한국어 형태소의 계량언어학적 연구
-신문 사설을 중심으로-

배희숙*, 시정곤**, 백혜승***, 최기선****
{*elle, ***hspaik, ****kschoi}@world.kaist.ac.kr
**shi@mail.kaist.ac.kr

QUANTITATIVE STUDY ON KOREAN MORPHEMES IN JOURNAL EDITORIALS

Bae Hee-Sook, Shi Jeong-Kon, Paik Haeseung, Choi Key-Sun
CS LAB, KAIST
Dept. of Humanities and Social Science, KAIST**

요약

말뭉치 기반 언어 연구에서 균형성은 매우 중요하게 대두되는 문제이다. 말뭉치의 균형성을 맞추려면 여러 유형의 말뭉치가 갖는 언어적 특성을 고려하여야 한다. 그러나 계량언어학적 방법으로 접근한 한국어 말뭉치의 유형별 언어 연구는 아직 미미하다. 본 연구는 언론 매체의 주요 부분인 신문의 사설을 말뭉치로 구성하여 그 언어적 특성을 살펴 보고자 한다. 계량언어학의 전형적 방법에 따라 계량화 작업을 먼저 다루고, 이어 신중한 계량화 작업으로 얻어진 자료를 조사 분석하였다.

1. 서론

한국어 형태소에 대한 계량적 연구는 이미 여러 번 소개된 바 있다.[1] 이러한 연구들은 자동으로 분석된 대량의 말뭉치를 기반으로 한 형태소 연구였다. 그러나 대량의 말뭉치를 대상으로 한 보편적 언어 연구와 더불어 개별 분야들의 분야 특성적 언어 연구도 이루어져야 한다. 근자에 들어 빈번히 논의의 대상이 되고 있는 균형 말뭉치 문제도 유형별 말뭉치들이 갖는 언어 특성이 계량적으로 연구함으로써 해결될 수 있는 것이다. 그럼에도 불구하고 계량언어학적 방법으로 접근한 한국어 말뭉치의 유형별 언어 연구는 아직 미미한 실정이다. 이러한 맥락에서 본 연구는 언론 매체의 주요 부분인 신문 사설을 말뭉치로 하여 그 계량언어학적 특성을 살펴 보고자 한다.

신문을 말뭉치로 선정할 때, 경제, 정치,사회, 문화 등의 특화된 분야를 구분하여 구성하거나 신문 전체를 대상으로 할 수도 있다. 여기서 사설을 우선적으로 선정한 것은 사설이 한 시기에 사회적으로 가장 커다란 이슈가 되는 문제를 특정 분야에 구애 받지 않고 언급한다는 장점을 갖는 만큼, 적은 양의 말뭉치로 이를 다루기에 적합한 것으로 판단하였기 때문이다.

논문의 1 장에서는 말뭉치의 계량화 작업으로, 형태소 단위에서의 문제점과 그 적용 규칙을 다룰 것이다. 이어서 2 장에서는 신중한 계량화 작업의 결과로 얻어진 언어 자료를 바탕으로 빈도에 대한 계량적 조사와 분석이 이루어질 것이다.

2. 말뭉치 구성

본 연구를 위해 1999년 9월 1일부터 9월 20일까지 게재된 동아일보와 조선일보의 사설로 말뭉치를 구성하였다. 두 신문의 사설은 같은 날짜의 것으로 구성되어 유사한 주제의 사건들을 다루고 있다. 사설의 대상이 된 주요 사건은 남북문제, 박지원 관련 대출 비리 사건, 최열의 사외이사 문제, 경제관련 등이다.

[1] 김홍규&강범모(1997), (2000).

3. 계량화(Quantification)

언어(langue)의 발현체인 담론(discours)은 "그것이 글말이건 입말이건, 길건 짧건, 문학적이건 일상적이건, 수 데이터로 전환될 수 있다" (Ch. Muller, 2000). 글말에 대한 작업에서는 언어의 내적 단위의 수가 대상 단위로 쓰일 것이다. 텍스트는 문장으로, 단어로, 형태소로, 음소로 분할될 수 있고, 분할된 단위들은 계산될 수 있다.

본 논문에서 다룰 대상 단위는 한국어 형태소이다. 한국어에서 형태소는 통사론, 형태론, 어휘론 분야에 광범위하게 걸쳐있는 단위로서, 한국어 처리를 위해서는 필수 불가결한 요소이다. 한국어 처리에 있어서 형태소가 이렇듯 중요한 단위임에도 불구하고 계량화 과정에서 발생하는 문제에 대한 국어학적 입장이 분명하지 못한 것이 사실이다. 본 논문에서 다루어지는 형태론적 문제는 계량화 과정에서 나타나는 것에 한정될 것이다. 계량언어학적 연구란 단위의 수에 대한 연구이고, 계량화 과정에서 부딪치는 문제에 대한 조정은 이 수에 영향을 미치므로, 본 연구의 대상 단위인 형태소 처리 규칙은 꼭 짚고 넘어가야 할 문제인 것이다. 그러나 계량화에서 중요한 것은 일관성 있는 규칙의 적용이지 규준 자체가 아님[2]을 염두에 두도록 하자.

3.1 한국어 형태소의 특성

인도유럽어에서 단어에 대한 정의는 말 그대로 공백이나 구두점으로 분리되는 문자의 덩어리이다. 또한 문장에서의 문법적 역할은 많은 경우 순서에 의해 정해진다. 그러나 교착어인 한국어에서 공백이나 구두점으로 분리되는 문자의 덩어리는 어절에 해당된다. 어절은 어휘형태소만으로 구성되거나 어휘형태소와 문법형태소의 결합에 의해 구성된다. 때로는 생략되어 버리기도 하지만 이 결합에서 문법형태소는 문장에서의 문법적 기능을 결정하는 핵심적 역할을 한다.

한국어가 부분 자유어순의 교착어라는 특성은 한국어 형태소 분석을 어렵게 하는 요인이 된다. 그리고 한국어 형태소가 다량의 이형태를 포함하고 있으며 그 문법 형태소들의 행태가 단일하지 않다는 사실은 문제를 더욱 배가시키고, 구문해석 과정에서 다양한 중의성 문제까지 야기한다.

구문해석 과정에서 중의성을 최대한 줄이기 위해서는 중의성을 발생시키는 어절에 대해 형태소 분석 과정에서 가능한 한 섬세한 정보를 제공해주는 것이 좋다. 그러나 한국어 형태소의 다양한 이형태들을 구문해석에 합당하게 형태소와 구문분석 과정에서 분류하고 배당해 주는 것이 결코 쉬운 일은 아니다.

본 장에서는 한국어 형태소에 관한 전반적 문제를 다루는 것이 아니라, 신문 사설 말뭉치의 형태소 단위 계량화 과정에서 실제적으로 나타나는 구체적 문제만을 다룰 것이다. 단적으로 말해 계량화에서 규칙을 다루는 것은 기초가 확실한 통계 결과를 얻기 위함이다.

3.2 형태 통사론적 규칙

3.2.1. 품사분류

한국어의 품사를 분류하는 데에는 크게 최현배를 중심으로 한 한글학회의 방법과 국립국어연구원 계열을 들 수 있다. 국립국어연구원의 『표준국어대사전』이 나온 이후 많은 국어학자들이 후자의 결과와 방법을 따르고 있다. 본 연구를 위해서는 원칙적으로 KAIST KORTERM의 품사분류를[3] 따르고, 『우리말 사전』과 『표준국어대사전』을 참고하였음을 밝힌다.

3.2.2 형태소 분할 경계

형태소 분석은 순수언어학적 접근과 자연언어처리를 위한 접근이 다소 차이가 있으나 본 연구를 위해 KAIST 품사분류를 따른 만큼, 다분히 자연언어처리의 입장에서 형태소분석을 말하게 될 것이다. 강승식(1999)에 따르면, 형태소 분석이란 단위 형태소를 분리한 후에 변형이 일어난 형태소의 경우 그 원형을 복원하고, 분리된 단위 형태소들로부터 단어 형성 규칙에 맞는 연속된 형태소들을 구하는 과정이다. 그러면 단위 형태소의 분할은 얼마나 용이한가? 본 연구를 위한 말뭉치에서 발견되는 문제점들은 대략 복합어, 용언 결합형, 자율적 명사 결합형, 약형, 접사 결합형, 결합 형태소 등으로 정리된다.

3.2.2.1 복합어

복합어의 어휘화 기준이 모호한데도 불구하고, 사전 편찬자들은 그 기준을 분명히 밝히고 있지 않다. 『표준국어대사전』에서 "놀이마당"은 하나의 단위로 다루면서 "놀이기구"는 왜 두개의 단위로 다루는가? "동의안"은 두개의 어휘로 다루면서 '동의서'는 왜 하나의 단위로 다루는가? 이 경우, '서'가 자율적으로 사용될 수 없는데 반해 '안'은 자율적이라는 사실을 떠올릴 수 있다. 그러면 "현대인"은 "현대"와 "인"으로 분리되어야 하는가? 아니면 "현대인"으로 보아야 하는가? 이런 종류의 문제를 모두 접미사나 접두사로

---

[2] "계량언어학적 조사를 위해 채택한 규준이 언어학자에게 만족스럽기는 대단히 어렵다. 그러나 이런 어려움을 과장해서는 안 된다. 왜냐하면 중요한 것은 (...) 400개의 음소 중에 모음이 170개인지 180개인지를 아는 것이 아니라 모음이 왜 174개로 계산됐는지 그 규준을 아는 것이다. 그 결과 다른 텍스트를 가지고 같은 작업을 할 때 얻어지는 결과치의 차이가 규준 자체의 유동성에서 비롯된 것이 아님을 확인할 수 있게 되는 것이다." (Ch. Muller, 2000)

[3] 윤준태&최기선(1999), 한영균 et al. (2000)

처리하는 데에는 문제가 있다.[4] '불교계', '종교계', '학계' 등의 표현에서 '_계'는 접미사로 다루어져야 하는가? '계'는 표제어로 다루어지지 않지만, 종교계는 표제어로 존재한다. "장외정치"는 "장+외+정치"인가? "장외+정치"인가? 이런 문제를 해결하기 위해서는 접미사와 접두사의 목록을 결정하는 것이 중요하다. 그러나 형태소를 너무나 섬세하게 분할 할 경우 그 섬세함의 한계 또한 문제가 될 것이다.

약어에 대해서는 어떠한가? 하나의 문서 내에서 "특별검사제"와 "특검제"가 동시에 사용되고 있다. 이를 같은 어휘로 다룰 것인가? 같은 수의 형태소로 처리할 것인가?

3.2.2.2 결합형태소

결합 형태의 문제는 결합형태소에서 더욱 복잡해진다. 문법형태소는 중복의 문제가 심하기 때문이다. '-ㄴ데/-ㄴ데도', '-에/-에다', '-는데/-는데도', '-조차/-조차도', '-며/-다며/-다면서', '-는지/-는지를' 등에서 그 예를 볼 수 있다. 『표준국어대사전』에 따르면 '-에는'은 표제어가 아니고 '-에서, -다는' 등은 표제어이다. 만일 이러한 결합형태소들을 각각 분석할 수 있는 단위로 결합되어있다는 사실에 근거하여 분할을 한다면, 어디까지 분할할 것인가?

3.2.2.3 접사의 한계

'영사관', '대사관', '자료관' 등의 단어들은 하나의 어휘인가? 『표준국어대사전』에 따르면, '자료관'은 표제어가 아닌데, '영사관'과 '대사관'은 표제어이다. '자료관'의 '-관'을 따로 분리할 경우, '관'은 '자료관'에서 쓰이는 의미로 쓰인 것이 아니다. 표제어로 나타나는 '관'은 역사적 장소의 이름으로만 다루어지고 있기 때문이다.

3.3 품사태깅

자동 분석은 아직도 많은 문제를 안고 있다. 왜냐하면 음성적, 형태론적, 어휘론적 정보들이 문장 분석의 모든 애매성을 해결해주지 못하고, 의미론적, 통사론적, 나아가 화용론적 정보까지 필요로 하기 때문이다. 본 논문을 위해서는 KAIST의 태거를 사용하여 자동 태깅한 후, 수동으로 일일이 수정하였다. 이 과정에서 나타나는 문제들은 다음과 같다.

3.3.1 통사적 애매성

3.3.1.1 명사와 부사의 경계

한국어 형태소 품사를 결정할 때, 종종 명사와 부사의 경계에서 망설이게 된다. 한 어휘가 같은 의미이면서 명사적으로도 부사적으로도 쓰여, 자동 태깅에서 애매성을 발생시키는 것은 문법형태소의 생략 현상에서 비롯되는 경우가 많다. 더구나 한자어로서 다른 어휘들과 얼마든지 결합이 가능하다는 것은 문제를 더욱 어렵게 한다. 이런 애매성은 시간의 개념을 갖는 어휘에서 더욱 두드러진다.

1) 지금, 요즘, 잠시

2) 오늘, 내일, 어제, 최근, 엊그제

1)에 나열된 어휘들은 두 가지 품사가 모두 인정되지만 2)의 명사들은 명사만 인정되고 있다. 이는 두 종류의 어휘군이 명사와 부사로 모두 쓰이는 것이 사실이지만, 두 번째가 부사로 쓰일 때는 부사격 조사가 생략된 경우이다. 따라서 첫 번째 것은 어휘가 사용된 문맥에 의해 구분하되 두 번째 것은 모두 명사로 다루었다.

또한 '그동안'은 『우리말사전』에는 표제어로조차 등재되어 있지 않지만 『표준국어대사전』에는 명사로만 쓰이는 것으로 다루어지고 있다. 또한 '대강', '대략', '대충' 등의 경우는 문맥에 의해 모두 두 가지 문법적 범주를 구분하였다.

3.3.1.2 수사, 명사, 대명사

"그 사례 중 하나이다"에서 '하나'의 문법적 범주는 무엇인가? 사전에 따르면, 이 어휘는 수사나 명사로 쓰인다. 그러나 이러한 표현에서 '하나'는 조응적이며, 이 위치에 '둘'이나 '셋'으로 대체할 수 없다. 사전은 이 어휘의 대명사적 용법을 다루고 있지 않다.

3.3.2 문법적 분석의 애매성

3.3.2.1 대등 연결어미와 종속 연결어미

"친절하며"의 '며'와 "예쁘다며"의 '며'는 의미적으로 다르다. 그러나 대등적 연결어미와 종속적 연결어미의 구분이 늘 쉽게 결정될 수 있는 것이 아니다. 사실, '면서'가 항상 대등적 연결어미가 아닌 만큼, '-지만'이 항상 명백히 종속적 어미인 것도 아니다.

1. 신문을 보면서 밥을 먹는다.
2. 모르면서 아는 척한다.

같은 형태이지만 이 예들을 관찰하면, "지만"과 의 '면서'와 "면서"의 의미가 4번만 제외하고 유사한 경우임을 의 '면서'는 다르다.

3.3.2.2 띄어쓰기 기준의 문제

"안되다"와 "안 되다"에서 전자는 '불쌍하다'의 의미로 쓰인 것이고 후자는 부사 '안'과 형용사 '되다'의 결합이다. 또한 "_대로"는 불완전 명사이지만 많은 경우 조사처럼 붙여 쓴다. 그러나 띄어쓰기는 종종 임의적으로 쓰이기 때문에 이러한 문제를 결정하기 위한 규준으로 삼기에는 어려움이 있다.

3.3.2.3 고유명사

"오늘 남과 북은"과 같은 표현에서 '남'과 '북'은 남한과 북한을 지시한다. 이들은 이렇게 따로도 쓰이지만 같은 문장에서조차 다시 "남북은"으로 붙여서 쓰이기도 한다. 이 형태소들에 어떠한 품사를 배당해야 할까?

[4] 『세종 전자사전』의 접사 목록 참조.

의미적으로 분명히 남한과 북한을 지시하므로 고유명사인가? 아니면 고유명사가 하나의 유일한 지시대상을 갖는다는 특성을 고려하여 보통명사 '남'과 '북'의 확장된 개념으로 다루어야 할 것인가?

3.3.2.4 외래어

한국어로 음차 표기 되지 않은 외래어의 경우, 모두 f로 다루어졌다. 그러나 외래어가 음차 표기가 된 경우에는 문맥에 따라 알맞은 품사를 배당하였다. 한편, '-하다'가 뒤따르지 않는 영어 명사는 보통명사로, 동사나 형용사화 접미사가 붙은 경우는 서술성 명사나 동작성 명사로 다루어졌다. 정리하면, '-하다'가 붙지 않는 영어명사는 보통명사로 처리되었고, 접미사가 붙은 경우는 동작성 명사와 서술성 명사로 다루어졌다.

품사의 수는 적용된 규칙에 절대적으로 영향을 받기 때문에 형태소 분할 과정에서 발생하는 이와 같은 문제들을 잘 정리하는 것은 매우 중요하다. 본 논문에서는 시스템 해석의 애매성 제거와 효과를 참작하여, 문장을 형태소 단위로 분할할 때 그 경계를 KAIST 품사분석규칙에 따라 결정하였으며, 일일이 수동으로 수정하였다.

4. 빈도에 대한 고찰

말뭉치의 형태소 단위 계량화를 통해 말뭉치를 구성하는 모든 형태소는 각각의 빈도를 갖추게 되었다. 이 빈도를 바탕으로 신문 사설이 가질 언어적 특성에 대한 우리의 직관이 실제로 나타나는 언어적 특성과 일치하는지 조사할 것이다.

사설이란 "신문이나 잡지에서 자기네가 주장하는 바를 써놓은 글"로, "사회의 관심이 쏠리는 문제에 대해 자기의 생각과 평가, 주장을 논리적으로 적은 글"(언어정보개발연구원편, 2001)이다. 사설은 논설문의 전형인 것이다. 그러면, 전형적 논설문인 사설의 언어적 특성에 대해 먼저 주장하는 글로서의 특성을 가정할 수 있다. 예를 들어, 강한 주장을 나타내는 서술어가 많이 등장할 것이고 형식적 문체로 구성될 것이다. 그러나 이러한 가정이 구체적으로 어떤 언어적 도구를 사용해 나타나는지 우리는 알지 못한다. 계량언어학적 연구가 이에 대한 정보를 제공하게 될 것이다.

4.1 전체적 빈도 분포

먼저 말뭉치를 구성하는 타입(type)과 토큰(token)의 전체적인 빈도 분포를 조사하자. 결과는 표1과 같다.

표1

	동아	조선
token	20661	19012
type	3388	3487

토큰수는 동아일보가 많으나 타입수는 조선일보가 많다면 일단 자세한 내부 문제는 접어두고 조선일보에서 더 풍부하게 어휘가 구사된 것으로 볼 수 있다. 하지만 정확하게 얼마나 의미있는 차이인지를 조사하려면 이항분포에 입각하여 이론값을 구하고 비교하는 방식으로 조사해야 할 것이다.

4.2 품사별 빈도 분포

품사 분포가 신문 사설 말뭉치에서 그리고 두 부분 말뭉치에서 어떻게 나타나는지를 조사하자.

표2 사설 부분말뭉치의 품사별 빈도

동아일보			조선일보		
품사	빈도	백분율	품사	빈도	
Ncn	3600	17.3	Ncn	3134	16.4
Ncpa	1774	8.5	Ncpa	1608	8.4
Etm	1260	6.1	Etm	1197	6.3
Pvg	1174	5.6	Pvg	1010	5.3
Ef	891	4.3	Jxd	850	4.4
Jxc	810	3.9	Ef	774	4.0
Jca	805	3.9	Jca	715	3.7
Ecx	735	3.5	Jco	643	3.4
Jco	729	3.5	Ecx	630	3.3
Nbn	656	3.2	Xsv	586	3.1
Xsv	651	3.2	Nbn	584	3.1
Nqq	625	3.0	Sf	533	2.8
Jcs	601	2.9	Paa	508	2.7
Sf	573	2.8	Jcs	478	2.5
Px	511	2.5	Mag	471	2.5
Paa	505	2.4	Nqq	452	2.4
Jcm	503	2.4	Jcm	444	2.3
Mag	438	2.1	Px	434	2.3
Xsn	373	1.7	Jp	368	1.9
Ep	362	1.7	Ecs	357	0.9
Ecs	345	1.7	Ep	316	1.7
Jp	327	1.6	Xsn	311	1.6
Ncr	230	1.1	Ncps	253	1.3
Ncps	194	0.9	Jcj/ncr	192*2	1.0*2
Jcj	176	0.8	Sl	187	1.0
Ecc	175	0.8	Sr	183	1.0
Nnc	165	0.8	Ecc	175	0.9
Sr/sl	137*2	0.7*2	Sp	146	0.8
Nbu	133	0.6	Mmd	140	0.7
Etn	130	0.6	Nnc	132	0.7
Sp	124	0.6	Npp	128	0.7
Xss	112	0.5	Xss	122	0.6
Mmd	108	0.5	Nbu	113	0.6
Npp	101	0.5	Etn	108	0.6
Maj	85	0.4	Maj	99	0.5
Jcr	81	0.4	Npd	97	0.5
Mma	80	0.4	Mma	81	0.4
Jcc	77	0.4	Jcc	62	0.3
Npd	60	0.3	Xp	60	0.3
Xp	53	0.2	F	51	0.3
Jct	35	0.2	Xsm	45	0.2
Xsa	32	0.1	Jcr	44	0.2
F	26	0.1	Xsa	38	0.1
Su	8	0.0	Su	2	0.0

N.B.품사기호는 KAIST TAG SET과 동일하다.

두 신문의 빈도순 품사 목록은 약간의 차이를 보이기는 하지만 대체로 그렇게 다르지는 않은 것 같다. 스피어만(Spearman) 상관계수로 그 차이가 의미있는 것인지 조사하면 결과는 다음과 같다.

$$-1 < \rho = 1 - \frac{6\sum d^2}{n(n^2-1)} < +1$$

$$r = +0.9857$$

이제, 두 신문 사설에서 품사의 빈도별 순서는 매우 유사하다고 말할 수 있다. 특히 상위 빈도에 있는 품사들은 두 집단 간에 거의 일치했다. 그러나 자세히 관찰하면 인용격 조사, 형용사, 부사, 고유명사, 불완전 명사, 명사전성어미의 분포가 특징적임을 알 수 있다. 구체적으로 가장 눈에 띄는 것은 관형사형 어미의 "과도한" 사용이다. 균형말뭉치에서 품사 분포를 보면, 관형사형 어미는 중반부 이후의 순위를 차지한다. 그러나 두 신문 모두에서 세 번째 순위에 올라와 있다. 이는 백과사전에 대한 언어 빈도 연구에서 얻은 결과와 유사하다.[5] 균형 말뭉치와 백과사전 말뭉치, 그리고 본 신문 사설 말뭉치에 대한 조사의 결과를 통해 한국어 말뭉치에서는 관형격 조사의 수가 형식언어류의 단서가 될 수 있다고 판단된다. 격조사의 경우, 그저 전체적 순서를 생각하면 신문 코퍼스에서의 조사 행태가 그다지 특이하다고 생각되지 않을 수도 있지만 빈도를 개별적으로 살펴보면 특수 보조사와 처격, 주격 조사의 빈도는 두 코퍼스 모두에서 매우 올라가 있고, 형용사와 부사는 상대적으로 낮게 나타난다. 다소 성급한 점이 있으나 이들 품사의 분포는 문체를 반영하는 지표가 될 수 있다고 판단된다.

4.3 형태소 빈도별 분포

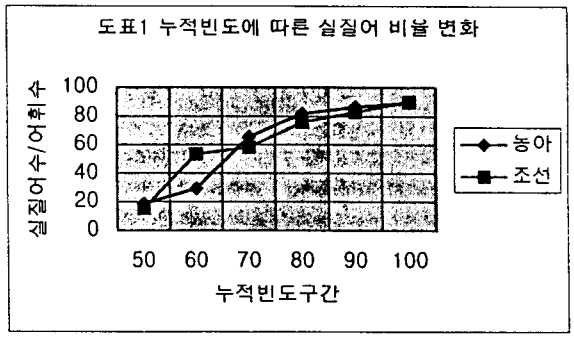
4.3.1 누적 빈도에 따른 실질어 비율

고빈도어휘에서 출발하여 저빈도 어휘로 내려가면서 누적 빈도 구간별로 어휘수에 대한 실질어의 비율을 계산하였다.

표3 실질어[6] 점유율

	동아			조선		
>50%	11	59	18.64	9	58	15.52
50-60	22	75	29.33	40	75	53.33
60-70	118	180	65.56	107	184	58.15
70-80	300	366	81.97	304	402	75.62
80-90	691	799	86.48	718	866	82.91
90<	1707	1909	89.42	1710	1902	89.91

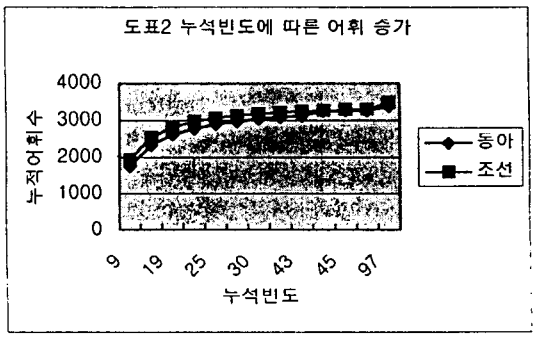
조사 결과 전체 토큰에 대한 실질어의 점유율은 동아일보에서 84.1%, 조선일보에서 82.8%를 보였다. 전체적 실질어의 비율은 동아일보에서 조금 더 올라가 있음을 알 수 있다. 아울러 각 신문의 구간별 어휘수에 대한 실질어 비율을 보면 다음 도표와 같다.



동아일보의 경우, 누적빈도 60%까지는 전체 어휘수의 약 70% 정도가 문법형태소로 점유되고 누적빈도 70%부터 문법형태소의 수가 급속도로 줄어드는 반면, 조선일보의 경우는 문법형태소의 수가 누적빈도 60% 이전에 급증하고 이후 완만한 증가를 보인다. 어쨌든 두 신문 사설 모두에서 누적빈도 70을 기점으로 문법형태소 증가가 완만해짐을 알 수 있다.

4.3.2 어휘의 증가율

토큰수가 증가함에 따라 어휘의 수도 증가할 것이다. 그러나 같은 어휘가 반복되는 비율도 있을 것이므로 어느 정도까지 증가 추세를 보이다가 그 증가율은 어떤 지점에서 완만해질 것으로 추정된다. 그러면 한국어 신문 사설을 대상으로 할 때, 그 완만해지는 지점은 어디인지 조사해보자. 저빈도 어휘에서 출발해서 토큰수가 누적될수록 서로 다른 형태소의 수가 어떻게 증가해 나가는지 도표를 통해 보자.



두 신문은 어휘증가 곡선의 양상에 있어서 매우 유사하다. 누적빈도 약 25%부터는 증가율이 매우 둔화되는 것을 알 수 있다.

이 두 그래프에서 보여주는 언어적 특성이 신문 사설 말뭉치가 갖는 고유한 특성인지는 아직 판단할 수 없다.

---

[5] 백혜승, 배희숙, 강영수, 최기선(2001) 참조.

[6] 본 조사를 위해 nb류와 np류를 제외한 체언류와 pvg, paa, mag를 실질어 집합으로 간주하였다.

4.3.3 고빈도 어휘

위의 도표에 의거하여 75% 상위 빈도 어휘를 고빈도 어휘라 하고, 이 구간을 구성하는 어휘들을 살펴보자. 동아일보의 경우, 여기에 해당하는 어휘들은 6번 이상 반복된 어휘들이고, 조선일보는 5번 이상 반복된 어휘들이 해당된다.

일반적으로 계량언어학에서는 최다빈도 어휘를 주제어라고 부른다[7]. 최다빈도 어휘들은 말뭉치의 주제를 매우 잘 반영하는 것이 사실이다. 잘 알려진 바와 같이 반복되어 나타나는 어휘인 문법형태소와 기본어를 제외하면 글의 주제어가 나타난다. 본 연구에서 전체 토큰의 50%에 해당되는 형태소 중에서 의미를 갖는 어휘형태소는 단 11개와 9개 뿐이다.

표3 고빈도 상위 50%

동아일보	조선일보
하다/pvg	하다/pvg
있다/paa	있다/paa
없다/paa	없다/paa
대하다/pvg	북/nqq
문제/ncn	정부/ncn
사건/ncn	우리/nqq
아니다/paa	문제/ncn
북/nqq	아니다/paa
기업/ncn	대하다/pvg
정부/ncn	
검찰/ncn	

'하다', '있다', '없다', '아니다'를 기본 서술어로 본다면, '문제'와 '사건'은 사설에 잘 나타나는 기본 명사로 볼 수 있을 것이고, '정부', '검찰', '기업', '북'은 말뭉치가 시기적으로 다룬 주제를 잘 보여주는 주제어에 해당한다. 50%-60% 사이에 분포한 형태소들을 살펴보면, 경제, 정치, 남북문제가 주제가 되고 있음을 알 수 있다. 상위 50%를 차지하는 형태소에는 두 신문 모두에서 단 5개의 용언이 있을 뿐이다. 흥미로운 점은 순서는 같지 않지만, 두 신문에서 이 용언들이 일치한다는 것이다. 한편, 75% 까지의 어휘 중에서 직업을 지시하는 명사류인 ncr'의 분포를 보면 '대통령, 지점장, 의원, 사외이사, 부행장, 비서, 달라이라마, 대장, 대표, 위원장, 비서관"을 꼽을 수 있는데, 이를 통해 다루어진 사건에 연루된 사람들의 직위나 직업이 잘 나타난다. 말뭉치에서 다루어진 주제는 남북문제, 은행 비리사건, 대기업 사외이사 문제 등이 아닌가?

4.4 신문사설 말뭉치의 언어특성

신문의 사설을 구성하는 언어적 특성에 대한 우리의 직관을 앞서 기술한 바 있다. 그 직관이 실제 나타나는 특성과 맞아 떨어지는지 구체적으로 형태소들을 특성에 따라 분류하여 조사하자.

4.4.1 대명사의 분포

표4를 통해 말뭉치에 나타나는 대명사 빈도를 볼 수 있다. 전반적으로 사설에는 직시소인 1인칭 단수형이나

표4

대명사	동아일보		조선일보	
우리	0.28	58	0.46	88
너/너희	0.00	0	0.016	3
당신	0.00	0	0.016	3
그/그들	0.20	41	0.21	38
그것	0.05	10	0.18	34
누구	0.01	1	0.04	7

2인칭 대명사는 잘 쓰이지 않는다. 누군가 논쟁의 대상에 연루된 사람들을 직접적으로 고유명사나 직위를 통해 지시하거나 3인칭 대명사로 지시한다. 이러한 사항은 직관적으로 알 수 있는 사항과 그다지 다르지 않다. 그러나 논설문에 1인칭 복수형 우리'가 3인칭을 통틀어 나온 수치보다 높다는 것은 의외이다.

두 신문을 비교하면, 언뜻 보기에도 조선일보가 훨씬 다양하고 풍부하게 대명사를 사용하고 있음을 알 수 있다. 특히 조선일보 사설에서 1인칭 대명사 복수형 우리'를 많이 사용한 것이 특징적이다. 이는 독자를 우리'라는 이름으로 묶어 한편으로 끌어들이는 효과를 얻을 수 있게 한다. 또한 1, 2인칭의 사용은 어투를 좀더 직접적이고 강하게 하는 요소로 작용할 것이다.

4.4.2 명사전성어미 기'와 음'

KAIST 균형말뭉치 분석에서 나타나는 명사전성어미의 수에 비추어 볼 때[8] 두 신문 사설에서 이 수치가 대단히 높은 것을 알 수 있으며, 두 신문 중에는 특히 동아일보에서 이 현상이 두드러진다. 한편, 관형격 조사 의'의 빈도를 조사하면 동아일보에서 14.67%, 조선일보에서 12.45% 나타난다. 한 마디로 유사한 경향을 보이고 있는 것이다. 명사전성어미의 수치나 관형격 조사의 수가 간결체의 지표가 될 수 있을지 아직 확언할 수 없으나 논설문이나 설명문에서 명사전성어미와 관형격 조사의 수가 올라가는 것은 일관된 현상이다.

표5

	동아	조선
기/etn	110	79
음	7	12
ㅁ	12	14

4.4.3 시제 선어말어미 분포

표7

	동아	조선
현재(는/ep)	17	14
과거(었/았)	232/56	205/52
과거회상(더)	31	24
미래추측(겠)	23	14

실제로 시제의 유형을 알려면, 조사된 선어말 어미 외에도 -'고 있는' -'중이다' 같은 표현도 고려해야 하지만, 여기서는 간단히 몇 가지 선어말 어미에 한하여 조사하였다. 결과에 따르면 두 신문 모두에서 과거형이 압도적인데 이는 사설에서 사회적 이슈가 되고 있는 사항에 대해 기술하고 의견을 펴나가는 과정에서 논증 자료를 과거 사건들을 동원하여 제시하는 데서 비롯한 것이라 생각한다.

[7] P. Guiraud(1953:155)는 빈도수가 높은 단어들을 주제어라 하고, 다른 단어들과 상대적으로 구분되며 텍스트 내용에 핵심적인 단어를 핵심어라 하였다. 이는 계량언어학에서 일반적인 것으로 받아들여지는 사항이다.

[8] KAIST 균형말뭉치에 대한 조사에 따르면 전체 어휘의 98.5%에 해당하는 65000 어휘 중에서 명사 전성어미는 단 157개로, 0.24%에 해당한다.

4.4.4 연결어미와 접속사 분포

우선 연결어미를 크게 등위와 종속으로 나누어 두 신문을 비교하여 보자.

표8

	동아		조선	
	종	수	종	수
등위	13	175	7	174
종속	53	361	60	356

얼핏보아 두 신문의 등위적 연결어미와 종속적 연결어미의 분포가 같은 양상을 보이는 듯하다. 그러나 수치적으로 확연하지 않으므로 확인을 위해 연결어미 빈도 내에서 간단히 카이제곱 검정을 해보면 $\sum(o-c)^2/c$는 0.0038로 1dd1에서 80%와 90% 사이에 위치한다. 가정에 어긋나지 않게 두 분포의 차이는 두 신문 사설에서 의미가 전혀 없으며, 전체 연결어미를 두 부류로 나눌 때 그 분포는 전적으로 임의적임을 알 수 있다. 여기서 중요한 것은 오히려 두 신문에서 일관되게 나타난 등위연결어미의 두 배 정도인 종속적 연결어미의 수가 신문 사설의 언어특성인가 하는 점이다.

연결어미를 의미를 고려하여 좀더 구체적으로 분류[9]하여 조사해 보자.

표9

	동아		조선	
	ecs	maj	ecs	maj
조건	119	1	84	11
이유	133	14	178	14
부가	16	30	15	24
역접	71	34	58	38
환기	22	7	21	7

전체 토큰수에 대한 종속적 연결어미의 비율은 동아일보에서 1.18%, 조선일보에서 1.19%를 나타내며, 접속부사의 경우 동아일보에서 0.42%, 조선일보에서 0.49%를 보인다. 전체적 품사분포상 연결어미와 접속부사의 수는 두 신문에서 같은 양상을 보인다. 그러나 표9에서의 분류에 따른 빈도 분포를 비교하면 대단히 흥미롭다. 접속부사의 경우, 카이제곱 검정을 하기에는 이론 빈도가 너무 적으므로 종속적 연결어미만을 테스트해 보자.

표10

	동아			조선		
	o	c	$(o-c)^2/c$	o	c	$(o-c)^2/c$
조건	119	102.208	2.759	84	100.792	2.798
이유	133	156.584	3.552	178	154.416	3.602
부가	16	15.608	0.010	15	15.392	0.010
역접	72	65.453	0.655	58	64.547	0.664
환기	22	21.650	0.006	21	21.350	0.006

$\sum(o-c)^2/c$는 14.062로, 자유도 4에서 0.01%에 못 미친다. 이러한 결과를 이끈 주 원인은 조건에 해당하는 연결어미와 이유, 목적, 근거를 나타내는 연결어미의 분포에서 이끌어지는 것으로, 동아일보에는 조건에 해당되는 연결어미가 과다하게 사용되고 이유나 목적, 근거에 해당하는 연결어미는 희박하게 사용된 데 반해, 조선일보에는 조건은 희박하게, 이유나 근거는 과다하게 사용되었다. 또한 부가, 역접, 환기에 해당하는 연결어미는 평이한 분포를 보였다.

직관적으로 판단하건대 사설과 같은 논설문에서는 먼저 현상을 기술하고, 주장을 펴고, 자기 주장에 대한 근거를 제시하는 방식으로 전개될 것이다. 이를 정리하면 다음과 같다.

현상 기술: -이다.

주장: -해야 한다, -해야 된다.

주장에 대한 정당화: -때문이다, -이유이다.

종속적 연결어미와 접속부사의 조사 결과로 보아, 이러한 판단은 동아일보에서 보다는 조선일보에서 전형적으로 나타날 듯하다. 사실, -'이다'형에 대한 빈도가 동아일보에서 326회인데 비해 조선일보에서는 367회나 나타나는 것이다.

4.4.5 빈도에 근거한 대표 형태소 결정

형태소가 여러 개의 이형태를 가질 때 그 대표형을 어떻게 결정할지 애매한 경우들이 있다. 이를 해결하는 하나의 방법으로서 빈도를 본다면, 다음의 도표는 하나의 정보로 간주될 수 있을 것이다.

표11

	동아일보	조선일보
가/이	244/351	197/279
을/를	451/274	414/227
은/는	270/206	280/225

조사된 주격 조사, 목적격 조사, 주제격 조사는 모두 두 신문 사설에서 같은 양상을 보였다. 가' 보다는 이'가, 을' 보다는 를'이, 는' 보다는 은'이 훨씬 높은 빈도로 사용되었다.

4.5 어휘 풍부성 평가

마지막으로 조선일보의 어휘는 이론값에 비해 얼마나 풍부한지 동아일보는 어떠한지 평가해보도록 하자. 그러면 길이가 다른 말뭉치들을 어떻게 비교할 수 있을까? 이에 대한 해결책은 다음의 공식으로 가능하다.

$$V_0=qVi+q^2V_2+q^3V_3+\cdots+q^nV_n=\sum q^nV_n.$$

표12

동아일보		조선일보	
f_1	1750	f_1	1950
f_2	578	f_2	599
f_3	288	f_3	261
f_4	169	f_4	160
f_5	110	f_5	95
f_6	81	f_6	71
f_7	65	f_7	48
f_8	46	f_8	43
f_9	40	f_9	27
f_{10}	24	f_{10}	24
$\nearrow f_{11}$	237	$\nearrow f_{11}$	209

이 공식에 대입될 말뭉치에서 뽑은 데이터는 다음 표와 같다. 텍스트의 길이가 짧은 것을 기준으로 할 때 p는 0.9202이므로 q는 0.0798이다. 이를 대입하여 계산하면 V_0는 139.65+3.6807+0.1464+0.0069+0.0020=143.486이다.

만일 동아일보가 조선일보의 길이에서 텍스트의 진행을 멈춘다면 3244.514개의 형태소를 가질 것이다. 따라서 조선일보는 텍스트를 구성하는 형태소들이 단일하게 분배되어 있다면 이론적으로 동아일보 보다 242.486개의 형태소를 더 포함하고 있는 것이 된다. 이 차이가 의미 있는 차이인지는 표준편차를 구해 알 수 있다. 1.96*표준편차(σ) =

[9] 표9의 분류를 위해 이희자&이종희(1999)를 참고하였다.

76이므로 이론적으로 조선일보의 형태소 수가 3411에서 3464 사이에 있으면 정상적 분포가 된다. 그러나 3487이므로 조선일보가 동아일보 보다 더 풍부한 어휘를 구사하고 있는 것으로 판단할 수 있다.

5 결론

지금까지 동아일보와 조선일보의 같은 날짜에 해당하는 20일분 사설을 말뭉치로 구성하여, 형태소의 계량적 특성을 조사 분석하였다. 전형적 논설문 형식인 사설의 언어적 특성은 먼저 품사의 빈도 분포에서 잘 나타났다. 가장 눈에 띄는 것은 관형사형 어미의 과도한 사용, 특수 보조사, 처격 조사, 주격 조사의 빈번한 사용, 그리고 형용사와 부사의 저조한 사용은 사설 말뭉치의 품사 빈도 특성이다. 이는 논설문의 객관성 표방과 관계 있는 것으로 추정된다.

형태소의 빈도 분포 조사에서 누적 빈도 구간별 전체 형태소 수에 대한 실질형태소 비율을 살펴본 결과 두 신문 모두에서 누적빈도 70%를 기점으로 문법형태소의 증가가 둔화되는 것을 알 수 있었다. 두 신문의 비교에서는 동아일보가 누적빈도 70%까지 꾸준히 문법형태소의 증가를 보인 반면, 조선일보에서는 누적빈도 60%까지 더욱 급격한 증가 양상을 보였다. 한편, 토큰수의 증가에 따른 서로 다른 형태소의 증가 양상을 그래프로 확인 해 본 결과 누적빈도 약 25%부터 어휘 증가치는 둔화되며 같은 형태소들의 반복이 심화됨을 알 수 있었다.

구체적 언어특성 연구에서 사설 말뭉치에는 1인칭 단수형이나 2인칭 대명사는 잘 쓰이지 않으며, 이는 당연한 결과로 해석되었다. 그러나 신문 사설과 같은 전형적인 논설문에서 1인칭 복수형 우리'가 3인칭의 총빈도보다 높다는 것은 의외의 결과였다. 특히 조선일보의 대명사 사용은 매우 풍부하고 다양했다. 1인칭 대명사 복수형 우리'의 빈번한 사용은 독자를 우리'라는 이름으로 묶어 한편으로 끌어들이는 효과를 얻을 수 있게 하며, 1, 2인칭의 사용은 어투를 좀더 직접적이고 강하게 하는 요소로 작용할 수 있다.

선어말어미만으로 조사한 시제 분포에서는 과거형이 압도적으로 많았으며, 종속적 연결어미 빈도 조사에서는 동아일보가 조건에 해당되는 연결어미를 과다하게 사용하고 이유나 목적, 근거에 해당하는 연결어미를 희박하게 사용된 데 반해, 조선일보는 조건은 희박하게 이유나 근거는 과다하게 사용하고 있었다. 논설문이 현상을 기술하고, 주장을 펴고, 자기 주장에 대한 근거를 제시하는 방식으로 전개된다고 가정한다면, 이 전형적 형식에 충실한 것은 동아일보 보다는 조선일보라고 추측하였다. 한편 대표 형태소를 결정하는데 빈도 정보를 주고자 한 조사에서 두 신문 사설은 같은 양상을 보였다. 가' 보다는 이'가, 을' 보다는 를'이, 는' 보다는 은'이 훨씬 높은 빈도로 사용되었던 것이다. 나아가 두 신문의 어휘의 풍부성이 이론적으로 기대되는 값에 비해 어떠한지도 평가하여 보았다.

서술어의 공기정보는 꼭 짚고 넘어가야 하는 부분이지만 여건상 밝히지 못한 점은 아쉬움으로 남는다. 공기 정보에 대한 발표는 다음 기회로 미루기로 하겠다.

6 참고문헌

[1] 이익섭, "국어학개설", 학연사. 2000.

[2] Makoto Nagao, "자연언어처리", 황도삼, 최기선, 김태석 공역, 1998.

[3] Charles Muller, "통계언어학", 배희숙 역, 2000.

[4] 백혜승, 배희숙, 강영수, 최기선, "인간 개체 정보 획득을 위한 백과사전의 계량언어학적 분석", ASIALEX, 2001.

[5] 시정곤, "국어의 단어형성원리", 한국문학사, 1998.

[6] 김홍규, 강범모, "한국어 사용빈도의 분석", 고려대학교 민족문화연구소, 1997.

[7] 강범모, 김홍규, 허명회, "한국어의 텍스트 장르, 문체, 유형", 태학사, 2000.

[8] Guiraud, P. 1954. *Les Caracteres statistiques du vocabulaire*. P.U.F. Paris.

[9] BAE Hee-Sook, Structures phonetiques, lexicales et syntaxiques dans deux pieces de J. Tardieu", Universite de Strasbourg, These de Doctorat, 1997.

[10] 배희숙, 어휘 풍부성 평가에 대한 계량언어학적 연구", 한국음성과학회, 2000, pp.139-149.

[11] 강승식, 다층 형태론과 한국어 형태소 분석 모델", http://dcenlp.chungbuk.ac.kr/thesis/94-h-html.

[12] 이희자&이종희(1999), 텍스트 분석적 국어 어미의 연구", 한국문화사.

[13] 윤준태&최기선, 한국어 품사부착 말뭉치에 대한 고찰", CS-TR-99-138.

[14] 한영균 et al., KAIST 품사태깅 매뉴얼", CS-TR, 2000.