통신언어의 일탈도 측정에 대한 탐색적 연구

A Research Note on Measuring the Delinquency Index of Internet Communication Language

김 재 준 (Kim, Jai-june) *
(E-mail: jjkim@kookmin.ac.kr)
김 바 우 (Kim, Ba-woo) **
(E-mail: grusin@empal.com)
김 재 범 (Kim, Jae-beom) ***
(E-mail: dreamie@empal.com)

논문접수일 : 2008년 11월 10일
논문심사일 : 2008년 11월 20일
게재확정일 : 2008년 12월 26일

* 학위취득대학: Princeton Univ.
현직: 국민대학교 경제학부 부교수
** 학위취득대학:
현직: 사이버사회과학연구소 연구원
*** 학위취득대학: Manchester Business School
현직: 성균관대학교 경영학부 부교수

통신언어의 일탈도 측정에 대한 탐색적 연구

<국문요약>

본 논문은 인터넷 통신언어, 특히 댓글을 이용하여 네티즌의 성향을 분석한 탐색적 연구이다. 본 연구에서는 댓글 작성에 투입되는 노력의 정도를 측정하기 위하여 바이트의 개념에서 출발하여 비규범적 언어의 사용을 걸러내는 측정(measurement)방법을 개발하여 적용하였다. 일탈성의 정도를 측정하기 위한 척도를 개발하기 위하여 규범적 문장의 기준을 산정하였고, 이를 바탕으로 세 가지 일탈도를 만들었다. 통신언어(CMC)가 출판언어와 어떻게 다른 방식으로 감성을 표현하는 지를 구체적인 수치로 계량화 하였다. 동 개념에 대한 탐색적 연구로서 사이버 정치에 이를 적용하여, 2007년 이회창 후보의 제17대 대선 출마에 대한 최다 댓글 텍스트를 분석하였다. 아이디 사용이 다른 집단 사이에 다른 일탈도를 보인다는 것을 보여 주었다. 본 논문은 일탈도의 보완과 발전을 통해 텍스트 분석에 기여할 수 있는 방법론의 개발에 기여하고자 한다.

[주제어] 댓글, 통신언어, CMC, 일탈, 텍스트 분석, 사이버정치

I. 서론

21세기는 기술과 감성이 공존하는 시대이다. 인간은 역사상 어느 때 보다도 뛰어난 기술을 개발해내었고, 그 기술을 감성을 다양하게 표현하는 도구로 사용하고 있다. 기술의 진보는 동일한 노동의 투입으로 더 많은 생산물을 얻게함으로써, 더 많은 여가를 소비할 수 있게 해주었고, 더 나아가 개인의 표현 방식과 더불어 예술의 영역도 확장시켜주었다. 특히 정보기술의 발전은 개개인의 생각을 표현하고 공유하는 기회비용을 획기적으로 낮춰주었고, 이는 삶의 질의 향상에도 긍정적인 영향을 주고 있다(한국전산원 1996). 다양한 연령층의 개인들이 상당한 정보기술을 체험함으로써 삶의 질이 향상되는 것은 분명하지만(이상주・배경희 2004; 조혜영 2004), 가상세계와 현실세계가 상호 충돌하는 측면도 부정할 수 없다.[1] 인터넷 공동체(internet community)에 대한 연구는 1990년대 중반 이후 꾸준히 진행되어 왔는데, 특히 최근 연구로는 전자이웃(e-neighbors)을 연구한 햄톤Hampton 2007)과 컴퓨터매개커뮤니케이션(Computer Mediated Communication) 환경에서의 공동체 연구들을 정리한 브로크맨(Bruckman 2006)등을 들 수 있다.

인터넷 커뮤니케이션의 주체는 자신이 아닌 또 다른 자아이다. 이 자아는 사이버 세계 속에 구현된 자신을 대신할 분신(avatar)이며 이것은 영상, 문자 혹은 기호 등으로 설정된다. 또 다른 나의 감정 표현은 현실 속의 커뮤니케이션(Face to Face communication)과는 다른 측면이 있다.[2] 따라서 현실의 자아와 분신과의 괴리에서 오는 인터넷 중독현상, 외계어 등의 통신언어를 통한 언어의 왜곡 현상으로 오는 폐해등이 발생하게 되는 것이다.

통신언어의 사회방언적 성질에 대한 논의는 차치하고서라도 과도

1) 수업시간에 문자 보내기, 친구와 전화하며 이메일 체크하기는 일상적인 것이 되었다(Sherry Turkle ,2007).
2) Donath, J. (1999) 참조.

한 통신언어로 인하여 사이버공간에서의 언어파괴가 이루어지는 현상을 볼 수 있다. 특히 언어파괴현상은 사이버공간의 익명성(anonymousness)과 결부되어 일부 개인이 다수에게 부정적인 영향을 끼치는 표현이나 행동을 하더라도 이에 대하여 명확한 가치판단을 하지 못하는 경우가 허다하다.

우리말에 있어서 통신언어의 정의는 관점에 따라 다양한 정의가 가능하지만, 현재까지의 논의를 살펴보면, 형식적인 측면에서 이정복(2000)은 "PC통신, 인터넷, 휴대폰 문자메세지등의 전자매체에서 문자로 표현되는 언어를 포괄하는 것"으로 정의하였으며, 신희삼(2004, 207-225)은 "화자와 청자가 발화 구성에 미치는 대화지향적인 입말이며, 메시지는 문자로 쓰여졌지만 입말로서 전달되도록 의도하고 쓰여진 입말 중심의 글말"이라 어휘적 의미 측면에서의 통신언어를 정의하였다. 한편, 사용공간의 측면에서 통신언어를 정의하면 광의의 통신언어는 PC통신, 인터넷, 휴대폰 문자메세지에서 쓰이는 문자언어 뿐 아니라 음성과 모든 사이버공간에서 일어나는 의사소통 행위에 사용된 방법을 포괄하는 것으로 볼 수 있다.

통신언어를 분류하는 기준은 학자마다 상이한 견해를 가지고 있지만, 시기에 따른 분류를 하면 1980년대 후반부터 1990년대 후반까지 PC통신 등에서 사용하던 통신언어, 2002년이후 인터넷 사이트 디씨인사이드(DC Inside)를 중심으로 발생한 이모티콘, 그리고 1990년대 후반부터 소수의 집단에서 강한 유대감을 느끼기 위해 사용하기 시작한 외계어로 나뉠 수 있다(강옥미 2004). 특히 본 연구의 분석 자료에 가장 높은 비중을 둔 통신언어는 이른바 두문자어[3](acronym)에 해당하는 통신언어인데, 이는 통신언어의 분석의 대상이 된 자료들에서 빈번하게 관측되는 통신언어의 하위문화이다.

통신언어의 사용이유와 관련하여, 이정복(2005, 37-79)은 통신언

3) 약성어(略成語), 두문자어(頭文字語)로 칭한다.

어의 기능을 자판의 글자 입력을 빠르게 함으로써 시간과 노력을 줄이려는 경제성, 문자로 전달하려는 표현성, 글자를 바꾸어 수수께끼나 암호와 같이 만들고 다시 해독하는 데서 재미를 느끼는 오락성, 통신 이용자들이 기호를 사용함으로써 동료 의식을 느끼는 유대성, 기존의 사회 규범에서 벗어남으로써 심리적 해방감을 느끼기 위해 의도적이고 적극적으로 언어 규범을 어기는 일탈성으로 분류하였다. 그리고 이러한 과정에서 통신 언어는 일상 언어에서 나타낼 수 없는 어휘를 포함하게 되었다. 특히 본 연구에서는 두문자어, 단자음과 단모음의 두드러진 사용과 이모티콘의 사용을 통하여 통신 언어의 형태가 부분적으로 규범 언어의 틀에서 벗어난 모습을 하고 있다는 점에 주목하게 되었고, 이 점에 착안하여 본 논문은 인터넷 댓글로 나타난 네티즌의 성향을 "텍스트의 분해 기법"을 통해 연구하였다.

기존의 인터넷 텍스트 분석[4]에 대한 연구들은 주로 인터넷 언어의 사용동기와 사례 위주로 이루어졌고, 계량적 측면에서 접근한 논문은 이모티콘 사용의 성별차이(Wolf 2000, 827-833)등 매우 제한적이다. 한국에서 또 하나 주목할 만한 현상은 댓글 달기와 이를 둘러싼 사회적 파장이다.[5] 댓글은 작성자에 의해 창작된 글이며, 댓글이라는 텍스트를 통하여 작성자의 성향과 작성 목적을 탐구하는 것은 인터넷상의 창작 활동 성향을 연구하는 것과도 관련이 있다. 광의의 UCC에서는 뉴스 기사에 달린 댓글부터 UCC에 속한다고 볼 수 있다(이영재 2006). 최근의 UCC에 관한 연구(성윤택 외 2007, 70-111)에서는 UCC의 동영상 콘텐츠와 댓글의 상관관계를 연구한바 있다.

통신언어가 일상생활에 미치는 영향에 관한 실증적 분석의 결과

4) 이모티콘을 포함한 통신언어에 대해서 Walther and D'Addario(2001, 324-347) Rivera, Cooke, Bauhs(1996, 99-100) 참조.

5) 2007년 한국대선에서 댓글을 통한 선거 캠페인은 소위 알바 논쟁으로까지 비화되었다.

는 상반되고 있다. PC통신 나우누리 사용자 4,681명을 대상으로 한 2000년의 설문조사에서, '통신 언어가 일상 언어에 영향을 주느냐?'는 질문에 대하여, 68.72%는 '그렇다'고 응답했고, 27.13%는 '아니다', 4.14%는 '잘 모르겠다'고 응답하였다(권연진 2000, 5-27).[6] 오히려, 5년이 경과한 2005년에 초등학생 297명을 대상으로 실시된 설문조사(신승용 2005, 75-102)에서는 '채팅에서 사용하는 말 때문에 국어 시험을 잘못 봤다고 생각한 적이 있는가?'라는 설문에 대하여 전체의 93.7%가 영향이 없다고 응답하여 대조를 이루고 있다.

한편, 통신언어의 사회방언으로써의 역할과 방향에 대해서 통신언어의 언어적 존재감은 다차원적이고, 통신언어의 경우 문자언어보다 이탈/몰입요인에 의해 강한 언어적 존재감을 느끼므로 문자언어와는 다른 역할을 하고 있다는 견해(김봉섭・이인희 2007)는 통신언어가 일상생활에 크게 영향을 미치지 않는다는 견해를 뒷받침하고 있다. 본 연구에서는 통신언어가 일상생활에 미치는 영향력보다는 현실적으로 나타난 사실(fact)의 실증적 분석에 집중하였다.

한편, 한글의 통신언어는 인터넷 이모티콘에서도 미국, 유럽과는 다른 이모티콘이 사용되는 등 독자적인 시스템이라고 할 수 있다. 그리고 통신언어의 경우 문자 파괴의 현상이 더 빈번하게 관찰되고 있는 것을 주목하였다. 이런 관점에서 본 연구에서는 인터넷 사용자의 언어 사용 현상에 대한 새로운 척도, 즉 일탈도(delinquency index)와 엄숙도(seriousness index)를 규정해 보고 이를 통해 네티즌의 자유롭고 비규범적 언어 사용의 정도를 측정하였다.

이전의 통신언어에 대한 실증적 연구는 대부분 통신언어에서 나타나는 언어파괴현상을 구별하는데 그쳤다. 즉, 통신언어의 언어 파괴현상을 시기나 종류에 따라 정의하였으나, 적당한 척도(measurement)가 개발되지 않아서 계량적 분석을 시도하기 어려웠다. 이에 본 논문은 언어파괴현상의 척도를 개발하였는데, 먼저 인터넷 언어로써 한

6) 재인용

글의 특이성에 집중하였다. 연구자들은 먼저 영어의 경우 컴퓨터로 입력할 때 개별 알파벳 단위로 입력이 되지만, 한글 입력 단계에서는 자음/모음 하나하나의 단위에서 초성/중성/종성이 통합된 새로운 체제로 바뀐다는 사실에 주목하였다. 자음과 모음으로만 구성된 분절된 단위의 경우 언어가 파괴된 두자어나 이모티콘의 형태로 볼 수 있으므로 일차적으로, 분절된 단위가 차지하는 비율을 언어파괴정도의 척도로 고려하였다. 아울러, 한글 띄어쓰기의 정규성(regularity)에 의거하여 의도적인 띄어쓰기의 생략이 존재하는 문장을 언어파괴현상 존재하는 문장으로 간주하여 그 정도를 계량화하였다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 음절의 숫자와 바이트를 이용한 텍스트 분석 방법을 소개하고, 단모음과 단자음, 이모티콘(emoticon)등의 제거를 통한 측정의 정확도를 높이는 방법을 논할 것이다. 제Ⅲ장에서는 이를 활용하여 만든 일탈도 와 엄숙도를 소개하고, 제Ⅳ장에서는 이를 실제상황에 적용하고, 각종 변수들이 일탈도와 글의 성향에 미치는 영향을 조사한다. 마지막으로 제Ⅴ절의 결론에서는 향후의 연구방향을 논의할 것이다.

Ⅱ. 음절공간과 바이트를 이용한 텍스트 분석

기존의 인터넷 텍스트 분석은 주로 인터넷 언어의 사회언어학적 연구(이정복 2006; 2007)나 음운, 표기의 변용에 대한 연구(이정복, 2002)등 국어학적 접근을 통해 이루어졌다. 본 연구에서는 텍스트 작성자의 노력과 태도에 초점을 맞추어서 작성자가 어떠한 노력을 통해 최적화된 행동의 결과로써 텍스트를 작성하였는지 연구하고, 그 노력의 정도를 측정하려 하였다. 컴퓨터상에서 텍스트의 입력-인식-저장은 다양한 방법으로 이루어질 수 있지만, 입력과정에서는 한글 자음/모음과 아스키 문자(ASCII-American Standard Code

for information interchange, 영어 알파벳, 숫자, 문장부호를 포함)만을 사용하여 입력한다. 텍스트 분석의 첫 단계는 텍스트가 몇 개의 음절 공간으로 이루어져있는 것인지를 분석하는 것인데, Microsoft사의 Excel의 함수를 이용하여 알아볼 수 있다.[7]

<그림1>

M1

	A	B	C	D	E
1	Navy Seal	=LEN(A1)	9	=LENB(A1)	9
2	강아지1234	=LEN(A2)	7	=LENB(A2)	10
3	강아지1234	=LEN(A3)	7	=LENB(A3)	10
4	눈오는날백구	=LEN(A4)	6	=LENB(A4)	12
5					

LEN(TEXT위치) 명령어는 해당 텍스트가 몇 개의 음절공간으로 이루어져있는지를 보여주고, LENB(TEXT위치) 명령어는 해당 텍스트가 몇 개의 바이트로 이루어져있는지를 보여준다. 바이트(byte)는 컴퓨터가 처리하는 정보의 기본단위로써 8개의 비트(bit)로 이루어져 있는데, 비트(bit)는 0과 1을 나타내는 기본단위를 뜻한다. 즉 2진법을 이용하는 컴퓨터의 1바이트는 $2^8 = 256$가지 정보를 표현할 수 있는데, 이는 26개의 영문과 10개의 숫자, 그 외의 문장기호들을 표현하는 데에는 무리가 없지만 자주 쓰이는 완성형 한글 2350자, 11172자에 이르는 현대 한글을 음절단위로 표현하기에는 턱없이 부족하다.

따라서 한글은 65536가지 정보를 표현할 수 있는 2바이트의 정보단위를 사용하여 표현하고, 한문역시 같은 방법을 통하여 표현되고 있다. 한글과 한문, 그리고 아스키문자의 범위를 벗어난 특수문자(■△▲▽▼→←등)의 입력에는 단자음, 단모음을 사용하지 않는 이상 한번 이상의 키보드 입력이 필요하다. 상기한 점을 고려할 때

7) 이후의 Excel 함수의 활용 과정에 대해서는 부록을 참고

바이트 수가 높은 텍스트의 경우 바이트 수가 낮은 텍스트보다 노력이 더 많이 들어 간 텍스트라고 볼 수 있다. 아울러, 같은 음절수로 이루어진 텍스트의 경우에도 한글을 비교적 많이 사용한 텍스트를 작성한 집단과 영어 알파벳, 숫자와 ASCII문자를 비교적 많이 사용한 집단간에 내용의 차이를 비교해 볼 수 있다.

그림1에 나타난 A1부터 A4까지의 텍스트는 2007년 11월 10일에 미디어 다음의 리플토론방에서 추출한 사용자 ID의 표본들이다. 이 표본들에서, '강아지1234'와 '눈오는날백구'를 비교해보면 '강아지1234'는 7개의 음절로, '눈오는날백구'는 6개의 음절로 이루어져있지만 전자의 경우 11번의 키보드 입력으로, 후자의 경우 16번의 키보드 입력으로 작성된다. 즉 단순히 공간(음절)을 사용한 노력의 측정에는 오차가 존재하며 그 오차는 2바이트 언어와 1바이트 언어를 구분하여 측정하는 것으로 줄일 수 있다.

오차의 조정을 위해 단모음과 단자음, 이모티콘을 포함하여 측정할 경우에 대해 살펴보면, 바이트 단위로 노력을 측정할 때의 문제점은 문장의 구성을 위해 사용된 1바이트와 의도적으로 자리를 차지하거나 습관적으로 쓰인 1바이트의 차이점을 구분할 수 없다는 것이다.

<표 1> 문자수와 바이트 문자수의 비교

문장	총 문자수	2바이트 문자수	1바이트 문자수
안녕? 멋진A반 친구들, 반가워!	18	11	7
글쎄=_=ㅋ별루안방가운데요;;;;	18	11	7

예컨대, <표1>의 첫번째 문장의 경우 1바이트 문자는 문장의 형식을 맞추거나 내용의 전달을 위해 사용되었지만, 두번째 문장의 경우 =_= 등은 감정을 표현하기 위한 이모티콘으로써 사용되었고,

연속된 ;;;; 역시 이모티콘으로 해석될 수 있다. 즉, 동일한 횟수로 2바이트 문자와 1바이트 문자가 사용되었더라도 그 의미의 해석에 있어서는 상이하게 접근해야 한다. 즉 바이트만으로 텍스트의 성향을 파악하는 것은 기본적 노력을 측정할 수는 있지만, 성향을 분석하기 위해서는 보완이 필요하다.

특히 최근 들어 단자음과 단모음, 그리고 이모티콘을 이용한 표현이 현저하게 증가하고 있는데, 먼저 이모티콘이란 이모션(emotion)과 아이콘(icon)의 합성어인데 컴퓨터 자판의 문자, 기호, 숫자 등을 적절하게 활용하여 미세한 감정이나 사물을 나타내는데 사용되는 일종의 언어이다.

이모티콘은 1982년 9월 당시 카네기 멜론대학의 학생이었던 스콧 펄만(Scott Fahlman)에 의해 처음 사용되었다. 당시 사용된 최초의 이모티콘은 :-) 이었지만, 그 후 몇 가지 표정의 이모티콘을 중심으로 종류가 다양해졌다. 이모티콘은 통신언어의 특수한 형태에 속하며, 그 기능으로는 자판의 글자를 빠르게 입력함으로써 시간과 노력을 줄이려는 경제성, 문자로 전달하려는 표현성, 글자를 바꾸어 수수께끼나 암호로 만들어 해독하는 데서 재미를 느끼는 오락성, 통신 이용자들간에 기호를 사용함으로써 동료 의식을 느끼는 유대성, 기존의 사회 규범에서 벗어나 심리적 해방감을 느끼기 위해 의도적 적극적으로 언어 규범을 어기는 일탈성 등을 들 수 있다 (이정복 2002).

다음과 같이 이모티콘과 단자음/모음은 감정의 표현으로서 이성적 표현의 반대급부로 사용된다고 이해할 수 있다. 따라서 한 문장 내의 이모티콘과 단자음/모음은 진지하게 글의 내용을 위해 사용되었다고 간주하기 어렵다. 반면에 이모티콘과 단자음/모음의 상대빈도는 작성자의 성향을 반영하고 있기에 연구할 가치가 있다고 판단하였다.

<표 2> 이모티콘의 활용 사례

즐거움을 나타내는 이모티콘	놀람을 나타내는 이모티콘	우울함을 나타내는 이모티콘
^^ ^_^ ^-^ ^_=	ㅇ.ㅇ 0.0 @.@	T_T ㅜ.ㅜ ㅠ.ㅠ
(^-^) (^.^) *^.^*	--; ㅡ..ㅡ -_ㅡ+	ㅜ..ㅜ =_= o(T^T)o
^ㅇ^ ^ㅅ^ ^ㅗ^	ㅡ..ㅡ耽 -_-+	(づ_T)
ㅇ(-▽-)o (^(oo)^)		

출처: 조상진, 2006, 인터넷 통신 언어의 연구.

<표 3> 단자음, 단모음의 활용 사례

단자음/단모음	의미	단자음/단모음	의미
ㅋㅋㅋ	크크 킥킥	ㄷㄷㄷ	덜덜덜
ㅎㅎㅎ	하하 호호	ㄴㄴㄴ	No, No(부정의 의미)
ㅉㅉㅉ	쯧쯧	ㅌㅌㅌ	텨텨(튀어 튀어)
ㅊㅊㅊ	축축(축하)	ㅠㅠ	슬픔, 우울함
ㄱㄱㄱ	고고(go go)	ㅇㅇ	응응(긍정)
ㅎㅇ	하이	ㅊㅋ	추카(축하)
ㅎㅇ	하악(경악)	ㅅㄱ	수고

출처: 조상진, 2006, 인터넷 통신 언어의 연구.

그 방법으로써 이모티콘이 차지하는 바이트와 단모음, 단자음이 차지하는 바이트를 측정하였는데, 이렇게 계산된 바이트를 전체 바이트에서 제거하면 순수하게 문장의 내용에 쓰인 바이트를 산출할 수 있다. 한편, 원하는 문자의 숫자를 세기 위해서는 텍스트 분석 프로그램이 필요하지만, 부록에 간단히 마이크로소프트사의 엑셀 함수를 이용하여 분석할 수 있는 방법을 소개하였다.

일련의 과정을 통해 문자열을 한 글자씩 분해한 뒤에, 단자음, 단모음, 이모티콘에 해당하는 바이트를 제외하면 순수하게 의미전

달을 위해 사용된 바이트를 산출할 수 있다. 여기서 계산된 바이트는 다음과 같이 표시할 수 있다.

<식1>

$$net\,byte = total\,byte - (emoticon\text{에 사용된 }byte + \text{단자음,모음에 사용된 }byte)$$

즉 순 바이트(net byte)의 비율이 높은 경우 더 진지한 태도로 글을 작성했다고 생각할 수 있다. 하지만 여전히 문장부호의 과도한 사용을 어떻게 판별해낼지에 대한 과제가 남아있다. 표정을 나타내는 이모티콘 -.- 의 경우 - 와 . 는 실제 문장에서 쓰이는 문장부호이지만, -.- 가 아닌 -..- 등의 응용된 형태를 일일이 조사하여 계산에서 제하는 것은 용이하지 않다. 게다가 어느 정도 문장부호를 사용하는 것이 진지한 것인지를 알려주는 객관적 기준도 존재하지 않는다. 본 연구에서는 그 대안으로 진지하게 작성되었다고 널리 받아들여질 수 있는 3대 종합일간지(조선, 중앙, 동아일보)와 이들 일간지와는 이념적으로 대척점에 있는 한겨레신문의 사설문장들을 표본으로 추출하여, 이 문장들에서 쓰인 문장부호의 글자당 평균빈도를 기준으로 하였다. 하지만 이 값들의 절대치를 따르는 것은 기대하기 어려운 만큼, 각각의 수치에 정규분포를 적용하여, 댓글에서 사용된 문장부호 빈도의 95% 신뢰구간을 조사하였다. 실제 관측치와 표준화된 구간을 비교하여 관측치가 구간의 상한보다 더 클 경우 관측치와 95% 신뢰구간의 상한의 차, 즉 상한보다 큰 값만을 진지하지 않은 문장부호의 값으로 받아들였다. 동일한 방법으로, 관측된 문장부호의 빈도가 진지한 문장에서 쓰인 문장부호의 빈도의 하한 보다 작은 경우 하한과 관측치의 차, 즉 하한보다 작은 값 역시 진지하지 않게 사용된 문장부호로 간주하였다.

<표 4> 진지한 텍스트 1000글자의 문장기호 사용 빈도

문장부호	구독점[,]	마침표[.]	물음표[?]	인용부호[“ , ”]	인용부호[‘ , ’]
평균빈도	3.78	19.20	0.02	0.99	0.69
문장부호	물결표[~]	퍼센트[%]	느낌표[!]	하이픈[-]	묶음부호[(,)]
평균빈도	0.11	0.58	0	0.28	5.54

출처: 조선, 중앙, 동아, 한겨레 2008년 4월 8일-15일자 사설 63788 음절의 평균값.

<표 5> 진지한 텍스트 1000글자의 문장기호 사용 빈도의 95% 신뢰구간

문장부호	구독점[,]	마침표[.]	물음표[?]	인용부호[“ , ”]	인용부호[‘ , ’]
상한	12.11	33.67	0.47	8.92	5.51
하한	0	4.53	0	0	0
문장부호	물결표[~]	퍼센트[%]	느낌표[!]	하이픈[-]	묶음부호[(,)]
상한	1.58	4.72	0	2.90	19.7
하한	0	0	0	0	0

출처: 표4와 동일

III. 일탈도와 엄숙도 지수의 개발

지금까지의 논의를 바탕으로 한 텍스트에서 얼마나 규범적이지 않은 언어를 사용하였는지를 측정하는 세 가지 ‘일탈도[8] (delinquency index)’들을 만들었고, 반대로 텍스트에 들어간 진지

한 노력의 정도인 '엄숙도(seriousness index)'들을 구하였다. 먼저 일탈도1의 계산식은 다음과 같다.

<식2>

$$deliquency\ index_1 = \frac{deliquent\ 2byte\ componants + deliquent\ 1byte\ componants}{Total\ byte}$$

$$= \frac{2(sinle\ consonant/vowel) + residual\ byte}{Total\ byte}$$

첫 번째 일탈도인 일탈도1(delinquency index1)은 단모음과 단자음, 그리고 숫자, 영문 알파벳, 문장부호, 띄어쓰기를 제외한 1바이트 아스키문자가 전체 바이트에서 사용되는 비중을 측정하였다. 일탈도1은 'ㅋㅋㅋ' 이나 ' :-) '등의 사용빈도가 글의 전체 길이에 비해 높을 경우 큰 숫자를 나타내고, 0부터 1까지의 값을 가진다. 여기서 사용된 잔여 바이트(residual byte)의 개념은 다음과 같다.

<식3>

$$residual byte = Total\ 1byte\ componants - byte\ of\ alphabets - byte\ of\ number - byte\ of\ pucntuations - byte\ of\ tab$$

잔여 바이트(residual byte)라는 명칭을 부여한 것은 전체 바이트 중 2바이트 문자(한글, 한문, 혹은 아스키 문자가 아닌 특수문자[9])가 차지하는 부분, 영어 알파벳이 차지하는 부분, 숫자와 문장부호

8) '일탈도'라는 것은 사회적으로 통용되는 어법에서 벗어난 정도를 측정하는 것으로, 청소년 비행의 정도를 측정하는 것과는 무관하다.

9) 아스키 코드에서 지원하는 특수문자는 키보드상에서 한 번의 입력으로 사용 가능하거나 Shift와의 조합을 통해 나타낼 수 있는 문자를 의미하고, 그 외의 특수문자는 보통 windows 환경에서 한글 단자음과 한자 키, 숫자키를 통해 사용 가능하다.

가 차지하는 부분이 아닌 & * $ 등 정규적으로 쓰이지 않는 아스키문자들이 차지하는 부분을 의미하기 때문이다.

<식4>

$$delinquency\ index_2 = \frac{\sum_{1}^{N} e_i}{Total\ byte}$$

두 번째 일탈도인 일탈도2(delinquency index2)는 문장부호의 비규범적 사용을 통계적으로 측정한 것이다. 일탈도2는 전체 바이트 중 초과 문장부호 바이트가 차지하는 비중을 측정하는 지수로 역시 0부터 1까지의 값을 가진다.[10]

여기서 사용된 초과 문장부호란 다음과 같다.

<식5>

$e_i = exceed punctuation,\ n_i = number\ of\ puctuation$

$s_i^* = avg\ puctuation\ of\ serious\ text,$

$s_{i\max}^* = \max[95\%\ confidence\ level]$

$n_{i\min}^* = \min[95\%\ confidence\ level]\ \ if\ \min < 0\ ,\ n_{i\min}^* = 0$

먼저 n_i는 관측된 각 문장부호의 숫자를 의미하고, s_i^* 는 3절에서 논의한 바와 같이 규범적인 문장의 63788음절을 대상으로 조사한 값을 기준으로 만들어낸 문장부호의 적정 수준이다.

여기서 주목할 것은 관측된 문장부호사용의 빈도가 통계적으로 동일한 길이로 작성된 규범적 문장의 그것과 상이한 경우, 그 차이를 단순하게 과부족으로 간주하기에는 무리가 있다는 것이다. 그러

10) 실제로 0.1이상의 값을 가지는 경우는 매우 드물다.

므로 95퍼센트 신뢰 구간의 상한 혹은 하한과의 차이를 초과된 문장부호의 사용으로 간주하여, 그 합이 전체 문장길이에서 차지하는 비중을 계산하였다. 기준 문장에서조차 낮은 빈도로 사용되는 문장기호의 경우 95% 신뢰구간의 하한이 음의 값을 가질 수 있는데, 이 경우 하한보다 낮은 빈도는 존재하지 않는다. 즉 음의 하한이더라도 그 값은 전혀 사용되지 않으므로 0으로 간주하였다.

<식6>

$$if\, e_i < n_i^*,\ e_i = n_i(number\ of\ puctuation) - n_{i\max}^*\, avg\ puctuation\ of\ serious\ text$$

$$if\, e_i > n_i^*,\ e_i = n_{i\min}^* - n_i$$

이렇게 구한 초과 문장부호는 다음과 같이 쓸 수도 있다.

<식7>

$$e_i = \left| n_i - n_i^* \right|,\ n_i^* = n_{i\max}^*\ if\ n_i > n_{i\max}^*,\ n_i^* = n_{i\min}^*\ if\ n_i < n_{i\min}^*$$

요약해 보자면, 지금까지 구한 일탈도는 두 가지 측면에서 계산되었다. 일탈도1(delinquency index1)은 단모음과 단자음, 그리고 숫자, 영문 알파벳, 문장부호, 띄어쓰기를 제외한 1바이트 아스키문자가 전체 바이트에서 사용되는 비중을 측정하였고, 일탈도2(delinquency index2)는 비규범적 문장부호의 사용 정도를 측정하였다. 그러므로 이 두 효과를 합하면 종합적인 비규범적 언어사용에 대한 정도를 측정할 수 있고, 이를 종합일탈도[11](total delinquency index)라 명명하였다.

11) 이하에서 index3로 표기.

<식8>

$$deliquency\ index_3 = deliquency\ index_1 + deliquency\ index_2$$

또한 일탈도1,2는 비규범적 문자의 사용과, 비규범적 문장부호의 사용 정도를 측정하므로, 1에서 일탈도1,2를 제하면 0에서 1사이의 값을 가지는 엄숙도(seriousness index)1,2를 구할 수 있다. 일탈도3의 경우 이론적으로 상한치가 2이므로, 2에서 일탈도를 제하고 2로 나누어주면 다른 엄숙도와 같은 범위를 가지도록 조정이 가능하다. 이렇게 계산된 엄숙도들는 0에서 1의 값을 가지며, 1에 가까울수록 정형화된 텍스트를 사용하는 사용자를 의미한다.

<식9>

$$seriousness\ index_1 = 1 - deliquency\ index_1$$

$$seriousness\ index_2 = 1 - deliquency\ index_2$$

$$seriousness\ index_3 = \frac{2 - deliquency\ index_3}{2}$$

Ⅳ. 지수의 적용 사례 연구

1. 주제와 연령에 따른 게시판의 댓글 성향 비교

이상에서 개발한 일탈도라는 지수를 인터넷 상에 적용해 보았다. 본 논문은 이 지수의 개발과 측정 방법에 대한 논문이기에 다른 성격의 집단들이 왜 다른 일탈도를 보이는가를 이론적으로 규명하는 것이 목표라고 할 수는 없다. 가령 국가투명성 지수의 개발을 예로 들어 보자. 국가 간의 부패의 정도를 숫자로 알아 볼 수 있게 되어 국제기구나 기업경영에 유용하게 쓰이는 지수라고 할 수 있다. 하지만 부패 지수가 높은 국가와 부패지수가 낮은 국가 사이에

왜 부패의 정도가 다른지는 별개의 문제라고 할 수 있고, 지수를 개발한 기관에 그에 대한 이론적 설명을 요구하지는 않는다. 마찬가지로 일탈도 라는 지수의 개발은 간편하게 인터넷 언어의 구사 방식에 대한 측정을 목적으로 이루어졌으며 영어 아이디를 쓰는 사람과 한국말 아이디를 쓰는 사람 간에 일탈도가 왜 다른가를 이론적으로 설명할 필요는 없다. 이는 또 다른 연구자 내지는 후속 연구의 몫으로 생각한다.

먼저 일탈도의 유효성을 검증하기 위해 상이한 연령대의 네티즌들이 사용하는 두 인터넷 게시판의 게시물에 달린 댓글을 텍스트 분석의 대상으로 삼아 분석하였다. 댓글의 정의는 다음과 같다. '댓글'은 하나의 게시물에 종속된 글로써, 게시물을 읽은 사람이 게시물에 대한 감상이나 의견 등을 추가로 적은 글이고, 관리자를 제외한 게시물의 작성자는 댓글에 대한 권한이 없다(정일권・김영석 2006). 즉 댓글은 게시물에 종속된 글이지만 그 속성은 댓글의 작성자에 의해서만 결정된다.

10대, 20대 사용자가 많은 디씨인사이드(dcinside.com)에 있는 6인조 여성 아이돌 그룹 소녀시대 갤러리(게시판)의 공지사항(2008/01/12; 2007/08/07일자)에 달린 댓글544개와, 연령대가 다소 다양하기는 하지만, 상대적으로 보다 높은 연령층들이 활동하는 다음의 리플토론방에 2007년 11월 2일자 주제인 '이회창 후보의 대선 출마, 어떻게 생각하세요?'에 달린 댓글 6561개, 그리고 일탈도를 위한 표본으로 사용한 신문사설묶음을 대상으로 일탈도 1,2의 기술통계량을 조사하였다.

<표 6>의 기술 통계량을 조사해 본 결과, 비규범적 문자의 사용을 측정한 일탈도1과 비규범적 문장부호의 사용을 측정한 일탈도2는 모두 사설묶음에서 가장 낮은 수치인 0.0001과 0.0029를 나타내었다. 이는 어느정도 예견했던 결과이나 그 차이가 매우 크다는 점은 음미해 볼 만하다. 정치 토론방에서는 일탈도2(0.0164)가 일탈도1(0.0139)보다 비교적 높게 나타났음을 주목할 수 있는데, 그 원인

을 살펴보면 정치 토론방에서는 잦은 구두점이나 말줄임표의 사용이 현저했기 때문이다. 반면, 소녀시대 갤러리에서는 일탈도 1(0.1068)이 매우 높게 나타나 다른 집단과의 차이가 현저함을 볼 수 있다. 그 원인을 살펴보면 소녀시대 갤러리에서는 단모음, 단자음의 사용이 두드러졌기 때문임을 알 수 있다.

<표 6> 게시판 댓글들의 일탈도 비교

집단	일탈도1				일탈도2			
	평균	표준편차	최대값	최소값	평균	표준편차	최대값	최소값
소녀시대	0.1068	0.2413	0.9744	0	0.0197	0.0707	0.8374	0
정치토론방	0.0139	0.0681	0.9565	0	0.0164	0.0425	0.8495	0
사설묶음	0.0001	0.0014	0.0202	0	0.0029	0.0042	0.0346	0

출처: 디씨인사이드, 다음, 네 개의 종합일간지 사이트

2. 이회창 후보의 17대 대선 출마 선언에 관한 댓글 분석

인터넷을 통한 여론 형성 더 나아가 인터넷을 통한 시민의 정치참여에 대한 기존 문헌들을 살펴보면, 그 개념부터 제도화 방안등을 고찰하였으며(김용철・윤성이 2005; 정일권・김영석 2006), 인터넷 참여에 대한 시각은 대개 긍정적 전망과 부정적 비관적 전망으로 나뉘어진다. 긍정론자들의 주장은 동원이론(mobilization theory)으로, 비관적 주장은 강화이론(reinforcement theory)으로 대표될 수 있다(조성대・정연정 2006, 29-62). 특히 긍정적인 입장에서, 온라인상에서의 토론 공간을 공론권의 확장과 숙의 민주주의의 실현의 측면에서 논의하여 왔다(김병철 2004). 특히, 공론권이란 시민들이 대화의 과정을 통해 자유롭고 자연스럽게 야론을 형성해 가

는 것으로 과거의 오프라인상에서의 측면만이 아니라 최근에는 온라인상에서의 공론권에 대해서도 논의가 진행되고 있다 (김병철 2004; 정일권・김영석 2006). 그러나 어떠한 계층이 참여를 하였는지를 차치하고라도 합리적이고 이성적인 심의적(deliberate)정치참여에서 표출적(expressive)정치참여로의 전환은 이미 한국에서 정형화된 사실(stylized fact)이다(이원태 2004). 그렇기에 인터넷 댓글을 통한 여론조사가 편의를 가질 수 있지만, 정치효능감이 높은 계층이 텔레비전 프로그램중에서도 참여 프로그램을 이용하였다는 뉴하겐(Newhagen 1994, 366-379)의 연구결과를 응용하면 컴퓨터 매개 정치참여 행동(박선희 1998)인 인터넷 댓글작성을 하는 집단은 비교적 정치효능감이 높은 계층이라 간주할 수 있고, 이들은 상당한 대표성을 지니고 있다고 할 수 있다.

김병철(2004)에 의하면 인터넷 뉴스 댓글의 상호작용성은 토론내용과 정방향의 상관관계를 지니고 있는데, 이러한 사실은 현대사회의 공론장은 이성적 측면과 탈이성적 측면이 존재하는 포스트 모던적 공간으로 새롭게 정립해야 한다는 견해[12]를 뒷받침하고 있다. 본 연구는 아이디의 패턴이 댓글 작성 패턴에 미치는 영향, 특정사실에 대한 찬/반 의견의 요인별 비교를 통하여 탈이성적 측면과 이성적 측면이 공존함을 조사하고자 하였다.

2007년 대선 관련 지지율 여론조사에 대한 국민적 관심이 고조된 바 있다. 하지만 유선전화와 휴대전화가 다른 여론조사 결과를 보여 주어 정확한 여론을 알기 어렵다는 지적도 상당하였다. 인터넷 댓글을 통한 의견은 그 의견이 편향되어 있다는 지적도 있으나 국민 여론 형성에 주요한 일부가 된 것도 사실이다.

우리는 대선 관련 인터넷 여론 조사와 관련하여 '이회창 전 한나라당 총재 출마 선언' 에 대하여 인터넷 포털사이트'다음'의 리플토론방에 7115 개의 리플이 달린 것에 주목하여 텍스트 분석을 통해

12) Dalgren(1995), Poster(1997) 참조

출마 지지와 반대의 여론을 조사해 보았다.

본 연구에서는 우선 사용자의 아이디의 언어적 구성이 댓글 작성의 성향에 어떠한 영향을 미치는지 독립 T 검정을 실시하였는데, 여기에서의 아이디는 다음과 같이 정의한다. 아이디란 어떠한 개인이 소통을 목적으로 하는 게시판, 혹은 포털 사이트에 들어가서 자신을 상대방에게 나타내는 이름을 의미한다. 즉, 아이디는 각 개인의 정체성(identity)을 나타내는 것이다. 접속 초기에 비밀번호와 입력하는 아이디가 아닌 별명 개념으로써의 아이디를 의미하고, 접속에 별다른 과정이 필요하지 않은 경우 별명으로써의 아이디와 접속할 때 입력하는 아이디는 동일할 수도 있다. 여기서의 아이디는 자신을 남에게 나타내는 수단으로 사용되기 때문에, 아이디의 구성과 댓글 작성의 성향은 상호 관계를 가지고 있다고 생각할 수 있다.

한국사회에서 영어의 비중은 세계 어떤 나라보다도 적지 않다. 심지어 최샛별(2003, 5-21)은 “한국사회에서 영어를 잘 한다는 것은 사회적 성공을 보장받을 수 있는 능력과 성품을 가지고 있다는 것”이라 평가하기도 하였다. 그러나 한국인에게 영어라는 외국어의 인식은 사람마다 다르다. 개개인의 언어 습득방법은 모두 독립적이지만 개인의 심리는 언어습득과정에 많은 부분 결정되고(Fodor, 1970), 같은 맥락에서 영어의 친숙도가 다른 두 집단은 다른 심리상태를 가지고 있다고 고려할 수 있기에 영어를 비교적 친숙하게 사용하는 집단(영어 알파벳이 아이디에 포함된 집단)과 그렇지 않은 집단을 구분하여 독립 t 검정을 실시하였다.

먼저 영어 알파벳이 아이디에 포함된 사용자의 경우 일탈도1에서만 유의한 차이를 보였는데, 영어 알파벳이 아이디에 포함된 사용자는 일탈도1이 더 낮은 경향을 보였다. 그 원인을 살펴보면 영어 알파벳을 쓰는 사용자의 경우 한글 단자음, 단모음을 통한 감정표현을 그렇지 않은 집단보다 덜 하기 때문으로 나타났다. 하지만 문장부호의 규범적 사용에 있어서는 유의하지 않은 차이를 보였다.

<표10> 영어 알파벳을 포함한 아이디에 의한 일탈도의 비교

일탈도1	평균	표준오차	표준편차
영어 알파벳을 아이디에 포함한 집단	.0159407	.0009185	.0652146
대조군	.0165113	.0019768	.0770462
차이	.0019943	-.0044801	-
t통계량	-0.2861	유의확률	0.6126
일탈도2	평균	표준오차	표준편차
영어 알파벳을 아이디에 포함한 집단	.0163322	.0006324	.0448975
대조군	.0138074	.00086	.0335189
차이	.0025248	.001245	-
t통계량	2.0280	유의확률	0.0213
일탈도3	평균	표준오차	표준편차
영어 알파벳을 아이디에 포함한 집단	.0322729	.0011192	.0794656
대조군	.0303187	.0021992	.0857138
차이	.0019542	.0023695	-
t통계량	0.8247	유의확률	0.2048

<표 11> 숫자가 포함된 아이디에 의한 일탈도의 비교

일탈도1	평균	표준오차	표준편차
숫자를 아이디에 포함한 집단	.016384	.0008769	.0693566
대조군	.009690	.001941	.0338986
차이	.006693	.0039946	-
t통계량	1.6755	유의확률	0.0469
일탈도2	평균	표준오차	표준편차
숫자를 아이디에 포함한 집단	.0157827	.0005448	.043086
대조군	.015026	.0016823	.0293796
차이	.0007567	.002495	-
t통계량	0.3033	유의확률	0.3808
일탈도3	평균	표준오차	표준편차
숫자를 아이디에 포함한 집단	.0321667	.00104	.0822523
대조군	.0247169	.0026605	.0464628
차이	.0074498	.0047465	-
t통계량	1.5695	유의확률	0.0583

숫자의 인지가 인간의 심리에 미치는 영향에 관해서는 국내에서 자세한 실험이 시도되지 않았으나, 숫자와 어휘표현에 있어서 인지

의 차이가 존재함이 실험을 통해 밝혀진 바 있다(안서원・도경수 2004, 299-316). 따라서 그 역방향으로의 상관관계 가능성역시 추론해 볼 수 있다.

숫자 아이디를 사용하는 집단의 경우 그렇지 않은 집단보다 일탈도2, 3이 낮은 것으로 조사되었는데, 이는 숫자 아이디를 사용하는 집단이 그렇지 않은 집단보다 규범적으로 행동하는 것을 말해준다. 숫자 아이디는 대부분 자신의 출생년도, 생년월일, 전화번호 등 주어진 환경을 표현하는 숫자와 자신의 이름, 이니셜, 지역 혹은 자신이 나타내고 싶은 고유명사의 조합으로 이루어져있다.

<표 12> 숫자가 포함된 아이디의 사례

이름 이니셜 / 생년월일 포함		표현하고 싶은 단어 / 숫자	
sls0822	tkstk90	kmysk50	twofish5
assa8914	gjh80	hytgo1004	art007
cky8585	parkjs3808	쉐인18	sik5733

숫자가 아이디에 포함된 많은 경우가 태어난 해를 포함하고 있고, 경우에 따라 사용하고 싶은 아이디를 다른 사용자가 선점한 경우 아이디에 숫자를 붙임으로써 만들어진 아이디도 자주 보인다(kim1, peter3등). 이는 숫자를 아이디에 사용하는 사람이 그렇지 않은 사람보다 사회 순응적 태도를 가지고 있음을 말해주고, 대조군과 비교하였을 때 유의한 수준에서 낮은 일탈도2, 3에서도 그러한 증거를 살펴볼 수 있었다.

<표 13> 영어 알파벳과 숫자가 들어간 아이디가
글 길이에 주는 영향

바이트로 측정한 글의길이(total byte)	평균	표준오차	표준편차
영어 알파벳을 아이디에 포함한 집단	175.266	2.085968	164.9762
대조군	153.9344	7.592283	132.5935
차이	21.3316	9.594365	-
t통계량	2.2233	유의확률	0.0131
바이트로 측정한 글의길이(total byte)	평균	표준오차	표준편차
숫자를 아이디에 포함한 집단	178.3347	2.413236	171.3398
대조군	160.7992	3.445463	134.2847
차이	17.53545	4.785883	-
t통계량	3.6640	유의확률	0.0001

아이디에 영어 알파벳과 숫자가 들어간 사용자들은 그렇지 않은 사용자들에 비해 더 짧은 글을 작성하는 것으로 나타났는데, 아이디에 숫자를 사용하는 집단은 그렇지 않은 집단에 비해 21.3바이트, 약12%가량 짧게 댓글을 쓰는 것으로, 아이디에 영어 알파벳을 사용하는 집단은 그렇지 않은 집단에 비해 17.5바이트, 약 10%가량 짧게 댓글을 작성하는 것으로 나타났다. 이를 종합하면 아이디에 영어 알파벳이나 숫자가 들어간 집단의 경우 그렇지 않은 집단보다 글을 더 짧게 쓰고, 대체적으로 단자음, 단모음의 사용이나 문장부호의 비정규적 사용이 상대적으로 낮다는 것을 알 수 있다.

Ⅴ. 결론

텍스트는 인간의 이성을 가장 잘 전달할 수 있는 도구이다. 그러나 감정 역시 가장 잘 표현할 수 있는 도구이다. 텍스트에 들어간 노력의 정도를 글의 길이, 바이트, 분절되지 않은 한글의 비중 등을 통해 측정하였고, 기준 지표로써 일탈도와 엄숙도를 만들었다.

그러나 띄어쓰기의 정도는 텍스트의 성격에 따라 큰 편차를 가지고 있기에 의도적인 띄어쓰기의 생략은 지금의 일탈도로는 측정하지 못하였고, 최근 인터넷에서 등장한 변형된 현대 한글 '뵑' 등에 대한 일탈성의 측정은 시간과 기술의 제약으로 접근하지 못하였다.

본 논문에서 개발한 일탈도라는 지수를 인터넷 상에 적용해 보았다. 본 논문은 이 지수의 개발과 측정 방법에 대한 논문이기에 다른 성격의 집단들이 왜 다른 일탈도를 보이는가를 이론적으로 규명하는 것이 목표라고 할 수는 없다. 그 이유를 국가투명성 지수의 개발을 예로 들어 본문에서 설명한 바 있다. 마찬가지로 일탈도라는 지수의 개발은 간편하게 인터넷 언어의 구사 방식에 대한 측정을 목적으로 이루어졌으며 영어 아이디를 쓰는 사람과 한국말 아이디를 쓰는 사람 간에 일탈도가 왜 다른가를 이론적으로 설명할 필요는 없었다. 하지만 어떠한 이론적 논의들이 학계에서 이루어졌는지에 대한 논의를 추가하였다. 이는 또 다른 연구자 내지는 후속 연구의 몫으로 생각한다.

사례 연구로써 디시 인사이드의 소녀시대 갤러리와 다음의 정치 게시판을 분석하였다. 특히 다음의 리플토론에 올라온 이회창 대선 후보의 17대 대선출마에 대한 댓글 분석에 있어서는 댓글 분석 결과가 기성 언론과 차이를 보이고 있지만, 중복된 글의 제거를 통하여 신뢰할 수 있는 샘플을 얻어낼 수 있었다. 또한 댓글의 길이에 영향을 주는 것으로 나타난 아이디의 숫자와 영문 포함 유무가 일탈도에도 영향을 미치는 것을 보임으로써 일탈도가 어떠한 통제변수에 의해 예측될 수 있는 가능성을 보였고 이에 대한 이론적 규명은 후속 연구의 몫이라고 할 수 있다.

인터넷 언어를 객관적 수치로 측정할 수 있는 방법론은 이제 시작이라고 생각한다. 한글이 총성, 중성, 종성으로 이루어진 특수한 체제인 것을 생각하면 서구의 이론이 아닌 한국식의 새로운 지수 개발이 이루어져야 한다. 특히 감성을 어떻게 객관적으로 측정할 수 있는가에 대해 더욱 다양한 시각에서 새로운 방법론이 개발될

필요가 있다고 생각한다. 이와 같은 노력을 통해서 대면조사, 전화조사 같은 기존의 방법 보다 더 정확하게 여론을 조사할 수 있는 새로운 수단이 개발될 가능성을 생각해 본다.

< 참고문헌 >

권연진, 2000, “컴퓨터 통신언어의 유형별 실태 및 바람직한 방안”, 『언어과학』, 제7권 2호.

김봉섭・이인희, 2007, “언어를 매개로 하는 실재감 구성요인 연구 : 통신언어와 문자언어를 중심으로”,『한국방송학보』, 제21-2호.

김용철・윤성이, 2005, “E-Governance 구축의 전략적 모색:정책결정과정의 관점에서”, 『한국정치학회보』, 제39-5호.

박선희, 1998, “시민적 관여가 매개 정치커뮤니케이션에 미치는 영향”, 서울대학교대학원 박사학위 논문.

성윤택・김영기・이세영, 2007, “동영상UCC 유형과 댓글에 관한 탐색적 연구 판도라TV를 중심으로”, 『사이버커뮤니케이션학보』, 통권 제 23호.

신승용, 2005, “사회방언으로서의 통신언어의 위상 재정립”, 『한민족어문학』, 제46호.

이상주・배경희, 2004, “인터넷 과잉 이용 학생과 보통 학생간의 자존감, 공격성, 우울 비교”,『청소년학연구』, 제11권 제3호.

신희삼, 2004, “인터넷 통신언어의 어휘적 의미”,『한국어 의미학』, 제15호.

안서원, 도경수, 2004, “불확실성에 대한 어휘 표현의 선별적 틀 효과”, 한국심리학회지 실험, 제16호.

이원태, 2004, “인터넷 정치참여에 관한 연구-2004년 한국의 제17대 총선정국을 중심으로”, 서강대학교 대학원 박사학위 논문.

이정복 외, 2000, 『바람직한 통신언어 확립을 위한 기초연구』, 문화관광부 연구보고서.

———, 2002, “전자편지 언어에 나타난 우리말 변용 현상”,『사회언어학』, 제10권 제1호.

———, 2005, “사회언어학으로 인터넷 통신 언어 분석하기-최근의 연구 현황과 과제”, 『한국어학』, 제27호.
———, 2006, “인터넷 통신 언어 자료에 나타난 대구 지역 고등학생들의 방언 사용 실태”,『우리말글』, 제38집.
———, 2007, “인터넷 통신 언어에서 보이는 방언 사용의 성별 차이”, 『어문학』, 제97집.
유승현・황상재, 2006, “포털미디어의 뉴스 프레임에 대한 탐색적 연구: 미디어다음, 조선일보, 한겨레신문의 비교를 중심으로”, 『사이버커뮤니케이션학보』, 통권 제 20호.
정일권・김영석 2006, “온라인미디어에서의 댓글이 여론에 미치는 영향에 관한 연구”, 한국언론학회 본 정기학술대회
조성대・정연정, 2006, “사이버커뮤니티와 정보접근, 그리고 정치참여: 17대 총선과정에 나타난 인터넷의 정치적 효과”, 『한국과 국제정치』, 제22권 제2호.
조혜영, 2004, “노인의 인터넷 사용과 삶의 질에 관한 연구”, 숙명여자대학교 대학원 석사학위 논문.
최샛별, 2003, “한국 사회에서의 영어실력에 대한 문화자본론적 고찰”, 사회과학연구논총, 제11호.
한국전산원, 1996, 『정보화와 삶의 질』, 서울: 한국전산원.
Bruckman, Amy, 2006, "A New Perspective on ‘Community’ and its Implications for Computer-Mediated Communication Systems", *In Proceedings of the 2006 ACM SIGCHI Conference on Human Factors in Computing Systems*, Extended Abstracts (pp. 616-621). Montréal, Québec.
Dahlgren, P., 1995, *Television and the Public Sphere*, Sage Publications.
Hampton, Keith, 2007, "Neighborhoods in the Network Society: The e-neighbors Study," *Information, Communication, and Society* 10(5).

Huffaker, D.A., S.L. Calvert, 2005, "Gender, identity, and language use in teenage blogs", *Journal of Computer-Mediated Communication* 10(2).

Jerry, Fodor, 1970, *Psychology and Language*, international society for rehabilitation of the disabled, New York, N.Y.

Newhagen, J. E., 1994, "Self-efficacy and call-in political television show use", *Communication Research* 21(3).

Poster, M., 1997, *Cyberdemocracy: Internet and Public Sphere in David Porter(ed.)*, Internet Culture, London:Routledge.

Rivera, K, N.J. Cooke, J.A. Bauhs, 1996, "The effects of emotional icons on remote communication" *Conference on Human Factors in Computing Systems* Vancouver, British Columbia, Canada.

Turkle, Sherry, 2007, "Can You Hear Me Now," *Forbes*, May 7, 2007 http://www.forbes.com/free_forbes/2007/0507/176.html

Walther, J.B, K.P, D'Addario, 2001, "The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication." *Social Science Computer Review* Vol. 19, No. 3.

Wolf, A., 2000, "Emotional Expression Online: Gender Differences in Emoticon Use." *CyberPsychology & Behavior* October 1, 3(5).

A Research Note on Measuring the Delinquency Index of Internet Communication Language

Kim, Jai-june (Kookmin University)
Kim, Ba-woo (CSSRI)
Kim, Jae-beom (Sungkyunkwan University)

Abstract

This study examines the language use patterns of the computer mediated communication(CMC) users. Comparing the CMC with the paper based text, a new concept of delinquency index(DI) is proposed. The DI measures how much the CMC is deviated from the newspaper text, considering the distinct features of the Korean "Hangeul" text such as the Asian emoticons and single syllable vowels and consonants. This index is applied to the text of the Korean Pop culture of DC-Inside and the Political discussion section of the internet portal site of Daum. The authors show that the different groups classified by the use of identification names reveal different levels of DI. Further research is needed to explain this behavioral differences.

Keywords : Reply, CMC language, Delinquency, Cyber politics, Text analysis

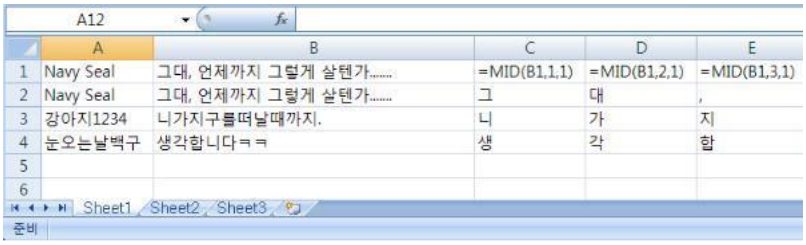
부 록

스프레드시트를 통한 텍스트 분석 기법

원하는 문자의 숫자를 세기 위해서는 텍스트 분석 프로그램이 필요하지만, 널리 보급된 Microsoft사의 Excel프로그램을 통한 방법을 소개하고자 한다.

먼저 긴 문자열을 한 글자씩 분해하는 작업을 수행하면 원하는 문자가 해당 문자열에 몇 개 있는지 조건식으로 계산할 수 있다.

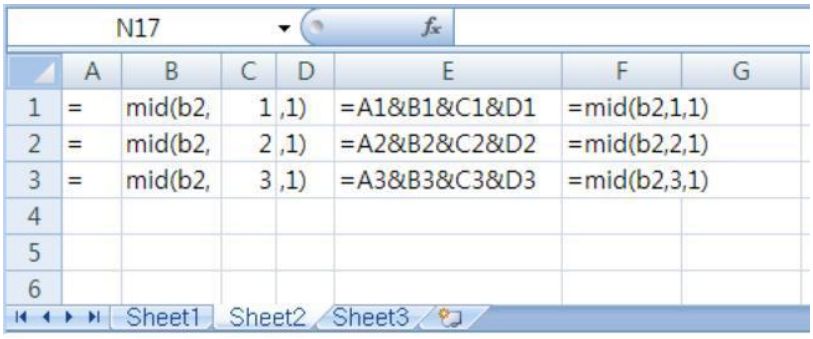
<그림1>



	A	B	C	D	E
1	Navy Seal	그대, 언제까지 그렇게 살텐가.....	=MID(B1,1,1)	=MID(B1,2,1)	=MID(B1,3,1)
2	Navy Seal	그대, 언제까지 그렇게 살텐가.....	그	대	,
3	강아지1234	니가지구를떠날때까지.	니	가	지
4	눈오는날백구	생각합니다ㅋㅋ	생	각	합
5					
6					

MID(텍스트 위치, 시작위치, 반환할 글자수) 명령어는 지정한 텍스트를 원하는 위치에서부터 반환하여주는데, 시작위치를 한 칸씩 증가하도록 설정하면 긴 텍스트를 한 칸씩 분해할 수 있다.[1] 또한 적당한 엑셀 기능을 사용하여 손쉽게 비슷한 명령어를 만들어낼 수 있다.

<그림2>



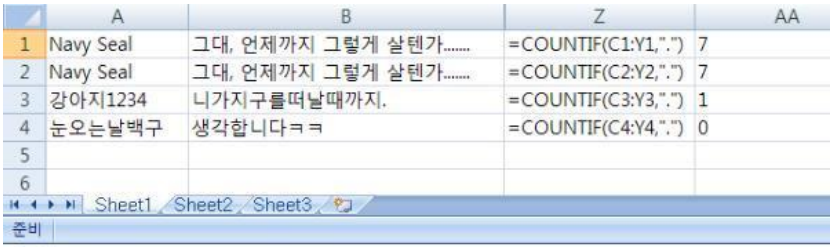
	A	B	C	D	E	F	G
1	=	mid(b2,	1	,1)	=A1&B1&C1&D1	=mid(b2,1,1)	
2	=	mid(b2,	2	,1)	=A2&B2&C2&D2	=mid(b2,2,1)	
3	=	mid(b2,	3	,1)	=A3&B3&C3&D3	=mid(b2,3,1)	
4							
5							
6							

1) Microsoft Excel의 2003버전 이하에서는 256개의 열만을 지원하므로 긴 텍스트에 대하여 본 작업을 수행하기 위해서는 행을 활용해야 한다. 한편 2007버전에서는 16384개의 열을 지원한다.

그림2은 엑셀의 자동 채우기를 활용하여 비슷한 함수를 쉽게 만들어 내는 방법을 보여주고 있다. 위와 같이 입력한 뒤에 A1부터 E3까지의 영역을 드래그하면 A1,B1,D1의 내용은 동일하게, C1의 내용을 순차적으로 증가하게 되고, F는 실제로 표시되는 E열의 값이다. E열의 값을 복사하여 값으로만 붙여 넣으면 원하는 값을 표시하게 된다.[2)]

긴 문자열을 한글자씩 분해한 뒤에는 COUNTIF(검색할 범위,"검색할 조건")명령어를 사용하여 원하는 문자가 지정한 범위 내에 몇 번 사용되었는지 알아볼 수 있다.

<그림3>



	A	B	Z	AA
1	Navy Seal	그대, 언제까지 그렇게 살텐가.....	=COUNTIF(C1:Y1,".")	7
2	Navy Seal	그대, 언제까지 그렇게 살텐가.....	=COUNTIF(C2:Y2,".")	7
3	강아지1234	니가지구를떠날때까지.	=COUNTIF(C3:Y3,".")	1
4	눈오는날백구	생각합니다ㅋㅋ	=COUNTIF(C4:Y4,".")	0
5				
6				

Sheet1 Sheet2 Sheet3

준비

2) 복사/붙여넣기의 단축키는 Microsoft windows와 대부분의 OS에서 제공하는 단축키와 같다. Ctrl+C는 복사를 수행하는 명령어이고, Ctrl+V는 붙여넣기를 수행하는 명령어이다. 하지만 값으로만 붙여넣기는 마우스 오른쪽 버튼(혹은 106 키보드의 드롭다운 버튼), S 혹은 붙여넣기 메뉴를 클릭한 뒤 값(value)을 선택한다. 또한 사용 환경에 따라서 값이 붙여진 다음에라도 값이 표시 안되는 경우가 있는데, 이 경우 수식 표시(Ctrl+`)를 이용하면 값을 볼 수 있다.