ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 September: 25(3): 89~115 http://dx.doi.org/10.13088/jiis.2019.25.3.089

텍스트 마이닝 기법을 활용한 고전 추리 소설 작가 간 문체적 차이와 문체 구조에 대한 연구

문석형
아주대학교 e-비즈니스학과
(*segreto123@ajou.ac.kr*)

강주영
아주대학교 e-비즈니스학과
(*jykang@ajou.ac.kr*)

본 연구는 고전 추리 소설 작가로 유명한 아서 코난 도일과 애거서 크리스티의 문체적 차이점을 데이터 분석을 통해 제시하고, 나아가 텍스트 마이닝에 입각한 문체 연구의 해석적 방법론을 제시하고자 시행되었다. 추리 소설의 핵심 요소인 사건과 인물에 더해 작가의 문법적인 집필 방식을 문체로 정의하고 분석을 시도하였다. 작가 별로 각 2권, 총 4권의 책을 선정하였으며 문장 단위로 텍스트를 나누어 데이터를 확보하였다. 각 문장에 따른 감성 점수를 부여한 뒤 페이지 진행에 따른 감성을 시각화하였으며, 페이지에 따라 토픽 모델링을 적용하여 소설 속 사건 진행 흐름을 파악할 수 있었다. 동시 발생 매트릭스(co-occurrence matrix)를 구성하고 네트워크 분석(Network Analysis)을 시행함으로써 사건이 진행되는 과정에서 인물들 간 관계의 변화를 확인할 수 있었다. 또한 전체 문장을 총 6가지 문체를 기준으로 문법적인 체계를 나누어 작가 간, 그리고 작품 간 집필 방식의 차이점을 확인하였다. 이러한 일련의 연구 과정은 문체에 대한 이해를 바탕으로 글 전체의 맥락을 파악할 수 있도록 도움을 줄 수 있으며, 나아가 기존에 개별적으로 진행되었던 문체 연구를 통합시킴으로써 문체 구조에 대한 이해를 도울 수 있다. 그리고 이러한 선행된 이해를 통해 온라인 텍스트를 비롯한 비정형 데이터 속 문체의 존재를 발견하고 구체화하는 작업에 기여할 수 있다. 뉴미디어를 포함한 온라인 텍스트를 심도 있게 분석하고자 하는 시도가 증가하고 있는 상황에서 해당 연구들과 연계를 통해 보다 의미 있는 온라인 텍스트 분석에 기여할 것으로 기대된다.

주제어 : 감성 분석, 토픽 모델링, 네트워크 분석, 문법, 문체

논문접수일 : 2019년 6월 19일 논문수정일 : 2019년 8월 20일 게재확정일 : 2019년 9월 17일
원고유형 : 일반논문 교신저자 : 강주영

1. 서론

소셜 미디어의 증가와 컨텐츠의 다양화는 사람들로 하여금 새로운 커뮤니케이션 방식에 참여하는 것을 활발하게 하였다. 소셜 미디어를 통한 커뮤니케이션은 단순히 의사를 전달하는 것을 넘어 정보를 교환하는 수단으로서 자리매김하였고, 이에 따라 기업은 자사의 제품과 서비스의 가치를 전달하는 새로운 마케팅 수단으로 소셜 미디어를 활용하게 되었다. Hwang (2013) 은 국내 100대 기업의 PR 커뮤니케이션을 위한 SNS 활용을 연구하면서 고객과의 접점을 형성하고 관계를 유지, 확대시키는 소셜 미디어의 기능을 논한 바 있다. 하지만 소셜 미디어에서 발

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2018-0-01424).

생하는 커뮤니케이션은 텍스트를 기반으로 함에도 정확한 문법적 체계가 아닌 커뮤니케이션의 상황과 사용자 개인의 성향, 즉 문체에 의해 좌우된다. 따라서 기업은 자사의 가치를 보다 효과적으로 전달하기 위해 마케팅 시장인 미디어와 대상인 사용자의 성향을 파악할 필요가 있으며, 이는 곧 텍스트에 담긴 사용자의 문체를 분석하는 텍스트 마이닝(Text Mining) (Kim et al., 2016) 활동으로 발전하였다.

텍스트 마이닝(Text Mining)은 언어학, 통계학, 기계 학습 등에 입각한 자연 언어 처리 기술을 통해 반 정형/비정형 텍스트 데이터를 정형화하고 분석하는 기법으로서, 주로 기업의 제품과 서비스에 대한 소셜 네트워크 상의 고객 리뷰 혹은 사회 이슈에 대한 대중의 의견을 분석하는 활동에 적용되었다 (Chae et al., 2015; Cho et al., 2017; Kim and Kang, 2018). 이러한 활동은 주로 텍스트 내에 존재하는 키워드를 도출해내거나 키워드 간 관계를 규명하는 것에 집중되었다 (Kim et al., 2016). 이는 문체를 탐구하는 연구 중 텍스트 언어학에 기반한 연구로, 텍스트의 고유한 구성 원칙을 특징적인 언어 실현 방식으로 파악하고자 하는 시도이다. 하지만 텍스트 차원에서 문체를 탐색하는 작업은 근원적인 어려움을 내포하는데, 특정 어법 및 표현 수단의 출현 빈도에 대한 정량적 분석이 문체 형성에 어떠한 기여를 했는지 추가적인 해석이 필요하기 때문이다. ‘문체’란 특정한 텍스트에서 느껴지는 전체적인 심적인 인상 혹은 이미지가 텍스트의 여러 층위에 속하는 언어 수단의 상호 작용에 의해 생겨난 것으로 정의된다. 따라서 문체 분석 또한 그들의 ‘문체 구조’를 밝혀내는 데 초점을 모아야 한다 (Cho, 2009).

본 연구는 텍스트 내에 숨겨진 언어 수단과 표현법을 관측하기 위해 고전 추리 소설의 대가로 명시되는 아서 코난 도일과 애거서 크리스티의 탐정 소설(Detective story) 2권, 총 4권의 텍스트 데이터를 수집했다. 탐정 소설이란 사건을 해결하는 탐정과 범행을 저지르는 범인으로 분명하게 역할이 구분되는 추리 소설 장르를 일컫는다. 두 작가의 탐정 소설을 비교 분석해 봄으로써 사건이 진행되는 양상과 그 속에 포함된 인물들 간 관계를 표현하는 작가의 문체적 특성을 파악하고자 했다. ‘문체 구조’를 파악하는 방법으로 추리 소설의 특징에 입각하여 ‘사건이 진행되는 양상’, ‘인물들 간 관계의 변화’, ‘문법적 표현 방법’ 3가지로 규정하였다. ‘사건이 진행되는 양상’은 각 문장 텍스트에 따른 감성 점수를 측정하여 감성 분석 그래프로 시각화한 뒤, 페이지별로 그래프 구간을 나누어 토픽 모델링(Topic Modeling)을 적용하였고, 이를 통해 사건의 진행 양상과 이를 표현하는 감정 표현의 일치성을 비교하였다 (Blei et al., 2003). ‘인물들 간 관계의 변화’를 측정하기 위해 나누어진 구간마다 동시 발생 매트릭스(co-occurrence matrix)를 구축하고 네트워크 분석(Network Analysis)을 수행함으로써 주요 인물들 간의 관계와 거리를 확인하였다 (Scott, 1988). 끝으로, ‘문법적 표현 방법’을 측정하기 위해 자체적으로 문법 기준을 설정한 뒤 문법 기준에 부합하는 6가지 문체(구어체, 문어체, 화려체, 건조체, 만연체, 간결체)를 정의하여 문장들을 분류하였다.

본 연구의 주된 목적은 진위가 텍스트의 여러 층위에 감춰져 복합적인 구조를 가지는 추리 소설을 텍스트 마이닝을 통해 분석해보고, 문법적 분석에 추가적인 문체 방식을 정의해 봄으로써 문체 구조의 기계적 분석이 가능함을 확인하는 것이다. 복합적인 텍스트에 내포되어 있는 인물들의 관계 변화를 사회 연결망으로 관측하고, 이

야기의 진행 흐름을 시각화하는 작업은 글 전체의 맥락을 파악하는 데 도움을 줄 수 있다. 또한, 기존에 연구되었던 분석 기법들을 하나의 주제로 산정하고 정의해 봄으로써 개별적으로 진행되었던 텍스트 언어학적 접근을 통합시키고 나아가 '문체 구조'에 대한 이해를 확립할 수 있다. 이러한 기계적 분석의 의의는 인간의 직관에 머무르는 문체를 구현함에 따라 장기적으로 능동적인 언어 표현을 요구하는 인공지능 개발 연구에 기여할 수 있을 것으로 판단된다. 다만, 후속 연구에 기여할 수 있는 분야가 분명함에도 언어학에서 정의하는 문체의 정의와 다양한 표현법을 본 연구를 통해 전부 담아내지 못하는 점이 한계점으로 고려된다.

2. 문헌 연구

2.1 문체 해석을 위한 연구

'문체'의 어원은 '끝이 뾰족한 필기구'라는 뜻의 라틴어 'stilus'에서 유래된 것으로, 필기구가 나타내는 글자의 모양에서 서법이나 어법의 방식을 나타내는 표현법으로 발전하였다 (Lee, 2006). 통상적으로 문체는 필자의 사상이나 개성이 글의 어구나 표현법을 통해 드러나는 글의 특색 내지는 체제로 정의가 되며 글의 종류에 따라 다른 양상을 띄는 것으로 받아들여진다. 따라서 '문체'에 대한 관련 연구 또한 텍스트의 형식과 표현법을 발견하고 이를 체계화하는 것으로 발전하였다. 아일랜드 소설가 조너선 스위프트의 서사 장르와 비-서사 장르에 따른 문체의 차이를 단어의 사용, 문장의 사용 등을 측정하여 비교한 연구 (Lee, 2019)나, 헤밍웨이의 단편 소설의 어휘, 문법의 사용을 측정하고 문체로서 미치는 영향을 규명한 연구 (Suh, 2018) 역시 이와 같은 맥락을 따르고 있다. 텍스트적 관점에서 벗어나 언어학적 관점에서 문체를 조명한다면 그러나, 동일한 행위를 서로 다른 발화를 통해 전달하는 것으로서 동일한 텍스트라도 발화자의 상황과 전달 방식에 따라 다른 의미를 갖출 수 있는 것이다. 이로 인해 문체는 언어 행위의 부분 양상으로 정의될 수 있으며, 행위들을 실현하기 위한 구체적인 방식뿐만 아니라 심적으로 내재된 의미와 개념을 드러내는 수단으로서 의의를 갖는다 (Cho, 2009). 따라서 문체에 대한 연구는 어법과 어휘에 기반한 의사 전달 방식과 더불어 표현 속에 내재되어 있는 주제의 흐름과 의미에 주목하는 것으로 발전되어야 한다. 일례로 서술의 시점 추이, 서술자의 시공간적인 위치에 따른 서술 문체의 특징과 효과를 살펴보고 해당 양상을 통해 텍스트의 심층에 깔린 주제와 의미를 분석하고자 시도한 연구 (Jeong, 2019)를 들 수 있다.

문체를 프로그램으로 구현하고 분석하고자 하는 시도는 분석 기법이 다양해지고 분석 툴 자체가 사용자 편의에 맞게 발전하면서 지속적으로 증가하고 있다. 문체를 총 4가지로 간주하여 (Lexical, Syntactic, Structural, Content-specific) 온라인상의 텍스트 문체가 소셜 미디어의 사용자 평판에 미치는 영향을 서포트 벡터 머신(SVM)과 랜덤 서브 스페이스를 결합한 앙상블 기법으로 관측하고 분류한 연구 (Suh, 2016) 가 있으며, semantic field(의미가 유사한 단어들의 집합)의 vector space를 구축하고 영미권 소설 작가들의 문체를 군집화(clustering) 한 연구 (Pavlyshenko, 2014) 가 있다. 또한 단어의 사용 현황을 분석하여 문체를 파악함으로써 텍스트의 표절을 밝혀내는 연구 (Oberreuter and Velásquez, 2013) 가 있다. 이러한 관련 연구들은 텍스트의 층위에 담겨

있는 내재적 의미의 흐름을 발견하고 정의하는 텍스트 마이닝 연구 라기보다는, 텍스트에 드러나는 어휘와 표현법을 체계화하고 보다 정확하게 분류하는 텍스트 마이닝 연구라고 할 수 있다 (Yang et al., 2018). 언어학에서 표명하는 문체의 개념이 다양하고 특정 유형을 산정하기 어려움에도 이를 포괄적으로 담아 내기 위해 기계적 분석을 시도한 사례를 찾아볼 수 없었다.

2.2 문체의 내재적 특성에 관한 연구

문서와 코퍼스 집단에서 관념적으로 이해할 수 있었던 추상적 개념을 자연 언어로 기술하고 해석하고자 하는 시도는 분석 툴이 발전하면서 더욱 확산되었다. 언어학에서 정의하는 언어는 정보를 전달하는 수단과 함께 '가변성'을 지니고 있어, 동일한 정보를 전달하는 과정에서도 표현 방법과 어휘의 사용에 따라 전달할 수 있는 방식이 달라진다. 따라서, 막연히 언어를 통해 표출되는 정보를 파악하는 것을 넘어 정보 간 연결을 통해 전달하는 방식과 그 속에 담긴 추상적 관념을 감지하는 것은 일련의 분석 과정보다 더 풍부한 해석을 도출해낼 수 있다. 이러한 공감대는 언어학자와 더불어 추상적 관념을 기술함으로써 확장적인 결론을 도출할 수 있는 연구 분야에서도 공유되고 있으며, 자연히 관련 연구가 활발히 진행되고 있다. 본 연구는 문체 속에 담긴 언어의 가변성을 감지하고 해석하기 위한 분석 기법으로 감성 분석(Sentiment Analysis) (Pang and Lee, 2008)과 토픽 모델링(Topic modeling)을 선정하였다.

감성 분석은 텍스트 마이닝(Text Mining) 및 오피니언 마이닝(Opinion Mining)의 분석 기법으로서, 텍스트 내에 포함되는 발화자의 주관적인 감정 및 태도를 발견하고 수치와 도식으로 표현할 수 있는 분석 기법이다 (Pang and Lee, 2008). 텍스트의 주제 혹은 구성과 같은 객관적 정보 추출을 목적으로 하는 타 분석 기법에 비해, 텍스트 발화자의 성향과 태도 등 주관적인 정보를 보다 정확하게 추출할 수 있어 트위터, 페이스북, 카카오톡과 같이 감성이 높게 관찰되는 소셜 네트워크에 대한 연구에서 활용되고 있다. 또한 제품과 서비스에 대한 소비자들의 의견과 태도를 수렴하거나 콘텐츠에 대한 리뷰의 전반적인 평가를 측정하는 용도로도 사용되고 있다 (Cho et al., 2017). 감성 분석은 텍스트에 포함되어 있는 어휘들의 극성 값(Polarity)을 측정하여 감성 점수를 부여함으로써 실행될 수 있고, 이를 통해 인지적 차원에서만 머물렀던 발화자의 감정과 성향을 프로그램이 인식할 수 있는 자연 언어로 기술하고자 하는 시도가 이루어지고 있다 (Cho et al., 2016).

앞서 언급했듯이 문체 연구는 어휘와 어법의 사용과 더불어 글 속에 내재된 필자의 심적 의미를 발견하는 포괄적인 접근을 통해 이루어져야 한다. 따라서 감성 분석은 글 속에 내재된 필자의 성향과 감정을 분석할 수 있다는 점에서 큰 범위의 문체 연구에 해당한다. 감성 분석을 활용한 연구로는 은어, 비속어, 이모티콘을 포함하는 소셜 데이터의 감성 분석 연구 (Jang, 2014) 가 있고, 감성 분석과 서포트 벡터 머신(SVM)을 활용하여 악성 댓글을 탐지하고 분류하고자 하는 연구 (Hong et al., 2016) 가 있다. 하지만 감성 분석을 문체 연구의 일환으로 간주하고 기계적 분석을 진행한 연구는 찾아볼 수 없었다.

토픽 모델링은 방대한 텍스트/언어 데이터를 의미 있고 해석 가능한 언어 단위로 결합하는 텍스트 마이닝 기법으로, 텍스트에 존재하는 맥락을 단서를 통해 발견하고 유사 단어들을 군집화

함으로써 실현될 수 있다. 이러한 특성으로 인해 문서 간 정보를 분류하거나 특정 사회 이슈를 구분하고 요약하는 연구에 주로 활용된다. Kang et al. (2013) 은 언론 매체가 가지는 정파성을 오피니언으로 간주하고, 대선 이슈에 대한 매체 간 입장을 분석하기 위해 토픽 모델링을 활용하였다. Lee and Lee (2014)은 소셜 미디어 상의 짧은 텍스트 데이터의 이슈를 추적하기 위해 댓글 그래프를 이용한 토픽 모델링을 시도하였다.

본 연구는 감성 분석을 통해 구분되는 텍스트 구간 별사건의 흐름을 파악하고 맥락적 구조를 이해하기 위해 토픽 모델링 중 빈번하게 사용되는 LDA(Latent Dirichlet Allocation)를 사용한다. LDA는 문서에 포함되는 키워드들이 특정 토픽에 포함될 확률을 계산하며, 문서는 단일한 토픽이 아닌 여러 토픽으로 표현될 수 있다 (Blei et al., 2003). 이러한 방법을 통해 각 구간 별사건의 진행 양상을 파악하고 작품들과 비교해 봄으로써 작가 간 맥락적 구조를 서술하는 방식과 문체적 특성의 유무를 파악하고 정리하였다.

2.3 소설 등장인물들의 관계에 대한 연구

문체가 드러나는 텍스트 중에서도 인물의 영향을 강하게 받는 것은 소설이다. 소설은 비단 사건이 전개되는 이야기의 흐름에만 지나는 것이 아니라, 사건 발생을 일으키고 중심에 서서 갈등을 유발할 수 있는 등장인물이 필수적이다. 각 등장인물은 고유한 개성을 지니고 있으며, 이러한 개성이 이야기의 흐름과 작가 자신의 집필 방식에도 영향을 주는 경우가 빈번하다. Choi and Yoo (2014) 은 스토리의 전개 란 등장인물들이 부여된 성격과 주어진 배경에 따라 이루어 가는 담화의 연결이라 표했다. 여기서 담화 란 명확한 사건의 발단과 해결을 의미하며, 하나의 이야기를 앞뒤 문맥에 맞게 연결하는 것을 말한다. 텍스트를 접하는 독자는 이야기의 진행을 추측하고 다음에 나올 스토리에 대해 흥미를 가지고 집중하게 된다. 따라서 소설 속 등장인물이 텍스트와 작가, 독자에게 미치는 영향이 상당하다고 볼 수 있으며, 이는 곧 작가가 등장인물을 통해 드러내고자 하는 고유한 문체 해석이 가능하다는 의미를 가진다. 본 연구는 소설 속 등장인물들의 관계를 규명하고 이를 통해 작가 간 서술 방식을 비교하고자 언어 네트워크 분석을 시도하였다.

네트워크 분석은 실제 사회에서 관찰될 수 있는 인간관계의 도식적 형태를 텍스트에서 발견하고자 하는 시도로서, 텍스트 데이터에 존재하는 키워드 간 관계(네트워크)를 그래프 혹은 수치로 표현하는 분석 기법이다 (Borgatti et al., 2009). 네트워크 분석은 오랜 기간 다양한 분야에서 활용된 만큼 정의를 내리는 개념 또한 다양하지만, 대체로 언어 네트워크에서 관계의 대상이 되는 중심을 노드(node), 그들 간의 연관성을 나타내는 선형 관계를 링크(link)라고 부른다 (Borgatti et al., 2009). 실제 인간관계와 같은 연결성을 텍스트에서 발견하기 위해서는 텍스트에 존재하는 데이터들의 자질(feature)을 평가하고 그들 중 관계의 중심이 될 수 있는 키워드를 선정함으로써 가능해진다. 이러한 키워드들 간 관계에서 링크의 방향성을 고려하는가에 따라 무방향 네트워크와 유 방향 네트워크로 구분되고, 또 관계의 가중치가 계산되었는가에 따라 이진 네트워크와 계량 네트워크로 구분된다 (Knoke and Kuklinski, 1991). 네트워크 분석은 해당 특성들로 인해 키워드 간 관계를 파악하기 용이하여 소설을 비롯한 사회 이슈에서 인물들 간 관계를 파악하는 연구에서 활용되고 있다. Park et al.

(2013) 은 소설 데이터 속 등장인물들의 거리를 계산하여 사회 연결망을 분석하였다.

네트워크는 텍스트 내에서 키워드들 간 동시 발생(Co-occurrence) 정도를 계산함으로써 구현 가능하다. 본 연구는 각 소설에 등장하는 핵심 등장인물(탐정, 조력자, 형사, 피해자, 용의자)을 선정하여 그들 간 텍스트 내 동시 발생 수를 매트릭스(matrix)로 계산한 뒤, 이야기 진행에 따른 관계 변화를 그래프로 시각화하였다.

3. 연구 방법

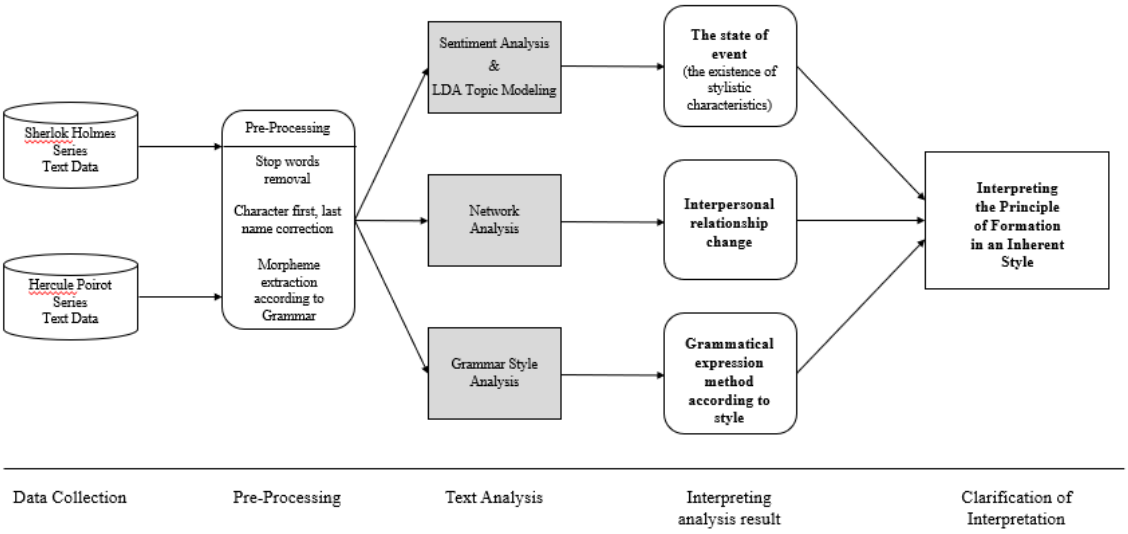
3.1 연구 문제

본 논문은 앞서 언급한 문헌 연구를 바탕으로 고전 추리 소설의 대가로 명시되는 아서 코난 도일과 애거서 크리스티의 작가 고유 문체를 판별하고 비교함으로써 문체의 종합적인 접근을 시도하였다. 작가 고유 문체를 판별하기 위해서는 동일 작가의 작품들 간 문체적 특성에서 공통점을 발견할 수 있어야 하며, 이러한 공통성이 타 작가와 차별성을 가져야 한다. 따라서 두 추리 소설 작가의 공통된 추리 소설 장르를 선택하여 분석하되, 총 세 가지 문체 접근법을 적용하여 스토리 진행 양상, 인물 관계 변화, 문법적 서술 방식의 차이점을 명시하였다. 두 작가의 공통된 장르로 '탐정 소설'을 선정하였으며, 비교의 통일성을 주고자 탐정 소설 시리즈의 데뷔작과 후작을 분석 대상으로 정했다. 이에 따라 아서 코난 도일은 셜록 홈즈 시리즈의 <주홍색 연구>와 <네 개의 서명>을, 애거서 크리스티는 에르퀼 푸아로 시리즈의 <스타일스 저택의 괴사건> 과 <골프장 살인 사건>을 분석 대상으로 선정하였다.

(연구 문제)

- 작가 간 이야기를 서술하는 방식은 무엇이며 어떠한 차이가 있는가?
- 작가 간 이야기 진행에 따른 인물 관계 변화를 어떻게 서술하는가, 그리고 어떠한 차이가 있는가?
- 작가 간 서술 상의 문법 체계는 어떠하며 구체적인 차이는 무엇인가?

문체 연구를 위해 두 고전 추리 소설 작가와 탐정 소설 장르를 선택한 이유는 다음과 같다. 고전 추리 소설 장르로 대표되는 탐정 소설은 사건을 해결하는 탐정과 이를 도와주는 조력자, 사건 해결에 문제를 일으키는 형사와 사건을 일으킨 범인, 그리고 범인에 의해 살해당한 피해자로 역할이 뚜렷하게 구분된다. 작가는 등장인물의 성격과 사건의 특성을 고려하여 이야기를 서술하기 때문에 동일한 시리즈는 작품 고유의 클리셰의 영향을 받을 확률이 높다. 이는 곧 독자가 직감적으로 인지하는 문체에 해당한다. 또한 추리 소설은 사건이 진행되는 텍스트의 층위 속에서 사건의 단서와 흔적을 남기며, 독자는 이를 발견하고 스스로 해석하는 묘미를 느끼게 된다. 결과적으로 독자가 특정 작가의 추리 소설을 지속적으로 접할 때, 단서의 출현과 흔적의 표현을 파악하는 것만으로도 범인이 누구인지 예측할 수 있게 된다. 이 또한 인지적 측면에서 문체의 발견으로 해석할 수 있다. 이러한 문체적 의의를 파악하기 위해 탐정 소설을 분석 대상으로 선정하였으며, 현대 추리 소설의 배경이자 대다수의 사람들이 친숙하게 접한 아서 코난 도일과 애거서 크리스티의 작품을 분석함으로써 해당 연구의 이해를 돕고자 하였다.



〈Figure 1〉 A research model of Text mining for detective stories' writing styles

3.2 연구 방법

전체적인 연구 과정은 <Figure 1>과 같이 진행한다. 두 작가의 추리 소설 시리즈를 분석하기 위해 각자 전처리를 수행한 뒤 감성 분석, 토픽 모델링, 네트워크 분석, 문법 체계 분석을 시도하여 텍스트를 통해 드러나는 언어 표현법을 관찰하였다. 그리고 각 표현법을 통해 제시될 수 있는 문체적 특성과 독자에게 전달 가능한 작가의 성향을 파악하고 해석함으로써 최종적으로 어떠한 내재적 문체 형성에 기여했는지를 확인하였다.

3.3 데이터 수집 및 전처리

3.3.1 데이터 수집

두 작가의 작품을 분석하기 위해 고전 소설을 전자책 형태로 무료로 제공하는 manybooks.net 사이트를 이용하였다. Manybooks는 약 5만 개의 무료 소설을 전자책을 비롯한 pdf, txt 파일 형태로 제공하고 있으며 전 세계 약 15만 명이 이용하는 도서 제공 사이트이다. 원하는 도서를 쉽게 검색할 수 있어 손쉽게 데이터에 접근할 수 있었다. 본 사이트에서 앞서 언급한 4개의 작품을 txt 파일로 다운로드하여 분석을 진행하였다.

3.3.2 데이터 전처리

분석 과정에서 중복적으로 계산될 것을 고려하여 공통적으로 등장인물들의 성과 이름을 하나의 이름으로 통일하는 과정을 거쳤다. 하지만 이후 분석 방법에 따라 전처리 방식을 달리하였는데, 우선 문법적 체계를 확인하기 위한 분석 방법의 경우 기준이 되는 전체 문장과 문법 체계를 구성하는 문장의 전처리 방식을 구분하였다. 기준이 되는 전체 문장은 영어 텍스트에 해당하지 않는 부분을 삭제하고 문자를 소문자로 전환하는 방식으로 전처리를 진행했다. 반면 문법 체

계를 구성하는 문장은 앞선 전처리 방식에 더해 각 문장을 형태소 단위로 구분하여 해당 문법 체계와 일치하는 품사만 남기고 전부 삭제하는 방식으로 전처리를 진행하였다. 이렇게 구분한 이유는 문법 체계를 구성하는 문장을 기존 텍스트에서 구분할 때 텍스트에서 가지는 인덱스를 기준으로 구분하기 때문이며 보다 정확하게 문법 체계를 구분하기 위함이다. 감성 분석을 실시하는 경우 보다 정확한 감성 판별을 위해 기준이 되는 전체 문장을 전처리 하는 경우와 동일한 전처리 방식을 수행하였다. 반면 토픽 모델링을 실시하는 동안 분명한 결과 도출을 위해 불용어를 제거하고 단어들의 원형을 추출했으며, 형태소 분석을 통해 명사, 동사, 형용사에 해당하는 품사만 문장이 포함할 수 있도록 하였다. 언어 네트워크 또한 키워드 도출로 인해 전체 문장과 마찬가지로 영어 텍스트에 해당하지 않는 부분을 지우고 소문자로 전환하는 방식으로 전처리를 수행하였다.

3.4 데이터 분석

3.4.1 감성 분석과 토픽 모델링을 이용한 소설 전개 관찰

본 연구에서는 두 작가의 이야기 전개 방식을 이해하고 차이점을 규명하기 위해 파이썬 영문 자연 언어 처리 패키지인 NLTK의 'SentimentIntensityAnalyzer'와 'vader_lexicon' 감성 사전을 이용하여 감성 분석을 실시하였으며 LDA(Latent Dirichlet Allocation) 토픽 모델링을 통해 감성이 변화하는 각 구간의 전개 내용을 확인하였다. 각 문장의 극성 값을 판단하여 긍정, 부정, 중립 총 세 가지 카테고리로 분류했으며, 긍정 카테고리 문장은 1점, 부정 카테고리 문장은 -1점, 중립 카테고리 문장은 0점으로 환산하여 점수를 누적해 각 페이지가 진행되는 동안 감정의 변화를 그래프로 시각화하였다. 본 그래프를 통해 페이지 진행에 따른 감정의 변화를 파악할 수 있었으며 사건 진행 양상이 변화하는 구간을 확인할 수 있었다. 이후 감정이 변화하는 구간 혹은 페이지의 일정 범위를 나누어 LDA 토픽 모델링을 적용해 사건 진행에 따른 핵심 키워드를 도출함으로써 사건 진행을 유추하였다.

3.4.2 네트워크 분석을 이용한 등장인물 관계 변화 관찰

소설의 핵심이 되는 스토리는 등장인물들 간 상호 작용과 갈등을 통해 형성될 수 있다. 범인은 피해자와의 갈등으로 인해 살인과 같은 사건을 일으키게 되고, 사건 발생에 따른 형사 간의 갈등은 탐정의 역할을 요구하게 된다. 탐정이 사건에 참여함으로써 이야기의 갈등은 첨예하게 변모하고 인물들 간에 벌어지는 갈등과 과거에 얽힌 사연을 탐정이 해석함으로써 사건이 마무리된다. 소설 속에서 이어지는 일련의 스토리는 등장인물들의 참여로 이루어지는 것이며, 그들의 관계를 파악하는 것만으로도 하나의 이야기가 형성될 수 있다. 따라서 본 연구는 앞서 구분된 사건 구간에 따른 등장인물 간 관계 변화를 파악하기 위해 네트워크 분석을 수행하였다. 각 소설에 등장하는 주요 등장인물들을 정리하여 구간에 따른 동시 발생 매트릭스를 구축한 뒤, 가장 많은 등장 빈도수를 차지하는 인물을 기준으로 빈도수를 나누어 가중치를 설정하였다. 이를 바탕으로 언어 네트워크 그래프를 형성함으로써 각 사건에 따른 인물 관계 변화를 확인할 수 있었다.

3.4.3 문법 체계의 관찰

텍스트의 문체를 구분하는 가장 기본적인 접근법은 텍스트에서 사용된 어휘와 어순 혹은 어법의 체계를 판단하고 분류하는 것이다. 정보 전달을 원칙으로 하는 언어의 특성상 어법에 대한 일정 체계가 갖추어지기 마련이며, 이를 반영하는 텍스트 역시 어법에 영향을 받는다. 따라서 텍스트에 담긴 어법을 파악하고 분류하는 것은 기본적인 문체 구분 활동에 구조를 제공할 수 있다. 이러한 표현되는 어휘와 어순, 혹은 어법을 반영하여 분류한 사례로는 국내 문체로 정의된 구어체, 문어체, 화려체, 건조체, 간결체, 만연체, 강건체, 우유체 등이 있으며, 해외 사례로는 Lexical(어휘 기반), Syntactic(구문론적), Structural (구조적), Content-specific(내용 중심적)이 있다. 다만, 이는 전반적인 문체를 표현하기 위한 구분에 지나지 않으며 이러한 문체들이 복합적으로 사용되거나 혹은 사례에 해당하지 않는 문체가 있는 등 문체가 가지는 범위는 방대하다.

본 연구는 종속 접속사와 등위 접속사의 사용 유무, 일정 기준 이상의 형용사적 표현, 부사 표현(to 부정사, 분사, 관계사, 전치사 구) 유무, 감탄사의 유무, 비유적 표현의 유무, 그리고 숙어 표현의 유무에 따라 구어체, 문어체, 화려체, 건조체, 간결체, 만연체를 구분하여 문법적 체계를 정리했다. 각 문법적 표현들의 어순에 따른 역할 구분을 위해 형태소 분석을 수행한 뒤 각 형태소들의 인덱스 값을 활용하여 조건문을 설정하였으며 전체 문장을 기준으로 해당 문법 체계를 가진 문장들의 비율을 계산하여 작가 간 문법 체계의 차이점을 정리하였다. <Table 1>은 문체 종류에 대한 정의와 문법적 체계에 대한 설명을 보여주고 있다.

- 접속사의 사용: 접속사의 사용 유무에 따라 문장은 단문, 중문, 복문으로 구분된다. 이때 종속 접속사를 사용하면 복문, 등위 접속사를 사용하면 중문으로 구분된다. 접속사의 사용 유무에 따라 간결체와 만연체를 구분하였고, 종속 접속사와 등위 접속사의 사용을 구분하여 구어체와 문어체를 나누었다. 종속 접속사를 빈번하게 사용하는 경우 문어체로 분류되며, 등위 접속사를 빈번하게 사용하는 경우 구어체로 분류될 수 있다.
- 형용사와 부사의 사용: 화려체와 건조체를 구분하는 기준으로 형용사와 부사의 사용을 고려하였다. 단, 같은 형용사라고 해도 문장에

〈Table 1〉 Type of writing style[1)]

Style	Meaning
Informal	a style in which expressions, such as those encountered in everyday conversation, are expressed in writing.
Literary	a style that is not often used in modern painting but can be expressed in writing.
Gorgeousness	a style that enhances the expressiveness of a sentence by using various modifiers.
Dry	a style that does not use a particular modifier.
Diffusion	a style that avoids compression and omission and uses many phrases to describe and detail the sentence itself.
Brevity	a style that minimizes and compresses the expression of a sentence.

1) 네이버 지식 백과 참조

서 사용되는 성분에 따라 필수 성분과 부속 성분이 구분될 수 있어, 형용사의 경우 한 문장에 3회 이상 사용한 경우, 부사의 경우 한 문장에 2회 이상 사용한 경우로 화려체와 간결체를 구분하였다.

- 형용사적 표현의 사용: 부정사, 분사의 사용, 그리고 문장에서 형용사의 역할을 할 수 있는 관계사와 전치사 구를 통해 화려체와 건조체를 구분하였다. 부정사의 경우 어순상 다음에 동사 원형이 오고, 앞선 어순에서 명사가 오는 To 부정사를 형용사적 표현으로 간주하였고, 분사의 경우 동사의 과거 분사와 구분하기 위해 앞선 어순이 동사 품사에 해당하지 않는 경우에 한해 형용사적 표현으로 간주하였다. 관계사는 앞선 어순에 명사가 오고, 다음 어순에서 명사가 제시되지 않을 경우 관계사의 형용사적 표현으로 간주하였다. 끝으로 빈번하게 사용되는 전치사를 정리한 뒤 이에 해당하며 앞선 어순에서 명사가 오는 경우 전치사 구의 형용사적 표현으로 간주하였다.
- 해당 문장이 숙어 표현을 포함하는 경우에 따라 구어체와 문어체로 구분하였다. 숙어 표현을 포함하는 경우 구어체의 성격을 강하게 띄는 것으로 간주하였다.
- 해당 문장이 비유적 표현을 포함하는 경우에 따라 구어체와 문어체를 구분하였다. 비유적 표현을 사용하는 경우 구어체의 성격을 강하게 띄는 것으로 간주하였다.

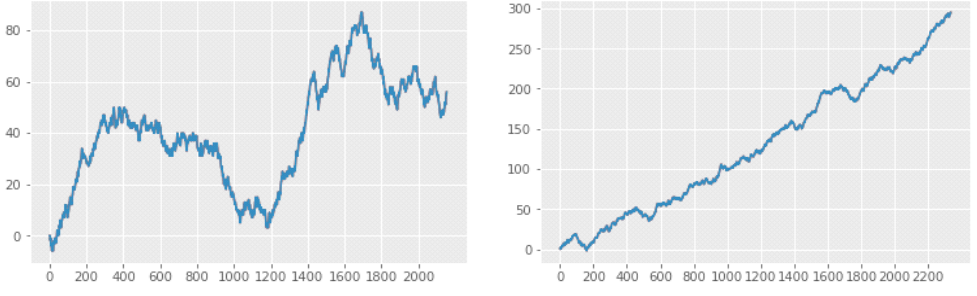
4. 연구 결과

문체 구조에 대한 세 가지 분석 방향을 설계하고 앞서 제시한 연구 문제를 확인할 수 있었다. 셜록 홈즈와 에르퀼 푸아로 시리즈의 사건 진행 양상을 감성 분석과 토픽 모델링의 비교를 통해 파악할 수 있었고 실제 사건과 이를 표현하는 표현법 간 차이를 발견함으로써 작가 간 문체적 특성이 존재하고 있음을 확인하였다. 또한 구간 변화에 따른 인물 관계 형성 방식과 글의 문법적 표현 방식을 분석함으로써 작가의 문체적 특성을 보다 구체적으로 정의하고 어떠한 내재적 특성에 기여하는지 해석할 수 있었다.

4.1 각 소설 사건 진행 양상

4.1.1 셜록 홈즈 시리즈

셜록 홈즈 시리즈의 두 작품을 문장 단위로 감성 점수를 부여한 뒤, 200문장(약 10페이지, 전체



〈Figure 2〉 Sherlok Holmes series sentiment analysis (Left: A Study in Scarlet, Right: The Sign of Four)

〈Table 2〉 〈A Study in Scarlet〉 Topics for each section

<A Study in Scarlet> Topic modeling	
First section (Page 0~20) **Daily Reasoning**	**Topic#1**: holmes, thing, test, stamford, round, hotel, london, night **Topic#2**: life, street, study, expression, chemical, laboratory, idea, theory **Topic#3**: case, crime, table, fact, laugh, style, book, earth **Topic#4**: house, morning, deduction, detail, class, trade, glance, article **Topic#5**: room, nothing, friend, question, work, point, month, hour **Topic#6**: knowledge, time, fellow, wall, marine, ground, detective, difficulty **Topic#7**: blood, doubt, stain, finger, face, something, smile, road **Topic#8**: hand, companion, thought, mind, matter, door, observation, train
Second section (Page 20~40) **A sign of case**	**Topic#1**: ring, door, hand, affair, mystery, nothing, matter, mind **Topic#2**: holmes, street, floor, doubt, foot, part, anyone, window **Topic#3**: woman, name, house, pocket, address, gate, sign, victim **Topic#4**: time, gregson, face, lestrade, rance, murder, body, dust **Topic#5**: wall, anything, word, voice, watson, finger, morning, glass **Topic#6**: companion, paper, fact, thing, manner, table, head, account **Topic#7**: case, night, gold, brixton, road, idea, apartment, detective **Topic#8**: room, round, blood, corner, feature, murderer, square, death
Third section (Page 40~60) **Case Investigation**	**Topic#1**: gregson, water, head, house, body, road, plain, yard **Topic#2**: time, drebber, death, word, door, exertion, boot, notice **Topic#3**: room, hand, charpentier, thing, daughter, chair, something, moment **Topic#4**: nothing, street, matter, gentleman, half, glass, brixton, pound **Topic#5**: holmes, pill, silence, voice, window, grey, work, river **Topic#6**: face, mother, course, detective, girl, year, statement, track **Topic#7**: lestrade, case, mind, instant, effect, affair, police, part **Topic#8**: stangerson, secretary, companion, crime, name, question, clock, morning
Fourth section (Page 60~80) **Case Reasoning**	**Topic#1**: face, daughter, ferrier, door, time, house, guess, farm **Topic#2**: hand, father, mormon, night, waggon, road, prophet, whip **Topic#3**: heart, stangerson, mountain, voice, name, fear, matter, course **Topic#4**: elder, year, companion, wanderer, drebber, horse, moment, hope **Topic#5**: lucy, head, wife, city, others, creature, religion, brother **Topic#6**: child, woman, none, farmer, gate, crowd, mind, nothing **Topic#7**: jefferson, girl, sight, horse, rock, faith, room, friend **Topic#8**: morning, land, plain, hand, foot, size, party, nevada
Last section (Page 80 ~) **Case Resolution**	**Topic#1**: jefferson, life, mountain, night, city, matter, rock, rifle **Topic#2**: drebber, hunter, death, head, thing, people, door, fire **Topic#3**: hour, murder, result, chance, side, window, blood, place **Topic#4**: time, stangerson, prisoner, ring, room, reason, holmes, mormon **Topic#5**: hand, heart, track, lucy, lestrade, foot, anything, poison **Topic#6**: house, horse, moment, mind, street, father, ravine, question **Topic#7**: face, point, road, pill, hotel, friend, station, step **Topic#8**: word, case, justice, work, police, murderer, garden, danger

〈Table 3〉 〈The Sign of Four〉 Topics for each section

<The Sign of Four> Topic modeling	
First section (Page 0~20) **Case Request**	**Topic#1**: time, father, quality, power, pocket, knowledge, business, earth **Topic#2**: case, watch, work, point, night, observation, london, sholto **Topic#3**: companion, manner, place, part, book, window, habit, word **Topic#4**: year, letter, friend, woman, morning, captain, officer, home **Topic#5**: matter, chair, mystery, mind, something, face, voice, feature **Topic#6**: morstan, hand, address, line, corner, side, pipe, experience **Topic#7**: holmes, street, fact, name, paper, brother, pearl, cecil **Topic#8**: house, anything, head, character, detective, cocaine, doctor, value
Second section (Page 20~40) **Case Occurrence**	**Topic#1**: father, word, morstan, something, sholto, case, year, paper **Topic#2**: face, window, side, thaaddeus, table, lamp, share, glass **Topic#3**: night, matter, time, place, floor, trace, sign, garden **Topic#4**: door, house, head, morstan, moment, ground, lodge, wall **Topic#5**: holmes, treasure, foot, chaplet, corner, work, fact, instant **Topic#6**: thing, thaddeus, heart, order, hour, woman, master, stair **Topic#7**: room, friend, brother, round, hole, part, roof, step **Topic#8**: hand, bartholomew, light, voice, mark, rope, skin, fashion
Third section (Page 40~60) **Case Investigation**	**Topic#1**: side, door, voice, corner, course, night, word, forrester **Topic#2**: step, room, jones, matter, force, presence, athelney, reason **Topic#3**: time, sholto, associate, look, point, stair, sergeant, difficulty **Topic#4**: treasure, face, ground, death, chart, theory, nothing, mean **Topic#5**: toby, place, house, window, fact, mind, sherman, wall **Topic#6**: holmes, hand, case, foot, street, barrel, water, thing **Topic#7**: scent, thaddeus, round, road, print, luck, hour, london **Topic#8**: name, business, smell, creasote, brother, morstan, head, sign
Fourth section (Page 60~80) **Case Reasoning**	**Topic#1**: holmes, voice, watson, treasure, something, thing, head, violin **Topic#2**: matter, anything, friend, message, companion, aurora, work, jonathan **Topic#3**: jones, street, baker, thought, forrester, stair, foot, word **Topic#4**: news, moment, sholto, breakfast, side, people, paper, clue **Topic#5**: time, night, nothing, mind, norwood, wiggins, sound, cigar **Topic#6**: face, hand, line, business, doubt, wooden, lady, fellow **Topic#7**: case, river, police, hour, athelney, thaaddeus, course, well **Topic#8**: launch, smith, door, boat, room, steam, clock, wharf
Last section (Page 80 ~) **Case Resolution**	**Topic#1**: thing, heart, story, something, rope, fellow, blood, sign **Topic#2**: side, friend, matter, front, voice, akbar, abdullah, justice **Topic#3**: fort, tonga, word, merchant, thought, jewel, name, iron **Topic#4**: holmes, sikh, face, chance, business, guard, doubt, gate **Topic#5**: time, sholto, boat, morstan, head, convict, country, troop **Topic#6**: place, hand, nothing, water, room, anything, part, death **Topic#7**: jones, life, round, moment, launch, aurora, course, yard **Topic#8**: night, treasure, work, light, half, river, agra, hour

텍스트의 10분의 1)을 기준으로 x 축의 범위를 산정해 시각화를 하였다. 이후 감정 변화가 나타나는 각 구간에 따른 사건 진행을 토픽 모델링으로 정리하고 사건 진행 전반을 유추하였다.

셜록 홈즈 시리즈의 두 작품에 토픽 모델링을 적용한 결과 시점은 다를 수 있으나 고정된 역할의 등장인물들이 동일한 순서로 등장한다는 것을 알 수 있었다. <주홍색 연구>의 토픽 모델링 결과인 <Table 2>를 살펴보면 주인공이자 탐정 역할의 'holmes'가 첫 번째 구간에서부터 등장하는 것을 확인할 수 있다. 뒤이어 형사 역할의 'lestrade'와 'gregson'이 두 번째 구간에서 등장하였고, 피해자 역할의 'drebber'와 제1용의자로 지목되었던 'charpentier', 사건 관계자인 'stangerson'과 형사 역할의 캐릭터가 세 번째 구간에서 동시에 등장하고 있음을 확인할 수 있었다. 그리고 또 다른 사건 관계자로 'ferrior'와 'lucy'가, 범인 역할의 'jefferson'이 네 번째 구간에서 등장하고 있음을 알 수 있었으며, 이후 사건 속 직접적인 갈등 관계에 놓였던 인물들('jefferson', 'lucy', 'stangerson', 'drebber')과 탐정이 마지막 구간에서 함께 등장하는 것으로 확인되었다. 이를 통해 <주홍색 연구>는 **탐정 등장(일상) -> 형사 등장(사건 발생) -> 피해자, 관계자, 제1용의자 발생(사건의 심화, 형사의 잘못된 추리) -> 추가 관계자, 범인 등장(사건 전말의 폭로) -> 갈등 관계에 놓였던 인물들과 탐정의 등장(사건의 해결)**과 같은 순서로 사건이 진행되고 있음을 유추해볼 수 있다.

이러한 유추는 <네 개의 서명>에서도 동일하게 적용해 볼 수 있다. <Table 3> 역시 인물들이 등장하는 구간은 다르나 순서가 동일하다는 것을 알 수 있는데, 비록 본 작의 첫 번째 구간에서 전작의 첫 번째 구간과 달리 'sholto'나 'morston'과 같은 인물 명사가 제시되거나, 전작에서 피해자와 제1용의자, 관계자가 세 번째 구간에서 등장한 것에 비해 본 작에서는 두 번째 구간에서 등장하는 등 사건의 내용에 따라 인물의 출현 시점이 다를 수 있으나 **탐정의 등장 -> 피해자(batholomew), 제1용의자(Thaddeus), 관계자(morstan) 등장 -> 범인(jornathan) 등장 -> 추가 관계자(Norwood, tonga)** 등장 순으로 <주홍색 연구>와 동일한 인물 등장 패턴을 보이고 있음을 알 수 있다. 따라서 <네 개의 서명> 역시 <주홍색 연구>와 유사한 사건 진행 흐름을 가지고 있음을 유추할 수 있다.

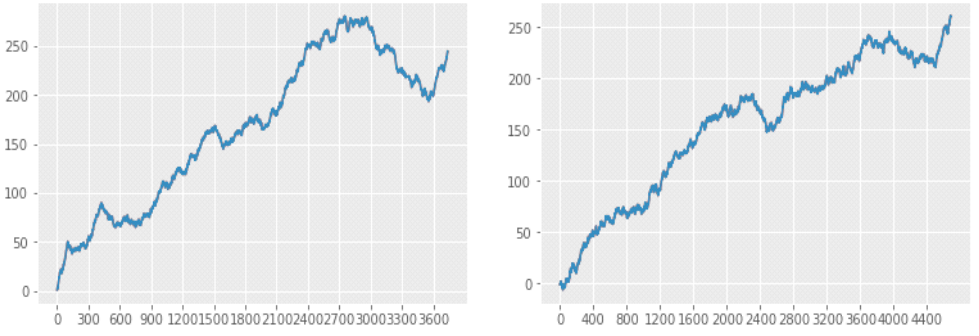
그러나 두 작품의 사건을 서술하는 감정 표현이 서로 상이한 것을 볼 수 있는데, <Figure 2>에서 드러나는 <주홍색 연구>의 감정 표현은 사건 진행의 흐름에 따라 굴곡을 보이고 있는 반면 <네 개의 서명>은 큰 굴곡 없이 상향하는 모습을 확인할 수 있었다. 유사한 사건 진행을 보이고 있음에도 이러한 감성 표현에서 차이가 발생하는 이유는 앞서 언급했던 문체의 '가변적 특성'이 작용한 결과로 해석할 수 있다. 실제로 토픽 모델링 결과 특정 감정을 표현하는 형용사를 찾기 어려웠으며 동시에 동사와 명사를 위주로 사용하여 사건을 묘사하는 모습이 두드러지게 보였다. 이는 특정 인물의 심리와 그들 간 갈등에 집중하기보다 상황을 묘사하는 방식을 주로 사용하는 특성이 있는 것으로서, 결국 유사한 사건 정보라 하더라도 이를 전달하는 방식이 다를 수 있음을, 그리고 이러한 차이를 통해 문체적 특성이 존재하고 있음을 유추해 볼 수 있다.

4.1.2 에르퀼 푸아로 시리즈

아서 코난 도일의 두 작품을 비교하였듯이 애거서 크리스티의 두 작품 역시 감성 분석과 토픽 모델링을 실시하였으며 두 작품 간 사건 진행 양상과 이야기 전개를 비교하였다.

애거서 크리스티의 에르퀼 푸아로 시리즈는 셜록 홈즈 시리즈와는 정반대되는 양상을 띄고 있는데, 두 작품 모두 초반부터 후반까지 대다수의 인물들이 등장하고 있으며 'piece', 'love', 'jealousy', 'happiness', 'kind', 'favour' 등과 같이 인물의 주관적인 정서나 감정을 표현하는 형용사가 사용되었음을 확인하였다. <Table 4>와 <Table 5> 모두 공통적으로 다수의 인물 명사를 전 구간에서 제시하고 있는 모습을 보이면서 사건을 유추하기 어려웠으나, <Figure 3>에서 확인할 수 있는 사건을 서술하는 감정 표현은 두 작품 간 서로 유사한 형태를 띠고 있다는 것을 알 수 있었다. 이는 셜록 홈즈 시리즈와 대비되는 문체적 특성이 반영된 결과로 해석할 수 있다. 애거서 크리스티는 아서 코난 도일에 비해 인물의 심리와 성향, 그들 간 갈등을 보다 세밀하게 묘사하는 서술 방식을 선호한 것으로 보인다. 핵심적인 사건 진행을 이루는 한편 그러한 사건 속에서 벌어지는 인물들의 이야기에 집중하는 모습을 보임으로써 독자로 하여금 사건 진행에 집중시키기보다 인물들의 시선에 몰입하도록 하였다. 이러한 특성으로 인해 인물의 심리를 표현하는 감정 형용사가 사용되었으며 셜록 홈즈 시리즈와 달리 사건 진행이 분명하지 않으나 공통된 감정 표현 방식이 사용된 것으로 보인다. 이 역시 크리스티의 고유한 문체적 특성이 존재하고 있음을 알 수 있는 부분이다.

그러나 사건의 특성에 따라 이를 표현하는 방식에 차이가 있음을 토픽 모델링을 통해 확인할 수 있는데, <Table 4>의 전 구간에서 'coffee'라는 명사가 공통적으로 관측됨으로써 이를 사건의 단서로 유추해 볼 수 있는 것에 반해 <Table 5>에서는 이러한 전 구간에서 공통적으로 발견할 수 있는 특정 명사를 발견하기 어려웠다. 또한 <Table 4>에서는 'favour', 'happiness', 'jealousy', 'love'등 다수의 감정 형용사를 발견할 수 있었지만 <Table 5>에서는 'interest', 'love'와 같이 감정을 표현하는 형용사가 보다 적게 사용된 것을 확인할 수 있었다.



〈Figure 3〉 Hercule Poirot series sentiment analysis
(Left: The Mysterious Affair at Styles, Right: The Murder on the Links)

〈Table 4〉 〈The Mysterious Affair at Styles〉 Topics for each section

<The Mysterious Affair at Styles> Topic modeling	
First section (Page 0~75) **Daily Life**	**Topic#1**: window, mother, cavendish, strychnine, clock, lady, kind, home **Topic#2**: hand, face, evelyn, head, woman, idea, manner, wilkins **Topic#3**: hastings, word, yesterday, fact, afternoon, wife, life, piece **Topic#4**: poirot, something, lawrence, voice, bauerstein, thing, question, hall **Topic#5**: john, moment, dorcas, paper, matter, boudoir, doctor, morning **Topic#6**: time, cynthia, case, coffee, mind, minute, letter, despatch **Topic#7**: inglethorp, friend, nothing, course, style, cocoa, tray, annie **Topic#8**: room, door, night, house, year, place, mistress, table
Second section (Page 75~135) **Case Occurrence / Investigation**	**Topic#1**: john, hall, poison, anything, anyone, arrest, gentleman, woman **Topic#2**: evelyn, time, idea, head, lawrence, death, miss, matter **Topic#3**: inglethorp, bauerstein, mind, murder, point, possibility, name, night **Topic#4**: strychnine, coffee, morning, mary, clock, hour, truth, cavendish **Topic#5**: hand, face, japp, something, house, case, village, question **Topic#6**: nothing, friend, evidence, moment, room, door, hastings, doctor **Topic#7**: coroner, course, voice, word, fact, wife, paper, afternoon **Topic#8**: poirot, thing, cynthia, dorcas, style, minute, instinct, crime
Third section (Page 135~150) **Case Reasoning**	**Topic#1**: finger, inglethorp, mark, anything, nothing, voice, smile, favour **Topic#2**: poirot, head, friend, moment, monsieur, link, proof, fact **Topic#3**: case, strychnine, kind, bauerstein, photograph, husband, room, doubt **Topic#4**: evidence, cynthia, defence, prosecution, wife, mean, court, police **Topic#5**: prisoner, lady, poison, question, afternoon, murder, witness, wardrobe **Topic#6**: lawrence, john, time, father, style, minute, hastings, something **Topic#7**: woman, thing, mother, parcel, hand, happiness, matter, part **Topic#8**: morning, mary, pride, jealousy, japp, forehead, discovery, coffee
Last section (Page 150 ~) **Case Resolution**	**Topic#1**: fact, prisoner, strychnine, bottle, drawer, powder, cross, book **Topic#2**:inglethorp, room, friend, cynthia, cavendish, quarrel, mademoiselle, something **Topic#3**: evelyn, room, door, house, night, table, miss, hour **Topic#4**: hand, nothing, suspicion, note, mind, matter, chemist, desk **Topic#5**: poirot, letter, crime, question, thing, person, clock, dorcas **Topic#6**: evidence, lawrence, style, monsieur, medicine, love, piece, brother **Topic#7**: paper, case, husband, bromide, mother, word, moment, sli **Topic#8**: john, time, idea, coffee, minute, wife, head, fragment

〈Table 5〉 〈The Murder on the Links〉 Topics for each section

<The Murder on the Links> Topic modeling	
First section (Page 0~60) **Case Occurrence**	**Topic#1**: poirot, moment, girl, villa, question, course, genevi, window **Topic#2**: monsieur, time, night, friend, crime, life, fact, method **Topic#3**: door, husband, room, house, point, hour, finger, people **Topic#4**: renauld, hand, commissary, hautet, woman, daubreuil, shoulder, marchaud **Topic#5**: head, doubt, face, thing, arrichet, lady, police, side **Topic#6**: magistrate, something, giraud, interest, idea, voice, piece, sister **Topic#7**: case, madame, letter, place, body, work, ground, money **Topic#8**: nothing, minute, morning, paper, detective, murder, footprint, kind
Second section (Page 60~100) **Case Investigation**	**Topic#1**: poirot, time, fact, surprise, word, madame, watch, morning **Topic#2**: hautet, nothing, moment, idea, lady, marchaud, year, husband **Topic#3**: head, gabriel, face, jack, hand, name, matter, table **Topic#4**: father, course, mother, mystery, bien, importance, term, half **Topic#5**: renauld, magistrate, juge, case, life, daubreuil, detective, question **Topic#6**: giraud, thing, voice, woman, match, letter, chair, cigarette **Topic#7**: friend, minute, room, villa, shoulder, round, back, spot **Topic#8**: door, girl, crime, place, house, hour, smile, body
Third section (Page 100~140) **Case Resolution**	**Topic#1**: renauld, hand, word, moment, voice, face, mind, mystery **Topic#2**: poirot, woman, heart, body, hair, villa, doctor, affair **Topic#3**: giraud, lady, story, monsieur, question, reason, name, marthe **Topic#4**: murder, girl, husband, interest, george, part, death, news **Topic#5**: friend, thing, case, course, detail, thought, police, paper **Topic#6**: time, dagger, train, hour, idea, jack, minute, magistrate **Topic#7**: crime, nothing, fact, hotel, door, hastings, murderer, anything **Topic#8**: beroldy, madame, head, marchaud, doubt, life, matter, look
Last section (Page 140 ~) **Case Resolution**	**Topic#1**: case, dagger, course, letter, morning, story, truth, train **Topic#2**: moment, friend, point, doubt, magistrate, bien, daubreuil, money **Topic#3**: jack, head, night, tramp, idea, father, question, name **Topic#4**: renauld, crime, madame, body, murder, house, gabriel, anything **Topic#5**: poirot, time, george, something, love, room, manner, foot **Topic#6**: face, girl, word, fact, bella, nothing, marthe, hair **Topic#7**: woman, poirot, hastings, villa, station, shoulder, door, child **Topic#8**: giraud, thing, hand, voice, mind, mademoiselle, mistake, chance

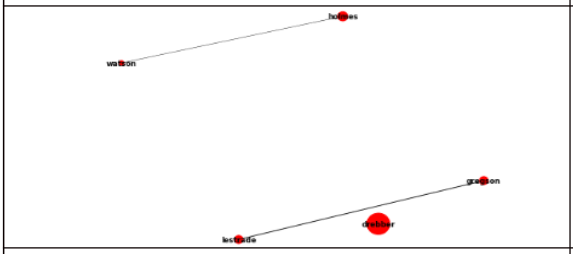
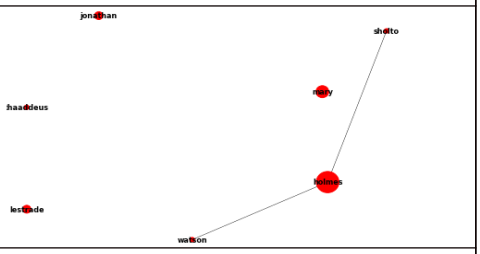
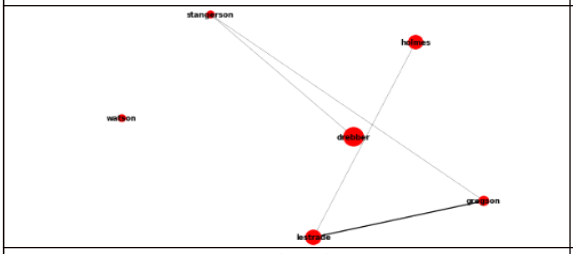
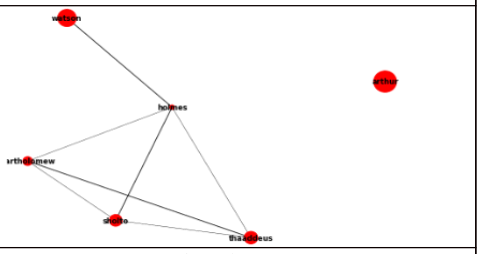
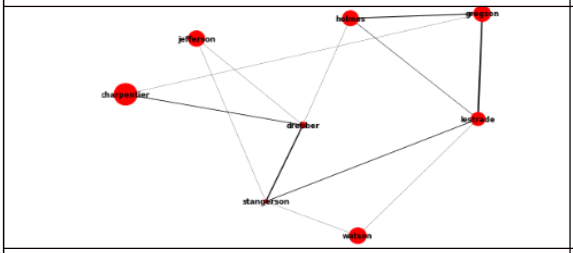
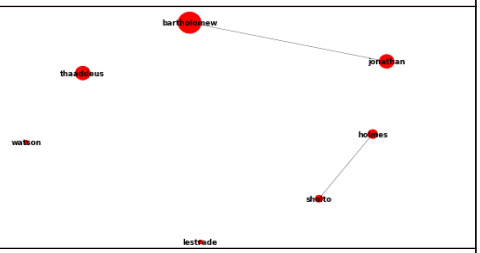
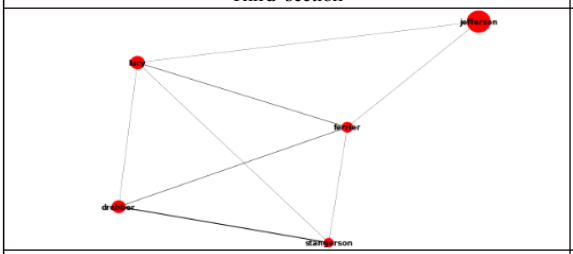
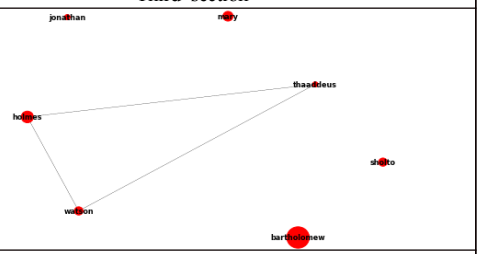
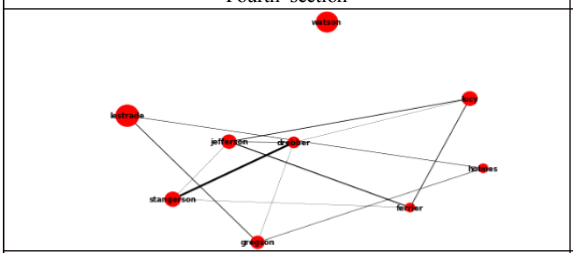
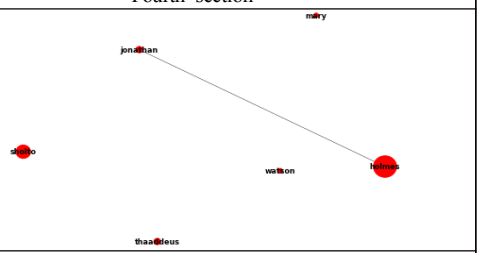
4.2 각 구간별 인물 관계 변화 양상

토픽 모델링을 통해 구분되는 각 구간에 대한 인물 관계 변화를 관찰하기 위해 각 구간별 등장인물의 네트워크 분석을 시도하였고 <Table 6>와 같이 도출된 결과를 확인하였다. 분석 결과 토픽 모델링과 마찬가지로 시점은 다를 수 있으나 고정된 역할의 등장인물들이 등장하는 순서가 동일한 것을 확인할 수 있는 반면, 두 작품 간 인물 관계 형성이 다를 수 있음을 확인하였다. 예를 들어 <주홍색 연구>는 시점이 변함에 따라 인물 관계가 좀 더 복잡해지고 인물들 간 얽혀 있는 관계의 심화 정도를 링크의 굵기를 통해 확인할 수 있지만, <네 개의 서명>은 사건이 진행됨에 따라 인물 관계가 복잡해지지 않으며 연결 또한 전작과 동일하지 않았다. 탐정이 등장하는 구간도 다른 것을 확인할 수 있으며 특히 마지막 구간에서 <주홍색 연구>는 범인 역할의 'jefferson'과 'holmes'가 간접적으로 연결되어 있는 것에 비해 <네 개의 서명>은 'holmes'와 범인 'jonarthan'이 직접 연결되어 있는 것을 확인할 수 있다. 이를 통해 두 사건은 등장하는 인물들의 순서가 고정되어 있어 사건이 진행되는 클리셰 혹은 패턴이 유사하지만 사건 자체가 가지는 내용은 다를 수 있다고 짐작할 수 있다. 실제로 두 작품은 탐정에게 사건이 의뢰되는 방식에서부터 범인의 등장과 사건 갈등이 상이한 것으로 알려져 있다.

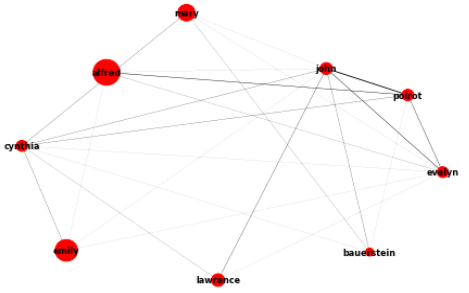
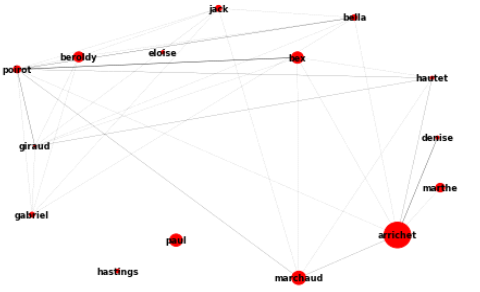
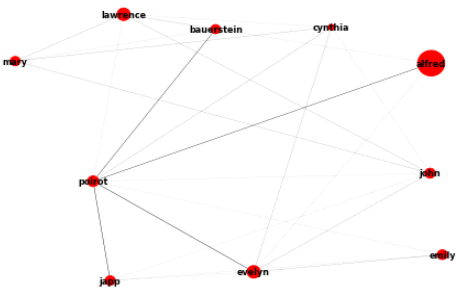
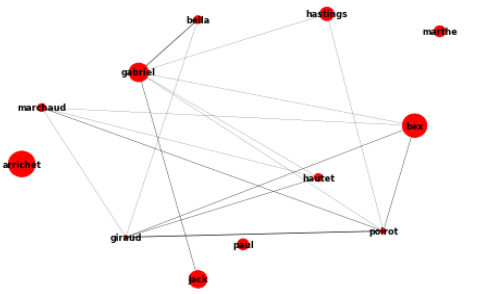
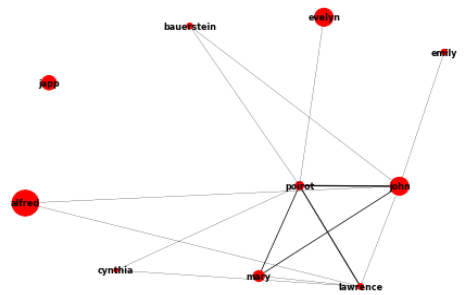
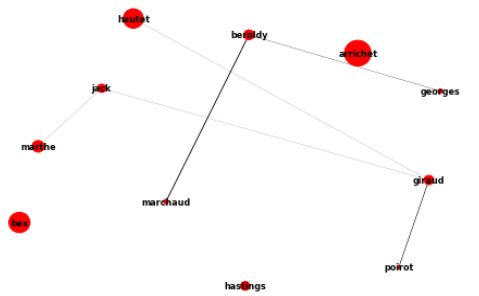
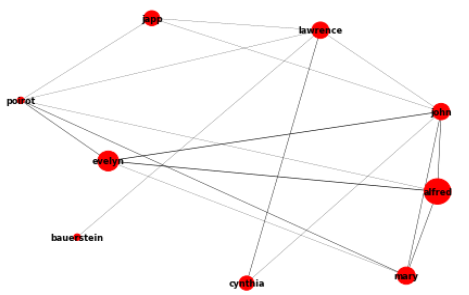
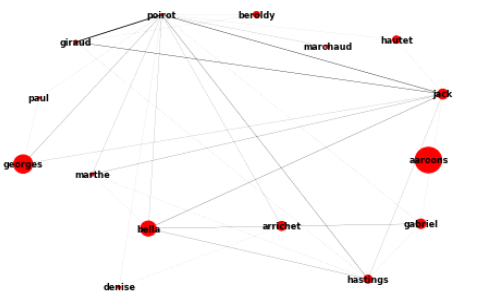
에르큘 푸아로 시리즈의 인물 관계 변화 양상은 토픽 모델링 결과와 동일한 것으로 드러났다. <Table 7>에서 드러나는 두 작품 모두 대다수의 인물을 네트워크에서 언급하고 있으며 연결되어 있는 링크 역시 복잡하게 얽혀 있음을 확인할 수 있다. 이를 통해 앞서 토픽 모델링과 감성 분석에서 언급하였던 애거서 크리스티의 인물의 심리와 갈등에 초점을 맞춘 문체적 특성이 반영된 결과로 해석해볼 수 있다. 이에 더해 셜록 홈즈 시리즈와 마찬가지로 에르큘 푸아로 시리즈 역시 두 작품 간 인물의 관계 형성이 다를 수 있다는 것을 확인하였는데, 예를 들어 첫 번째 작품인 <스타일스 저택의 괴사건>의 경우 범인인 'alfred'와 탐정인 'poirot'이 직접적으로 연결되어 있는 네트워크를 형성한 것과 달리 두 번째 작품인 <골프장 살인 사건>은 범인인 'marthe'가 탐정과 직접적으로 연결되어 있지 않은 것을 확인할 수 있었다. 이는 범인에게 접근하는 탐정의 방식 혹은 사건을 대하는 탐정의 행동과 정보의 양이 다를 수 있는 것으로, 이를 통해 에르큘 푸아로 시리즈의 두 작품 역시 사건의 내용이 다르기에 생긴 차이라고 해석할 수 있다.

앞서 연구에서 아서 코난 도일과 애거서 크리스티의 소설 진행 양상이 다를 수 있음을 언급하였다. 아서 코난 도일은 셜록 홈즈와 왓슨이라는 핵심 인물들을 사건 중심으로 던져 그들 스스로 사건과 부딪히고 갈등을 해결할 수 있도록 구조를 설계한다. 이러한 그의 집필 방식은 독자들로 하여 등장인물들의 말과 행동에 몰입하게 만들어 마치 그들 또한 핵심 인물들과 함께 사건을 파헤치며 모험을 떠나는 기분이 들도록 한다. 이런 특성으로 인해 셜록 홈즈 시리즈를 읽은 독자는 '모험 소설'로서 소설을 평가하는데, 그러한 평가의 근거를 본 분석을 통해 확인할 수 있었다. 앞서 셜록 홈즈 시리즈를 토픽 모델링과 네트워크로 분석하면서 각 구간에 등장하는 인물 유형이 유사하게 드러나고 있음을 확인하였다. 인물이 드러나는 순간이 순서에 따라 정리가 되면서 그들 간에 얽혀 있는 관계를 추측할 수 있었으며 나아가 사건 진행의 흐름을 유추할 수 있었다. 감성 분석을 통해 도출된 그래프는 셜록

〈Table 6〉 Sherlok Holmes series Network Analysis for each section

<A Study in Scarlet>	<The Sign of Four>
	
First section	First section
	
Second section	Second section
	
Third section	Third section
	
Fourth section	Fourth section
	
Last section	Last section

〈Table 7〉 Hercule Poirot series Network Analysis for each section

<The Mysterious Affair at Styles>	<The Murder on the Links>
	
First section	First section
	
Second section	Second section
	
Third section	Third section
	
Last section	Last section

〈Table 8〉 Author's inherent style of writing

	Arthur Conan Doyle	Agatha Christie
Inherent style	A description of events according to the character's role and timing of appearance. Encourage immersion and unity in character's behavior	A description of events centered on the inner psychology of a character and the conflict between them. Inducing immersion in the emotion of a character

홈즈 시리즈의 두 작품이 담고 있는 사건의 내용이 다를 수 있음을 보여주었다. 하지만 언급한 대로 토픽 모델링과 언어 네트워크를 분석해본 결과 아서 코난 도일은 애거서 크리스티에 비해 '모험 소설'이라는 평가에 걸맞게 인물들의 배치에 따른 클리셰를 포함하는 모습을 보였다. 따라서 <Table 8>의 정리에 따라 아서 코난 도일은 애거서 크리스티에 비해 사건 진행에 따른 등장인물의 관계와 유형을 중심으로 사건을 서술하는 문체적 특성을 보이는 것으로 판단된다.

반면 애거서 크리스티의 작품은 '퀴즈 소설' 혹은 '로맨스 소설'로도 평가받을 정도로 셜록 홈즈 시리즈와 궤를 달리한다. 앞서 언급하였듯이 복잡한 인물 관계에서 펼쳐지는 갈등과 그에 따른 내면 심리를 집중적으로 묘사하는 방식을 즐겨 사용하였으며, 이러한 감정 묘사와 갈등 속에 사건의 실마리를 숨겨 사건 끝에서 독자에게 반전을 선사하는 집필 방식을 주로 사용하였다. 따라서 그녀의 소설을 접한 독자는 그녀의 풍부한 감정 묘사에 놀라는 한편 셜록 홈즈 시리즈와 다른 사건 접근 방식으로 마치 어려운 퀴즈를 푸는 듯한 인상을 받는다. 에르퀼 푸아로 시리즈의 두 작품에 대한 감성 분석 결과는 유사한 형태를 띠고 있었다. 이는 인물들의 역할과 유형에 따라 사건을 진행하는 것이 아닌, 그들 간에 발생하는 갈등과 내면 심리를 통해 사건을 진행하는 방식으로 서술되기 때문이다. 언어 네트워크와 토픽 모델링을 통해서 확인할 수 있듯이 사건 진행에 따른 등장인물의 유형이 구분되었던 셜록 홈즈 시리즈와 다르게 초반부터 다양한 인물들에 집중하는 모습을 확인할 수 있었다. 따라서 <Table 8>에서 언급하듯이 그녀는 사건 진행에 앞서 인물들 간의 관계와 그들의 내면적 심리를 묘사하는 문체적 특성을 가진 것으로 판단된다.

4.3 문법 체계 분석

앞서 언급한 대로 문체 간 특징을 정립하고 독자적인 문법 체계를 설정하여 이를 포함하는 문장들을 구별하였다. 이후 해당 문장들이 전체 문장에서 차지하는 비율을 계산하여 작가의 문법 체계를 정리하였다. 본 연구를 진행하기 위해 마찬가지로 파이썬 영문 자연 언어 처리 패키지인 'NLTK'를 사용하였으며, 형태소 분석을 통해 각 어휘들의 품사를 구분한 뒤 그들 간 어순을 고려하여 문법 체계를 정립하였다.

아래는 각 작품에서 문법적 체계를 가진 문장들의 비율을 작품 간 비교하고 확인하기 위해 표로 정리한 것이다. 해당 문법적 체계가 전체 문장에서 30% 이상을 차지하는 경우 핵심 문체를 사용한 것으로 간주했으며, 두 작품 간의 비율 차이와 작가 간 비율 차이가 평균 5% 이내인 경우 두 작품과 작가는 유사한 문법적 체계를 가지는 것으로 간주하였다.

〈Table 9〉 Grammar sentence ratio according to each work

Grammar	<A Study in Scarlet>	<The Sign of Four>	<The Mysterious Affair at Styles>	<The Murder on the Links>
Subordinating conjunctions	65.8%	60.5%	55.7%	52.0%
Coordinate conjunctions	54.5%	47.0%	36.5%	36.7%
Conjunctions occurrent use	39.8%	33.2%	26.1%	24.2%
Use adjectives more than three	18.0%	17.7%	12.0%	8.5%
Use adverb more than two	28.4%	28.9%	29.0%	23.3%
Prepositional adjective expression	56.6%	49.8%	40.8%	39.2%
Relative adjective expression	25.9%	20.5%	14.7%	13.3%
Particle adjective expression	49.5%	38.7%	37.7%	34.9%
To-Infinitive adjective expression	6.6%	10.2%	5.8%	6.2%
Idiomatic expression	12.4%	12.6%	8.6%	7.0%
Figurative expression	50.7%	48.2%	43.8%	40.1%
Use interjection	0.1%	0.3%	0.1%	0.3%
All adjective, adverb expression	83.8%	79.3%	69.4%	67.6%
Difference between works	mean 4.16%		mean 2.19%	
Difference between authors	mean 7.91%			

분석 결과, 두 작가 모두 유사한 문법 체계를 사용하는 것으로 드러났다. 가령 <Table 9>에서 드러나는 것처럼 두 작가 모두 종속 접속사와 등위 접속사의 사용이 전체 30% 이상을 차지하며 두 가지 접속사를 비중 있게 사용하는 것으로 확인되었다. 뿐만 아니라 전치사 구와 분사를 포함한 형용사적 표현의 빈번한 사용, 숙어 표현의 빈번한 사용 역시 공통된 문법 체계로 확인할 수 있다. 이에 더해 두 작가의 작품 간 문법 체계의 차이가 평균 5% 이내를 나타내면서 역시 같은 시리즈 간 문법 체계가 일정하게 유지되고 있다고 산정할 수 있었다. 그러나 두 작가의 문법 체계 사용 정도가 구체적으로 다를 수 있음을 확인하였는데, 한 문장에서 종속 접속사와 등위 접속사를 동시에 사용하는 비중이 셜록 홈즈 시리즈가 더 많거나, 전반적으로 애거서 크리스티에 비해 아서 코난 도일이 공통된 문법 체계를 사용하는 경우가 더 많았다. 이러한 차이로 인해 두 작가의 문법 체계 차이는 평균 7% 이상으로 드러난 것으로 보이며, 두 작가는 동일한 문법 체계를 사용하나 작품에서 드러나는 문체로서의 차이점이 존재하고 있음을 유추해볼 수 있다.

결국 두 작가는 종속 접속사와 등위 접속사를 비중 있게 사용함으로써 구어체와 문어체를 유사하게 사용하는 집필 방식을 사용하고 있으며, 형용사적 표현의 사용, 숙어 표현의 사용 등을 통해 만연체와 화려체를 중심으로 소설을 묘사하고 있음을 확인할 수 있다. 그러나 두 작가의 문체 종류는 같을 수 있으나 사용하는 정도가 다르다는 것을 차이를 통해 인지할 수 있었다.

5. 결론

본 논문은 감성 분석과 토픽 모델링, 언어 네트워크와 자체적인 문법 체계를 통해 두 작가 간 문체적 특성에 대해 연구하였다. 감성 분석과 토픽 모델링, 언어 네트워크는 어휘와 어법의 구조를 파악함으로써 문체를 규명하고자 했던 기존 연구에서 나아가 스토리의 진행에 따른 사건의 전개 양상과 인물 간의 관계를 확인하는 방식으로 보다 내재된 작가의 문체 성향을 확인하고자 시도되었다. 분석 결과, 아서 코난 도일은 작품 간 역할로 구분되는 등장인물들의 등장 순서와 그들 간 관계를 고정시킴으로써 스토리를 전개하는 내재적 문체 특성을 보였다. 애거서 크리스티에 비해 사건 자체의 특성과 그들의 심리 묘사에 집중하기 보다 그들의 등장과 역할에 집중하는 모습을 보여줌으로써 독자들로 하여금 사건의 중심에 선 인물들의 특성과 역할에 몰입할 수 있도록 하였다. 반면 애거서 크리스티는 대다수의 인물을 작품 초반부터 후반까지 언급하며 인물들의 심리와 감정 묘사에 집중하는 모습을 보였다. 특히 정서적 감정을 드러내는 어휘를 사용하는 모습을 보였는데, 이러한 그녀의 집필 특성은 독자들로 하여 등장인물의 역할과 사건의 진행에서 벗어나 등장인물 자체가 가진 성격과 배경 스토리에 몰입하도록 할 수 있다. 또한 문법적 체계를 통해 구조적 형태의 문체를 분석해 본 결과, 두 작가는 공통적으로 구어체와 문어체, 그리고 화려체와 만연체를 주로 사용하는 것으로 나타났다. 그러나 같은 문체적 특성이라 하더라도 이를 표현하는 방식에서 차이가 있을 수 있음을 확인하였다.

본 연구는 발전하는 소셜 네트워크와 뉴 미디어, ICT 산업을 통한 마케팅 분야에서 각 플랫폼의 성향에 따른 광고 메시지 전달 효과를 측정하고 기획하는 용도로 사용될 수 있다. 기업마다 전달하고자 하는 제품과 서비스의 가치는 다양하며, 이에 따른 목표 고객 역시 다양할 수 있다. 그들이 자주 접하고 이용하는 네트워크 채널이 다를 수 있기 때문에 그들의 성향과 채널의 성향을 복합적으로 구성하는 마케팅 전략이 요구된다. 특히 국내 시장을 넘어 해외 시장을 목표로 하는 다국적 기업들은 각 나라의 문화에 대한 이해를 바탕으로 마케팅 전략을 수립할 수 있으며, 그들의 대외적 이미지 형성을 위해서도 접근성이 높은 SNS를 통해 기업 커뮤니케이션을 시도하게 된다 (Sung and Cho, 2016). 이에 더해 직관적인 언어 표현을 통한 엔터프라이즈 자동화를 추구하는 인공지능 개발 환경에서 본 연구는 언어가 가진 다양성과 내재된 의미에 대한 아이디어를 제공하며 자연어 알고리즘 개발에 기여할 수 있을 것으로 기대된다.

그러나 이러한 의의에도 불구하고 다음과 같은 한계점을 가진다. 문체적 특성을 파악하기 위해 감성 분석과 토픽 모델링을 사용하였으나, 추론을 기반으로 하였기에 다양한 해석 역시 가능하다는 한계점이 있다. 또한 해당 텍스트가 영문 텍스트임에 비해 한글 텍스트에 적용되는 문체 특성을 적용한 점, 문법 체계를 나누는 기준에 있어 보다 구체적이고 다양한 문법 기준을 대입하지 못한 점이 한계점으로 고려된다.

본 연구는 추후 소설 텍스트를 벗어나 SNS에서 사용되는 채팅 데이터의 문체를 분석함으로써 실제 기업이 목표로 하는 SNS 플랫폼과 사용자의 성향을 분석하고 전략적 방안을 제시할 수 있을 것으로 기대된다. 또한 문학적 범주에 머물렀던 정통 문법 체계와 더불어 채팅 데이터에서 사용되는 고유한 문법적 체계를 정립할 수 있을

것이다. 이러한 후속 연구에 대하여 플랫폼의 특성과 사용자의 성향에 따라 SNS를 분류하고자 하는 관련 연구에 참고할 수 있을 것이다.

참고문헌(References)

Blei, D. M., Ng, A. Y. and Jordan, M. I., "Latent Dirichlet Allocation", *Journal of machine Learning research,* Vol. 3, No. Jan, 2003, 993-1022.

Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G., "Network Analysis in the Social Sciences", *science,* Vol. 323, No. 5916, 2009, 892-895.

Chae, S. H., Lim, J. I. and Kang, J., "A Comparative Analysis of Social Commerce and Open Market Using User Reviews in Korean Mobile Commerce", *Journal of Intelligence and Information Systems,* Vol. 21, No. 4, 2015, 53-77.

Cho, H. J., Kang, J. and Jung, D. Y., "An Exploratory Study on Mobile App Review through Comparative Analysis between South Korea and U.S.", *Journal of Information Technology Services* Vol. 15, No. 2, 2016, 169-184.

Cho, H. J., Kim, S. G. and Kang, J. Y., "An Empirical Analysis of Doppelganger Brand Image Effects: Focused on the Internet Community", *The Journal of Information Systems,* Vol. 26, No. 1, 2017, 21-51.

Cho, K., H, "Texsorte Und Stil -Eine Analyse Der Textsorte "Leserkommentar" Und "Kommentar" Im Deutschen", *Koreanische Zeitschrift fuer Deutschunterricht,* Vol. 46, No. 1, 2009, 61-82.

Choi, S. R. and Yoo, J. W., "Present of the Analysis Method of the Validation between the Story Proceeding and the Character - by the Generative Trajectory of Meaning with Greimass and Enneagram", *Journal of Digital Design,* Vol. 14, No. 2, 2014, 139-147.

Hong, J., Kim, S., Park, J. and Choi, J., "A Malicious Comments Detection Technique on the Internet Using Sentiment Analysis and Svm", *Journal of the Korea Institute of Information and Communication Engineering,* Vol. 20, No. 2, 2016, 260-267.

Hwang, S., "How Korean Top 100 Companies Use Social Network Services: An Analysis of Relationship Cultivation Strategies, Message Topics, and Posting Types", *Studies of Broadcasting Culture,* Vol. 25, No. 1, 2013, 235-273.

Jang, P.-S., "Study on Principal Sentiment Analysis of Social Data", *Journal of the Korea Society of Computer and Information,* Vol. 19, No. 12, 2014, 49-56.

Jeong, E. G., " The Characteristic and Meaning of Narrative Style Based on a Point of Viɛw of a Novel", SOONGSILOHMUN, Vol. 24, 2010, 39-68.

Kang, B., Song, M. and Jho, W., "A Study on Opinion Mining of Newspaper Texts Based on Topic Modeling", *JOURNAL OF THE KOREAN SOCIETY FOR LIBRARY AND INFORMATION SCIENCE,* Vol. 47, No. 4, 2013, 315-334.

Kim, S. G., Cho, H. J. and Kang, J. Y., "The Status of Using Text Mining in Academic Research and Analysis Methods", *Journal of Information Technology and Architecture,* Vol. 13, No. 2, 2016, 317-329.

Kim, S. G. and Kang, J., "Analyzing the Discriminative Attributes of Products Using Text Mining Focused on Cosmetic Reviews", *Information Processing & Management,* Vol. 54, No. 6, 2018, 938-957.

Knoke, D. and Kuklinski, J. H., *Network Analysis: Basic Concepts*. Markets, Hierarchies and Networks: The Coordination of Social Life. SAGE, 1991.

Lee, J. O., *Stylism*. Seoul: Salim. Seoul: Salim, 2006.

Lee, S. Y. and Lee, K. M., "A Reply Graph-Based Social Mining Method with Topic Modeling", *Journal of Korean Institute of Intelligent Systems,* Vol. 24, No. 6, 2014, 640-645.

Lee, H. S. and Jeon, M. G., " A corpus stylistic analysis of Jonathan Swifts writing style in Gullivers Travels and A Tale of a Tub", *Korea Journal of English Language and Linguistics,* Vol. 19, No. 1, 2019, 120-141.

Oberreuter, G. and Velásquez, J. D., "Text Mining Applied to Plagiarism Detection: The Use of Words for Detecting Deviations in the Writing Style", *Expert Systems with Applications,* Vol. 40, No. 9, 2013, 3756-3763.

Pang, B. and Lee, L., "Opinion Mining and Sentiment Analysis", *Foundations and trends in information retrieval,* Vol. 2, No. 1-2, 2008, 1-135.

Park, G.-M., Kim, S.-H. and Cho, H.-G., "Analysis of Social Network According to the Distance of Characters Statements", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION,* Vol. 13, No. 4, 2013, 427-439.

Pavlyshenko, B., "Clustering of Authors' Texts of English Fiction in the Vector Space of Semantic Fields", *Cybernetics and Information Technologies,* Vol. 14, No. 3, 2014, 25-36.

Scott, J., "Social Network Analysis", *Sociology,* Vol. 22, No. 1, 1988, 109-127.

Suh, J. H., "Comparing Writing Style Feature-Based Classification Methods for Estimating User Reputations in Social Media", *SpringerPlus,* Vol. 5, No. 1, 2016, 261.

Suh, Y. H., " An Analysis of Style in Hemingways Short Story A Canary for One with Special Focus on the Function of Repetition", *The Journal of Linguistics Science,* Vol. 87, 2018, 329-344.

Sung, M. and Cho, J., "Corporate Communication Management on Social Networking Sites : Analysis of Communication Strategies on Corporate Facebook Pages", *Journal of Communication Science,* Vol. 16, No. 4, 2016, 41-82.

Yang, N.-Y., Kim, S.-G. and Kang, J.-Y., "Researcher and Research Area Recommendation System for Promoting Convergence Research Using Text Mining and Messenger Ui", *The Journal of Information Systems,* Vol. 27, No. 4, 2018, 71-96.

Abstract

A study on detective story authors' style differentiation and style structure based on Text Mining

Seok Hyung Moon* · Juyoung Kang**

This study was conducted to present the stylistic differences between Arthur Conan Doyle and Agatha Christie, famous as writers of classical mystery novels, through data analysis, and further to present the analytical methodology of the study of style based on text mining. The reason why we chose mystery novels for our research is because the unique devices that exist in classical mystery novels have strong stylistic characteristics, and furthermore, by choosing Arthur Conan Doyle and Agatha Christie, who are also famous to the general reader, as subjects of analysis, so that people who are unfamiliar with the research can be familiar with them.

The primary objective of this study is to identify how the differences exist within the text and to interpret the effects of these differences on the reader. Accordingly, in addition to events and characters, which are key elements of mystery novels, the writer's grammatical style of writing was defined in style and attempted to analyze it. Two series and four books were selected by each writer, and the text was divided into sentences to secure data. After measuring and granting the emotional score according to each sentence, the emotions of the page progress were visualized as a graph, and the trend of the event progress in the novel was identified under eight themes by applying Topic modeling according to the page. By organizing co-occurrence matrices and performing network analysis, we were able to visually see changes in relationships between people as events progressed. In addition, the entire sentence was divided into a grammatical system based on a total of six types of writing style to identify differences between writers and between works. This enabled us to identify not only the general grammatical writing style of the author, but also the inherent stylistic characteristics in their unconsciousness, and to interpret the effects of these characteristics on the reader. This series of research processes can help to understand the context of the entire text based on a defined understanding of the style, and furthermore, by integrating previously

* e-Business, School of Business, Ajou University
** Corresponding Author: Juyoung Kang
e-Business, School of Business, Ajou University
206 Worldcup-ro, Yongtong-gu, Suwon, 16499, Korea
Tel: +82-31-219-2910, Fax: +82-31-219-1616, E-mail: jykang@ajou.ac.kr

individually conducted stylistic studies. This prior understanding can also contribute to discovering and clarifying the existence of text in unstructured data, including online text. This could help enable more accurate recognition of emotions and delivery of commands on an interactive artificial intelligence platform that currently converts voice into natural language.

In the face of increasing attempts to analyze online texts, including New Media, in many ways and discover social phenomena and managerial values, it is expected to contribute to more meaningful online text analysis and semantic interpretation through the links to these studies. However, the fact that the analysis data used in this study are two or four books by author can be considered as a limitation in that the data analysis was not attempted in sufficient quantities. The application of the writing characteristics applied to the Korean text even though it was an English text also could be limitation. The more diverse stylistic characteristics were limited to six, and the less likely interpretation was also considered as a limitation. In addition, it is also regrettable that the research was conducted by analyzing classical mystery novels rather than text that is commonly used today, and that various classical mystery novel writers were not compared. Subsequent research will attempt to increase the diversity of interpretations by taking into account a wider variety of grammatical systems and stylistic structures and will also be applied to the current frequently used online text analysis to assess the potential for interpretation. It is expected that this will enable the interpretation and definition of the specific structure of the style and that various usability can be considered.

Key Words : Sentiment Analysis, Topic Modeling, Network Analysis, Grammar, Writing Style

Received : June 19, 2019 Revised : August 20, 2019 Accepted : September 17, 2019
Publication Type : Regular Paper Corresponding Author : Juyoung Kang

저 자 소 개

문석형
현재 아주대학교 e-비즈니스학과에 학부생으로 재학 중이다. 주 연구 관심분야는 빅데이터 분석, 텍스트 마이닝, 자연어 처리 등이다.

강주영
현재 아주대학교 경영대학 e-비즈니스학과 교수로 재직 중이며, 포항공과대학교 컴퓨터공학과에서 학사, 서울대학교 컴퓨터공학과에서 석사, 한국과학기술원 경영공학전공에서 공학박사학위를 취득하였다. 주요 관심분야는 빅데이터, 텍스트마이닝, 시맨틱 웹, 지능형 정보시스템 등이다.