu k-średnich, jednakże należy zaznaczyć, że w miarę postępu badań nad klasyfikacją postaci inne miary podobieństwa, bardziej podkreślające różnice, powinny być uwzględnione wraz z rozważeniem zastosowania innych algorytmów, jak również bardziej wyrafinowanych postaci k-średnich.

Następnie należy przyrzeć się stabilności współczynnika przy ilości klastrów od 2 do 6 oraz przy liczbie 9 i 12. Powyżej tych ilości widać wyraźne wahania, co sugeruje utratę stabilności grupowania.

Z analizy wynika, że liczbą klastrów, która pozostaje bardzo stabilna i jest najbardziej obiecująca, jeśli chodzi o wielkość wskaźnika, jest 12 – jest to średnie maksimum wskaźnika, zostaje też zachowana duża stabilność grupowania.

Do analizy WCSS zastosowano funkcję `fviz_nbclust` z pakietu `factoextra` (The R Foundation for Statistical Computing 2024: `fviz_nbclust`). Użyto kilku ziaren, w wyniku czego uzyskano bardzo zbliżone wyniki. Poniżej przedstawiono reprezentatywny wykres:



Wykres 3. Optymalna liczba klastrów; metoda łokcia dla ziarna 100 000. Źródło: opracowanie własne

„Łokieć” nie jest bardzo wyraźny, jednak daje się zaobserwować, że wartości zaczynają stabilizować się w podziałach złożonych z między 9 a 14 klastrów. W połączeniu z analizą współczynnika Silhouette na potrzeby tego badania możemy przyjąć, że najodpowiedniejszym założeniem dotyczącym liczby grup będzie 12.

W ramach przeprowadzonego badania zbudowano pięć modeli k-średnich z parametrem $k = 12$, przy ustawieniu maksymalnej liczby iteracji algorytmu na 1 000 000, kolejno rozpoczynając od ziarna: 10, 100, 1000, 10 000, 100 000

(dalej: modele I–V). Następnie porównano stabilność grupowań za pomocą matrycy współczynnika ARI, którego wartości wynosiły od 0,9366 do 1,0000 ze średnią 0,9746 – co oznacza, że klastry są bardzo stabilne i algorytm powinien zwracać zbliżone wyniki niezależnie od początkowego ziarna. Matryca dokładnych wyników współczynnika ARI prezentowana jest w tabeli.

	V1	V2	V3	V4	V5
1	1.0000000	0.9365794	1.0000000	1.0000000	1.0000000
2	0.9365794	1.0000000	0.9365794	0.9365794	0.9365794
3	1.0000000	0.9365794	1.0000000	1.0000000	1.0000000
4	1.0000000	0.9365794	1.0000000	1.0000000	1.0000000
5	1.0000000	0.9365794	1.0000000	1.0000000	1.0000000

Tabela 1. Współczynnik ARI dla porównań między modelami I–V. Źródło: opracowanie własne

Daje się zauważyć, że modele I, III, IV i V są identyczne. Model II natomiast różni się od nich w niewielkim stopniu – sześć obserwacji zostało zaklasyfikowanych inaczej.

Następnie przystąpiono do opisowej analizy dopasowania wyników, aby sprawdzić, czy są interpretowalne i czy błędne dopasowania stanowią znaczący problem.

Omówienie wyników

Zbiór danych podzielono na 12 klastrow w każdym z modeli. Poniżej zaprezentowano wyniki analizy z proponowaną nazwą dla klastra i krótkim omówieniem. Zaznaczono, w którym miejscu różnice w klasyfikacji w modelu II wpływają na interpretację klastra. W nawiasach na początku podano numery bajek z analizowanego zbioru oraz liczbę obserwacji, jakie znalazły się w danym klastrze. Zmienne funkcji fabularnych nie uczestniczyły w procedurze klasyfikacji.

1. Klaster Śmierci (18A–D, 19; pięć obserwacji)

Śmierć można scharakteryzować jako mającą ludzką postać kobietę o cechach trupich (wszystkie obserwacje). Często jej pojawienie się zwiastuje czyjś zgon, lecz funkcjonuje ona również jako bohater kulturowy: przekazuje zdolność leczenia. Przejawia cechy pozytywne związane z pracą (rzetelność), a także negatywne (agresja). Klaster Śmierci pokrywa się z postaciami określonymi *explicitie* jako Śmierć.

2. Klaster upiora (1A–F; sześć obserwacji)

Związany z grupą wariantów bajki numer 1 klaster to ludzka postać mężczyzny, którego pochodzenie jest jednoznacznie określane jako tajemnicze bądź nieznane. W toku opowieści okazuje się, że związany jest z obszarami granicz-

nymi, takimi jak zaczarowana chata na uboczu czy cmentarz. Początkowo opisywany jako elegancki, bogaty, okazuje się trupojadem. Choć czasem nazywany diabłem, jest częściej duszą osoby zmarłej, pokutującej bądź zakłętej. Reaguje negatywnie na przedmioty związane z sacrum chrześcijańskim. Różne próby konfiguracji modeli oraz wariantów kodowania „diabelskości” wykazały, że to grupowanie jest odporne na drobne zmiany czy błędy.

3. Klaster prototypowej kobiety (1A–F, 3A, 7C-var, 10–14, 18D, 20; 17 obserwacji)
Klaster ten jest dość szeroki i zawiera postacie, które intuicyjnie mogłyby sugerować sensowność dalszego podziału. Należą do niego głównie kobiety, które często przynależą do prototypowej klasy społecznej (7) i mają związek z przędzeniem (6). Równie liczne są postacie o pozytywnych cechach charakteru (6). Za błędne należy uznać przypisanie bociana z tekstu 18D do tej grupy.

4. Klaster pięknej kobiety (1C, 1D, 2, 13, 15, 17; sześć obserwacji)
Zapewne należy patrzeć na ten klaster jako na podgrupę kobiecych bohaterek, jednak został wyodrębniony z uwagi na zespół cech związanych ze szczególną urodą (piękno charakteryzuje wszystkie obserwacje). Bywa przy tym, że kobiety te są istotami nadprzyrodzonymi (4), a także często bywają wybredne w stosunkach międzyludzkich (4). Czasem funkcjonują nocą, w okolicach północy (3). Charakterystyka ta wydaje się sugerować pewien związek z demonicznym wyobrażeniem latawicy, aczkolwiek jest to hipoteza śmiała, wymagająca dalszej weryfikacji.

5. Klaster prototypowych mężczyzn (1A, 1C, 3A–E, 5–9, 11–20; 39 obserwacji + jedno błędne przyporządkowanie w modelach I, III, IV, V)
W tym klastrze co do zasady mamy po prostu do czynienia z żywym mężczyzną. Czasami taka postać odgrywa rolę protagonisty, ale bywa też, że pełni inne funkcje fabularne, jak np. inicjowanie wydarzeń. Męska postać wielokrotnie należy do klasy prototypowej (12), choć zróżnicowanie zawodów jest tu duże. Może być zarówno bogaty (8), jak i biedny (6), przy czym protagonista bywa raczej biedny, przynajmniej na początku opowieści. Może mieć pozytywne (7), i negatywne (12) cechy charakteru. W modelach I, III, IV, V do tej grupy trafiły strachy z kościoła obecne w tekście 17, co wydaje się błędnym przyporządkowaniem.

6. Klaster męskiego bohatera (2, 3C, 4, 12, 13, 17, 18A–C, 20; 11 obserwacji)
Tę grupę należy uznać za wyodrębnioną podgrupę prototypowych mężczyzn. Charakteryzuje ją przede wszystkim dysponowanie magicznymi przedmiotami (wszystkie obserwacje). Stosunkowo często przejawiają pozytywne cechy charakteru (4), a wykonują zawody takie jak lekarz (3) czy ksiądz (3).

7. Klaster nadprzyrodzonego antagonisty (1C–E, 3D, 4, 7A–C-var, 7E, 8, 9, 13, 18A, 20; 16 obserwacji, ale z różnicami)

Grupa ta składa się z mężczyzn (16) w ludzkiej postaci (14 lub 16), będących istotami nadprzyrodzonymi (15), często diablami (12). Bohaterowie ci przejawiają negatywne cechy charakteru (7 lub 8), przy czym charakteryzuje ich naiwność (7). Bywa, że są rolnikami (6), czyli przynależą do klasy prototypowej. Są związani z motywem ekonomicznej rywalizacji chłopca z diabłem. Co ciekawe, do takich antagonistów w badanych tekstach zalicza się także Bóg. W modelach I, III, IV i V do tej grupy zostały przyporządkowane diable z tekstów 9 i 18A, które nie wystąpiły w modelu II, gdzie z kolei pojawiły się Antychryst i Tyłynda z tekstu 20. Zapewne wszystkie cztery postacie są przypadkami granicznymi z innymi „diabelskimi” kategoriami.

8. Klaster diabła leśnego (3A–C, 3E, 5, 6, 10, 12, 16; dziewięć obserwacji)

Diabły leśne mają postać mężczyzn i charakteryzuje je nadludzka siła, która umożliwia im niezwykle wydajną pracę (wszystkie obserwacje). Zazwyczaj zamieszkują lasy bądź po prostu drzewa (7). Często wchodzi w rolę służących protagonisty (6). Ze względu na swoją charakterystykę odgrywają ambiwalentną rolę fabularną: z jednej strony pomagają bohaterem się wzbogacić, a z drugiej stanowią w pewnym sensie ich rywali. Ich pomoc bywa niebezpieczna.

9. Klaster arystokracji piekielnej (2, 3A, 3C, 3E, 4, 7D; siedem–osiem obserwacji)

Do tej grupy należą męskie diable (wszystkie obserwacje), które można scharakteryzować jako swego rodzaju klasę wyższą (6) czy władców (6), co nie może dziwić, skoro znaczną część grupy stanowią różne przedstawienia Lucyfera-Antychrysta (5). Co ciekawe, w badanych tekstach postacie te wykazywały się poczuciem sprawiedliwości (5). W modelu II grupa ta jest rozszerzona o „Najstarszego wojownika” z tekstu 3E – jest to dopasowanie dosyć trafne, zapewne do pewnego stopnia przypadek graniczny z inną grupą.

10. Klaster różnych diablów i strachów (wspólne: 11, 17, 20; modele I, III, IV, V: 3E, 20; model II: 9, 17, 18A; siedem obserwacji, klaster wyraźnie mniej stabilny niż pozostałe)

Grupa, którą należy uznać za zbiór różnych postaci demonicznych lub granicznych. W przypadku siedmiu obserwacji w modelach I, III, IV, V mamy identyczny skład grupy, jednak w modelu II aż trzy obserwacje są inne. Mimo to członków tej grupy można scharakteryzować jako diable, często związane z kolorem czarnym. W modelu II dodatkowymi postaciami były diable z tekstu 9, strachy z kościoła obecne w teście 17 oraz diable z tekstu 18A. Zabrakło natomiast „Najstarszego wojownika” z tekstu 3E oraz Antychrysta i Tyłyndy z tekstu 20, które występują w ramach tej grupy w pozostałych modelach. Być może w miarę rozwoju bazy danych obserwacje z tej grupy będą bardziej jednoznacznie przypisywane do innych zbiorów albo zostanie utworzona odrębna, łatwiejsza do scharakteryzowania grupa.

11. Klaster dusz zmarłych (2, 12, 20; cztery obserwacje)

Mała, ale łatwa do interpretacji grupa, zawierająca dusze zmarłych. Co interesujące, często przyjmują one postać roślinną (3): jablek lub snopków.

12. Klaster pokutującego ojca (17; jedna obserwacja)

Prócz wskazanych powyżej klastrow została wyodrębniona jedna, dosyć złożona postać dziadka z pałacu z tekstu 17. Mamy do czynienia z duszą pokutującą, ale na tyle złożoną, że wyraźnie różni się od innych dusz. Trudo powiedzieć, czy przy powiększeniu bazy danych można spodziewać się rozwinięcia tej grupy.

Konkluzje

Najważniejszym wnioskiem, jaki płynie z przeprowadzonego badania, jest to, że nawet niewielki zbiór danych poddany podstawowej metodzie analizy skupień, jaką jest algorytm k-średnich, i przy zastosowaniu najprostszych metod walidacji modelu (Silhouette i WCSS) oraz stabilności grupowań (ARI) daje łatwo interpretowalne wyniki. Co prawda stworzone w ten sposób charakterystyki postaci są powiązane z funkcjami fabularnymi, ale nie redukują się do nich.

Należy się spodziewać, że powiększenie zbioru danych i zastosowanie bardziej zaawansowanych metod poskutkuje uzyskaniem kolejnych interesujących wyników. Do wyzwań należy z pewnością poprawa parametrów walidacji modelu w kierunku jaśniejszego grupowania (wyodrębnienie archetypów kulturowych). W przypadku niepowodzenia w jednoznacznym określeniu grupowań należałoby uznać, że postacie tworzą w istocie kontinuum wyobrażeń, a ich klasyfikacja daje tylko orientacyjną mapę dla danego zespołu tekstów. Warto podkreślić, że istnieje szereg alternatywnych algorytmów klasyfikujących (wydaje się, że godne uwagi byłyby zwłaszcza algorytmy hierarchiczne; przegląd zob. Ezugwu i et al. 2022; Saxena i in. 2017) oraz inne miary odległości⁶, zwłaszcza podobieństwo Jaccarda, które należałoby zastosować z myślą o uzyskaniu zbliżonych wyników, aby potwierdzić poprawność klasyfikacji. Z pewnością interesującym kierunkiem weryfikacji proponowanej metody byłoby również przeanalizowanie, na ile możliwe jest wykorzystanie innych uzasadnionych formularzy kodowania oraz czy taki zabieg znacząco wpływa na zmianę grupowania.

BIBLIOGRAFIA

- Atran, S. (2013). *Ewolucyjny krajobraz religii*. Zakład Wydawniczy Nomos.
- Czeremski, M. (2011). *Oswojenie bajki. Morfologia Władimira Proppa*. W: W. Propp, *Morfologia bajki magicznej* (s. VII-XXXVII). Zakład Wydawniczy Nomos.

6 Algorytm k-średnich w standardowej postaci działa na odległościach euklidesowych. Jednocześnie najbardziej odpowiednią miarą podobieństwa między grupami zmiennych kategoryalnych wydaje się podobieństwo Jaccarda, którego zastosowanie w algorytmie k-średnich nie jest banalne. Zob. próby integracji rzeczonych metod: Shameem, Ferdous 2009 oraz Lu i in. 2014.

- Czeremski, M. (2021). *Mit w umyśle*. Wydawnictwo UJ.
- Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, Ch. I., Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110, 104743. <https://doi.org/10.1016/j.engappai.2022.104743>
- Finlayson, M. A. (2017). ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory. *Digital Scholarship in the Humanities*, 2, 284–300.
- Grundkiewicz R., Gralinski, F. (2011). How to distinguish a kidney theft from a death car? Experiments in clustering urban-legend texts. In P. Nakov, Z. Kozareva, K. Ganchev, J. Hobbs (eds), *Proceedings of the RANLP 2011 Workshop on Information Extraction and Knowledge Acquisition* (pp. 29–36). Association for Computational Linguistics.
- Lévi-Strauss, C. (2001). *Myśl nieoswojona* (przeł. A. Zajączkowski). Wydawnictwo KR.
- Lévi-Strauss, C. (2010). *Surowe i gotowane* (przeł. M. Falski). Wydawnictwo Aletheia.
- Libera, Z. (2021). Etnografia jest praktykowaniem semiotyki (przykład z polskiej etnografii XIX wieku i później). *Prace Etnograficzne*, 3, 185–200.
- Liszka, J. J. (1989). *The Semiotic of Myth. A Critical Study of the Symbol*. Indiana University Press.
- Lu, M., Qin, Z., Cao, Y., Liu, Z., Wang, M. (2014). Scalable news recommendation using multi-dimensional similarity and Jaccard-Kmeans clustering. *Journal of Systems and Software*, 95, 242–251.
- Michalec, A., Niebrzegowska-Bartmińska, S. (red.) (2019). *Jak chłop u diabła pieniądze pożyczał. Polska demonologia ludowa w przekazach ustnych* (s. 21–85). Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Weiping, D., Lin, Ch.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267, 664–681.
- Shameem, M., Ferdous, R. (2009). An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. *Proceedings of the First Asian Himalayas International Conference on Internet*, Kathmundu, Nepal. <https://doi.org/10.1109/AHICI.2009.5340335>
- The R Foundation for Statistical Computing (dostęp: 2024, 28 maja). *RDocumentation* [dokumentacja oprogramowania]. <https://www.rdocumentation.org>
- Wierchoń, S., Kłopotek, M. (2017). *Algorytmy analizy skupień*. Wydawnictwo Naukowe PWN.

Aneks 1. Dodatkowe uwagi o kodowaniu i lista zmiennych

Poszczególne zmienne są kodowane w systemie zerojedynkowym, gdzie 1 oznacza obecność cechy, a 0 jej nieobecność. Oznaczenia X_1 itd. odnoszą się do kodowanych zmiennych (ten sam system zastosowano w aneksie 2). Zmienne, w których występuje konieczność doprecyzowania charakterystyki, są zbudowane hierarchicznie, a oznaczenia zmiennych szczegółowych podano w nawiasach (zob. np. pierwszy przypadek: X_2 : zwierzę [X_3 : ptak, X_4 : kot, X_5 : szczur])⁷:

7 Są kodowane tak, żeby suma wartości w najogólniejszej kategorii wynosiła 1, ale jednocześnie, by móc rozróżnić cechy szczegółowe w mniejszym stopniu niż ogólne. Jako przykład podajmy czarownicę wyprawiającą się na nocne hulanki i stracha pojawiającego się o północy. Postacie te łączy nocne bytowanie. Pod zmienną „nocne bytowanie” zakodowane więc będzie 0,5 oraz po 0,5 pod kolejno „nocne hulanki” i „związek z północą”. W ten sposób różnica między nimi wyniesie tylko 0,5, a z postaciami niewykazującymi cech nocnych: 1 w obrębie tej zmiennej.

- kategorie ontologiczne⁸:
 - człowiek (X₁),
 - zwierzę (ptak, kot, szczur) (X₂ [X₃, X₄, X₅]),
 - roślina (jabłko/jabłoń, kwiat) (X₆ [X₇, X₈]),
 - postać kamienna (X₉),
 - trup (X₁₀);
- cechy biofizyczne:
 - mężczyzna (X₁₁),
 - kobieta (X₁₂),
 - aktor zbiorowy (X₁₃),
 - motyw dwóch przybyszów (X₁₄),
 - niejednoznaczność płciowa (X₁₅),
 - niepełnosprawność (X₁₆),
 - uroda i elegancja, piękny wygląd (X₁₇),
 - brzydota (X₁₈),
 - czerń (X₁₉),
 - biel (X₂₀),
 - znikomość (np. chudy, lekki) (X₂₁),
 - gigantyzm (np. wysoki) (X₂₂);
- relacje społeczne (w tym pokrewieństwo):
 - związki z nadprzyrodzonym (np. jest dzieckiem niespodzianką, „ma diabła”) (X₂₃),
 - samotność (np. staropanieństwo, jedynactwo) (X₂₄),
 - doświadczenie nadmiaru (np. wielodzietność) (X₂₅),
 - transgresja norm kulturowych (np. posiadanie kochanka) (X₂₆),
 - władza bezpośrednia (np. rodzic, przywódca grupy) (X₂₇);
- klasa społeczno-ekonomiczna:
 - klasa prototypowa (kupiec, fryzjer, żołnierz, lekarz, garncarz, prądką, kowal, praca wykonywana w lesie, rolnik)⁹ (X₂₈ [X₂₉–X₃₇]),
 - klasa wyższa (władca, szlachta, sędzia) (X₃₈ [X₃₉, X₄₀, X₄₁]),
 - klasa niższa, służba (służący, strażnik) (X₄₂ [X₄₃, X₄₄]),
 - kapłan (X₄₅),
 - poza systemem klasowym (czarownik/czarownica, bandyta, banita): (X₄₆ [X₄₇, X₄₈, X₄₉]);

8 Postać właściwa danej kategorii kodowana jest jako 1. Aktor podlegający transformacjom będzie miał kilka cech zakodowanych w ten sposób. Pojedyncze cechy kontrintuicyjne kodowane są jako 0,5 (np. rogi, ogon) i nie sumują się.

9 Zawody te przynależą głównie do trzeciego stanu w kategoriach feudalnych, ale wyraźnie daje się również zauważyć obecność zajęć będących elementem doświadczenia XX w., w tym zawierających opowieść o awansie społecznym przez pracę (zwłaszcza bajki o lekarzu i śmierci).

- status społeczno-ekonomiczny:
 - wolny (bogaty, biedny) (X50 [X51, X52]),
 - zniewolony (pokutujący/potępiiony, uwięziony) (X53 [X54, X55]);
- pochodzenie i sposób istnienia:
 - żywy (wszystkie postacie intuicyjnie identyfikowane jako żywe w zwyczajny sposób – tzn. funkcjonujące zgodnie z umysłowym modulem „potocznej biologii”) (X56),
 - istoty nadprzyrodzone (X57)¹⁰,
 - » zmarły (ofiara morderstwa, dusza pokutująca/potępiiona) (X58 [X59, X60]),
 - » zaklęty (X61),
 - » „diabeł” (Lucyfer/Antychryst)¹¹ (X62 [X63]),
 - » Bóg (X64),
 - » Śmierć (X65),
 - tajemnicze pochodzenie – czyli określone *explicite* przez informatora jako nieznane/tajemnicze (mogą to być obcy albo istoty nadprzyrodzone, co zresztą niekiedy się przenika) (X66);
- powiązania przestrzenne:
 - ekumena (X67),
 - obszary graniczne: „magiczne domostwo”, trumna/cmentarz, rozstaje/gościniec/wędrowni, las/drzewo (X68 [X69–X72]),
 - anekumena, czyli rozmaite odludzia, bagna itp. (X73),
 - zaświaty (X74);
- atrybuty:
 - zima/zimno (X75),
 - nadprzyrodzony środek transportu (komin, ożóg) (X76 [X77, X78]),
 - ogień (X79),
 - skarby (X80),
 - zwierzęta (węże/gady, kruki) (X81 [X82, X83]),
 - hałas (X84),

10 W tym przypadku uszczegółowienie kategorii musiało być dwustopniowe. Zachowano zasadę, w której suma z kategorii pierwszego rzędu musi wynosić 1. Na przykład: istota nadprzyrodzona: 0,5, zmarły: 0,25, ofiara morderstwa: 0,25.

11 Przyjęta konwencja: jeśli tylko diabelska natura, to X62 = 1, jeśli Lucyfer/Antrychryst, to: X62 = 0,5 i X63 = 0,5. W przypadku pojawiających się w bajce ludowej diablów problematyczne jest stwierdzenie, na ile są to ogólne określenia demonologiczne (termin, którym informatorzy mogliby opisywać wyobrażenia o różnym charakterze nadprzyrodzonym), a na ile określenia specyficznego typu bytów związanych z mitologią chrześcijańską. Przetestowano grupowania w obu przypadkach i zaobserwowano, że wybór założenia nie wpływa w znaczący sposób na przyporządkowanie tych postaci do poszczególnych grup. W artykule prezentowane są wyniki uzyskane przy założeniu drugiego przypadku, tj. kiedy informator używa słowa „diabeł” do nazwania danej postaci, to zostaje ona zakodowana jako diabeł.

- nadprzyrodzone przedmioty (np. święte albo magiczne, w tym zioła) (X85),
 - wódka (X86),
 - muzyka/instrumenty (X87),
 - pieczęć: wypalona ręka (X88),
 - cyrograf (podpisywany krwią, dziecko-niespodzianka¹²) (X89 [X90, X91]),
 - uzbrojenie (X92),
 - tajne imię (X93);
- magiczne zachowania i cechy:
- nadzwyczajny sposób posilania się (trupojad, nadludzki apetyt) (X94 [X95, X96]),
 - reakcja na sacrum (antypatia, obłąskawienie) (X97 [X98, X99]),
 - możliwość zjednania sobie postaci (przez przedmioty sakralne, kradzież ubrania) (X100 [X101, X102]),
 - brak wstępu do ekumeny/domu (X103),
 - nocne bytowanie, życie w ciemności (związek z północą, znikanie o poranku, nocne hulanki/wyprawy) (X104 [X105–X107]),
 - zwiastun śmierci (X108),
 - motyw duszenia (X109),
 - motyw wniesienia czegoś/kogoś do ekumeny (X110),
 - nadludzka siła/praca (X111),
 - przewidywanie przyszłości (X112);
- cechy charakteru i właściwości moralne¹³:
- pozytywne (związane z pracą: rzetelność/pracowitość itp.; spryt, odwaga; gościnność; sprawiedliwość; mądrość) (X113 [X114–X119]),
 - negatywne (związane z pracą: lenistwo/niekompetencja itp.; naiwność; agresja, w tym mściwość i chęć mordy; psotliwość; niesprawiedliwość; wybredność w stosunkach międzyludzkich) (X120 [X121–X127]),
 - ciekawskość X128);
- podstawowa funkcja fabularna:
- protagonista (X129),
 - antagonistą (X130),
 - donator (postać przekazująca protagoniście przedmioty potrzebne do wykonania zadania) (X131),

¹² Motyw oddania diabłu tego, co się zostanie w domu, a o czym się nie wie.

¹³ Cechy tej makrokategorii dzielą się głównie na pozytywne i negatywne. Zasadą jest, że w takich przypadkach koduje się wartość 0,5 przy zmiennej oznaczającej cechy pozytywne lub negatywne oraz 0,5 przy szczegółowym określeniu, jaka jest to cecha: spryt, pracowitość, lenistwo itd. W ten sposób uzyskuje się sumę 1 dla każdej cechy. Jednakże możliwa jest także sytuacja, w której dana postać ma więcej niż jedną cechę pozytywną lub negatywną. W tym przypadku wartości w zmiennej ogólnej sumują się w taki sposób, że np. postać o trzech cechach pozytywnych zostanie zakodowana z wartością 1,5 (3 × 0,5) w zmiennej „cechy pozytywne” oraz po 0,5 przy każdej szczegółowej cesze.

- fałszywy donator (potencjalny donator, którego dary są albo odrzucane, albo okazują się zbyteczne/szkodliwe) (X132),
- strażnik zaświatów/dusz (X133),
- trickster (X134),
- bohater kulturowy (postać przekazująca „dobra kulturowe”, np. rolnictwo, sztukę lekarską) (X135),
- narzeczona/narzeczony protagonisty/protagonistki (X136),
- postać inicjująca przygodę/fabulę. (X137),
- sędzia/rozdawca dóbr (postać nagradzająca protagonistę albo przywracająca ład) (X138).

Kodowanie odbywa się na każdym poziomie poniżej makrokategorii. Przykładowo, pies będzie różnił się od kota tylko o jeden punkt (oba zwierzęta mają zakodowane 1 pod podstawową kategorią ontologiczną: zwierzę, więc różnią się w jej ramach), natomiast o dwa punkty od człowieka (inna kategoria ontologiczna).

Korpus badanych tekstów zawierał również postacie nienadające się do scharakteryzowania, z których ujęcia zbiorze danych zrezygnowano. Na przykład jeśli bajka rozpoczyna się od stwierdzenia, że młodszy brat został wygnany przez starszych, przy czym ci starsi nigdy później nie są wspomniani, to nie ma potrzeby kodowania ich jako *dramatis personae* bajki. Podobnie, w bajce 18D pojawiają się rycerz i wieśniak, którzy nie zgadzają się zostać kumami dziecka bohatera, ale poza przynależnością do klasy społecznej nie mamy o nich żadnych informacji, więc nie można ich sklasyfikować. Co ciekawe, w wariacie 18A w tych rolach pojawia się Bóg i diabeł, przy czym bohater odrzuca ich propozycje ze względu na kolejno niesprawiedliwość i psotność. W tym przypadku mamy poszerzone charakterystyki, które dają się kodować i klasyfikować.

Aneks 2. Dane

Dane zaprezentowane zostały w następującym formacie: [oznaczenie tekstu w analizowanej próbie] : [nazwa identyfikująca postać] : [lista zmiennych charakteryzujących daną postać]. Zmienne przyjmują wartości oznaczone symbolami objaśnionymi w aneksie 1.

- [1A]: [diabeł] : [X1;X11;X17;X66;X130]
- [1A]: [dziewczyna-kwiat] : [X1;X12;X56;X73;X85;X129]
- [1A]: [królewicz] : [X1;X11;X17;X56;X136]
- [1B]: [diabeł] : [X1;X11;X57;X62;X66;X103;X130]
- [1B]: [dziewczyna] : [X1;X12;X56;X85;X129].
- [1C]: [dziewczyna-kwiat] : [X1;X12;X17;X56;X73;X85;X129]
- [1C]: [diabeł 1] : [X1;X11;X17;X57;X62;X66;X103;X130]
- [1C]: [diabeł 2] : [X1;X57;X62;X11]
- [1C]: [hrabia] : [X1;X11;X56;X136]
- [1D]: [dziewczyna] : [X1;X12;X17;X56;X129]

- [1D]: [dwóch panów] : [X1;X11;X14;X17;X66;X73;X136;X138]
[1D]: [nieboszczyk] : [X1;X10;X11;X73;X130]
[1E]: [dziewczyna] : [X1;X12;X24;X56;X73;X129]
[1E]: [kawaler-królewicz] : [X1;X11;X17;X66;X130;X136]
[1E]: [czarownik] : [X1;X11;X136]
[1F]: [kawaler] : [X1;X11;X66;X130;X136]
[1F]: [dziewczyna] : [X1;X12;X56;X129]
[2]: [diablica] : [X1;X12;X17;X57;X62;X74;X129;X136]
[2]: [chłopak] : [X1;X11;X56;X85;X129]
[2]: [ksiądz] : [X1;X11;X45;X56;X85;X131]
[2]: [brat diablidy] : [X1;X11;X57;X62;X74]
[2]: [ojciec diablidy] : [X1;X11;X57;X62;X74;X131;X133;X138]
[2]: [dusze] : [X13;X74]
[3A]: [gospodarz] : [X1;X11;X56;X129]
[3A]: [gospodyni] : [X1;X12;X56;X136]
[3A]: [Lucyfer] : [X1;X11;X57;X62;X74;X138]
[3A]: [dziedzic] : [X1;X11;X56]
[3A]: [diabeł] : [X1;X11;X57;X62;X111;X131;X138]
[3B]: [chłop] : [X1;X11;X56;X129]
[3B]: [diabeł] : [X1;X11;X57;X62;X86;X111;X130;X131;X138]
[3C]: [gospodarz i gospodyni] : [X1;X13;X56;X85;X129]
[3C]: [diabeł] : [X1;X11;X57;X62;X111;X130;X131;X138]
[3C]: [Antychryst] : [X1;X57;X62;X11;X74;X138]
[3C]: [dziedzic] : [X1;X11;X56;X131]
[3D]: [chłop] : [X1;X11;X56;X129]
[3D]: [diabeł] : [X1;X11;X57;X62;X130;X131]
[3E]: [Bogacz/gospodarz (dwie postacie przenikające się w narracji)] :
[X1;X11;X56;X129]
[3E]: [diabeł pustelnik] : [X1;X11;X57;X62;X111;X130;X131;X138]
[3E]: [najstarszy wojownik] : [X11;X57;X62;X130;X138]
[4]: [człowiek] : [X1;X11;X56;X85;X129]
[4]: [diabeł] : [X1;X11;X57;X62;X86;X131;X138]
[4]: [Lucyfer] : [X1;X11;X57;X62;X74;X138]
[5]: [człowiek] : [X1;X11;X56;X129]
[5]: [diabeł] : [X1;X11;X57;X62;X111;X131;X138]
[5]: [zbójnicy] : [X1;X11;X56]
[6]: [kował] : [X1;X11;X56;X129]
[6]: [diabeł] : [X1;X11;X57;X62;X111;X130;X135]
[7A]: [chłop] : [X1;X11;X56;X129]
[7A]: [diabeł] : [X1;X57;X62;X11]
[7B]: [diabeł] : [X1;X57;X62;X11]
[7B]: [Bóg] : [X1;X11]
[7C]: [chłop] : [X1;X11;X56;X129]
[7C]: [diabeł] : [X1;X57;X62;X11]

- [7C-wariant 1]: [baba] : [X₁;X₁₂;X₅₆;X₁₂₉]
 [7C-wariant 2]: [diabeł] : [X₁;X₅₇;X₆₂;X₁₁]
 [7D]: [Lucyfer] : [X₁;X₁₁;X₅₇;X₆₂;X₇₄;X₁₃₀;X₁₃₈]
 [7D]: [chłop] : [X₁;X₁₁;X₅₆;X₁₂₉]
 [7D]: [diabeł] : [X₁;X₁₁;X₅₇;X₆₂;X₇₄;X₁₁₂;X₁₃₀]
 [7E]: [chłop] : [X₁;X₁₁;X₅₆;X₁₂₉]
 [7E]: [diabeł] : [X₁;X₅₇;X₆₂;X₁₁]
 [8]: [chłop] : [X₁;X₁₁;X₅₆;X₁₂₉]
 [8]: [diabeł] : [X₁;X₁₁;X₅₇;X₆₂;X₁₃₀;X₁₃₁;X₁₃₈]
 [9]: [bogaty gospodarz] : [X₁;X₁₁;X₂₃;X₅₆;X₁₃₀;X₁₃₈]
 [9]: [ojciec i syn] : [X₁;X₁₁;X₅₆;X₁₂₉]
 [9]: [diabeł/diabły] : [X₁;X₁₁;X₁₉;X₅₇;X₆₂;X₁₃₁;X₁₃₈]
 [10]: [wdowa] : [X₁;X₁₂;X₂₇;X₅₆;X₁₃₇]
 [10]: [córka wdowy] : [X₁;X₁₂;X₅₆;X₁₂₉;X₁₃₈]
 [10]: [Rumpalski] : [X₁;X₁₁;X₁₇;X₅₇;X₆₂;X₇₄;X₉₃;X₁₁₁;X₁₃₀;X₁₃₁;X₁₃₈]
 [11]: [dziedzic] : [X₁;X₁₁;X₅₆;X₁₃₀]
 [11]: [wdowa] : [X₁;X₁₂;X₅₆;X₁₂₉]
 [11]: [diabły] : [X₁;X₁₁;X₁₇;X₁₉;X₅₇;X₆₂;X₈₈;X₁₃₈]
 [11]: [sędzia] : [X₁;X₁₁;X₅₆]
 [12]: [kupiec] : [X₁;X₁₁;X₅₆;X₁₃₇]
 [12]: [diabeł] : [X₁;X₁₁;X₅₇;X₆₂;X₁₁₁]
 [12]: [syn kupca] : [X₁;X₁₁;X₂₃;X₄₅;X₅₆;X₈₅;X₁₂₉;X₁₃₅]
 [12]: [diabły] : [X₁₃;X₅₇;X₆₂;X₇₄;X₁₃₃]
 [12]: [zbójnik] : [X₁;X₁₁;X₅₆;X₁₃₇]
 [12]: [kobieta z chatki zbójnika] : [X₁₂;X₅₆;X₁₃₁]
 [12]: [jabłka-gołębie] : [X₁₃]
 [13]: [głupi brat] : [X₁;X₁₁;X₅₆;X₈₅;X₁₂₉;X₁₃₄]
 [13]: [bracia] : [X₁;X₁₁;X₅₆;X₁₃₇]
 [13]: [czarnoksiężnik] : [X₁;X₂;X₁₁;X₅₆;X₁₃₁;X₁₃₇]
 [13]: [żona czarnoksiężnika] : [X₁;X₂;X₁₂;X₂₇;X₅₆;X₁₃₁]
 [13]: [córka czarnoksiężnika] : [X₁;X₁₂;X₁₇;X₂₆;X₅₆;X₁₂₀;X₁₃₀;X₁₃₆]
 [13]: [trzy diabły] : [X₁;X₁₁;X₅₇;X₆₂;X₈₅;X₁₂₀;X₁₃₁]
 [13]: [kochanek córki czarnoksiężnika] : [X₁;X₁₁;X₅₆;X₁₃₀]
 [14]: [córka wdowca] : [X₁;X₁₂;X₅₆;X₁₂₉]
 [14]: [córka wdowy] : [X₁;X₁₂;X₅₆;X₁₂₀;X₁₂₁;X₁₂₉]
 [14]: [Baba Jaga] : [X₁;X₁₂;X₅₆;X₁₃₁;X₁₃₈]
 [14]: [wdowa] : [X₁;X₁₂;X₂₇;X₅₆;X₁₃₀;X₁₃₇]
 [14]: [wdowiec] : [X₁;X₁₁;X₂₇;X₅₆;X₁₃₇]
 [15]: [fryzjer] : [X₁;X₁₁;X₅₆;X₁₁₃;X₁₂₈;X₁₂₉]
 [15]: [stary człowiek] : [X₁;X₁₁;X₅₆;X₁₃₇]
 [15]: [strach] : [X₁;X₁₂;X₁₅;X₁₇;X₁₈;X₁₂₀]
 [15]: [książę] : [X₁;X₁₁;X₁₈;X₅₆]
 [16]: [głupi syn] : [X₁;X₁₁;X₅₆;X₁₁₁;X₁₂₉]
 [16]: [ojciec] : [X₁;X₁₁;X₂₇;X₅₆;X₁₃₇]

- [16]: [strach] : [X1;X11;X111;X130;X138]
[17]: [Janko] : [X1;X11;X24;X56;X85;X87;X92;X128;X129]
[17]: [ksiądz] : [X1;X11;X56;X137]
[17]: [hrabia] : [X1;X11;X56;X137;X138]
[17]: [strachy z kościoła] : [X1;X11;X19;X56;X84]
[17]: [dziadek z pałacu/ojciec Janka] : [X1;X10;X11;X16;X27;X56;X87;X88;X112;X131]
[17]: [strachy z pałacu] : [X11;X19;X84]
[17]: [Marysia] : [X1;X12;X17;X19;X56;X84;X85;X136]
[18A]: [gospodarz] : [X1;X11;X27;X56;X137]
[18A]: [Bóg] : [X1;X11;X132;X138]
[18A]: [diabeł] : [X1;X11;X57;X62;X19;X132]
[18A]: [Śmierć] : [X1;X10;X12;X108;X131;X135]
[18A]: [chłopak lekarz] : [X1;X11;X25;X56;X85;X129]
[18B]: [chłop] : [X1;X11;X27;X56;X129]
[18B]: [Śmierć] : [X1;X10;X12;X20;X75;X108;X110;X131;X135;X137]
[18B]: [dziecko lekarz] : [X1;X11;X56;X85;X129]
[18C]: [chłop] : [X1;X11;X25;X27;X56;X85;X129]
[18C]: [Śmierć] : [X1;X10;X12;X108;X109;X135;X138]
[18D]: [człowiek] : [X1;X11;X25;X27;X56;X129]
[18D]: [bocian] : [X56;X137]
[18D]: [Śmierć] : [X1;X10;X12;X108;X109;X120;X135;X138]
[19]: [gospodarz] : [X1;X11;X25;X56;X79]
[19]: [Śmierć] : [X1;X10;X12;X110]
[20]: [gospodarz] : [X11;X56]
[20]: [Panek (diabeł)] : [X11;X17;X19;X57;X62]
[20]: [ksiądz (syn gospodarza)] : [X11;X23;X45;X56;X85;X129]
[20]: [staruszka] : [X12;X27;X56]
[20]: [Madej] : [X9;X11;X56;X137]
[20]: [Antychryst] : [X11;X57;X62;X129]
[20]: [Tyłynda] : [X11;X16;X57;X62;X131]
[20]: [diabły] : [X11;X19;X57;X62;X129]
[20]: [bawełna-jabłoń] : [X13]