Rola analizy sieci społecznych w odkrywaniu narracyjnej struktury fikcji literackiej

A. JARYNOWSKI[1, 2, 3], S. BOLAND[4]
andrzej.jarynowski@sociology.su.se

[1]Zakład Teorii Układów Złożonych, Instytut Fizyki, Uniwersytet Jagielloński w Krakowie
[2]Pracownia Technik Wirtualnej Rzeczywistości, Centralny Instytut Ochrony i Bezpieczeństwa Pracy – Państwowy Instytut Badawczy w Warszawie
[3]Instytut Socjologii, Uniwersytet w Sztokholmie (Szwecja)
[4]Instytut Filologii, Queen Mary, Uniwersytet w Londynie (Wielka Brytania)

Narzędzia matematyczne rozwinięte w celu opisu układów złożonych z powodzeniem są używane w naukach społecznych oraz coraz częściej znajdują zastosowanie w humanistyce. W prezentowanym interdyscyplinarnym projekcie pragniemy wykorzystać metody analizy sieciowej, aby lepiej zrozumieć sposób kreacji oraz przedstawienia świata przez autorów utworów literackich. Jednakże percepcja takiego świata zależy od subiektywnej wizji czytelnika, więc zwróciliśmy szczególną uwagę na różne sposoby ekstrakcji sieci powiązań społecznych z fikcyjnej rzeczywistości. Celem badań było odczytywanie różnych interakcji społecznych w tekście przez porównanie sieci otrzymanych przez algorytmy przetwarzania języka naturalnego (ang. *natural language processing NLP*) z tymi odtworzonymi na podstawie kwestionariuszy wypełnionych przez czytelników. Sieci dialogów, czy uczestnictwa w tej samej scenie zostały już opisane przez naukowców z amerykańskich uniwersytetów Stanford i Columbia, ale wciąż brakowało analizy relacji na ogólniejszym poziomie (interakcje międzyludzkie nie ograniczają się jedynie do dialogów bądź przebywania w tym samym miejscu). Zaproponowaliśmy kilka metod NLP w celu detekcji tych interakcji i skonfrontowaliśmy je z ludzkim postrzeganiem. Przy okazji odkryliśmy obszary teorii literatury, w których nie da się wykorzystać analizy sieciowej (np. interakcje nawiązujące do fabuły nie tworzą klasycznego trójkąta z punktem kulminacyjnym znanym z teorii literatury).

Słowa kluczowe: analiza sieci społecznych, przetwarzanie języka naturalnego, percepcja narracji.

1. Wprowadzenie

1a) Ilościowe badania tekstów

W związku z rozwojem metod komputerowych w przetwarzaniu informacji pojawiła się możliwość wykorzystania procedur automatycznych w analizie literackiej. Proste czynności zliczeniowe (ang. *counting*), które dało się zautomatyzować, zostały zaadaptowane do warsztatu badacza literatury. Jednak prawdziwy przełom został spowodowany wykorzystaniem narzędzi, które wykazały relacje między różnymi jednostkami analizy. Umożliwiły to techniki przetwarzania języka naturalnego (ang. *Natural Language Processing* – NLP), które pozwoliły analizować treści strukturyzowane składniowo i gramatycznie w ramach określonego kodu językowego. Metody i techniki NLP zostały opracowane przez informatyków w celu ułatwienia przetwarzania informacji zapisanych w sposób zrozumiały dla człowieka, a nie dla systemów analitycznych (takich jak komputery). Dopiero przetransformowanie takich informacji pozwala analitykom na zautomatyzowanie swojej pracy. Początkowo techniki NLP służyły dokonywaniu analizy niestrukturyzowanych danych w postaci zwykłego tekstu. Na przykład bez konieczności czytania tekstów, dokonywano oceny nacechowania emocjonalnego wypowiedzi na podstawie doboru słów oraz ich struktury. Techniki NLP znalazły wiele komercyjnych zastosowań i są coraz silniej rozwijane w celu jak najdokładniejszej klasyfikacji tekstów.

Nie jest to jednak ich jedyne zastosowanie, gdyż badania ilościowe w literaturze zagościły już na dobre w literaturoznawstwie czy językoznawstwie. Najlepszym przykładem takiego podejścia jest analiza Zipfa [1], badająca rozkład częstości występowania słów w tekście. Okazuje się, że na podstawie obserwacji częstości występowania słów można odróżniać pisarzy czy gatunki literackie. Ponadto wielu badaczy [2] zajmuje się analizowaniem dodatkowych połączeń między słowami (ich umiejscowienia w zdaniu), które można przedstawić w postaci sieci.

1b) Analiza sieciowa

W niniejszym artykule analiza sieciowa jest podstawowym narzędziem metodologicznym. Początków teorii sieci złożonych (niezwykle dynamicznie rozwijającej się dziedziny nauki) można szukać w pracach [3, 4] o grafach przypadkowych. Istotną zaletą podejścia sieciowego jest ogromne pole jego potencjalnych zastosowań: od układów fizycznych, przez biologiczne, po społeczne. Można ich używać właściwie wszędzie, gdzie istnieją zależności między elementami. Na pograniczu nauk ścisłych oraz społecznych wyodrębniła się bardzo popularna technika badawcza: analiza sieci społecznych (ang. *Social Network Analysis* – SNA). Struktura powiązań między ludźmi ma istotny wpływ na wiele czynników, takich jak przepływ informacji lub stosunek władzy. Położenie jednostki w sieci również warunkuje wiele cech, jak choćby pozycja społeczna. W ramach teorii SNA wyróżniamy sieci binarne (połączenie między elementami istnieje bądź nie) oraz ważone (połączenia mogą mieć różne wagi, odzwierciedlające różne poziomy intensywności interakcji). W naszym przypadku intensywność relacji będzie miała znaczenie w dokładniejszym opisie interesującej nas społeczności. Wyróżniamy ponadto sieci skierowane (połączenie między elementami ma ustalony kierunek) oraz nieskierowane (połączenia nie mają kierunku, a rejestrowana jest tylko relacja wzajemności).

1c) Percepcja sieci w literaturze

Wiedząc, jak ważne są sieci społeczne w rzeczywistości, warto podjąć się oceny roli takiej sieci wśród bohaterów świata fikcyjnego. Najpierw należy zwrócić uwagę na sam proces percepcji sieci oczami czytelnika postawionego przed konstruktem narracyjnym utworzonym przez autora. Warto podkreślić raz jeszcze, że niniejszy artykuł dotyczy właśnie przede wszystkim kwestii postrzegania sieci w różnych kontekstach czytelniczych, w porównaniu z komputerowymi metodami automatycznej detekcji. Rozważania zawarte w tekście są wynikiem badań eksploracyjnych, nie spotkaliśmy się bowiem wcześniej z analizą sieci społecznych w literaturze pod tym kątem. To znaczy, że spodziewanym wynikiem naszych analiz będą jedynie hipotezy badawcze dotyczące pogłębionego studium percepcji sieci przez czytelnika. Dlatego też chcielibyśmy skoncentrować się na interpretacji sieci społecznych, odczytywanych subiektywnie przez czytelnika, napominając tylko o innych kontekstach, które zapewne warte są gruntownego przebadania.

1d) Ekstrakcja sieci

Warto jednak umiejscowić nasze badania w szerszym kontekście dotychczasowej wiedzy na temat sieci społecznych w literaturze. Prym w tej dziedzinie wiodą trzy amerykańskie ośrodki: grupa NLP z Uniwersytetu Columbia w Nowym Jorku, „Literary Lab" z Uniwersytetu Stanford w Kalifornii oraz grupa „Google Books" z Uniwersytetu Harwarda w Bostonie. Tego typu badania zaliczają się do cyfrowej humanistyki (ang. *Digital Humanities*). Każdy z poprzednio wspomnianych zespołów koncentruje się na innym aspekcie badań, niemniej jednak ich wspólnym mianownikiem pozostaje problem automatycznej ekstrakcji sieci za pomocą technik przetwarzania języka naturalnego. Pierwsze prace dotyczyły analizy utworów dramatycznych, gdzie połączeniem między bohaterami (postaciami) było ich występowanie w tej samej scenie. Pierwsze właściwości takich sieci zostały pokazane na przykładzie dzieł Szekspira [5] i następnie porównane z dramatami chińskimi. Jedną z najważniejszych poczynionych obserwacji była znacząca różnica w ilości połączeń społecznych (liczba znajomych) centralnych postaci z dzieł europejskich, wokół których koncentruje się cała sieć w porównaniu z chińskimi, gdzie postacie miały bardziej równomiernie rozłożoną liczbę znajomych w społeczności. Naukowcy z Harvardu, wśród nich Martin Nowak [6], szukali asocjacji pomiędzy różnymi hasłami pojawiającymi się w zbiorze książek zdigitalizowanych przez Google. Powstały mapy semantyczne ukazujące np. natężenie występowania pojęć typu „grypa hiszpanka" w okresie następującym po pandemii. Jednak punktem wyjścia do niniejszej analizy są prace grupy z Uniwersytetu Columbia, w których m.in. wyekstrahowano sieci kontaktów konwersatoryjnych bohaterów dziewiętnastowiecznych powieści wiktoriańskich [7]. W tym celu skonstruowano algorytm oparty na uczeniu maszynowym, który na początku ustala listę bohaterów. W związku z przyjętą definicją sieci kontaktów, poprzez sieć konwersatoryjną uznawane są sytuacje dialogowe w postaci rozmowy oraz tekstu akapitowego zarówno w formie mowy niezależnej, jak i zależnej. Tak utworzone sieci pokazują np. zasadniczą różnicę między pierwszo- a trzecioosobową

narracją. W relacji trzecioosobowej powiązania w sieci są w miarę równomiernie rozłożone, w przeciwieństwie do koncentracji na „ja" w relacji pierwszoosobowej. Uznaliśmy jednak, iż zawężenie interakcji społecznych do konwersacji, bardzo sensowne z punktu widzenia automatycznej ekstrakcji, niestety wymaga uzupełnienia i weryfikacji poprzez porównanie z wizją czytelników. Tak wąsko rozumiane sieci korelowano tylko z wynikami kwestionariuszy, wypełnionych przez osoby, którym podano wytyczne zaznaczania interakcji zgodnie z algorytmem automatycznej anotacji. W związku z tym człowiek anotował interakcję według podobnych reguł, którymi również posługiwał się algorytm, w związku z tym wysoka zgodność między metodami była spodziewana.

2. Wybór materiału i koncepcja badania

Znając dotychczasowy stan wiedzy w dziedzinie cyfrowej humanistyki, wybraliśmy praktycznie nieporuszany temat percepcji sieci samej w sobie (czyli abstraktu niezależnego od nadawcy i odbiorcy). W tym celu postanowiliśmy przeprowadzić eksperyment, w którym porównywane byłyby komputerowe algorytmy ekstrakcji sieci z subiektywną percepcją uczestników badań. Samo zróżnicowanie charakterologiczne czytelników i ich opinii jest również istotne dla badania, a nawet najważniejsze z perspektywy tego artykułu. Do pracy nad tekstem wybraliśmy tom opowiadań Sherwooda Andersona *Winesburg, Ohio*, należący do kanonu literatury anglosaskiej. Wybór ten argumentowany jest treścią utworu: mamy tu do czynienia z opisem życia obyczajowego społeczności małego miasteczka, to właśnie interakcje społeczne budują więc fabułę opowieści (co próbowaliśmy później weryfikować ilościowo). Warto wspomnieć także o dużej liczbie bohaterów, którzy tworzą zamknięty zbiór postaci, pojawiających się w całym zbiorze opowiadań Andersona [8]. Utwory tego twórcy są też na tyle krótkie (średni czas czytania pojedynczego opowiadania, oszacowany na podstawie badania pilotażowego, to jedna godzina), aby nie stanowiło to bariery dla respondentów (odpowiedź na pytania wymagała oczywistej znajomości tekstu).

Konkretne metody komputerowe oraz sposób zbierania informacji o perspektywach czytelniczych zostały opracowane tak, aby można było powtórzyć badanie na każdym tekście, jednakże ze względu na powyższe zasady zostały one przeprowadzone właśnie na opowiadaniach o Winesburg.

3. Postrzeganie działania społecznego

3a) Operacjonalizacja badania

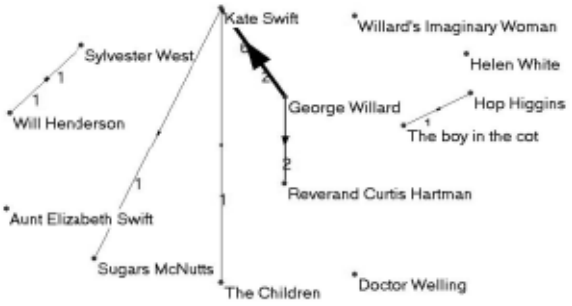
Przed ustaleniem ostatecznej wersji ankiety wysłanej do respondentów, rozważaliśmy różne sposoby rozumienia interakcji. Wynikiem był kompromis między chęcią zebrania jak największej ilości informacji a naukowymi standardami sprawdzalności oraz porównywalności. W związku z eksploracyjnym charakterem badań (nie wiedzieliśmy do końca, czego możemy się spodziewać), przedstawiamy kilka kontekstów, które również powinny zostać uwzględnione. Ostatecznie do naszych szczegółowych badań wybraliśmy sieci, tak jak widzi je czytelnik, bez dodatkowych kontekstów. W tym celu zostały przygotowane dwa zadania. Pierwsze, nazywane później komunikacyjnym, polegało na zaznaczaniu interakcji w trakcie czytania tekstu, a drugie – interakcyjne – na wypełnieniu macierzy interakcji wartościami wskazującymi na ważność relacji (znowu rozumianą tak, jak postrzega ją czytelnik). Pomimo ostatecznego wyboru najszerszej definicji interakcji, postanowiliśmy zasygnalizować w tym artykule również inne aspekty.

3b) Podejście komunikacyjne

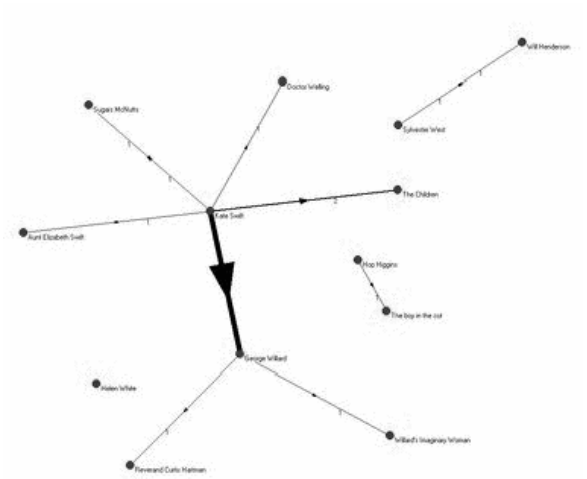
Zaczęliśmy od teorii informacji i schematu nadawca – komunikat – odbiorca [9]. Oznacza to istnienie kierunku w kanale komunikacji, choć czasami trudne wydaje się ustalenie, kto jest kim w tej konstrukcji, przy pewnym rygorze jest to możliwe (mniej lub bardziej obiektywnie). Należy mieć na uwadze, że komunikat nie jest tylko wypowiedzią ustną bądź pisemną. Rozróżniamy również komunikaty niewerbalne, a także różne fizyczne i niefizyczne formy interakcji. Założyliśmy, że nadawca to osoba, która ma wpływ na odbiorcę. Przykładowe zdarzenie: jeśli myślę teraz o mojej dziewczynie i jej słowach wypowiedzianych poprzedniego dnia, to ma ona wpływ na mnie w tym momencie. Należy rozumieć „myślenie" jako komunikację, w procesie której odbieram informacje, uprzednio wysłane przez moją dziewczynę. Konsekwencją będą tu interakcje bezpośrednie *explicite* i pośrednie *implicite*. Tak więc ja, myśląc o mojej dziewczynie, będę odbierał wiadomość *implicite*,

w przeciwieństwie do komunikatu *explicite*, wysłanego przez nią wczoraj.

Przykładowa sieć ukazująca interakcje rozumiane w kontekście komunikatów (rysunek 1) została wygenerowana przez jedną osobę, która notowała interakcje w trakcie czytania opowiadania *The Teacher*. Grubość połączenia jest sumą wszystkich zauważanych interakcji, a wielkość strzałki wskazuje na liczbę interakcji w danym kierunku. Utwór opowiada o romansie tytułowej nauczycielki z miejscowym dziennikarzem, jest to reprezentowane poprzez grubość połączenia między nimi. Natomiast poprzez kierunek interakcji można zauważyć, że to nauczycielka jest aktywną stroną w związku. W tym kontekście, pojedyncza interakcja *explicite* zaistniałaby, gdyby np. dwoje ludzi rozmawiało ze sobą. Ukryte działanie zostałoby oznaczone jako *implicite*, np. gdy jedna z postaci coś sobie przypomniała. Rozróżnienie między zdarzeniami *explicite* i *implicite* zostało usunięte z pełnego badania pilotażowego, pomimo że pozwalałoby na postawienie hipotezy, że interakcje, które możemy „zobaczyć", są oceniane jako ważniejsze. Przeprowadzenie pogłębionych badań w tym zakresie wymagałoby przyjęcia innej metodologii, mocniej akcentującej czynnik humanistyczny. Pozostawiliśmy więc tę sprawę otwartą.



Rys. 1a. Interakcje *implicite*

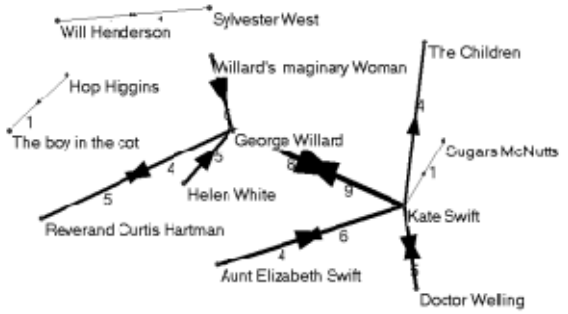


Rys. 1b. Interakcje *explicite*

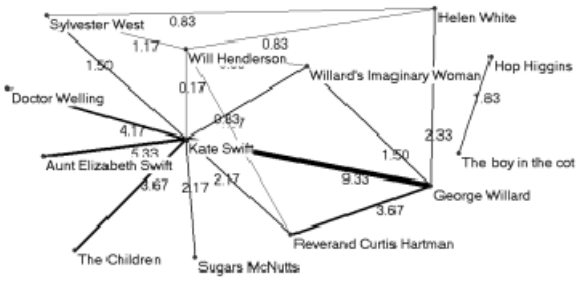
3c) Podejście komunikacyjno-interakcyjne

W celu uzupełnienia informacji, czytelnik miał zaznaczyć wagę połączeń między postaciami. Zestawione sieci (rysunek 2) dla zadania komunikacyjnego i interakcyjnego wyraźnie się różnią (są to zebrane wyniki badania pilotażowego).

W ostatecznej wersji zrezygnowaliśmy z sieci skierowanej, a komunikat został zrzutowany na interakcję. Było to spowodowane brakiem możliwości odczytania kierunku przesyłania komunikatu przez opracowane przez nas algorytmy.



Rys. 2a. Komunikaty z badania pilotażowego



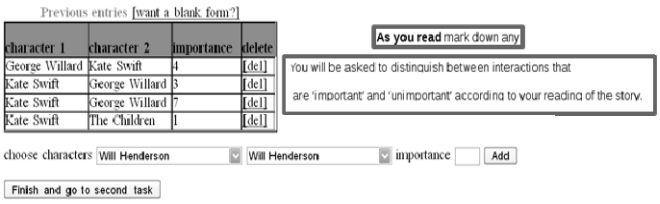
Rys. 2b. Interakcje z badania pilotażowego

4. Komputerowe metody ekstrakcji sieci

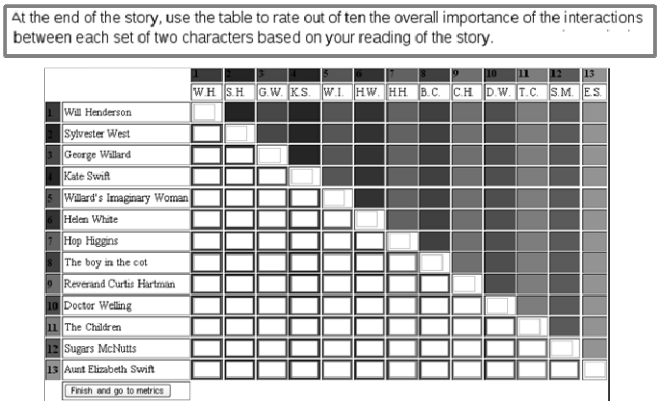
W implementacji algorytmów staraliśmy się skupić na jak najbardziej „ludzkim" sposobie wykrywania interakcji, czyli tak, jak widziałby je człowiek. Jak wyjaśniono wcześniej, interakcja jest to coś więcej niż rozmowa, może być fizyczna, psychiczna itp. Pojawienie się dwóch bohaterów w określonej jednostce tekstu (akapit, zdanie) wydaje się prostym i dobrym wskaźnikiem szeroko zdefiniowanej interakcji między postaciami. Znacznie trudniejszym zadaniem było wprowadzenie istotności czyli wagi połączenia. Jest to nowy element nie pojawiający się we wcześniejszych publikacjach. To, co dla człowieka wydaje się banalne, nie daje się łatwo zaimplementować. Ostatecznie wagę interakcji uzyskuje się przez pomnożenie częstotliwości nazw postaci występujących w danym fragmencie (jednostce analizy). Kolejnym uproszczeniem było określenie, kto jest postacią, a kto nie. Przykładowo: czy wyimaginowaną przyjaciółkę dziennikarza uznać za bohaterkę? W końcu sami przygotowaliśmy arbitralnie listę bohaterów z aliasami. Należy pamiętać, że algorytm może pracować w dwóch trybach: zdaniowym (jednostkami analizy są poszczególne zdania) oraz akapitowym (jednostkami analizy są poszczególne akapity).

5. Specyfikacja badania

Badanie zostało przeprowadzone w postaci ogólnodostępnej ankiety. Respondenci zostali poproszeni o wypełnienie czterech stron: formularza zgody, pierwszego zadania (rysunek 3), drugiego zadania (rysunek 4) i metryczki. W zadaniu pierwszym i drugim istotność interakcji była oceniana w zakresie 0–10. W metryczce pytaliśmy o różne czynniki demograficzno-społeczne, np. wiek, płeć, rodzaj wykształcenia. Do głównego badania wybraliśmy opowiadanie *The Philosopher*, w którym opisywane są perypetie młodego lokalnego dziennikarza. Osią fabuły są tu jego rozmowy z przyjacielem, lekarzem. Analiza wyników była anonimowa, ale badani mogli zostawić nam adres kontaktowy. Ostatecznie ankietę w całości wypełniło trzydzieści sześć osób (niestety, mieliśmy bardzo duży odsetek odrzuceń, gdyż ankieta była rozpoczynana ponad dwieście razy), głównie z Wielkiej Brytanii i USA. Badanie zostało przeprowadzone w języku angielskim, dużą grupę respondentów tworzyli jednak nienatywni użytkownicy języka angielskiego (studenci).



Rys. 3. Formularz do wypełniania zadania pierwszego komunikacyjnego z najważniejszymi wytycznymi

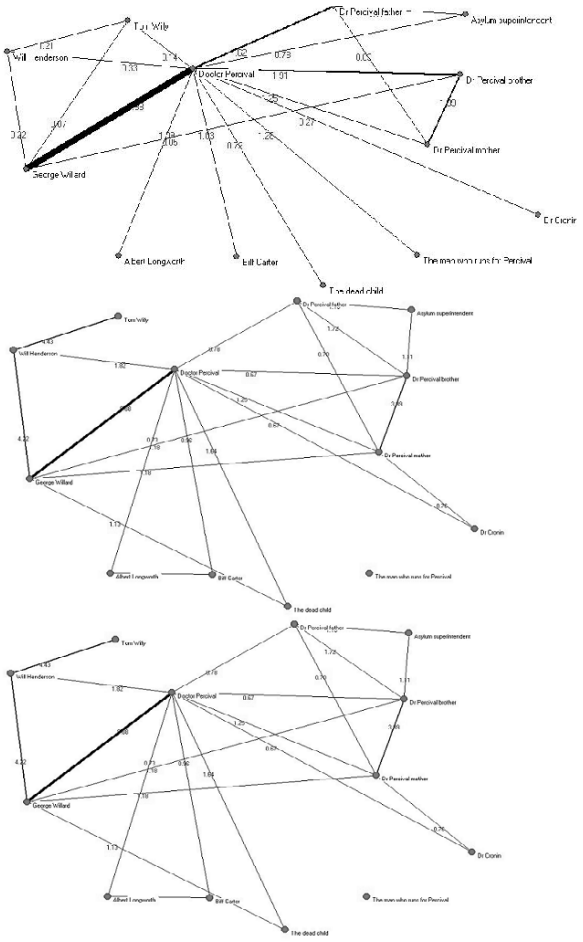


Rys. 4. Formularz do wypełniania zadania drugiego interakcyjnego z najważniejszymi wytycznymi

6. Analiza danych

6a) Budowa sieci

W celu porównania powyższych metod (dwóch opartych na odpowiedziach respondentów: pierwszej komunikacyjnej i drugiej interakcyjnej oraz dwóch komputerowych: akapitowej i zdaniowej) dokonano normalizacji wyników (ponieważ każda metoda posiadała inną skalę). Uzyskane sieci można przedstawić graficznie (rysunek 5). Widać, że wyglądają one podobnie. Najważniejsze połączenie z punktu widzenia fabuły (pomiędzy dziennikarzem a doktorem) w każdej metodzie jest uwypuklone. Zauważalną różnicą jest za to różna ilość połączeń w sieciach. Zadanie pierwsze – komunikacyjne wykazuje najmniej połączeń (16), a drugie – interakcyjne najwięcej (23), podczas gdy algorytm akapitowy daje wynik pośredni (21).



Rys. 5. Uzyskane sieci społeczne opowiadania *The Philosopher* z zadania pierwszego – komunikacyjnego (u góry), z zadania drugiego – interakcyjnego (u dołu) oraz algorytmu akapitowego (w środku)

Tab. 1. Statystyki sieci społecznych uzyskane z opowiadania *The Philosopher*

	Zadanie pierwsze – komunikacyjne	Zadanie drugie – interakcyjne	Algorytm akapitowy
Liczba połączeń	16	24	21
Gęstość sieci	0,103	0,153	0,135
Średnia krotność wierzch.	2,45	3,69	3,23

6b) Korelacja między metodami [10]

Najważniejszym zagadnieniem projektu było porównanie wyników uzyskanych od respondentów oraz za pomocą algorytmów (tabela 2). Zauważmy, że algorytm zdaniowy lepiej koreluje z zadaniem komunikacyjnym, a algorytm akapitowy lepiej z zadaniem interakcyjnym. Mimo to, ostatecznie w zestawieniu ze zdaniowym, lepiej wypada algorytm akapitowy. Oznacza to konflikt między holistycznym ujęciem nauk przyrodniczych a paradygmatem czynnika humanistycznego w naukach społecznych. Mianowicie, uzależnienie skuteczności metody algorytmicznej od sposobu zadania pytania respondentowi nie pozwala na opracowanie uniwersalnego narzędzia komputerowego odwarzającego sieci powiązań między postaciami.

Tab. 2. Macierz korelacji pomiędzy sieciami otrzymanymi od respondentów i za pomocą algorytmu

	Algorytm akapitowy	Algorytm zdaniowy
AD 1 (komunikacyjna)	**0,84**	**0,91**
AD 2 (interakcyjna)	**0,70**	**0,58**

6c) Korelacje pomiędzy odpowiedziami respondentów

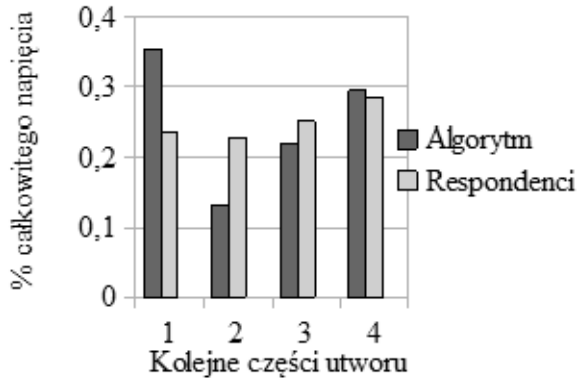
Porównajmy teraz korelacje pomiędzy sieciami uzyskiwanymi przez tych samych respondentów w zadaniu pierwszym – komunikacyjnym i drugim – interakcyjnym. Zgodność odpowiedzi spróbowaliśmy wyjaśnić za pomocą zmiennych niezależnych, podanych przez respondenta w metryczce za pomocą regresji logistycznej zmiennych o wymiarze nominalnym (tabela 3). Okazało się, że tylko rodzaj wykształcenia ma istotny statystycznie wpływ na te korelacje: tylko respondenci o wykształceniu ścisłym odpowiadali w wyraźnie podobny sposób. I tak: osoba z wykształceniem ścisłym wykazuje się większą powtarzalnością w obu zadaniach.

Tab. 3. Regresja logistyczna zmiennych nominalnych pokazująca relację między wykształceniem ścisłym a zgodnością w uzupełnianiu zadania pierwszego – komunikacyjnego oraz drugiego – interakcyjnego

Wykształcenie	Parametr	p-Value
Artystyczne i humanistyczne	0,05	0,37
Nauki społeczne	– 0,03	0,71
Nauki ścisłe i medyczne	0,16	< 0,01

6d) Testowanie tworzenia trójkąta Freytaga (punkt kulminacyjny)

Ostatnim etapem naszej analizy było zbadanie zdolności predykcyjnych dynamicznych charakterystyk sieci do odzwierciedlenia kulminacji napięcia charakterystycznego dla większości utworów literackich. Polegało to na zbadaniu intensywności i znaczenia interakcji w czasie. Nasza hipoteza zakładała, że suma i intensywność interakcji będzie odpowiadać budowaniu napięcia w narracji. W tym celu sprawdziliśmy, czy charakterystyczny trójkąt Freytaga [11] może być otrzymany automatycznie. Dzięki temu uzyskaliśmy metodę ilościową, ale wynik jakościowy. W tym celu wykorzystaliśmy dane z zadania pierwszego – komunikacyjnego oraz wyniki algorytmu akapitowego. Podzieliliśmy zebrane dane na cztery partie zgodnie z chronologią czasową, z tym że dla metody komputerowej ten podział miał miejsce wedle liczby wierszy, a z danych ankietowym wydzieliliśmy równoliczne kwartyle. Po utworzeniu wykresu z danych (rysunek 7), okazało się, że nie otrzymaliśmy oczekiwanej formacji. Możemy interpretować to na różne sposoby, np. że komputer znajduje wiele interakcji między postaciami na początku, ponieważ autor przedstawia większość bohaterów właśnie w tym miejscu. Respondenci za to w miarę jednorodnie oznaczają intensywność interakcji, z tym że obserwujemy niewielką tendencję wzrostową wraz z upływem czasu.



Rys. 7. Konstrukcja antytrójkąta na podstawie interakcji zebranych od respondentów w zadaniu pierwszym – komunikacyjnym oraz wygenerowanych przez algorytm akapitowy

7. Wnioski i spekulacje

Odkrywanie sieci społecznych w literaturze okazało się pasjonującym tematem, w którym wiele jest jeszcze do zrobienia, ponieważ nadal jest to pionierski obszar badań. Opracowaliśmy metody pozyskiwania sieci na podstawie ankiet czytelników, które mogą zostać wykorzystane do analizy dzieł innych niż *Winesburg, Ohio*, choć już na podstawie uzyskanych wyników możemy wyciągnąć pewne wnioski i postawić nowe hipotezy badawcze. W naszych badaniach uprościliśmy sposób zbierania informacji, w celu zachowania możliwości porównania z również wypracowanymi przez nas metodami komputerowymi, m.in. dlatego nie umieściliśmy respondenta w żadnym konkretnym kontekście czytelniczym. Już przedwstępna analiza pokazała, że sposób zadawania pytań ma znaczenie. Zamiast pytania o interakcje odczytywane przez czytelnika (ang. *reader world*), można rozważyć znaczenie interakcji dla fabuły (ang. *story world*). W ten sposób percepcja czytelnicza zostałaby ukierunkowana na świat przedstawiony w utworze, nie zaś na jego subiektywną wizję. Być może zmieniłoby to uzyskane wyniki i zbliżyłoby metody ankietowe do komputerowych. Pozostaje to jednak otwartym zagadnieniem, podobnie jak pytanie o związek między interakcjami typu implicite i explicite. Jednak, odnosząc się do uzyskanych przez nas wyników, najważniejsze wydają się korelacje między rezultatami ankiet oraz algorytmów (tabela 2). Niezmiernie ciekawym aspektem jest inny poziom korelacji pomiędzy odpowiednimi metodami komputerowymi a różnymi zadaniami czytelniczymi, co wskazuje na niejednoznaczność w sposobach opisywania interakcji i wrażliwość na czynniki zewnętrzne. Oznaczać to może, że nie istnieje jedna uniwersalna metoda ukazująca sieć społeczną w utworze literackim, co jest paradygmatem wciąż rozwijającej się dziedziny nauki, jaką jest NLP. Interesujące wydaje się również zaprezentowanie (tabela 3) istotnego statystycznie związku między sposobem postrzegania sieci a wykształceniem (reprezentanci nauk ścisłych budują sieci bardziej powtarzalne niż reszta). Na koniec sprawdziliśmy, że mierząc intensywność interakcji, nie możemy przełożyć tego bezpośrednio na budowanie napięcia w utworze.

Podsumowując, wydaje nam się, że poruszyliśmy wiele wątków, jakie niesie ze sobą wykorzystanie analizy sieci społecznych w literaturze, która ze względu na charakter ilościowy – może być narzędziem uzupełniającym w warsztacie współczesnego humanisty. Opisana metodologia może pomóc w weryfikowaniu hipotez odnoszących się do konkretnych tekstów, autorów czy też ogólnie – literatury. Opisane tu narzędzie jest na tyle

uniwersalne, że może zostać użyte – po dokonaniu niewielkich modyfikacji – w analizie innych tekstów literackich, tworzonych w różnych językach (choć po polsku tylko częściowo, gdyż cyfrowy korpus języka polskiego dopiero od niedawna daje się stosować w badaniach stricte ilościowych [12, 13, 14]).

8. Podziękowania

Opisane badanie zostało zrealizowane w Centrum Analizy Układów Złożonych przy Uniwersytecie w Yorku (Wielka Brytania), w ramach szkoły letniej [15]. Duży wkład w projekt mieli również pracownicy Centrum: Dan Franks, Elva Robinson, Richard Walsh oraz John Forrester.

9. Bibliografia

[1] G.K. Zipf, *The Psychobiology of Language*, Houghton-Mifflin, Oxford, 1935.
[2] I. Grabska-Gradzinska, A. Kulig i in. ”Complex Network Analysis of Literary and Scientific Texts”, *International Journal of Modern Physics C* 23: 50051 (2012).
[3] R. Albert, A.L. Barabási, ”Statistical mechanics of complex networks”, *Reviews of Modern Physics*, Vol. 74, 47–97 (2002).
[4] D. Watts, S. Strogatz, ”Collective dynamics of small worlds networks”, *Nature*, 393–440 (1998).
[5] F. Moretti, ”Network Theory, Plot Analysis”, *New Left Review*, No. 68, 1–64 (2011).
[6] J.M. Babtist i in., ”Quantitative analysis of culture using millions of digitized books”, *Science*, No. 331 (6014), 176–182 (2011).
[7] D.K. Elson, N. Dames, K.R. McKeown, ”Extracting Social Networks From Literary Fiction”, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, 2010.
[8] W.L. Phillips, ”How Sherwood Anderson Wrote Winesburg, Ohio”, *American Literature*, Vol. 23, No. 1, 7–30 (1951).
[9] R. Walsh, ”Emergent Narrative in Interactive Media”, *Narrative*, Vol. 19, No. 1, 72–85 (2011).
[10] A. Buda, A. Jarynowski, ”Network Structure of Phonographic Market with Characteristic Similarities between Artists”, *APhysPolA*, Vol. 123, No. 3 (2013).
[11] G. Freytag, *Die Technik des Dramas*, 1863.
[12] K. Głowacka, *Seks w 2005 miał kryzys, ale poprawia wyniki...*, *czyli jakich słów używa polska prasa*, TOK FM, 28.05.2012.
[13] A. Jarynowski, A. Rostami, ”Reading Stockholm Riots 2013 in social media using text-mining”, *Proccedings of 6th Language & Technology Conference* , Poznań, 2013.
[14] G. Szostek, M. Jaszuk, A. Walczak, „Automatyczna budowa semantycznego modelu objawów chorobowych na bazie korpusu słownego”, *Biuletyn Instytutu Systemów Informatycznych*, Nr 9, 35–43. (2012).
[15] http://www.york.ac.uk/yccsa/activities/summerschool/

Social networks analysis in discovering the narrative structure of literary fiction

A. JARYNOWSKI, S. BOLAND

In our paper we would like to make a cross-disciplinary leap and use the tools of network theory to understand and explore narrative structure in literary fiction, an approach that is still underestimated. However, the systems in fiction are sensitive to reader’s subjectivity and attention must to be paid to different methods of extracting networks. The project aims at investigating into different ways social interactions are *read* in texts by comparing networks produced by automated algorithms-natural language processing (NLP) with those created by surveying more subjective human responses. Conversation networks from fiction have been already extracted by scientists, but the more general framework surrounding these interactions was missing. We propose several NLP methods for detecting interactions and test them against a range of human perceptions. In doing so, we have pointed to some limitations of using network analysis to test literary theory (e.g. interaction, which corresponds to the plot, does not form climax).

Keywords: social network analysis, natural language processing, narration.