인지과학, 제20권 제2호
Korean Journal of Cognitive Science
2009, Vol. 20, No. 2, 225~241.

빈도 정보를 이용한 한국어 저자 판별*

한 나 래†
고려대학교 민족문화연구원

본고에서는 빈도 정보를 이용한 저자 판별 (authorship attribution) 기법을 한국어에 적용한 연구를 소개한다. 그 대상으로는 정형화된 장르인 신문 칼럼을, 구체적으로는 조선일보에 연재 중인 4인 칼럼니스트들의 각 40개 칼럼, 총 160개 칼럼 텍스트를 선정하였다. 이들에 대하여 어절, 음절, 형태소, 각 단위 2연쇄 등의 다양한 언어 단위들의 빈도 정보들을 이용한 저자 판별을 시도한 결과, 형태소 빈도를 기반으로 하여 최고 93%를 넘는 높은 예측 정확도를 얻을 수 있었다. 또한, 저자 개인 문체간의 거리도 빈도 정보로써 계량적 표상이 가능함을 보일 수 있었다. 이로써 빈도 분석과 같은 통계적, 계량적 방법을 통하여 한국어 텍스트에 대한 성공적인 저자 판별과 개인 문체의 정량화가 가능하다는 결론을 내릴 수 있다.

주제어 : 저자 판별, 전산 문체론, 정량적 문체론, 형태소 빈도

---

* 이 논문은 2007년 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원(KRF-2007-361-AL0013)을 받아 수행된 연구이다. 연구의 발단과 진행에 많은 지지를 해 주신 민족문화연구원 한국문화연구단 여러분, 또 연구단 내 전자인문학팀의 팀원들께 감사를 드린다. 특히 여러 차례 토론을 통해 큰 도움을 주신 강범모 선생님, 김홍규 선생님과 홍정하 선생님, 또 형태소 분석기를 지원해 주신 이도길 선생님께 깊은 감사를 드린다.

† 교신저자: 한나래, 고려대학교 민족문화연구원 HK 연구교수, 연구 분야: 전산 언어학, 코퍼스 언어학
E-mail: naraehan@korea.ac.kr.

서 론

저자 판별(authorship attribution)이라 함은 작자가 무기명(anonymous)으로 되어있거나 작자의 진위가 논쟁이 되고 있는 저작물에 대해 저자를 할당하는 작업이다. 텍스트에서 추출한 자질들을 통계적으로 분석하여 저자 예측에 사용하는 정량적(quantitative) 방법은 이미 19세기부터 쓰이기 시작하여 이제는 저자 판별의 중심적인 기법으로 자리 잡았다([11], [9]). 실제 상황에서 저자의 진위가 논란이 되는 텍스트에 대해 저자 판별이 요구되는 경우는([2], [8], [13]) 흔하지는 않으나 연구 대상으로서 또 과학적 정밀성이 요구되는 실제 작업으로서 큰 무게를 가진다. 다른 한편으로는, 저자가 확정적인 텍스트들에 대하여 여러 테크닉과 방법론을 적용하여 저자 판별을 시험, 이들 기법들의 성능을 평가하고 개선하려는 노력들도 저자 판별 연구의 큰 줄기의 하나이다([3], [5], [6], [7]). 영미권에서는 카이스퀘어 테스트 (chi-square test) 등을 이용한 정량적 저자 판별 연구가 활발히 이루어져 왔으며, 최근에는 자연언어 처리 테크닉과 기계학습 기법을 사용한 전산언어학 분야에서의 연구도 활성화 되고 있으나([4], [10], [14], [12], [15]), 한국어에 대해 이러한 방법론을 적용한 연구는 아직 발표된 바가 없다. 본고에 소개할 연구는 이러한 정량적 방법론을 한국어 텍스트의 저자 판별에 적용한 것이다.

대 상

연구 대상으로 쓰인 텍스트는 현재 조선일보에 연재 중인 4인 칼럼니스트들을 지정, 그들 각각에 대하여 40개의 칼럼을 취합한 칼럼니스트 코퍼스이다. 신문 칼럼 장르는 이미 몇 다른 언어를 대상으로 한 연구에서 사용되어 ([6], [4]) 저자 판별 작업에 적합한 종류의 텍스트로 확인 받은 바 있다. 저자 판별 작업을 수행하는 데 있어서 판별의 변별성이 순수한 개인 문체의 차이에 의거하도록 하기 위해서는, 판별 대상이 되는 텍스트들에서 개인 문체 외의 다른 변인들을 제한하는 것이 중요하다. 이런 면에서 신문 칼럼은 대단히 정형화된 장르로서, 여기에는 매체, 장르, 시대, 주 독자층, 주제, 저자의 사회적 배경 등의 다양한 변인들의 영향이 최

소로 반영되며 따라서 텍스트 내에서 개인 문체에서 비롯하는 특질들의 영향을 최대로 표상 가능하게 한다는 점에서 저자 판별 연구에 최적의 텍스트라 할 만하다.

칼럼니스트 코퍼스는 김창균, 김대중, 류근일, 양상훈 4인 칼럼니스트들이 저술한 각 40개 칼럼, 총 160개 칼럼들로 구성되어있다. 출판 연대는 작가마다 조금씩의 차이는 있으나 2006년에서 2008년 사이이며, 각 전자 파일은 조선일보 온라인 에디션 사이트에서 내려받은 것이다. 크기는 전체가 73,454 어절로, 칼럼당 평균은 459어절이다. 조선일보에 정기적으로 기고하는 칼럼니스트들은 이들 이외에도 여러 명이 더 있으나, 이들 4인은 모두 중년의 남성이며 또한 정치, 사회와 경제를 논설의 주된 주제로 한다는 공통점이 있어 이들을 선정하였다. 실제로 이들의 칼럼들을 대상으로 '이명박', '박근혜', '경제', '정치', '미국', '북한', '안보' 등의 키워드를 샘플링해 본 결과, 이들이 공통적으로 많은 칼럼들에 나타나 칼럼니스트들 개인이 선호하는 주제와 내용 사이에 차이가 현저하지 않음을 미루어 짐작해 볼 수 있었다.

방법론

카이스퀘어 테스트

두 텍스트, 또는 하나의 텍스트와 텍스트 군을 계량적으로 비교하기 위해서는, 이들을 먼저 정량적으로 표상하는 것이 필요하다. 이러한 정량화는 텍스트 전체라는 가장 큰 단위로부터 추출한 더 작은 단위의 정보에 기반하여 이루어지는데, 그 단계적 범위는 텍스트 > 단락 > 문장 > 구, 절 > 어절 > 형태소 > 음절 > 음소가 된다. 이들 단위로부터 추출한 상대적 빈도를 벡터화 한 것이 한 텍스트의 정량적 표상이 되는 것이다. 일단 두 빈도 벡터가 얻어지면, 이들 사이의 유사도를 여러 통계적 테스트를 통하여 측정할 수 있다. 저자가 알려져 있는 하나의 텍스트(x)에 대해서 저자 예측의 성공 여부를 시험하기 위해서는, 다음과 같은 과정을 거친다.

1. 타겟 텍스트(x)의 실제 저자(A)를 포함한 저자 후보군(A, B, C, D)을 선정한다.
2. 저자 후보군 내의 각 저자에 대해 충분한 양의 저작물 집합(a, b, c, d)을 확보한다. 실제 저자(A)의 경우, 저자 판별 타겟인 텍스트(x)는 물론 저작물 집합(a)에서 제외한다.
3. 타겟 텍스트(x)를 표상(x')화 한다.
4. 각 후보의 저작물 집합(a, b, c, d)을 같은 방법으로 정량적 표상(a', b', c', d')화 한다.
5. 타겟 텍스트의 표상(x')과 각 후보 저작물의 정량적 표상(a', b', c', d') 사이의 유사도를 산출한다.
6. 가장 근접한 유사도를 보이는 후보 저자를 타겟 텍스트의 저자로 예측한다.
7. 예측된 저자가 실제 저자와 일치할 경우 저자 판별이 성공한 것으로, 이외의 경우는 실패로 판정한다.

본 연구의 경우, 4명의 저자가 후보군이 되며, 저자 예측 시험은 저자당 40개, 모두 160개의 텍스트에 대해서 각각 시행하였다. 빈도 벡터 측정의 단위를 달리하여 여러 차례에 걸쳐 실험할 때마다 총 160개의 예측이 이루어졌다. “무기명”으로 임시 지정한 타겟 텍스트를 4개 저자 칼럼 군집들과 비교할 때에는, 실제 저자가 아닌 다른 저자에 의한 칼럼의 군집은 40개 칼럼을 모두 포함하지만, 실제 저자의 칼럼 군집은 실험 대상 텍스트를 제외한 39개 텍스트의 집합이 사용된다. 단일 텍스트가 아닌 텍스트 집합의 표상은 각 텍스트들의 빈도 벡터들의 평균으로 이루어지므로, 이런 칼럼 군집의 크기 차이(39개 vs. 40개)는 39개 빈도 벡터의 평균과 40개 빈도 벡터의 평균이 가지는 차이로서 유의미한 수준이 되지 못한다.

두 빈도 벡터 사이의 유사도 측정은 카이스퀘어(chi-square) 테스트를 이용하였다([6], [15]). 카이스퀘어 테스트는 샘플에 대해 관측된 수치들이 기대치를 대변하는 특정한 분포에 얼마나 잘 부합하는지를 측정하는 통계 테스트로, 다음의 공식에 따른다.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

(***O*** observed, 관측치. ***E***: expected, 기대치)

<표 1> 어절 출현 상대빈도와 카이스퀘어를 이용한 유사도 측정

어절	O	E	$(O_{+1}-E_{+1})^2/E_{+1}$
가능성이	0	0.0001401	1.96280100000245e-09
가능한	0.0021598	0	4.66473604000019e-07
것이다.	0.0043196	0.0069356	6.84345599999975e-07
...	...	...	...
못하고	0.0021598	0.0004376	2.96597284000041e-07
못하면	0	0.0001517	2.3012889999993e-09
...	...	...	...
한나라당	0.0064794	0.0007987	3.22703524899985e-06
한나라당을	0	0.0001892	3.57966399999582e-09
χ^2			**0.005088377**

두 수치 분포들이 유사할수록 카이스퀘어 수치는 0에 가깝게 된다. 카이스퀘어 테스트는 기본적으로 기대치가 항상 0 이상이어야 한다. 그러나 언어적 사용의 상대 빈도는 0일 수 있으므로 (예를 들어, 특정 어휘가 텍스트에서 전혀 나타나지 않았을 때) 이러한 경우를 잃지 않기 위해 테스트를 보완할 필요가 생긴다. 이를 위해서, 본 연구에서는 관측치와 기대치로 상대 빈도에 1을 더한 수치를 이용한다. 예를 보이자면 표 1은 한 텍스트와 한 저자에 의한 텍스트군을 각각에서 추출한 어절 상대 빈도와 카이스퀘어 테스트를 사용하여 비교한 결과이다. 단일 텍스트가 관측치(O)가 되며, 비교 대상인 저자 텍스트군이 기대치(E)가 된다. 어떤 어절들은 (예: “것이다”) 두 텍스트 양쪽에서 출현하며, 어떤 어절들은 어느 하나에서만 나타나는 것을 (예: “가능성이”, “가능한”) 볼 수 있다. “가능한”의 경우, 상대 빈도의 관측치 0.0021598과 기대치 0에 각각 1을 더한 두 수치가 카이스퀘어 값의 계산에 실제로 쓰이게 된다. 각 어절에 대하여 상대 빈도는 그 어절이 관찰된 횟수를 텍스트의 전체 어절 수로 나눈 것이다. 단일 텍스트가 아닌 텍스트군집을 대상으로 하는 상대적 어절 빈도는, 각 구성 텍스트에 대해 측정한 상대 빈도를 전체 텍스트의 개수로 나눈 평균 상대 빈도로 나타낸다.

빈도 측정의 언어적 단위

앞서 잠시 언급했듯이, 판별에 사용할 빈도 자질로는 어절 외에도 다양한 언어적 단위를 사용할 수 있다. 어떠한 언어적 단위가 저자 판별에 가장 효과적인가를 탐색하기 위해, 본 연구에서는 다음 10개의 언어적 단위를 사용하였다.

(a) 음절
(b) 형태소
(c) 대표형 형태소
(d) 비주제특정 형태소
(e) 품사
(f) 어절
(g) 음절 2연쇄
(h) 형태소 2연쇄
(i) 대표형 형태소 2연쇄
(j) 비주제특정 형태소 2연쇄

(f) 어절은 띄어쓰기를 경계로 자른 단위이며, (a) 음절은 한글 음절 단위, 한자 글자 단위, 기호를 포함한다. 음절보다 큰 단위로 형태소가 사용되고 있는데, 형태소 분석 작업은 확률 기반 자동 태거인 KomaTagger([15])를 사용하였다. 형태소 (b) 단위로는 개개의 형태소와 해당 품사 태그를 함께 고려하였다. (c) 대표형 형태소라 함은 문법적 형태소들의 이형태를 (예: “었”, “았”, “从”) 대표형(예: “었”)으로 통합한 것을 일컫는다.

(d)에 제시된 “비주제특정 형태소(non-topic-specific morphemes)”라는 단위는 형태소에서 품사에 의거해 동사, 일반명사, 고유명사, 외국어, 한자어를 제외하고 조사, 어미, 조동사, 의존명사, 형용사, 부사, 접속부사, 기호 부류만을 남긴 것이다. 전자의 부류는 글의 내용을 반영하는 정도가 크지만, 후자는 그렇지 않다. 즉, 비실질 문법 형태소 부류와 일부 제한적인 실질 형태소 부류가 이 “비주제특정 형태소”에 속하는데, 형용사와 부사, 접속부사 등을 포함한 것은 담화 표지와 수식어 등이 저

<표 2> "국정(國政)은 노무현 여파로 아예 몸살을 앓았다"로부터 추출된 단위 자질들의 예시

	어절	음절	형태소	대표형 형태소	비주제특정 형태소	품사	형태소 2연쇄
1	국정(國政)은	국	국정/NNG	←	(/SS	NNG	BEGIN+국정/NNG
2	노무현	정	(/SS	←	)/SS	SS	국정/NNG+(/SS
3	여파로	(	國政/SH	←	은/JX	SH	(/SS+國政/SH
4	아예	國	)/SS	←	아예/MAG	SS	國政/SH+)/SS
5	몸살을	政	은/JX	←	을/JKO	JX	)/SS+은/JX
6	앓았다.	)	노무현/NNP	←	았/EP	NNP	은/JX+SPACE
7		은	여파/NNG	←	다/EF	NNG	SPACE+노무현/NNP
8		노	로/JKB	으로/JKB	./SF	JKB	노무현/NNP+SPACE
9		무	아예/MAG	←		MAG	SPACE+여파/NNG
10		현	몸살/NNG	←		NNG	여파/NNG+로/JKB
11		여	을/JKO	←		JKO	로/JKB+SPACE
12		파	앓/VV	←		VV	SPACE+아예/MAG
13		로	았/EP	었/EP		EP	아예/MAG+SPACE

자의 개별 문체 특성을 반영할 것이라는 판단에 의한 것이다. 따라서 이 "비주제 특정 형태소" 단위의 목적은 글의 내용과 주제에서 비롯되는 빈도 특성을 최소화하고 저자의 개인 문체에서 비롯되는 빈도 특성을 부각시키는 데 있다.

마지막으로, (g-j)에 제시된 단위들은 일부 기본 단위들의 2연쇄(bigram)이다. 이들의 사용은 빈도 측정 단위를 좀 더 넓은 관찰 영역으로 확장함으로써 저자 판별에 도움이 되는지를 탐색해 보기 위한 것이다. 이들 단위 자질들에 대한 예시가 표 2에 나타나 있다.

저자 판별 결과

실험은 10개의 각 빈도 자질에 대해 160개 각각의 텍스트에 대해서 저자 판별

을 시도하는 것으로 진행되었다. 표 3은 단위 자질별 실험 결과이다. 전체를 통틀어 (b) 형태소, (c) 대표형 형태소, (d) 비주제특정 형태소의 세 형태소 기반 자질들이 93.1%과 93.7%의 정확도로 저자 판별에 가장 강력한 자질들임을 알 수 있다. 가장 약한 자질은 어절 빈도로, 160개 텍스트 중 132개를 정확히 예측하여 82%의 정확률을 구현하였다. (g-j)의 2연쇄 자질들은 복합적 결과를 보여준다. 음절 단위에 대해서는 2연쇄를 취하는 것이 성능의 향상을 가져오지만 (85% vs. 88.7%), 단일 단위로서 좋은 성능을 보여주었던 세 형태소 단위에 있어서는 2연쇄화는 오히려 성능을 저해하는 결과를 낳았다(93.1-93.7% vs. 84.4-93.1%). 그 원인을 다음과 같이 분석해 볼 수 있다. 개개 음절은 지나치게 작은 단위로서 저자 판별에 크게 변별적이지 않으나, 음절 2연쇄는 크기 면에서 음절과 단어 내지는 형태소 사이쯤에 위치하여 단일 음절보다 정보적이므로 저자 판별에 보다 유용하다. 반대로, 형태소 단위는 이미 충분히 변별력을 가지는 단위이며, 형태소 연쇄 각각은 단일 형태소보다 더욱 정보적일지 모르나 전반적으로는 연쇄화가 필연적으로 야기하는 분포의 희박성(sparseness)이 성능에 불리하게 작용하는 결과를 낳는다. 분포의 희박성이 저자 판별 성능에 부정적 요인으로 작용하는 또 한 예는 어절 자질의 성과 부진이

<표 3> 자질별 저자 판별 결과

빈도자질	김창균	김대중	류근일	양상훈	성공판별계	성공률(%)
어절	33	38	36	25	132	82.5
음절	30	38	38	30	136	85
형태소	34	40	39	36	149	**93.1**
대표형 형태소	35	40	39	36	150	**93.7**
비주제특정 형태소	36	40	39	35	150	**93.7**
품사	36	38	37	33	144	90
음절 2연쇄	31	39	39	33	142	88.7
형태소 2연쇄	34	40	40	33	147	91.9
대표형 형태소2	36	40	40	33	149	93.1
비주제특정 형태소2	30	36	38	31	135	84.4

다. 어절은 형태소보다 큰 단위로, 개개 단위가 보유하는 정보의 양은 더 크지만, 이를 빈도 벡터로 표상했을 때는 분포의 희박화가 불가피하다.

하나 주목할만한 점은 기대에 반해 비주제특정 형태소 단위가 일반 형태소 단위에 비해 통계적으로 유의미한 만큼의 뚜렷한 성능 향상을 가져오지 않았다는 것이다. 먼저, 양 단위간의 대등한 성능에 주목하면, 이 결과는 주제특정 형태소, 즉 대부분의 실질 어휘를 저자 판별 실험에서 전적으로 제외하는 것이 판별 정확도의 저하를 야기하지 않는다는 것을 입증하여, 범주적으로 비실질 형태소의 공헌이 실질 형태소의 공헌보다 더 크며, 실질 형태소의 공헌은 미미하다는 점을 시사한다. 다음으로, 비주제특정 형태소가 예상과는 달리 성능의 향상을 야기하지 못한 데에 대해서는 두 가지 해석이 가능하다. 먼저, 이 결과가 비주제특정 형태소라는 제한된 집합 자체가 주제특정, 비주제특정 형태소를 포괄하는 모든 일반 형태소와 비교해서 문체 구분에 있어 특별한 이점을 갖지 않는다는 결론을 지지하는 것으로 일반화한 결론이 가능하다. 아니면 이와는 달리, 본 연구의 대상으로 선정한 조선일보 4인 칼럼니스트 코퍼스가 디자인 과정에서 의도한 대로 글의 주제라는 변인이 이미 잘 통제되어 있기 때문에 주제에 특정한 어휘가 저자 판별에 큰 장애로 작용하지 않고 있어서 이들의 제거가 직접적인 성능 향상으로 이어지지 않았을 것이라는 추측을 해 볼 수 있다. 만일 후자의 경우라면, 저자들 사이에 특정한 주제와 내용에 대한 선호도의 차이가 크게 나타나는 새로운 저자 코퍼스를 연구 대상으로 했을 때 비주제특정 형태소가 일반 형태소 단위에 비해 저자 판별에서 좋은 성능을 보일 것으로 예상된다. 이는 후속 연구를 위한 과제로 남겨두기로 한다.

개인 문체간 거리의 계량화

표 4는 저자별 판정 결과인데, 이는 저자별로 10개의 빈도 단위별 성공률을 평균화한 것이다. 김대중과 류근일이 판별 성공률이 높은 저자이며, 김창균과 양상훈은 상대적으로 판별 정확성이 낮은 것을 볼 수 있다. 이를 좀 더 자세히 살펴보기 위해, 저자 예측 패턴을 혼동 매트릭스(confusion matrix)화 하면 표 5와 같다. 이 매트릭스에서 왼쪽 위에서 오른쪽 아래 방향의 대각선 축은 성공한 예측을, 여기서

<표 4> 저자별 저자 판별 결과

저자	평균 정확도	최고 성공률	최저 성공률
김창균	83.7%	90%	75%
김대중	97.2%	100%	90%
류근일	96.2%	100%	90%
양상훈	81.2%	100%	62.5%

<표 5> 저자 예측의 혼동 패턴

예측 \ 원저자	김창균	김대중	류근일	양상훈	계
김창균	**335**	2	2	51	390
김대중	10	**389**	13	22	434
류근일	8	4	**385**	2	399
양상훈	47	5	0	**325**	377
계	400	400	400	400	1600

벗어난 셀들은 실패한 예측을 나타낸다. 여기서 가장 먼저 주목할 점은 김대중과 류근일의 칼럼들이 다른 저자의 것으로 오인되는 일이 적다는 것이다. 그 둘 사이에서는 류근일 칼럼은 김대중 칼럼으로 오판되는 경우가 다소 (13개) 보인다. 예측은 김대중에 몰리는 편이며 (434개) 따라서 김대중으로 예측된 케이스 중에서 오판의 빈도(10, 13, 22개)는 류근일 예측 케이스(8, 4, 2개)보다 높다. 원저자 축과 예측 축을 함께 고려하면, 류근일이 단연 다른 저자들과 뚜렷이 구분되는 것이 된다 (원저자 축 오류 2, 13, 0, 예측 축 오류 8, 4, 2). 다른 주목할 만한 점으로는, 김창균과 양상훈 사이의 혼동이 심하다는 점이다 (51개, 47개). 또한, 양상훈은 김창균뿐 아니라 김대중의 저작으로 오판되는 경우도 상당하다 (22개).

이러한 개별 문서의 저자 판별 작업에 있어서의 오류 양상은 두 저자의 문체 사이의 거리를 반영하는 것일 수 있다는 점에 착안, 이번에는 개인 문체간의 거리에 대한 계량화를 시도해 보았다. 이는 두 저자의 개인 문체를 빈도 벡터로 표상하여 그 사이의 거리를 카이 스퀘어 값으로써 계량한 것이 된다. 단, 기대치와 관

<표 6> 저자 문체간 거리

랭크	저자쌍	문체간 거리
1	김창균 -- 양상훈	$0.42 * 10^{-3}$
2	김대중 -- 양상훈	$0.79 * 10^{-3}$
3	김대중 -- 류근일	$0.87 * 10^{-3}$
4	김창균 -- 김대중	$1.10 * 10^{-3}$
5	김창균 -- 류근일	$1.46 * 10^{-3}$
6	류근일 -- 양상훈	$1.63 * 10^{-3}$

측치 사이에서 비대칭적인 카이스퀘어 테스트를 보완하기 위해, 두 저자를 번갈아 가며 관측치(***O***)와 기대치(***E***)로 둔 두 카이스퀘어 값의 평균가를 기준으로 한다. 이 실험에는 10개의 모든 자질을 다 시험하지 않고 형태소 빈도만을 사용하였다. 개인 문체를 표상하기 위해서는 40개 칼럼 각각의 빈도벡터의 평균값이 사용된다. 6개의 저자쌍에 대한 문체간 거리의 수치는 표 6과 같다.

결과는 기대했듯이 개별 문서의 저자 판별에서 보인 혼동 양상과 부합하고 있다. 먼저, 앞서 상호간 혼동 정도가 크게 나타났던 김창균과 양상훈 사이의 문체 거리가 가장 작은 것으로 나타났다. 앞서 예측이 김대중에 몰려 있음을 보았는데, 김대중을 낀 세 저자쌍이 유사도에서 2, 3, 4위를 차지했다. 또한, 개별 텍스트 판별에 있어 가장 성공적이었던 류근일을 낀 김창균-류근일과 류근일-양상훈 쌍이 5위와 6위로 저자간 문체 거리가 가장 먼 것으로 나타났다. 이는 개별 텍스트의 저자 판별에 있어서의 오류 양상이 저자들의 개인 문체간의 거리를 반영하고 있음을 보여준다.

개별 언어 자질의 판별 공헌도

정량적 방법이 아닌, 정성적 기법을 사용하는 저자 판별에서는 저자 개인이 즐겨 쓰는 어휘나 표현과 같은 특정한 언어 자질에 집중하는 경향이 있다. 정량적

기법을 사용할 경우에도 개별 언어 자질에 초점을 맞추어 이들이 판별에 기여하는 공헌도를 산정하는 것이 가능하다. 그렇다면, 도대체 어떠한 언어 자질들이 저자 판별에 크게 기여하며, 그들의 공헌도는 어떻게 따져 볼 수 있을 것인가?

표 7은 류근일 칼럼 1개와 김창균 저작 전체의 형태소 빈도를 카이스퀘어로 비교한 결과를 공헌도가 큰 순서에 따라 정렬한 것이다. 공헌도는 맨 오른쪽 칼럼의 값으로 대표될 수 있는데, 이들 값을 합산한 것이 맨 아래줄에 보인 카이스퀘어 값이며, 따라서 이 칼럼의 값이 클 수록 카이스퀘어 값에 크게 기여하여 카이스퀘어 값으로 대변되는 두 텍스트간의 거리(distance)에도 역시 큰 기여를 한 것이 된다. 한 형태소에 대해 때로는 류근일 칼럼에서의 출현 빈도(***O***)가 높으며, 때로는 김창균 저작에서의 출현 빈도(***E***)가 높다 (진한 폰트가 높은 값).

먼저 눈에 띄는 것은 기호와 문법 형태소가 상위를 차지하고 있다는 점이다. 실제로 상위 30개 형태소 자질 중 실질 형태소는 “후보/NNG”, “대통령/NNG”, “이념

<표 7> 개별 빈도 자질의 판별 공헌도

	형태소	***O***	***E***	$(O_{+1}-E_{+1})^2/E_{+1}$
1	'/SS	**0.0363**	0.0030	0.0011072
2	./SF	0.0210	**0.039**	0.0003342
3	다/EF	0.0201	**0.335**	0.0001752
4	것/NNB	**0.0210**	0.0080	0.0001670
5	"/SS	**0.0134**	0.0021	0.0001261
6	ㄴ/ETM	0.0287	**0.175**	0.0001234
7	의/JKG	0.0239	**0.156**	6.8092173e-0
8	는/ETM	0.0201	**0.120**	6.4994469e-05
9	를/JKO	0.0038	**0.119**	6.4661862e-05
10	았/EP	0.0027	**0.102**	5.4079175e-05
...	...	...	...	...
2426	때문/NNB	0.0009	**0.0009**	6.0016808e-10
χ^2				**0.0038720**

/NNG"의 셋에 불과했으며, 비실질 형태소가 리스트의 위쪽에 주로 분포하는 것을 확인할 수 있었다. 다만, 성공적인 저자 판정은 절대적 카이스퀘어 수치에 의해 좌우되는 것이 아니라 여러 카이스퀘어 수치들 사이의 상대적인 순위에 의해 결정되므로 저자 판별에 있어서 이들 개별 언어자질의 공헌도는 보다 간접적인 것으로 보아야한다는 점에 주의할 필요가 있다. 그러나 이런 고려를 하더라도, 기호와 비실질 형태소가 이처럼 큰 무게를 가지는 점은 내용을 반영하는 실질 어휘보다도 비실질적, 문법적 어휘와 문장 부호의 사용이 개인 문체 판별에 보다 큰 변별력을 가짐을 어느 정도는 입증하는 것이라고 볼 수 있다. 또한, 이는 앞서 4절에서 언급한, 저자 판별에 있어서 범주적으로 비실질 형태소의 공헌이 실질 형태소의 공헌보다 더 크다는 결론과도 부합한다.

비교: 영어에서의 저자 판별 (Grieve, 2007)

본 연구의 결과와 직접적으로 비교 가능한 선행 연구로 영어를 대상으로 한 Grieve(2007 [6])에서의 저자 판별 연구 사례가 있다. Grieve(2007 [6])는 영국의 텔레그래프(Telegraph)지에 기고하는 40명의 칼럼니스트들을 대상으로 저자당 40개의 칼럼을 사용하여 진행한 연구로서, 여러 빈도 자질의 저자 판별에 대한 공헌도를 시험하는 것과 저자 후보군의 크기에 따라 성공률이 어떻게 달라지는지를 탐색하는 두 목적을 가지고 진행되었다. 빈도 자질로는 단어, 단어 2, 3, 4그램, 문자와 같은 자질 외에도 상대적 위치를 고려하는 "문장내의 첫 6 단어와 마지막 6단어", 단어와 문장 길이의 빈도 등과 같은 보다 다양한 자질들이 고려되었다. 저자 후보군의 크기가 미치는 영향을 탐색하기 위해서는 7회에 걸쳐 40명, 20명, 10명, 5명, 4명, 3명, 2명의 각 다른 저자 그룹 크기에 대해 저자 판별의 성공률을 조사하였다.

표 8은 4명 이하의 저자후보군에 대해 가장 성과가 우수했던 빈도 자질들이다. 4명에서 3명, 또 2명으로 저자 후보군의 크기가 작을수록 저자 판별의 성공률은 향상된다. 본 연구에서와 같이 4명 후보군을 기준으로, 영어에서의 성능은 전반적으로 본 연구에서 보인 한국어에서의 저자 판별 성능보다 떨어짐을 볼 수 있다. 영어에서 가장 성공적이었던 자질은 단어와 문장부호 빈도 프로파일(word and

<표 8> 영어에서의 저자 판별 성능 (Grieve, 2007)

빈도자질	4명(%)	3명	2명
word and punctuation profile	89	92	95
character 2-gram profile	88	91	94
character 3-gram profile	88	91	94
character 4-gram profile	85	89	93
grapheme* and punctuation mark	84	87	93
first and last 6 words in sentence	82	86	92
word profile	80	85	88

punctuation profile)로서, 89%의 성공률을 보여 93.7%의 성공률을 보였던 (기본, 비주제특정) 형태소를 사용한 한국어의 케이스에 못 미친다. 한국어의 4명 후보군의 성공률은 영어에서 3명 후보군의 성공률보다 오히려 높아서, 3명 후보군과 2명 후보군 성과의 중간쯤에 위치하고 있다.

자질 면에서 살펴 보자면, 한국어에서 가장 성공적인 자질이었던 형태소에 가장 근접한 영어의 단위를 단어와 문장부호라고 할 수 있겠는데, 그 빈도 프로파일(word and punctuation profile)이 영어에서도 역시 가장 성공적임을 확인할 수 있으며, 이는 저자 판별에 있어서의 두 언어 사이의 공통점이 된다.

또 하나 주목해야할 점으로는 영어에서의 캐릭터 단위 자질 연쇄(2-, 3-, 4-gram)의 선전이다. 한국어에서 음절 단위 자질의 연쇄가 성능 향상을 가져왔듯이, 영어에서도 캐릭터 단일 단위라고 볼 수 있는 “grapheme and punctuation mark profile”에서보다 캐릭터 2-, 3-, 4-연쇄가 더 좋은 성과를 내고 있다. 다만, 한국어의 형태소에 근접하는 영어 단위인 “word and punctuation”에 대한 연쇄화 실험은 이루어진 바가 없어 이의 연쇄가 성능의 향상 또는 한국어에서처럼 오히려 성능의 저하로 이어지는지는 확인할 수 없다.

* 26개의 영어 알파벳을 가리킨다.

결 론

이로써 빈도 정보와 카이스퀘어 테스트를 이용한 계량적 저자 판별 기법이 한국어의 저자 판별에 성공적으로 적용될 수 있음을 보였다. 한국어에서는 형태소가 저자 판별에 성공적으로 작용하는 언어적 단위이며, 이를 바탕으로 4인의 저자 후보군을 놓고 저자를 판별하는 데에 93%를 넘는 높은 정확률을 달성할 수 있었다. 또한, 저자들의 개인 문체간의 거리를 계량화하는 데에도 같은 기법이 성공적으로 사용될 수 있다는 점을 보였다.

앞으로의 연구는 더욱 큰 저자 후보군을 대상으로 진행할 예정이며, 저자 후보군의 크기에 따른 판별 성공률도 탐색해 볼 계획이다. 또한, 형태소 연쇄나 비주제 특정 형태소의 연쇄와 같은 자질들에 대한 실험을 여러 종류의 장르를 포함하는 또다른 저자 코퍼스에 대해 시행하는 것 역시 흥미로운 연구 방향이 될 것이다.

참고문헌

[1] 이도길. 2005. 한국어 형태소 분석과 품사 부착을 위한 확률모형. 고려대학교 박사학위 논문.

[2] Brinegar, C. S. 1963. Mark Twain and the Quintus Curtius Snodgrass Letters: a statistical test of authorship. *Journal of the American Statistical Association*, 58: 85--96.

[3] Chaski, C. E. 2001. Empirical evaluation of language-based author identification techniques. *Forensic Linguistics*, 8: 1--65.

[4] Diederich, J., Kindermann, J., Leopold, E., and Paass, G. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19: 109--123.

[5] Forsyth, R. S. and Holmes, D. 1996. Feature-finding for text classification. *Literary and Linguistic Computing*, 11: 163--74.

[6] Grieve, J. 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, 22: 251--270.

[7] Hirst, G. and Feiguina, O. 2007. Bigrams of syntactic labels for authorship

discrimination of short texts. *Literary and Linguistic Computing*, 22(4).

[8] Holmes, D. and Forsyth, R. 1995. The Federalist revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, 10: 111--127.

[9] Juola, P. 2008. *Authorship Attribution*. Now Publishers, Inc.

[10] Keselj, V., Peng, F., Cercone, N. and Thomas, C. 2003. N-gram based author profiles for authorship attribution. In *Proceedings for Pacific Association for Computational Linguistics*.

[11] Love, H. 2002. *Attributing Authorship: An Introduction*. Cambridge University Press.

[12] Luyckx, K. and Daelemans, W. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics*.

[13] O'Brien, D. P. and Darnell, A. C. 1982. *Authorship Puzzles in the History of Economics: A Statistical Approach*. London: Macmillan.

[14] Peng, F., Schuurmans, D., Keselj, V. and Wang, S. 2003. Language independent authorship attribution using character level language models. In *Proceedings for Tenth Conference of the European Chapter of the Association for Computational Linguistics*.

[15] Uzuner, O. and Katz, B. 2005. A comparative study of language models for book and author recognition. *Lecture Notes in Computer Science*. Volume 3651/2005, pp. 969. Springer-Verlag.

1 차원고접수 : 2009. 2. 16
2 차원고접수 : 2009. 4. 26
최종게재승인 : 2009. 4. 28

(*Abstract*)

Authorship Attribution in Korean Using Frequency Profiles

Na-Rae Han

Institute of Korean Culture, Korea University, Seoul, Korea

This paper presents an authorship attribution study in Korean conducted on a corpus of newspaper column texts. Based on the data set consisting of a total of 160 columns written by four columnists of Chosun Daily, the approach utilizes relative frequencies of various lexical units in Korean such as fully inflected words, morphemes, syllables and their bigrams in an attempt to establish authorship of a blind text selected from the set. Among these various lexical units, "the morpheme" is found to be most effective in predicting who among the four potential candidates authored a text, reporting accuracies of over 93%. The results indicate that quantitative and statistical techniques in authorship attribution and computational stylistics can be successfully applied to Korean texts.

Keywords : authorship attribution, Korean, computational stylistics, morpheme frequency