어휘 풍부성 평가에 대한 계량언어학적 연구
(프랑스어 텍스트를 중심으로)

A Quantitative Linguistic Study for the Appreciation of the Lexical Richness

배 희 숙*
(Bae Hee-Sook)

ABSTRACT

Studying language by the quantitative linguistic method is not a recent development. Lately however, the interest in the quantitative linguistics has increased according to the demand on communication between human and human or between human and machine. We are required to transfer the system of the natural language onto machine. This requires the study of quantitative linguistics because we are unable to seize the characters of the tiny linguistic units and their structure in an intuitive way. In fact, the quantitative linguistics treats the internal structure of the language by the relation between the linguitic units and their quantitative characters. It is natural then that there is this growing interest in quantitative linguistics. In addition, Korean linguists take interest in the quantitative linguistics, although quantitative linguistics in Korea is not advanced by the level of the statistical analysis. Therefore, this present study shows how statistics can be applied in the field of linguistics through the two texts written in French: *Lovers of the Subway* and *Our life's A. B. C.*

Keywords : quantitative linguistics, lexicon, frequency, quantification

1. 서 론

언어에 대한 계량적 연구는 오래 전부터 시도되어 왔다. 특히 최근에는 컴퓨터의 발달과 함께, 인간의 언어체계를 컴퓨터에 인식시켜 인간과 인간, 인간과 기계 사이의 자유로운 의사소통을 위한 연구가 활발해지고 있다. 이와 더불어, 인간이 직관적인 방법으로 포착할 수 없는 미세한 언어 단위들의 특성이나 단위들 간의 관계를 통해 언어의 내적 구조를 연구하는 계량언어학에 대한 관심도 높아지고 있다. 국내에서도 이런 연구들이 이루어지고 있으나 구체적으로 얻어진 데이터를 어떻게 통계학적으로 풀어내는지에 대해서는 연구가 미미하다. 따라서 본고에서는 언어학에 통계학이 어떻게 접목될 수 있는지 프랑스어로 쓰여진 텍스트

* 한국과학기술원 음성연구실

를 예[1]로 들어 소개하고자 한다.

2. 텍스트의 계량화 작업

계량언어학이란 다양한 언어 현상들의 내적 구조를 양적으로 분석하는 방법이다. 텍스트를 구성 단위의 빈도를 통해 분석하고자 할 때, 그 출발은 분석의 대상이 되는 텍스트의 정보처리, 즉 텍스트의 말뭉치화이다. 언어(langue) 자체는 질적인 성질을 지니지만 구체화된 담화(discours)는 음소, 형태소, 단어, 구와 같은 구성요소로 이루어지므로, 이러한 단위들의 빈도수에 의해 양적 자료로 전환될 수 있다. 계량화된 텍스트는 언어학적, 문체론적, 테마적, 혹은 전산학적 연구의 중요한 자료가 된다.[2] 정확하고 신뢰할 만한 자료가 되기 위해서 텍스트의 말뭉치화는 가능한 한 엄밀하게 이루어져야 한다. 그러나 언어란 정체된 것이 아닌 만큼, 언어 단위의 계량화 작업은 언어학적으로 많은 문제점을 안고 있게 마련이다. 본 논문의 1장에서는 텍스트의 계량화작업 과정에서 제기되는 문제에는 어떤 것들이 있는지 해결방안보다는 문제 제기에 중점을 두고 소개하도록 하겠다.

2.1 전처리

텍스트를 계량화하려면 텍스트를 입력파일로 만들고 앞으로 사용하게 될 프로그램에 맞게 정리해주는 작업이 필요하다. 이는 언어처리 중 전처리 작업에 해당된다. 구체적으로 예를 들면 인쇄기법, 구두점 처리, 대문자와 소문자 구분, 숫자처리, 의성어나 의태어, 특수문자 등의 문제들을 어떻게 처리할 것인지 규칙을 정해서 미리 입력파일에서 조정해주는 작업이다.

2.2 단위분할

텍스트를 원하는 단위로 분할하는 작업은 해당 단위에 대한 엄격한 정의에서 출발한다. 어휘분포를 연구하고자 할 때 선택되는 단위는 단어(mot)이다. 단어는 공백이나 구두점으로

---

1) 본고에서 소개할 예들은 BAE Hee-Sook(1997)에서 발췌한 것이며, 사용된 텍스트는 현대 프랑스 극작가 장 따르디유(Jean Tardieu)의 두 희곡작품 『지하철의 연인들 *Les Amants du Métro*』(이하 Métro로 약칭)과 『인생의 A.B.C. *L'A. B .C. de notre Vie*』(이하 Abc로 약칭)이다.

2) 텍스트의 양화작업과 대량의 단위 빈도를 여러 방향으로 활용하기 위한 작업은 대단히 복잡해서 아무리 신중을 기해도 오류가 발생하기 마련이다. 이러한 오류를 줄이기 위해서라도 컴퓨터를 통한 작업의 자동화는 필수적이지만 이를 자동화하는 것이 요원한 것이 현실이다. 그러므로 본인은 이러한 작업들을 단계별로 나누어 컴퓨터 프로그램을 사용함으로써 조금이나마 단순화시켰다. 텍스트를 단어로 분할하는 데에는 Mots.pas, 알파벳순으로 정리하는 데에는 Classe.pas, 어휘정리를 하는 데에는 Lemme.pas, 또 이들을 품사에 따라 분류하거나 통계처리를 하는 데에는 Explagao.pas라는 프로그램을 이용하였다. 이 프로그램들은 Marc HUG에 의해 Turbo-Pascal 언어로 짜여졌다. 물론 연구소에서 개인적으로 짜여진 프로그램을 사용할 수도 있겠으나, 이런 프로그램이 적용하는 규칙들이 만족할만한 언어처리 규칙에 의거하지 않기 때문에 사용하지 않았음을 밝힌다.

다른 문자연쇄와 분리되는 문자연쇄이다. 그러나 이러한 사전적 정의는 실제로 텍스트를 단어 단위로 분할하고자 할 때 많은 문제를 야기한다. 중요한 것은 항상 일관성 있는 규준을 적용하는 것이다. 예를 들어 복합어가 어휘화된 것인지 일시적 연합 형태인지 판단하기 위해서는 복합어를 구성하는 요소들이 항상 한 단위처럼 붙으려는 경향이 있는지, 고유한 의미를 만들어내는지, 일관성 있는 형태를 보이는지, 단위를 구성하는 요소들이 본질적으로 원래의 의미를 보유하고 있는지, 구성요소들의 대체가 가능한지, 혹은 다른 요소의 삽입이 가능한지 등의 여부가 규준이 될 것이다.

2.3 품사태깅

텍스트가 단어로 분할되면 문맥을 고려하면서 각 단어에 해당되는 품사와 성, 수 그리고 대표형(lemme)을 입력한다. 이는 나중에 품사와 품사와의 관계, 품사별 行態 등을 연구하는데 기초가 된다. 이때 제기되는 가장 대표적인 문제는 동음이의어나 다의어, 혹은 분사형에서 오는 동사와 형용사간의 경계선 문제, 혹은 명사와 형용사의 경계선 문제 등이 있는데, 이를 중재하는 데에는 고도의 언어학적 식견이 요구된다.

2.3.1 동음이의어와 다의어

동음이의어와 다의어는, 그 정의에 있어서는 구분이 분명하나 실제로 이 둘을 구분하기는 매우 어렵다. 사전은 이를 어원학적 기준에 의해 처리하지만, 동음이의어와 다의어 문제는 어원학적 기준만으로 해결될 수 있는 간단한 문제가 아니다. 이 문제에 대한 처리 방법이 사전에 따라 다르기 때문에 사전을 전적으로 의존하기는 어렵다. 동음이의어와 다의어 문제는 해당 어휘가 동음이의어로 간주되면 개별적으로 사전 항목을 부여하기 때문에 어휘수에 직접적인 영향을 주게 되고, 따라서 계량언어학에서는 매우 중요한 문제이다. 최근에는 여러 의미간에 그 연관성이 인정되면 다의어, 그렇지 않으면 동음이의어로 간주하는 것이 일반적이며, 이 연관성에 대한 판단은 어원학적, 의미론적, 형태론적, 통사론적 측면에서 두루 이루어져야 한다. 의미론적으로는 의미장을 통해 연관성을 판단하고, 통사적으로는 앞뒤에 오는 문장성분이 유사한 구조 보이는지 여부로 연관성 판단하고, 형태론적으로는 파생어 계열을 찾아 분류하여 계통이 분리되는지 여부로 그 연관성을 판단한다.

2.3.2 분사형

동사의 분사형이 항상 동사에 귀속될 수 있을까? 그 형용사적 쓰임이 고정되어 더 이상 동사로 분류하기 어려울 때가 빈번하다. 그런데, 사전 처리는 동음이의어나 다의어 경우와 마찬가지로 일치를 보이지 않는다. 학자에 따라서는 현재분사와 과거분사를 따로 분류하는 경우도 있으나, 가능하면 규칙의 계층화를 통해 구조화하려는 노력이 필요하다. 현재분사의 경우, 문맥을 살펴서 성수일치가 일어나야 하는 경우는 형용사로 처리하고, 그렇지 않은 경우는 동사에 귀속시키면 된다. 과거분사의 경우는 형용사형이 어휘화 되었는지 여러 측면에서 판단해야 하므로 더욱 까다롭다.

2.3.3 명사와 형용사의 경계

명사와 형용사의 경계선은 종종 애매하다. 이는 "형용사의 성수 일치 형태가 명사와 유사하여, 형태적·통사적으로 두 품사간의 중복이 생길 수 있기 때문이다"(F. Brunot et C. Bruneau, 1969, s.v. *l'adjectif*). 이 문제 역시 사전마다 그 처리 기준이 달라 사전에 의거해서 해결할 수 없다. D. Cressels(1979: 140)는 이 문제에 대한 처리에 있어서 문법학자들을 신랄하게 비판하지만, 그 자신 역시 좋은 대안을 제시하지는 못한다. E. Buyssens(1975: 60)는 명사로 쓰인 형용사에 또 다시 명사를 덧붙일 수 있는지 여부로 그 품사를 결정할 수 있다고 주장한다. M. Hug(1989: 53)는 형용사와 명사의 용법에 있어서 기초가 되는 것은 형용사임에 근거하여, 해당 어휘가 의미나 통사적 기능에 있어 형용사로 다루기에 문제가 없다면 모두 형용사로 다룬다.

2.3.4 Voilà, voici, revoilà, revoici

전통적으로 이 어휘들은 전치사로 다루어졌으며 *Le Petit Robert*사전 역시 아직까지도 전치사를 고수하고 있다. 그러나 오래 전부터 많은 언어학자들이 문제를 제기하고 있는 것이 사실이다. Grevisse(*Le G.L.F.*)는 부사로, Brunot & Bruneau(1969: 210), R. L. Wagner & J. Pinchon(1993: 528), M. Riegel(1993: 453) 등은 지시사로, 심지어 Kr. Nyrop는 감탄사로 분류하기까지 한다. 그러나 G. Moignet는 앞선 분류들이 어떻게 잘못됐는지 분포적 기준으로 설명하면서 이를 동사로 분석해야 한다고 주장한다.[3] 사실, 주어와 동사가 도치될 때 모음충돌을 피하기 위해 사용되는 "-t"가 "Voilà-t-il"에서 나타나는 사실은 이를 뒷받침한다.

2.4 어휘정리작업(lemmatisation)

연구 대상이 어휘에 관계된 경우, 사전처럼 인덱스 작업이 이루어져야 한다. 이때 필요한 것이 어휘정리작업이다. 어휘정리작업이란 단어들을 각 대표형에 귀속시켜 분류 정리하는 것으로, 사전작업에서는 필수적이지만 이 과정에서 제기되는 많은 문제들 때문에 학자에 따라서는 이 작업을 생략한 채 단어 형태들(forme graphique ou type)을 그대로 연구할 것을 주장하기도 한다.[4] 그러나 단어형태를 그대로 연구하더라도 어휘정리 작업은 중요한 정보를 제공할 수 있으므로 이를 병행하는 것이 바람직하다.

지금까지 나열된 문제들은 엄밀한 언어학적 규칙 위에서 해결되어야 한다. 타당한 데이터로 분석된 결과만이 신뢰성을 가질 수 있기 때문이다. 사실상, 언어란 끊임없이 변화 발전하는 성질을 지닌다는 점에서 그 규칙이 고정되어 아무 문제를 제기하지 않으리라 기대하는 것은 무모하다: 중요한 것은 이를 인정하고 가능한 한 일관성 있는 기준에 의해 규칙을 세워나가는 것이다.

---

3) 이 주제에 관해서는 G. Moignet(1989) 참조.

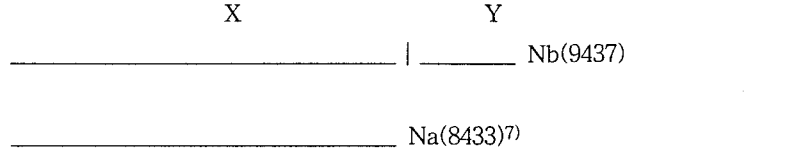
4) 이 논쟁에 관해서는 BAE Hee-Sook(1997:33-38) 참조.

3. 어휘의 풍부성 평가

지금까지 제기된 문제들에 대해 일관성 있는 규준을 마련하고, 이에 근거하여 텍스트가 계량화되면 각 어휘소와 단어형태들은 빈도수를 갖추게 된다. 빈도수를 갖춘 어휘소와 단어형태들은 적합한 통계작업을 거쳐 다양한 분야의 기초 자료가 될 수 있다. 그러나 본 논문에서는 어휘의 풍부성을 계량언어학적으로 분석하는 방법을 소개하는데 만족하도록 하겠다. 어휘의 풍부성을 판단하는 가장 간단한 평가 방법은 Charles Muller의 ≪어휘 반복지수 *f*≫이다. "어휘 반복지수란 전체 단어수(N)를 어휘수(V)로[5] 나누는 간단한 방법이다. 이때 *f*의 값이 클수록 전체 어휘수(V), 즉 단 한번 나오는 어휘(V_1)의 수는 작아지고, 반대로 *f* 값이 작아지면 어휘수는 많아진다"(배희숙, 2000: 95). 이 방법은 매우 간단해서 모든 계량언어학적 연구에서 간단히 실행해 볼 수 있는 장점을 갖는다. 그러나 *f*지수가 텍스트의 길이에 영향을 받을 수 있으므로, 많은 연구 결과를 통해 *f* 지수에 대한 평가 기준이 마련되어야 더욱 유용하게 사용할 수 있다. 이제 이항분포에 입각해서 어휘의 수를 비교하는 방법들에는 어떤 것들이 있는지 구체적으로 텍스트에 적용해 보면서 살펴보자.

3.1 길이가 다른 텍스트들의 어휘수 비교

작가는 『지하철의 연인들』과 『인생의 A. B. C.』 중[6] 어느 작품에서 더 풍부한 어휘를 사용하였을까? 이를 알기 위해서는 두 작품을 비교해야 한다. 그렇다면 길이가 다른 두 텍스트의 어휘수를 어떻게 비교할 수 있을까? 두 가지 방법을 가정할 수 있다. 길이가 더 긴 텍스트가 짧은 텍스트의 길이까지만 진행되고 멈췄을 때를 가정하고, 그때까지의 어휘수를 추정하여 계산하는 방법과, 짧은 텍스트가 긴 텍스트의 길이까지 늘어났을 때의 어휘수를 가정하여 조사하는 방법이 그것이다. 이에 대해 Ch. Muller와 N. Menard는 첫 번째 방법을 택한다. 왜냐하면 첫 번째 방법이 통계학적으로 훨씬 간단하기 때문이다. 이 방법은 N. Menard의 작업을 통해 그 합당성을 확인할 수 있으며, 이를 도식화하면 다음과 같다.



이와 같은 계산방법을 도입하기 위해서는 어휘분포가 텍스트의 전 부분에 균등하게 분포되어 있으며, 잘려 나간 텍스트의 부분은 단순한 삭제일 뿐이라는 귀무가설(l'hypothèse nulle)을 받아들여야 한다. 이런 가설은 다소 무리가 있어 보일 수 있다. 왜냐하면 텍스트

5) 국내 계량언어학에서는 이를 총어와 개별어로 표현하기도 한다.
6) Métro는 작품 『*Les Amants du Métro*』를 가리키고 Abc는 『*L'A.B.C. de notre Vie*』를, DialM은 Métro의 대사를, DidaM은 Métro의 지문을, DialA는 Abc의 대사, DidaA는 Abc의 지문을 의미한다.
7) Na(8433)은 Métro의 총 단어수이고 Nb(9437)은 Abc의 총 단어수이다.

란 현실적으로 단순한 단어의 연쇄가 아니고, 잘려 나간 부분도 단순한 삭제가 아니기 때문이다. 그러나 통계학적 테스트는 텍스트가 단순한 양적 성질의 총체로 가정되어야만 유효하다. 이러한 귀무가설 위에서, 텍스트 Abc의 잘려나간 Y부분이나 남은 X부분에서 이론적으로 기대되는 어휘수를 구해야 한다.

Abc에서 X부분에 나타날 어휘의 확률이 p라면, p는 전체 Abc에 대한 X부분의 단어수로 구한다 (p = Na/Nb = 8433/9437 = 0.8936 ; q = 1-0.8936 = 0.1064). X부분의 이론적 어휘수를 V'라 하고 X부분에 나타나지 않을 어휘수, 즉 Y부분에만 나타날 어휘수를 Vo라 하면, 전체 Abc의 총 어휘수(V)는 V'와 Vo의 합(合)이다. 확률 p를 통해서 직접 V'를 구하는 것은 대단히 복잡하다. 우선 q를 통해서 Vo를 구하고, Vo를 V에서 빼는 방법이 훨씬 간단하다. 이를 식으로 나타내면 다음과 같다.

$$V_0 = qV_1 + q^2V_2 + q^3V_3 + \ldots + q^nV_n = \Sigma q^nV_n.$$

이 공식에 준비된 데이터를 대입시키어 도표1을 얻는다. 도표1에 따르면, E(V0) = 97이고, 따라서 E(V') = 1601-97 = 1504이다. Abc가 Métro의 길이에서 진행을 멈췄을 때 Abc는 이론적으로 1504개의 어휘를 갖게된다. 다시 정리하면, 두 작품의 길이가 똑같이 8433 단어에서 조사했을 때, Métro는 1549의 어휘를 갖는데 반해 Abc는 1504의 어휘를 갖는다는 말이다. 따라서 이 두 작품의 총 어휘수는 45개이다. 그러나 이 결과를 가지고 즉시 두 텍스트의 어휘를 비교·해석할 수는 없다. 먼저 이 45개라는 어휘수의 차이가 통계학적으로 의미있는 것인지, 아니면 단순한 우연에 근거한 차이인지 테스트를 해보아야 한다. 기대되는 어휘수 E(V')를 평균으로 간주하고 이론적 표준편차를 공식($\sigma=\sqrt{npq}$)[8]에 의해 구한다.

도표 1

fn	q^n	V_n	$q^n V_n$
1	0.1064^1	886	94.2704
2	0.1064^2	247	2.7963
3	0.1064^3	128	0.1542
4	0.1064^4	77	0.0099
5	0.1064^5	263	0.0036
$\Sigma\, q^n V_n = 97.2344$			

정규 분포에 따르면, 분산은 평균 양쪽으로 표준편차 1.96까지 모든 변수의 95%가 분포된다. 비교된 두 작품의 어휘수 차이가 m±1.96σ의 범위밖에 있다면, 이 차이는 우연에 근거하지 않은 것으로 인정되고 귀무가설은 기각된다. 계산에 의하면, 이론적으로 Métro의 어휘수는 1485.2898에서 1522.7102 사이에 있어야 한다. 그러나 실제 어휘수는 그 한계를 넘어선 1,549개이다. 따라서 두 작품의 어휘수의 차이는 우연에 따른 것이 아니며, 작가는 Abc보다는 Métro에서 조금 더 풍부하고 다양한 어휘를 구사하고 있다고 말할 수 있는 것이다.

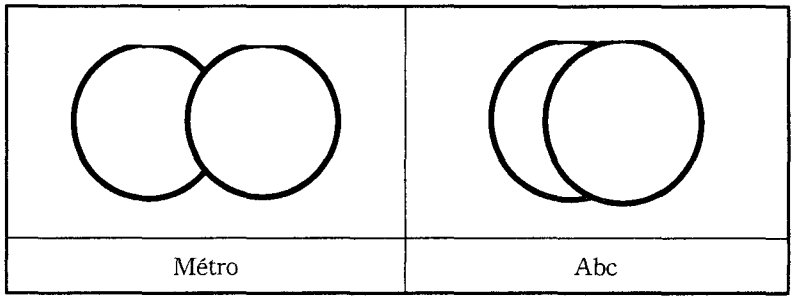
이 결과에 대한 적합한 해석을 내리기 위해, 같은 방법으로 부분집단(DialM/DialA, DidaM/DidaA, DialM/DidaM, DialA/DidaA)들을 대립시켜 조사하면 결과는 다음과 같다.

8) 여기서 p는 Abc의 X부분에서의 이론적 어휘비율이고 q는 그 보수이며, n은 작품의 총 어휘수이다.

Métro > Abc ; DidaM=DidaA ; DialM≒DialA.

얼핏보기에 이 결과는 모순을 보여주는 것 같다. 두 작품의 지문과 대사를 비교해서 어휘가 비슷하다는 조사결과가 나오는데, 어떻게 두 작품 전체를 비교하면 Métro의 어휘가 더 풍부한 것으로 나오는 것일까? 하나의 추정이 가능하다. 대사와 지문의 문체가 Métro보다는 Abc에서 좀 더 많은 공통점을 갖고 있다는 추정이 그것이다. 이를 벤다이어그램으로 나타내면 다음과 같다.

도표2



지문의 문체가 대사의 문체보다 좀더 묘사적·설명적이고, 대사의 문체는 좀더 친숙한 일상 대화체라는 사실을 인정한다면, 이런 두 부분의 문체의 차이는 Abc보다는 Métro에서 더욱 분명한 것이 사실이다. 이 추정이 설득력 있는 것인지는 품사분포 연구와 음소분포 연구에서 확인될 수 있다.[9]

3.2 이론적 어휘수와 실제 어휘수 비교

어휘빈도를 서로 직접적으로 비교하지 않고, 각 집단의 어휘를 이론적 모델과 비교함으로써 평가하는 방법을 살펴보자. 물론 구체적인 검사에 들어가지 않고 직접적이고 단순한 방법으로 두 작품의 지문과 대사에 쓰인 어휘의 풍부성을 비교할 수 있다.

도표3

	DialM		DidaM	DialA		DidaA
N	4224	>	4209	4869	>	4568
V	882	<	938	962	<	983

도표3에 따르면, 텍스트의 범위는 지문보다 대사가 더 크지만 어휘수는 지문에 더 많은 것으로 나타난다. 따라서 어휘가 대사보다는 지문에서 더 풍부하게 구사되었다고 말할 수 있다. 그러나 단지 이러한 비교를 통해서는 어떤 점에서 두 표본집단 사이의 어휘분포가 정규분포를 벗어나는지, 그리고 이 대립의 원인은 무엇인지를 알 수 없다. 이를 파악하기 위해서는 이론적 모델을 세워 비교하는 좀더 구체적인 테스트를 해야 한다.

이론적 모델은 이항법칙에 근거해 세워진다. 이 모델과 실제의 수치를 비교하고, 만약 차

9) 이 문제에 관한 자세한 설명은 BAE Hee-Sook(1997: 167-178)을 참조할 것.

이가 있다면, 그 차이가 통계학적으로 의미 있는 것인지 변환편차(l'écart réduit)[10]를 통해 조사한다. 이론적 모델은 두 프라그망에 어휘가 임의적으로 분포되어 있으며 관찰된 빈도수와 이론적 빈도수 간의 차이는 오직 우연에 근거한다는 귀무가설에 입각해서 세워진다. 이 모델에 따라 얻어지는 이론적 빈도수는 어휘가 작품의 전 부분에서 규칙적으로 분배될 때 얻게 되는 값인 것이다. 두 집단의 이론적 어휘수는 따로 계산된다. 계산 과정은 Métro의 대사를 예로 들어 도표 4에서 제시된다.

도표 4

f_n	q^n	V_n	$q^n V_n$
1	0.4991	854	426.2314
2	0.2491	231	57.5421
3	0.1243	114	14.1702
4	0.0621	60	3.7260
5	0.0310	48	1.4880
6	0.0155	40	0.6200
7	0.0077	21	0.1617
8	0.0038	24	0.0912
9	0.0019	15	0.0285
≥10	0.0009	142	0.1278
$\sum q^n V_n$=504.1869			

확률 p는 0.5009(4224/8433)이고, q는 0.4991(1-0.5009)이다. 이론적으로 기대되는 어휘수는 1045(V'=V-$\sum q^n V_n$=1549 - 504)이고, 실제 어휘수는 882이다. 이 차이에 대한 변환편차는 8.8395이다.

$$z=\frac{163}{\sqrt{npq}}=\frac{163}{1549\times0.6746\times0.3254}$$

(여기서 p는 $\frac{1045}{1549}$이다).

정규분포표에 따르면 이 어휘분포가 우연에 근거할 확률은 거의 0이다. 따라서 Métro의 대사에 어휘가 이론적으로 기대되는 어휘수에 비해 매우 빈약하다는 결론을 내릴 수 있다. 이와 같은 방법으로 나머지 표본 집단에 대해 조사하면 다음의 결과를 얻는다.

DidaM (z=5.6885) ; DialA(z=7.0344) ; DidaA (z=3.5787).

어휘의 실제 수는 DialM과 DidaM에서 이론적 모델에 비해 대단히 적으며, 이 수적 열세는 통계학적으로 의미가 있다. 그러나 이러한 결과는 어느 정도 예상 가능한 것이었다. 이 두 집단 하나 하나가 전체에 비해 빈약하다는 결과는 어휘 전체가 임의적으로 분포되어 있다는 가정 위에서 계산되었지만, 그것은 사실이 아니기 때문이다. 사실 어떤 어휘는 대사에만 나타나고, 또 어떤 어휘는 해설에만 나타나게 마련인 것이다. 중요한 것은 가장 큰 결핍이 해설에서보다는 대사에 나타난다는 사실이다. 이러한 결과는 어휘반복지수 f에 근거한 추측을 확인시켜준다. 작가는 Abc보다는 Métro에서 어휘를 더욱 다양하게 사용했고, 대사보다는 지문에서 더 풍부한 어휘를 구사하였다. 그리고 이러한 대립은 Abc보다는 Métro에서 더욱 분명하게 나타났다.

계속해서 1막과 2막을 구분시켜 어휘수를 평가해 보자. 그러나 이번에는 작품 하나에 대해서만 조사하자. 도표5는 Métro의 대사와 지문에서 조사된 어휘수를 모델과 비교한 결과이다. 전체적으로, 대사와 지문의 어휘분포는 모델과 거리가 있다. 1막의 지문만이 예외적으로

10) 변환편차=$\frac{\text{절대편차}}{\text{표준편차}}$, 표준편차(écart-type), $\sigma=\sqrt{npq}$이다.

도표5

	편차	z
DialM1	-113	13.8580
DidaM1	+8	0.6084
DialM2	-120	7.7849
DidaM2	-80	5.0132

모델과 일치한다. 어휘분포에 있어서, 1막의 지문이 모델에 가장 부합되는데 반해, 1막의 대사는 모델과 가장 큰 차이를 보인다. 얼핏보기에 오직 DidaM1만을 제외하고 나머지 세 표본집단은 비정상적으로 어휘가 결핍되어 있는 것 같지만, 앞서 말한바와 같이 마이너스 편차는 예견된 것이었다. 역설적이지만 모델과 비슷한 분포를 보인 DidaM1의 어휘분포가 바로 비정상적인 것이다. 즉, DidaM1의 어휘가 예외적으로 풍부한 것이다. 결론적으로, 어휘가 2막보다는 1막에서 더욱 비정상적으로 분포되어 있고, 대사와 지문 사이의 대립 역시 1막에서 더욱 분명하다는 것을 알 수 있다. 작가는 1막 지문에서 가장 다양한 어휘를 보여 준 반면에, 1막의 대사에서 가장 빈약한 어휘를 구사하였다. 또한 Métro의 지문이 가장 불규칙한 어휘 분포를 보인 것은 1막에서의 어휘의 풍부함과 2막에서의 빈약함으로부터 오는 불균형에 기인한 것이다.

이러한 현상은 1막의 지문에 사용된 어휘가 2막의 지문에서 광범위하게 되풀이되는 현상에서 오는 것일까? 이에 대한 대답은 각 막 사이의 어휘수 증가분이 정상인지를 테스트하여 찾을 수 있으며, 막 사이의 어휘수 증가분을 계산하는 방법을 여전히 Métro를 통해 살펴보겠다.

3.3 Métro 1막과 2막의 어휘 증가분

Métro의 1막에서 어휘수는 799이고 전 작품의 어휘수는 1,549이다. 따라서 1막에서 2막으로의 어휘의 증가분은 750이다. 그러나 이 숫자는 직접적으로 막 사이의 어휘분포를 파악하는 데 사용될 수 없다. 왜냐하면 2막의 길이가 훨씬 길기 때문이다. 따라서 막 사이의 어휘 증가분에 대한 이론값을 구해야 한다. 확률[11] p는 0.3361(2834/8433)이고 q는 0.6639이다. 계산하면 Vo, 즉 이론적 증가분은 727이고 실제는 750이다. 이 차이는 통계학적으로 의미가 없다. 즉 증가한 어휘수가 정상인 것이다. 그러나 대사와 지문을 구분하여 같은 테스트를 하면, 대사의 경우 증가는 실제가 이론보다 훨씬 더 크다.

이 결과는 Métro의 대사에서 1막과 2막의 상호침투가 이론적 모델보다 훨씬 작다는 것을 말해준다. 그것은 2막의 독립성을 보여준다. 사실, 이 작품에서, 두 주인공을 제외하면 1막에 나오는 등장인물들이 다시 2막에 나타나는 경우가 없다. 게다가, 이 등장인물들의 대사는 진정한 의미의 대화가 아니다. 대화란 일반적으로 화자와 청자의 언어적 교환이다. 그리고 발화는 하나이건 여럿이건 상대방에게 의사를 전달하기 위해 이루어진다. 그러나 그들은 직접적으로 청자와의 의사소통을 전제하지 않는 듯 말을 던진다. 말의 파편들은 그 수신자

11) "한 어휘가 이론적으로 1막에 나올 확률을 p라고 하면 q는 그 어휘가 2막에 나오지 않을 확률이다. 그러나 비교 대상이 이 두 막밖에 없고 또 이 두 막 사이에는 순서가 있음에 따라서 q는 오직 2막에만 나오게 되는 확률이 된다. 어휘들이 나타날 가능한 경우는 (o, o), (o, ×), (×, o), (×, ×), 네 경우이다. 여기서 'o'는 어휘가 나타날 경우를 상징하고 '×'는 나타나지 않을 경우를 상징한다. 1막에서도 2막에서도 안나타나는 어휘는 조사할 수 없으므로 네 번째 경우는 배제된다. 첫 번째와 두 번째 경우는 p로 나타날 것이고 세 번째 경우가 q이다."(배희숙, 2000: 89)

에게 드리워지지 않고 마치 메아리처럼 허공을 향해 던져진다. 그들의 대사는 "벽처럼 단단한", 그리고 "바다처럼 망망한" 개인의 세계를 강하게 반영한다. 사실, 주인공을 제외한 나머지 인물들은 사랑을 찾아 다가가려는 Lui를 방해하는 장애물이다. Lui에게 있어 그들은 벽처럼 단단하여 도저히 그들을 뚫고 그녀에게로 다가가는 것이 불가능해 보일 정도로 단단한 장애물이다. 그 개인적 세계는 "un plus un(하나 더하기 하나)"의 합일 뿐이어서 그들 하나 하나와 대화함으로써 해체될 수 있는 것이다.

대사에 비해 지문에서는 막 사이의 공통부분이 큰 편이다. 이는 제스처를 지시하기 위해 사용되는 단어들이 다소 되풀이되는 경향이 있기 때문일 것이다. 지문의 경우, 테스트의 결과는 $z = 0.7895$이다. 실제 관찰된 수치와 이론적 수치의 차이는 단 12이다. 통계표에 따르면, 이 z에 해당하는 확률은 기각수준을 넘지 않는다. 어쨌든 지문에서, 막 사이의 어휘분포는 대략 정상적이다.

Métro에 대해 행해진 막 사이의 어휘증가분 조사 결과는 다음과 같이 요약된다.

1) 대사와 지문을 통합하여 실시한 테스트에서, 1막에서 2막으로의 어휘 증가는 전적으로 모델에 부합되었다. 그러나 이 결과는 상대적으로 풍부한 어휘를 갖는 2막의 대사와 1막의 지문이 빈약한 어휘의 1막 대사와 2막 지문을 보충하기 때문이다.
2) 대사와 지문을 구분하여 실시한 테스트에서, 대사의 어휘수는 1막에서 2막으로 비정상적으로 증가되었으나, 지문에서는 그 증가치가 정상적이었다. 그것은 2막의 테마가 두 주인공 외의 다른 등장인물에 의해 결정되기 때문이다.

지금까지, 직접적으로 두 텍스트를 비교하기도 하고, 간접적으로 모델과 비교하기도 하면서, Métro와 Abc에 어휘가 풍부하게 사용되었는지 혹은 빈약하게 사용되었는지 비교하고 분석하는 방법을 소개하였다. 아울러 통계의 결과를 어떻게 해석해야 하는지도 설명하였다. 계량언어학의 다양한 활용 사례는 소개하지 못했으나, 통계학이 어떻게 언어학에 접목되는지 구체적으로 보여주기 위해 계산 방법을 비교적 상세하게 제시하였다. 이를 통해 언어 단위의 빈도수에 입각한 계량적 연구가 텍스트 분석에 대한 또 하나의 방향을 제시할 수 있다는 것도 알 수 있었다.

4. 결 론

텍스트의 어휘분포 연구에 통계학이 어떻게 활용될 수 있는지를 텍스트의 자료화와 그 활용이라는 두 단계로 나누어 간단하게 소개하였다. 본고에서는 문학작품을 텍스트로 사용한 예를 보여주었지만, 문학작품 외의 다양한 담론들, 예를 들면, 심리 상담문, 정치담화문, 신문이나 잡지 등을 텍스트군으로 하여 각종 언어 현상을 연구할 수 있다. 특히 이들 텍스트군에서 어휘 통계를 통한 테마어(mot-thème)나 핵심어(mot-clé) 찾기는 문학적 연구뿐만 아니라 한국어 정보처리에서 매우 효과적이고 필요한 연구이다. 통계학을 활용한 언어학적 작업의 또다른 예는 과도기적 양태를 보이는 언어 현상들[12]을 앙케이트를 통해 조사하거나, 한국어 대용량 말뭉치 구현과 이를 통한 한국어 개별 음소의 빈도수 구하기, 자음과 모음에

대해서 조합형을 구성하여 2음소 3음소의 확률 구하기 등이 있다. 이러한 통계작업의 결과가 있어야 확률모델에 의한 철자 오류 정정이나 잘못된 음운의 수복에 대한 작업도 제대로 이루어질 수 있는 것이다.

이러한 작업이 요구하는 정밀성을 충족시키고 통계처리의 정확성을 높이기 위해서는 방대한 양을 다루어야 하는데 이러한 다량의 자료를 처리하기 위해서, 컴퓨터의 활용은 필수적이다. 그러기 위해서, 즉 이러한 방법론이 제 기능을 다 할 수 있기 위해서는, 언어학자와 전산학자 상호간에 긴밀한 연계 작업이 활발히 이루어져야 하며, 언어의 체계화를 위한 규칙 정립에 많은 노력을 기울여야 할 것이다.

참 고 문 헌

[1] BAE Hee-Sook. 1997. *Structures lexicales, syntaxiques et phonétiques dans deux pièces de J. Tardieu*, Thèse de Doctorat, Strasbourg.
[2] 배희숙. 2000. "문체의 지수를 찾아서", 『프랑스문화예술연구』 제2집, 83-99.
[3] E. Buyssens. 1975. *Les catégories grammaticales du français*, Bruxelles Ed. de l'Université, 94.
[4] D. Cressels. 1979. *Unités et Catégories grammaticales: Réflexion sur les fondements d'une théorie générale des descriptions grammaticales*, Grenoble, PUF, 210.
[5] M. Grevisse & A. Goosse. 1993. *Le bon Usage, Gramaire française*, 13e éd. Paris, Duculot.
[6] G. Moigent. 1989. "Le verbe voici/voilà", in *TraLiLi*, vol. VII, 189-202.
[7] Ch. Muller. 1992. *Initiation aux méthodes de la statistique linguistique*, Paris, Champion.
[8] M. Pernier. 1986. *Le mot*, Paris, PUF, 126.
[9] M. Hug. 1989. *Structures du syntagme nominal français*, Paris-Genève, Champion -Slatkine, 496.
[10] M. Riegel, J.C. Pellat, R. Rioul. 1994. *Grammaire méthodique du français*, Paris, PUF.

접수일자: 2000. 7. 28.
게재결정: 2000. 9. 3.

▲ 한국과학기술원 인공지능센터 음성언어연구실
대전시 유성구 어은동 한빛 아파트 101동 202호(우: 305-333)
Tel : +82-42-861-2965 (H), +82-42-869-8720 (O)
H/P : 016-361-2965
e-mail : hsbae@bulsai.kaist.ac.k

12) 프랑스어의 경우 단어 *donc*의 발음이 그 예가 될 수 있다.