인지과학, 16권 제2호, 2005, 93~121.

언어적 특성을 이용한 '심리학적 한국어 글분석 프로그램(*KLIWC*)' 개발 과정에 대한 고찰*

The Review about the Development of Korean Linguistic Inquiry and Word Count

이 창 환** (Chang H. Lee)　심 정 미 (Jung-Mi Sim)　윤 애 선*** (Aesun Yoon)

요 약 최근 심리학 연구에서 *LIWC*(Linguistic Inquiry and Word Count)라는 '심리학적 영어 글분석 프로그램'을 사용하여 사람들이 사용하는 언어적 양식을 종속측정치로 사용한 연구들은 괄목할 만한 성과를 거두었다. 본 연구는 이러한 영어분석 프로그램을 원형으로 한국어의 특성과 문화를 반영한 '한국어 글분석 프로그램(*KLIWC*)'을 개발하기 위하여 실시되었다. 형태소 태깅을 통하여 다수의 형태소가 교착된 어절을 분석하는 기능을 추가하였고 기본형 사전과 활용형 규칙을 구축하였다. 또한 체면, 한국적 정서와 관련된 단어를 분석 변인에 포함시켰다. 이러한 한국어 분석프로그램의 개발과정과 특성을 고찰하였고 프로그램의 추후 개선방향에 대하여 논의하였다. (*KLIWC* 제공 웹사이트: www.k-liwc.net)

주제어 한국어 분석 프로그램, 한국어, 프로그램 개발

Abstracts Substantial amounts of research have been accumulated by the attempt to use linguistic styles as the dependent measure in conducting psychological research. This research was conducted to develope a Korean text analysis program (KLIWC) based on the English text analysis program, LIWC(Linguistic Inquiry and Word Count), and the program reflects the Korean linguistic characteristics and culture that is related with language. We made it possible to analyze agglutinative phrase of many morphemes by linguistic tagging, and basic form dictionary and inflection rule were built. In addition, the face-saving words and emotional words were included as the analysis variables. The process of development and characteristics of Korean text analysis have been reviewed, and future direction for the improvement of the program has been discussed.

Keywords *Korean Linguistic Inquiry and Word Count, Korean, Program Development*

---

* 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-042-H00010).
한국어 분석 프로그램 개발과정에서 도움을 준 김동희, 이미영, 정휘웅, 최미영(부산대)에게 감사를 드린다.

** 부산대학교 심리학과, 연구세부분야: 언어심리학,
부산시 금정구 장전동 산30, 전화: 051-510-2134, E-mail: chleehoan@pusan.ac.kr
*** 부산대학교 불어불문학과, 연구세부분야: 컴퓨터 언어학
부산시 금정구 장전동 산30, E-mail: asyoon@pusan.ac.kr

실제로 사람을 만나지 않더라도 그 사람이 쓰는 글(또는 말)의 방식과 내용에만 근거하여서 글쓴이의 나이, 성별, 교육수준 등의 인구학적인 변인은 물론 글쓴이의 의도, 성격, 심리상태나 건강상태까지 직관적으로 추론할 수 있는 경우가 많다. 또한 사람들 간의 대화된 텍스트나 실제 대화의 녹음된 말의 내용과 방식만을 바탕으로도 대화자 간의 인구학적 변인과 관련 심리학적 변인을 알 수 있고 그 대화가 대화자 간에 서로 교감을 이루면서 원만하게 이루어졌는 지의 여부를 추론할 수 있다. 이와 같이 일상생활이나 관련된 실험상황에서 사람의 글이나 말이 의식적인 수준에서도 주요 심리학적 변인을 반영할 수 있음에도 불구하고 일부 분야를 제외하고는 심리학의 주요 종속변인으로 사용되지 않아 왔다. 이러한 가능성을 초기 심리학자들도 생각해온 것이 사실이었지만(예: Freud, 1901), 프로토콜 분석(Protocol Analysis)이나 일부 임상, 상담 장면에서 언어분석을 통해 진단의 보조도구나 상담의 질을 평가하는 등의 언어와 제한된 관심변인 간의 관련성만을 밝히는 성과가 있었을 뿐 광범위한 심리학적 변인과의 관련성을 밝히지는 못하였다.

글을 분석하는 물리적인 단위는 문장이나 단락에 근거하여 이루어질 수 있지만 언어 의미의 기본단위인 단어수준에서 주로 이루어져 왔다. 이러한 몇몇 심리학적 글분석(Text Analysis) 프로그램을 간략히 살펴보면 다음과 같다. '일반 탐색자(*General Inquirer*)'는 단어들을 바탕으로 글을 분석하는 최초의 컴퓨터 글분석 프로그램이라 할 수 있다(Stone, Dunphy, Smith & Oglive, 1966). 프로그램에서 쓰인 단어들은 하버드 심리사회사전(Havard III Psychosocialogical Dictionary)에서 추출되었으며, 심리학적 의미(Meaning)를 가지는 단어들이었다. '*TAS/C*'라는 컴퓨터 분석 프로그램은 임상장면에서 심리치료의 결정적인 장면을 포착하기 위해 감정적인 어조(Emotional Tone)와 추상화(Abstraction)의 두 가지 언어차원을 사용하였다(Mergenthaler, 1996). 감정적인 어조는 감정단어가 텍스트 내에서 얼마나 집중적으로 분포되느냐에 따라 포착되며, 단어의 리스트에는 즐거움, 인정, 그리고 애착과 관련된 단어들이 포함되었다. 추상화는 단어의 추상화와 관련 있는 접미사(예: -ness, -ity, -ing, -ion 등)를 포함하는 단어들의 수로써 계산하였다. 최근의 글분석 프로그램으로는 정치인의 연설, 선거 광고 등을 분석하여 정치인이 쓰는 용어(예: 선동, 비난, 부정 등)들을 분석하기 위해 개발된 DICTION (Hart, 2001)이 있다.

최근 들어 널리 사용되고 있는 '*LIWC(Linguistic Inquiry and Word Count)*'는 사람들이 쓴 글을 72개의 언어 관련 목록의 비율에 따라 분석하는 컴퓨터 프로그램이다(Pennebaker & King, 1999; Pennebaker, Mehl, & Niederhoffer, 2003). 부록에서 보는 바와 같이 LIWC의 변인들은 크게 나누어 언어학적 차원(Standard Linguistic Dimension), 심리학적 과정(Psychological Processes), 상대성(Relativity), 개인관심사(Personal Concerns) 범주로 나뉜다. 이러한 4가지 범주 내에 72개의 하위 변인들이 해당 범주에 속해 있는 분석구조이며 변인에 해당되는 단어의 비율을 계산하는 단순한 방식이 심리학적 변인과 기제를 반영할 수 있을 것으로 가정하였다.

상술한 기존의 글분석 프로그램은 제한된 심리장면(예: 심리치료)에서 사람들이 사용하는 내용단어(Content Word)[1]를 중심으로 언어

를 분석하였다. 예를 들어, 감정적인 단어의 비율이나 욕구 및 성취 관련 단어와 같이 특정한 심리적 의미를 가지는 단어의 비율을 계산하는 데 초점을 맞추었다. 이에 비하여 *LIWC*는 일상생활에서 자주 쓰는 3000여 개의 내용 단어들을 바탕으로 다양한 내용 단어 변인을 포함하였을 뿐만 아니라, 단어길이, 관사, 전치사, 단어 수 등과 같이 단어와 단어, 또는 문장과 문장을 연결하는 기능 단어 변인과 1인칭, 2인칭, 3인칭 대명사 관련 변인도 포함하였다(Pennebaker, Francis & Booth, 2001). 이러한 기능 단어들과 대명사 관련 변인은 전체 글의 언어구조(Linguistic Structure)를 만드는 데 사용되는 단어로서 언어사용의 상당 부분을 차지하고 있음에도 불구하고 이전의 글분석 연구에서는 간과되어 왔다. *LIWC*는 기능단어(Functional Word)들과 대명사 관련 변인을 포함함으로써 기존 글분석 프로그램에서는 구별할 수 없었던 임상, 사회, 성격, 발달, 생리, 인지 등 주요 심리학 분야의 다양한 개인차 변인을 추출해 내는 연구결과를 축적할 수 있었다.

기능 단어들을 언어학적 변인이라 부르는데 이들이 글이나 말에서 차지하는 역할은 단어와 단어, 문장과 문장, 심지어 덩이 글과 덩이 글을 연결하는 기능적 역할을 하여 전체 글의 언어구조(Language structure)를 만든다. 또한 대명사도 이전 문장이나 덩이 글에서 언급된 사람이나 사물을 참조하는 역할을 하기에, 이의 사용을 위해서는 많은 무의식적인 인지적 용량이 필요하다. 반면에 내용단어들은 글이나 말을 사용하면서 의식적으로 사용을 통제할 수 있기에 실험상황에서 설정한 특정 상황, 집단, 대화상대자에 따라 사용의 편차가 없을 수 있다. 잠정적인 결론으로 *LIWC*가 기존의 내용단어를 바탕으로 한 언어분석프로그램과는 달리 괄목할 만한 성과를 거둘 수 있었던 것은 앞서 언급한 주요 심리학적 변인들의 대부분은 무의식적이고 자동적인 기제를 바탕으로 하고 있기에 이의 작용을 포착할 수 있는 언어학적 변인과 대명사 관련 변인을 LIWC의 주요 변인으로 설정한 것이 주효했을 것으로 판단된다.

LIWC를 사용하여 추출한 주요 심리학적 변인을 살펴보면 다음과 같다. 우선, 임상심리학 장면에서 LIWC를 사용하여 축적한 연구를 살펴보면 크게 나누어 글쓰기와 건강 간의 관계를 밝힌 연구와 글쓰기와 우울증 등 정신병리 간의 관계를 밝힌 연구로 나눌 수 있다. 글쓰기와 건강의 관계에 관한 연구의 요점은 사람들이 과거의 부정적인 감정과 연관된 사건이나 인물에 대해 감정적인 글쓰기(Emotional Writing)를 하게 되면 그 사건이나 인물과 관련된 심리적 고통이 경감함과 동시에 신체적 건강까지도 개선된다는 것이다. 구체적으로 교통사고나 가족상실과 같은 중대한 외상(Trauma)으로 인하여 심리적 고통뿐만 아니라 관련 신체적 질병을 갖는 집단은 그 사건의 원인과 당시의 느낌에 대하여 일련의 글쓰기

1) 언어학에서 어휘나 형태소를 생산력, 수, 역할 등의 특성에 따라 내용 단어와 기능 단어(functional words)로 구분할 수 있다. 내용단어는 명사(대명사), 동사, 형용사, 부사, 감탄사 등과 같이 문장의 의미를 구성하며, 어휘의 수가 많고 생산력이 높아 신조어를 만드는 기반이 된다. 이에 비해 기능단어는 전치사, 접속사, 한정사(관사)처럼 내용단어의 통사적인 기능을 표시하며, 그 수가 극히 제한되며, 신조어를 만드는 생산력이 매우 낮다. 이 중 대명사의 경우, 언어학에서는 내용단어로 분류되나, *LIWC*에서는 기능단어로 분류한다.

를 하는 것은 일반적인 일상의 주제에 대하여 글쓰기를 하는 것보다 그 외상에 대한 이해를 증진시킬 뿐만 아니라 질병과 관련된 병원 의존도나 혈중 면역지수와 같은 생리적 지표까지도 개선된다는 것이다(Lepore, 1997; Murray & Segal, 1994; Pennebaker, Mayne, & Francis, 1997; Petrie, Booth, Pennebaker, Davison & Thomas, 1995). 이러한 연구에서 밝혀진 또 다른 중요한 사실은 사람들이 쓰는 단어의 종류와 비율은 감정적 글쓰기와 관련된 심리적 변화와 건강 변화를 반영한다는 것이다. 감정적 글쓰기 후에 건강지표가 개선된 집단은 일상사의 원인과 통찰과 관계된 인지적인 단어(예: because, hence, insight 등)의 사용이 기준선보다 늘어나게 되고 긍정적 정서단어와 부정적 정서단어의 사용이 동시에 늘어나게 된다. 또한 앞서 언급한 대명사 관련 변인이나 문장당 단어 수가 늘어나게 되어 글의 구조가 상대적으로 복잡하게 된다. 감정적인 글쓰기에 따른 외상 극복과 관련된 심리적 기제에 대한 이론적인 설명은 다양하지만 감정적인 글쓰기를 실시한 임상환자들은 감정 단어를 사용함으로써 억제되었던 감정적인 상처에 대해 직면하게 되고 외상과 관련된 사고에 대하여 인지적으로 합리적 재해석 했을 것으로 추론된다(Pennebaker et. al., 1997; Pennebaker & Graybeal, 2001). 이러한 과정에서 트러마와 관련된 사고의 원인을 규명하려는 노력과 사고에 대한 객관적인 통찰을 얻어 글을 쓰기 이전의 혼란된 상태에서 상당 부분 벗어난 것으로 보고 있다.

글쓰기와 우울증 등 정신 병리와 관련된 연구로 Rude, Gortner, 그리고 Pennebaker(in press)는 우울증 집단과 정상 집단이 쓴 글을 *LIWC*로 분석해본 결과, 우울증 집단이 정상 집단보다 일인칭 단수 대명사(I, me, my, mine)의 사용빈도는 높지만, 복수대명사나 타인 지칭 대명사(예: we, he, she, they 등)의 사용빈도는 낮은 경향을 보고하였다. 또한 우울증 집단은 단어와 단어를 연결하는 기능 단어의 사용이 원활하지 않아 정상집단보다 글의 구조를 복잡하게 만들지 못한다는 결과를 보고하였다. 역사적으로 우울증에 빠져 자살했던 문학계의 수많은 시인의 시를 *LIWC*로 분석한 연구에서 자살시인은 동시대의 일반시인보다 일인칭 단수 대명사의 사용빈도는 높지만, 복수대명사나 타인 지칭 대명사의 사용빈도는 낮다는 것을 발견하였다(Stirman & Pennebaker, 2001). 또한 자살 시인들의 작품을 시간에 따라 분석한 결과는 자살의 시점에 가까워질수록 단수 대명사의 사용이 증가함을 보여 주었다.

*LIWC*를 사용하여 사회 심리학적 변인과 글쓰기 양상을 연관시킨 연구성과도 괄목할 만하였다. 이들 연구의 기본적인 논리는 사람들이 쓰는 말이나 글은 사람들이 처한 사회적 상황이나 대화 상대자가 누구냐에 따라 달라진다는 것이다(Goffman, 1959). Mehl과 Pennebaker(2003)와 Stone과 Pennebaker(2002)는 각각 9.11 뉴욕 테러사건과 영국 황태자비 Diana 사망사건과 같은 해당 국가의 위기상황에 따라 언어를 분석(인터넷 공간 포함)을 한 결과, 많은 *LIWC*의 언어 변인이 심리상태의 변화를 반영한다는 것을 보고하였다. 주요 결과로 위기 상황에서는 문장당 단어 수가 줄어들며, 감정적 단어의 사용이 증가하게 되고, 일인칭 단수 대명사보다는 복수 대명사(we, our, us, ours)의 사용이 많아지며 짧은 단어의 사용이 증가한다. 이는 사회집단의 상황이 개인의 언어사용에까지 영향을 미칠 수 있다는 실제적 증거를 제

시한 것이다. 또한 사회심리학에서 허위진술(Deceptive confession)은 단어의 사용을 분석하는 전통적인 주제(예: Knapp, Hart & Dennis, 1974; Buller, Burgoon, Buslig & Roiger, 1996)에 대하여, *LIWC*를 사용해 허위진술문과 진위진술문(True Statement)을 분석하였다. Newman, Pennebaker, Berry, 그리고 Richards (2003)는 *LIWC*를 사용하여 진위진술문은 허위진술문보다 "but," "without," "except" 등과 같이 범주를 구별하는 단어들을 상대적으로 많이 사용하고, 일인칭 대명사의 사용도 상대적으로 많다는 사실을 발견하였다. 이는 진위진술이 의미적으로 복잡한 문장의 사용과 자기 지향적인 진술이 보다 용이함을 의미한다.

*LIWC*를 사용하여 임상, 사회심리와 관련된 변인 외에도 발달심리적 변인, 생리심리적 변인, 성격심리적 변인 등 주요 심리학적 변인과의 관련성을 밝혔다.

종합적으로 약술하면 *LIWC*를 사용하여 언어사용과 주요 심리학적 변인 간의 관계성을 밝힌 연구는 심리학의 다양한 분야에서 이루어져 언어사용을 심리학 연구의 종속변인으로 사용할 수 있는 가능성을 제시하였다. 지금까지는 언어사용과 심리학 변인 간의 관계성을 증명하는 현상의 발견에 주력해 왔는데 축적된 자료와 향후 연구를 바탕으로 심리학 변인에 대응되는 심리 언어학적 지도(Psycho-Linguistic Map)의 작성이 가능할 것으로 기대된다. 이를 통하여 언어사용 방식이 어떤 심리적 기제에 의하여 작동되는 지를 밝힐 수 있을 뿐만 아니라 언어사용을 심리학 연구의 종속변인으로 사용할 수 있는 이론적 토대를 제공할 것이다.

본 연구에서는 상술한 *LIWC*의 방대한 연구 성과에 근거하여 기능단어라는 언어적 특성을 이용한 글분석이 심리학적 변인을 반영할 수 있는 새로운 연구방법이 될 수 있다고 가정하고, *LIWC*을 모형으로 한 한국어 글분석 프로그램(*KLIWC*)을 개발한 과정과 프로그램의 특징들을 고찰하고 변인에 대한 신뢰도 검증을 하였다.[2)]

*LIWC*의 연구 성과가 괄목할 만하나, 한국어 분석에 단순 차용할 수 있는 것이 아니며 *KLIWC*를 개발하기 전에 선행해야 할 연구내용이 방대하였다. 첫째, 영어와 한국어 간 언어적 이질성을 충분히 고려하여 언어 변인을 도출해야 한다. *LIWC*에서 중요한 언어적 변인으로 고려하는 요소는 이른바 기능단어인데, 하나의 형태소가 독립된 어휘로 나타나는 비율이 높은 영어와는 달리 한국어는 여러 개의 형태소가 교착하여 한 어절을 구성한다(권혁철, 1990; 이상복, 1989). 과거나 미래를 표현하는 선어말어미, 격조사, 양태(modality) 표현, 부정 표현, 접속어 등이 모두 어간에 교착하여 음절이나 음소 단위로 나타난다(고진숙, 1987; 김재훈, 이공주, 2003). 영어와 한국어 간 형태 · 통사적 이질성 이외에도 사회언어학적 차이도 존재한다. 예를 들어 *LIWC*에서 중요한 역할을 하는 대명사의 경우, 한국어에도 존재하나 존칭 어법에서는 2, 3인칭 대명사의 사용이 극히 억제되며 명사구의 반복이나 논항의 삭제가 빈번하다. 또한, 한국어에는 1개 어간이 갖는 활용형의 수가 매우 많으므로 어절 단위에서 나타나는 동형이의어(homonym)의 비율이 아주 높다(김기찬, 1998; 최덕수, 배해

2) '언어적 특성을 이용한 한국어 글분석 프로그램'의 명칭인 *KLIWC*은 'Korean Linguistic Inquiry and Word Count'의 약자이다.

수, 김광해, 1998). 따라서 *KLIWC*에서는 *LIWC*에 비해 매우 정교한 형태소 분석이 필요하다. 둘째, 영어와 한국어의 어휘가 사용하는 의미장이나 함의(connotation)가 다르므로 *LIWC*의 심리학적 변인으로 사용하는 어휘에 1:1로 대응하는 한국어 어휘를 찾기 어렵다. 심리학적 변인으로 사용할 수 있는 한국어 어휘 리스트에 대한 선행연구가 거의 수행되지 않았으므로 본 연구에서 도출할 어휘 리스트가 신뢰성을 가지려면, 영어와 한국어 간 어휘의 의미장에 관한 언어학적 연구 성과를 이용하였다(이기용, 이종민, 홍정하, 2002; 이동영, 2002). 셋째, *LIWC* 프로그램을 이용한 연구 성과는 눈에 띄나 소프트웨어 개발 관점에서 *LIWC*는 여러 가지 문제점을 가진다. 우선 자료의 갱신이 자주 이루어져야 하고, 다양한 자료 간 상호 관계가 복잡하게 연결되어 있는데 이를 제어 및 관리할 개발자용 프로그램이 없다. 또한 단순 사용자용 프로그램 역시 관계형 DB를 이용한 자료의 one-stop 통제를 하지 못하고 일일이 사전 갱신 등과 관계 표현을 수동으로 해주어야 한다. 또한 StandAlone 기반으로 이루어진 프로그램이므로 여러 지역 간 공동 작업 등이 불가능하다. 따라서 *KLIWC*은 *LIWC*의 이러한 단점을 보완하는 데 그치지 않고 더욱 지능적인 프로그램으로 설계하였다.

KLIWC 개발과정에서 반영된 한국어의 언어적 특성

한국어와 영어는 매우 이질적이며, *LIWC*에서 중요 변인으로 작용하는 기능 단어의 형태 및 통사적 형태가 매우 상이하다. 따라서 *KLIWC*의 개발은 *LIWC*의 단순한 평면적 이식(porting)이 아니라, 한국어 분석 및 두 언어에 대한 비교 연구를 토대로 이루어져야 한다. *LIWC*에서 사용하는 언어적 분석에 필요한 요소를 추출하기 위해서 *KLIWC* 개발과정에서 다음과 같은 한국어의 특성을 반영하였다.

1. 한국어의 '언어적 변인' 후보 분석

*LIWC*에서 가중치를 부여하고 있는 언어적 구성성분을 비교할 때 영어와 한국어는 다음과 같은 큰 차이를 보인다.

첫째, 언어계통을 분류하면, 영어는 게르만어나 라틴어와 같은 굴절어(inflectional language)에 속하나 굴절의 특성을 대부분 잃고 고립어적인 특성을 보유한다. 이에 반해 한국어는 대표적인 교착어(agglutinative language)이다. 이러한 차이는 두 언어의 형태 및 통사적 지배구조에 반영되어 있다. 특히 *LIWC*에서 사용하는 주요 변인인 기능단어의 형태와 기능에서 큰 차이를 낳는다(김기찬 1998; 정달영 1991). 기능단어란 문장 내에서 내용단어의 통사적 기능을 나타내는 어휘 또는 형태소로서, 영어에서는 '전치사, 접속사, 관사' 등의 품사가 이 부류에 속한다. 한국어에서는 '기능표지, 기능변환, 의미한정, 접속, 서술보조' 기능을 하는 '조사, 어말어미, 선어말어미, 보조사, 접속사' 등이 이에 속한다. 영어의 기능 단어는 독립된 어휘로 존재하며, 그 수가 매우 제한되어 있다(김용전 1988). 이에 비해 한국어의 기능단어는 용언이나 체언의 뒷부분에 교착하는 조사와 어미 같은 경우 음소 내지 형태소 단위로 나타난다. 형태소 자체의 수는 매우 제한적이나, 기능 단어 간 결합이 표 1에서 볼 수 있는 것과 같이 비교적 자유로워 한 어절을 구성하는 어미 및 조사 결합은 2천 개를

<표 1> 한국어 조사 및 어미 결합(발췌)

결합의 예	형태 분석
ㄴ가보다라고	ㄴ가 + 보다 + 라고
ㄹ뿐일거면	ㄹ뿐 + 이 + ㄹ거면
ㅁ직하더라도	ㅁ직 + 하 + 더 +라도
ㅂ디까라면서	ㅂ디까 + 라면서
게까지로밖에는야	게 + 까지 + 로 + 밖에 + 는 + 야
고나서부터는	고 + 나서 + 부터 + 는
댄다더라고	대 + ㄴ다 + 더 + 라고
라잖니	라 + 지 + 않 + 니
습니다마는	습니다 + 마는
으라고까지도	으라고 + 까지 + 도
으로써만이야말로	으로써 + 만 + 이야말로
으면서부터도	으면서 + 부터 + 도
처럼만이라도	처럼 + 만 + 이라도
한테서부터는커녕	한테 + 서 + 부터 + 는 + 커녕

상회한다(권혁철, 김민정 1992; 김민정, 1997; 김영택 외, 2001; 남기심, 1982; 채영숙, 최성필, 서정현, 2002).

둘째, 한국어에도 표 2와 같이 인칭대명사가 존재하지만, 사회언어학적 이유에서 그 사용이 매우 제한적이다. 인칭대명사는 언행(speech act) 상황에서 각각 화자, 청자 및 제3자를 지칭하는 경제적 표현이다. 하지만 언어의 경제성의 원칙와 대립되는 원칙인 '존대법'이 발달한 한국어에는 다음 표와 같이 3인칭 대명사의 형태가 '관형어+명사'의 구조를 가진 명사구에 가깝고, 2인칭 대명사의 사용은 청자가 화자보다 사회적 지위가 명백하게 높지 않은 경우에만 사용된다(Yoon, 1989). 따라서 격식을 갖춘 언행 상황이나, 청자 또는 제3자가 화자보다 우월적인 지위를 가져 경칭을 사용해야 하는 경우, 2인칭 및 3인칭 대명사의 사용은 지양되며 표 3과 같은 다양한 종류

<표 2> 한국어 인칭대명사의 종류(단수형)

인칭 / 화계	1(화자)	2(청자)	3(제3자)	
			사람	사물
겸양/존칭	저	-	당신	-
비존칭	나	너, 자네, 그대, 임자, 당신	그, 그이	이것, 저것, 여기, 저기

<표 3> 한국어 호칭어의 종류

기원	언행 상황	예
직업명	공식적 관계	선생(님), 교수(님), 사장(님), 사모(님), 스(님), 목사(님), ...
직위명	공식적 관계	장관(님), 과장(님), 팀장(님), 장군(님), 국장(님), ...
친족어	사적 관계	언니, 오빠, 누나, 형, 아저씨, 아주머니, 어머니, 아버지, ...

의 명사구가 호칭어의 기능을 하게 된다. 예를 들어 '선생님, 사장님, 사모님' 등은 존칭을 해야 하는 청자나 제3자 지칭어로 사용되는 대명사적인 표현이며, 손위 관계를 지칭하는 친족어 역시 친족관계 이외에서 친근한 경칭으로 사용된다(서정수, 1996; Yoon, 1989).

셋째, 논항(argument) 생략이 극히 제한되는 영어와는 달리 한국어 문장에서는 상황 또는 문맥이 허용하는 한 주어, 목적어, 보어 등 문장을 구성하는 필수적인 논항이 쉽게 생략될 수 있다(박민경, 김민정, 권혁철, 1997; 소길자, 권혁철, 2001; 심광석, 2000). 표 4는 필수적인 논항이 생략된 전형적인 예시로써, 5개의 단문과 2개의 복문에서 8개의 논항이 생략되었으나, 일상생활에서 흔히 접하는 정상적인 한국어 문장이다. 따라서 *KLIWC*를 개발할 때, 논항이 명시적으로 드러난 표면 구조만을 고려할 것인지 아니면 논항을 복원시켜야 할지 그 분석 범위를 설정해야 한다.

넷째, 한국어 어휘의 70% 이상은 한자어에서 유래하므로 영어에 비해 동형이의어의 비율이 높으며, 표 5와 같은 예에서 볼 수 있듯이 한 개의 형태소 또는 어휘에 해당하는 동형이의어의 수도 많다(김진형, 1996; 허윤영, 권혁철, 1994). 또한 한국어는 반자유어순어(semi-free order language)이므로 논항의 위치가 해당 논항의 통사적 기능을 나타내지 않는다. 통사적 기능은 '이/가/께서, 을/를, 에, 에서, 에게/한테/께, 은/는, 의, 와/과' 등과 같은 조사에 의해 표현되나, '나 거기 갈게.'에서처럼 조사

<표 4> 한국어 문장에 실재하는 논항 생략과 복원 가능성

기사원문	A: "어떻게 나왔냐?" B: "훈방되었어요. 새벽에 즉심판사가 와서 돌 던졌냐고 묻기에 안 던졌다고 했죠. 서류를 보더니 증거가 없으니 나가라고 하더라고요." A: "그 판사 이름 뭐냐?" B: "기억 안 나요. 근데 여자예요."
생략된 논항의 복원	A: "((네가)) 어떻게 나왔냐?" B: "((저는)) 훈방되었어요. 새벽에 즉심판사가 와서 ((저더러)) 돌 던졌냐고 묻기에 ((저는)) 안 던졌다고 했죠. ((그 사람이)) 서류를 보더니 증거가 없으니 ((제게)) 나가라고 하더라고요" A: "그 판사 이름 뭐냐?" B: "((저는)) 기억 안 나요. 근데 ((그 사람은)) 여자예요"

<표 5> 한국어 활용형 어절의 중의성의 예

순	'가는'의 예문	품사 분석	기본형(동사/형용사)
1	길을 **가는** 사람을 불러세워	동사+관형형 어미	가다
2	잠잘 때 이를 **가는** 사람은		갈다
3	침대 시트 **가는** 것을 게을리하면		
4	허락없이 남의 밭을 **가는** 행위는		
5	매우 **가는** 빗방울을 맞으며	형용사+관형형 어미	가늘다
6	**가는** 양보다 낮은 점수이다	명사+조사	
7	사전에서 **가는** 나보다 앞에 등재된다		

역시 생략되는 경우가 많다. 이러한 특성은 형태 및 통사 분석 시 중의성을 증가시켜, *KLIWC*에 필요한 태깅 시스템의 구현에 매우 큰 걸림돌이 된다.

2. 품사 태깅을 위한 형태 · 통사적 분석

영어는 좌우에 공백을 가진 한 단어 단위로 품사 정보를 갖는다. 한국어에서 띄어쓰기는 '어절' 단위로 이루어지며, '어절'은 1개 내지 십수 개의 형태소로 구성되므로 영어와 같이 품사 빈도 추출이 단순하지 않다. 예를 들어 "아이가 감기에 든 것은 어머님께서 머리를 찬 물에 감기셔서였잖았습니까라고만은 어머님께 직접 말하기 힘들어."라는 예문에서 '감기셔서였잖았습니까라고만은'이라는 1개의 어절은 다음과 같이 12개의 형태소로 분석된다. 이러한 어절은 실제 영어에 '구(phrase)'에 해당되는 통사적 복잡성을 갖는다.

더욱이 국어학계에서 한국어의 품사 체계 구분에 대한 논의가 계속되고 있다. 1990년 문교부에서 제시한 학교 문법에는 한국어 품사범주를 '명사, 대명사, 수사, 동사, 형용사, 관형사, 부사, 감탄사, 조사'로 구분한다. 학교 문법은 최현배의 품사 분류를 근간으로 하는데, ① 독립된 형태로 존재할 수 있는 어휘만을 품사로 구분하였으며, ② 다양한 형태 · 통사적 특성을 가진 기능 단어를 '조사'라는 한 품사로 범주화하였고, ③ 교착어의 특성이며 중요한 기능 단어인 다양한 어미의 존재가 분류되지 않는다는 문제점을 안고 있다. 따라서 *KLIWC*를 위한 품사 태깅을 위해서는 필요한 품사 세부 분류표를 설정하여 이를 검증하여야 한다(서정수, 1996; 정영자, 1991; 조세형, 2001; 한인석, 1982).

3. 기본형 사전 및 활용형 규칙 구축

언어학적 변인과 심리학적으로 유의미한 변인은 분석되는 텍스트에서 활용된 형태로 사용된다. 따라서 기본형 또는 어간의 형태로만 구성된 변인 시소러스의 구성요소를 그대로 텍스트에서 찾아낼 수 없다. 굴절어의 특성을 거의 상실한 영어가 극소수의 활용 형태를 가진 것에 비해, 한국어는 어간에 교착하는 조사 및 어미의 수가 2만 개에 이르며, 어간의 일부 음절이 변화하는 불규칙의 종류 역시 다양하다. 이렇게 상이한 특성 때문에 효율적인

<표 6> '감기셔서였잖았습니까라고만은'의 형태 분석

형태소	의 미
감	동사 '감다'의 어간
기	피동형 선어말어미
시	객체 존칭 선어말어미
어서	접속 어미
였	과거형 선어말어미
지	접속 어미
않	부정 형태소
았	과거형 선어말어미
습니까	청자존칭 어말어미
라고	접속어미
만	한정어
은	주제화 조사

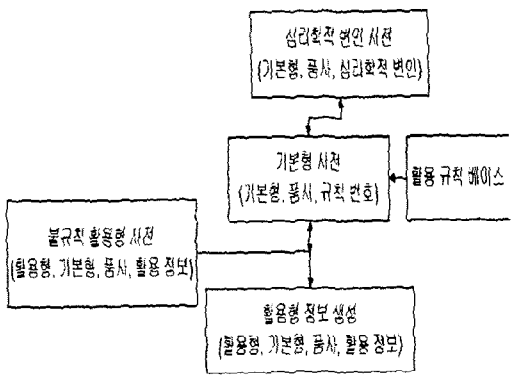
검색과 자료 갱신을 위해서는 두 종류의 자료 구축 방식이 사용되어야 한다(윤애선, 2001; 윤애선, 권혁철, 2001; Yoon & Kwon., 2000). 즉 *KLIWC*의 개발에 필요한 기본형 사전을 구축하고, 이 기본형의 활용 가능 형태를 규칙으로 생성할 수 있어야 한다(채영숙, 최성필, 서정현, 2002; 최덕수, 배해수, 김광해, 1998; Yoon & Kwon, 1997).

분석 프로그램 개발

이상과 같이 상술한 한국어의 언어학적 특성을 최대한 반영하기 위해 '한국어 품사 태깅 시스템(K-GraCaT: Korean Grammatical Category Tagger)'이 *KLIWC* 내에 장착되었다. 또한 한국어에 있어서 영어에 비하여 인칭 대명사가 적게 사용되거나 논항이 삭제되는 부분은 한국어 특정적 현상으로 인정하고 논항을 복원시키지 않기로 하였다. 품사 태깅 시스템과 개념 및 어휘 구축 시스템 개발과 관련한 자세한 내용은 다음과 같다.

우선 한국어 품사 태깅 시스템은 언어학 분야에서 이미 개발된 부산대학교 한국어정보처리 연구실의 형태소 분석기(권혁철, 김민정, 1992)를 기반으로, *LIWC*의 품사를 포함시켜 개발하였다. 이 분석기는 학교 문법에서 제시하는 9개의 범주를 단계별로 세분화하여 의미-통사적(semantico-syntactic) 특성에 따라 47개의 범주로 구분하고 있다(Chung, Park, Kwon & Yoon, 2000; Kim & Kwon, 1997). 현재 개발된 *KLIWC* 1.0판의 한국어 자동 품사 태깅 시스템은 이러한 분석기와 같은 47개 범주를 바탕으로 구현되었다(부록 2 참조).

*KLIWC*를 위한 기본형 사전 및 활용형 규칙 구축과 관련하여서는 그림 1과 같이 언어학적 분석을 위한 기본형 사전이 구축되고 각 기본형 사전은 부산대학교 한국어정보처리 연구실의 한국어 활용형 규칙베이스와 불규칙 사전을 이용하여 실시간(real-time)으로 활용형을 분석할 수 있는 기반을 마련했다. 또한 표제어



(그림 1) KLIWC의 활용형 분석을 위한 사전 및 규칙 베이스의 구성

정보로 기본형만을 수록한 심리학적 변인 사전 역시 기본형 사전과의 연동을 통해 활용형 정보를 생성/분석할 수 있도록 한다.

프로그램 개발의 중요한 목적 중의 하나는 영어권 LIWC 연구에서 얻어진 많은 자료와 *KLIWC*를 사용하여 얻어질 자료와의 문화 간 언어행동이 비교되어야 할 것이다. 또한 언어의 차이에 따른 글쓰기의 양상을 비교하여 일반적 글쓰기 심리 모형 제시를 하기 위해서는 한국어의 언어학적 특성을 반영하더라도 어느 정도 양 글분석 프로그램 간 차별을 두지 않아야 할 필요성이 있다. 이런 필요성으로 *KLIWC* 개발과정은 *LIWC*의 개발과정을 참조하되, 한국어와 문화의 특성을 적극적으로 고려하는 방식을 취하며, 프로그래밍도 *LIWC*의 기능을 이식하거나 문제점을 보완하는 정도에 만족하지 않고, 웹 플래트폼과 데이터베이스 연동을 통해 더욱 지능적인 시스템으로 개발하였다.

*LIWC*에 기본적으로 제공되는 Internal Dictionary라는 사전 시스템에 의해 글분석 처리를 수행하며 실험 주제 및 대상에 알맞은 사용자 사전을 사용자가 직접 등록하여 사용할 수 있도록 구성되어 있다. 전산에 대한 기본지식이 없는 심리분석 전문가와 같은 일반 사용자가 개별적으로 사전 시스템의 구성에 대한 전문적인 지식을 갖추기 위해서는 많은 어려움이 있으며 무엇보다 StandAlone 방식의 사전 시스템으로는 사전 데이터를 동적으로 공유할 수 없어 글분석에 있어 시간적 · 경제적인 효과를 극대화 할 수 없다. *KLIWC*는 주로 상하위 개념어로 구성되는 개별 사전 데이터 구축을 위해 프로그래밍에 관련된 문법이나 별도의 구분자(delimiter)를 사용하는 번거로움이 없도록 쉽게 사용할 수 있는 User Interface의 사전 시스템을 개발하고 On-Line으로 액세스 할 수 있도록 구현함으로써 효율적인 데이터의 갱신 및 공유가 이루어졌다.

*LIWC*에서는 기본 정보, 언어적 차원, 심리학적 과정 등 사용자가 필요로 하는 항목만을 선택하여 글분석을 할 수 있도록 구성되어 있다. *LIWC*에서 글분석의 처리 이전에 거치는 항목 설정 과정은 글분석 작업의 중복을 초래한다는 점에서 효율성이 떨어진다고 할 수 있다. 기본적으로 각 차원(Dimension)이나 변인별 Data는 서로 많은 연관성을 갖고 있으며 모든 항목을 적용한 글분석에 많은 시간이 소요되지 않으므로 모든 항목을 적용한 글분석을 수행한 후 분류나 검색, 인쇄를 차원이나 변인별로 수행할 수 있도록 제공하는 것이 더 효율적인 데이터 처리 방법이 될 것이다.

마지막으로 *LIWC*에서는 분석의 대상이 되는 텍스트를 선택한 후 분석 결과 생성될 파일은 이진파일(.dat) 형식으로 저장되며 텍스트 포맷(.txt)이나 엑셀 포맷(.xls)으로도 저장이 가능하다. 또한, 글분석 결과 파일 생성과 동시에 글분석 결과가 72변인에 걸쳐 화면에 횡으로 제시된다. 글분석 결과를 바탕으로 DB Architecture를 손쉽게 구현할 수 있도록 파일 포맷에 쿼리언어(.sql)를 추가하고 앞서 언급했듯이 *LIWC*의 화면 출력 방법은 수많은 변인들을 모두 횡으로 제시하여 가독성 및 분석 정보의 파악에 어려움이 많으므로, *KLIWC*에서는 차원(Dimension)별 분류 및 검색, 인쇄가 가능하도록 구현하였다.

KLIWC 1.0판은 현재 인터넷 상에서 이용할 수 있도록 웹 기반 글분석 시스템을 완료하였다(주소: www.k-liwc.net).

한국어 심리 단어집 구성과 변인에 단어할당

*LIWC*를 살펴보면 4차원 중 정서, 인지, 감각, 사회과정을 포함하는 심리학적 과정 차원이 있다. 다른 차원들에 비하여 이들 차원에 해당하는 단어들을 찾기 위해서는 영어권에서의 경우와 같이 심리 단어집과 같은 데이터베이스가 필요하다. 이러한 단어집이 국내에서는 개발되지 않은 상태이므로 영어권에서 사용된 단어와 한국문화에서 자주 쓰이는 관련 단어들을 찾아내어 한국어 심리 단어집을 구성하였다. 특히, 정서적인 단어인 경우 서구문화와는 달리 정서와 관련된 단어가 상대적으로 다양하므로(안신호, 이승혜, 권오식, 1993) 이와 관련된 하위변인을 만들었다. 구체적으로 한국 문화권에서는 정서와 관련된 단어가 다양하다. 안신호 등(1993)은 국어사전에서 감정 단어라고 판단되는 총 3582개의 단어를 뽑아 1340개로 간추렸다. 이들 단어가 감정 상태를 의미하는지를 확인하여 성격, 신체, 감각적 의미가 강하고, 정서적 의미가 약한 단어는 모두 제외하는 과정을 통하여 최종적으로 213개의 단어를 선정하였다. 이들 단어를 대상으로 군집분석을 통해 유사성을 평정하였는데 단어들은 쾌 단어(예: 즐거움, 행복 등)와 불쾌 단어(예: 아쉽다, 원통하다 등)로 나뉘었다. 이렇게 범주화된 단어들을 *KLIWC*에서 각각 긍정적 정서단어와 부정적 정서단어로 간주하였다.

부록은 기존 영어프로그램 *LIWC*에서 사용하고 있는 72개의 언어변인과 전형적인 단어들의 예와 단어숫자이다. 한국어와 한국문화의 특성을 반영하는 언어변인을 설정하기 위해 관련 언어학자와 심리학자의 자문과 연구팀의 토의를 통해 추가, 삭제, 수정되어야 할 변인을 확정하였다. 전형적인 예로, 한국문화에서 중요한 언어변인이 될 수 있는 것은 존칭어이다. 존칭어의 종류도 여러 단계로 나누어지므로 이를 감안하여 존칭어 변인을 추가하였다(예: "하세요", "하십시오", "하시옵소서").

단어를 선정하는 데이터 베이스로 한국어단어 빈도 조사집(연세대학교, 1989)에 근거하여 빈도 5 이상의 단어들과 국어사전을 바탕으로 단어별로 검토하면서 심리변인에 해당 단어들을 할당하였다. 이차적인 단어할당 작업으로 각종 서적, 신문, 잡지, 미디어 등 일상생활에서 자주 접하는 인쇄물과 방송언어들을 수집하여 빈도 조사집과 국어사전에 근거한 단어 선택 작업에서 간과하였던 자주 쓰는 단어들을 찾아내었다. 이 작업은 약 100,000단어 이상에 대해 사용 여부를 결정하고, 15,000단어 정도의 프로그램사용 단어를 결정해야 하는 작업이었다. 이 과정은 방대하고 시간이 소요되는 작업으로 연구팀의 토의를 거쳐 해당 심리학적 변인의 목록과 대응되는 단어를 선정하여 심리 단어집을 만들었다.

그리고, 기존의 영어권 연구에서 신뢰도가 너무 낮거나 전혀 개인차를 구별하지 못했던 변인들을 삭제하여, 보다 간결한 언어변인을 구성하는 것이 중요하였다. 구체적으로 개인관심사 차원(부록 1의 네 번째 차원)의 두 번째 하부 차원인 여가활동의 어떤 변인도 기존 연구에서 정보를 제공하지 못했다. 또한, 공간(Space)과 관계된 변인인 위치를 나타내는 두 변인(즉, Up과 Down)도 정보력이 약한 변인이었다.

근본적으로 기존 영어권 프로그램인 *LIWC*의 괄목할 만한 성과가 연구동기를 제공하였

기 때문에 *LIWC*의 개발과정을 최대한 참조하는 것이 필요하였다. 그러므로 영어권 연구자의 자문을 받아 *LIWC*의 변인 선정과정과 단어 할당과정을 *KLIWC* 개발과정에 반영하였다. 이는 *KLIWC*이 개발된 이후에 연구 결과에 대한 양 언어 간 대응 비교를 할 수 있는 근거를 제공한다고 볼 수 있다.

글분석 프로그램 변인 신뢰도 검증

글분석 프로그램을 다양한 심리학적 변인을 탐지하고자 하는 연구 도구로 사용하고자 한다면 사람들이 쓰는 언어 자체가 상황이나 글을 쓰는 시기에 따라 일관되어야 한다. 다시 말하면, 앞서 언급한 언어학적 변인이나 대명사 관련 변인의 단어사용이 글을 쓴 시차에 따라 글쓰기에서 일정한 비율로 나타나야 한다. 이와 같이 시차는 달리하지만 반복적인 글쓰기에서 얼마나 많은 *KLIWC*의 변인들이 일정한 비율로 나타나느냐가 분석 프로그램의 신뢰성을 결정한다.

신뢰도를 검증하기 위해서 필요한 자료는 한 사람이 시점을 달리하여 쓴 일정한 양의 글들이다. 이러한 글들을 얻기 위한 실험으로 몇 차례에 걸쳐 실험실에서 글쓰기를 실시하고 이에 대응되는 자료를 얻는 방법이 있다. 본 연구에서는 수차례의 반복적인 글쓰기 실험을 통하여 *KLIWC*의 변인들에 대한 신뢰도 검증을 실시하였다.

글쓰기 실험

피험자

실험심리학 또는 심리학개론을 수강하는 학부생 139명을 대상으로 교과과정의 일부분인 실험 학점을 이수하는 방식으로 실험에 참가시켰다. 실험의 성격이 글쓰기와 성의 있는 참여자세를 요구하기 때문에 실험에 참여하기를 원치 않는 피험자는 참가시키지 않고 다른 실험으로 학점을 이수하게 하였다. 글을 쓰는 것이 부담이 된다거나 글쓰기 장애가 있다고 호소하는 학생도 실험에 참여시키지 않았다.

실험 방법

사람들의 글쓰기가 글을 쓰는 시점에 따라 일정한지를 검증하는 객관적인 방법은 사회적 상황변인이 통제된 실험실에서 다양한 주제로 글쓰기를 실시하고 얻어진 자료를 컴퓨터 프로그램으로 분석하는 것이다. 참가자들은 2번(의식의 흐름에 대한 글쓰기, 감정적 글쓰기 [Pennebaker & King, 1999])의 글쓰기를 한다. 참가자들에게 처음에는 지금 현재 마음속에 흐르고 있는 의식에 대해 적으라 하였고, 그 다음에는 대학에 첫 입학했을 때의 감정관 느낌에 대해 적으라고 하였다. 의식의 흐름에 대한 글쓰기는 현재 생각의 변화에 따라 글이 어떻게 변화하는지를 알아보고자 한 것이다. 또한 의식의 흐름 글쓰기는 생각하고 있는 것을 자연스럽게 적기 때문에 자기의 사고에 대해 솔직하게 적을 수 있다. 두 번째 글쓰기는 학교에 첫 입학했을 때 혹은 첫 학기가 시작했을 때 감정과 느낌에 대해 적었다. 학교 신입생 때에 대한 글쓰기의 목적은 현재 자신의 대학 생활을 돌아보게 하여, 신입 때와 현재의 자기 모습을 반성하게 하는 감정적 글쓰기를 하기 위함이었다. 6주 뒤에 같은 주제로 동안 글쓰기를 하였다. 학생들을 5-10명의 소그룹 별로 실험실로 모이게 실험 절차에 관한

지시하였고 일 회당 글 쓰는 시간은 20-30분 정도를 할당하며 A4용지 한 장 분량을 넘지 않도록 제한하였다. 4회에 걸쳐 얻어진 글쓰기 자료를 컴퓨터 텍스트 파일로 옮긴 다음 KLIWC 시험판으로 분석하였다. 글분석 결과로 얻어지게 될 KLIWC의 100여 개 변인의 단어사용 비율이 신뢰도 산출을 위한 최종 수치가 되었다. 단어 신뢰도의 수치는 4회간에 걸친 언어사용의 상관관계에 바탕을 둔 Kronbach 알파계수로 구하였고 시차에 걸친 계수뿐만 아니라 주제 간에 걸친 계수도 산출하였다.

실험결과

139명의 참가자들 중 15분 동안 글을 쓰지 않은 학생과 무성의하게 쓴 글을 제외하고 총 126명의 글을 분석하였다. 각 참가자들의 글은 글 속의 잘못된 철자와 띄어쓰기를 수정한 뒤 컴퓨터 텍스트 파일로 변환시켜 *KLIWC* 프로그램으로 분석되었다(부록 Ⅳ는 *KLIWC* 프로그램을 통해 분석된 결과이다.). 글분석 결과들은 Cronbach's α 계수로 신뢰도가 분석되었다. 이 연구는 주제를 달리하여 글을 쓰고, 시간 간격을 두어 동일한 주제를 두 번 쓴 글쓰기를 반복측정 하였을 때, 각 반복측정치들 사이에 일관성이 있는지를 알아보고자 한 것으로, 내적 일관성 정도를 검증하기 위해 Cronbach's α 계수를 구하였다.

부록 4의 표에서는 각 변인들 중 신뢰도가 높은 변인들과 그 변인들의 전체 평균과 표준편차를 제시하였다. 의식의 흐름에 대한 글쓰기와 감정적 글쓰기의 상관관계를 비교해보면, 의식의 흐름에 대한 글쓰기에서 문장 당 어절 수와 문장 당 형태소 수에서 .71과 .66으로 감정적 글쓰기(.50과 .40)보다 조금 더 높은 상관이 나타났다. 반면, 조사와 어미에서는 두 글쓰기 모두에서 높은 상관이 나타났다(.90 ~ .72).

이는 Pennebaker와 King(1999)의 연구에서 환자 그룹과 학부생 하계 수련회 그룹을 대상으로 한 언어적 구성의 신뢰도(α)가 .83과 .88로써, 시간과 주제가 달랐음에도 기능 단어들을 안정적으로 썼던 것과 비슷한 결과이다. 이는 *LIWC*의 주요 변인들에서 높은 상관관계를 보인 것과 비슷한 결과가 *KLIWC*에서도 나타났음을 보여준다. 그러나 심리적 범주의 차원들에서는 Pennebaker와 King(1999)의 연구보다는 다소 낮은 상관을 보였다.

[표 4]를 살펴보면 선어말어미와 형이상학적 이슈와 관련된 단어들을 제외하고는 모두 .50 이상의 높은 신뢰도를 보이고 있다. 그리고 모든 102개의 변인들에 대한 자세한 신뢰도의 패턴을 살펴보면 대다수의 변인들에서 높은 상관관계를 나타냈다. 어떤 변인이 삭제되었을 때 α가 높아지는 변인 없이 각 변인들의 안정적인 신뢰도가 구해졌다. 참고로 전체적인 상관을 볼 때, 감탄사 · 관형사 · 수사에서 .20 이하의 낮은 상관관계가 있고, 고유명사 · 일반 고유명사(고유명사의 하위 범주) · 지시대명사 · 양수사 · 서수사 · 자동사 · 수관형사 · 감탄사 · 접두사 · 접미사 · 한자 · 영어 · 외래어에서 .20 이하의 낮은 상관관계를 나타낸다. 심리학적 변인에서는 감정 및 정서적 과정, 인지적 과정, 감각과 지각적 과정, 인지적 과정, 사회적 과정, 자기활동의 상관이 .40 이상의 상관관계를 나타내고 있다. 형이상학적

이슈는 의식의 흐름에 대한 글쓰기와 감정적 글쓰기, 1회 차 글쓰기에서 .30 이상의 상관을 나타냈지만, 2회 차 글쓰기에서는 .12로 전체 변인과 낮은 상관을 나타냈다. 감정 및 정서적 과정의 긍정적 정서 변인의 하위 변인인 긍정적 느낌과 긍정적 정서에서는 .42 ~ .16의 상관을 나타낸다. 부정적 변인의 하위변인인 슬픔 또는 우울, 불안 변인들은 .28 ~ .09의 낮은 신뢰도를 나타냈다. 특히 화 변인은 .04 ~ .00으로 신뢰도가 나타나지 않았다. 자기활동 변인 내 학교 변인을 제외한 하위 변인들에서는 .30 이하의 신뢰도를, 형이상학적 변인 내 종교 변인도 .12의 신뢰도를 나타내었다. 돈 & 재정 차원과 신체적 상태와 기능 차원 전체에서는 α계수가 .20 이하의 신뢰도를 나타내었다. 마지막으로 체면과 관련된 단어들은 기대와는 다르게 .28의 낮은 총평균 신뢰도를 나타내었다. 이는 체면과 관련된 단어들이 문화적으로 민감한 변인이 아니거나 언어 사용에는 문화특정적인 변인이 크게 영향을 미치지 않는 것으로 해석될 수 있다.

*KLIWC*의 미래

*KLIWC*의 개발과 신뢰성 검증은 단어사용과 같은 언어행동을 새로운 심리학 연구의 종속변인으로 사용할 수 있는지의 여부를 알고자 함이다. 다양한 심리 검사법이나 면접과 같은 기존의 연구 도구는 검사를 받는 사람의 왜곡된 보고와 검사자의 주관적인 해석이 가미될 수밖에 없는 한계점을 가지고 있다. 그에 반하여 글분석은 글쓴이가 통제할 수 없는 무의식적인 글의 방식을 반영할 수 있으며 방대한 양의 글에 대한 측정치를 효율적이고 신속하게 얻어낼 수 있다. 그러므로 심리학 연구에서의 *KLIWC*와 같은 언어분석 기법이 사용될 수 있는 근거가 제공된다면 언어분석은 주요 심리학 연구에서 기존 연구도구와 함께 수렴적인 심리 측정치를 제공하는 데 광범위하게 사용되는 성과가 기대된다.

이러한 *KLIWC* 개발의 거시적인 목적 외에 우리의 문자언어, 음성언어, 대화의 인지적 처리과정과 표상 체계를 밝히는 학문적인 목적이 있음을 상술하였다. 실용적인 목적으로 간과할 수 없는 것은 임상장면에서의 진단, 상담 장면에서의 상담의 질 평가, 상담자의 수준과 유형에 따른 언어 분석, 언어훈련프로그램 개발, 영역별 인터넷 공간에서의 언어사용 양상, 범죄 수사 장면에서의 용의자 자술서 진위 여부 분석(예: 언어지문검사[Linguistic Finger Printing])을 비롯하여 글쓰기와 관련된 학교 상황에서의 교육적인 진단의 목적이 있다.

세계적으로 영어권 및 유럽의 다수 국가에서 학문 및 상업적 필요로 인해 언어분석 컴퓨터 프로그램 연구 및 모형 개발과 이의 응용이 상당한 수준으로 이루어지고 있음에도 불구하고, 한국어를 대상으로 한 심리학적 언어분석 프로그램이 국내에 전무하다는 현실은 이러한 연구의 의의를 보다 분명히 한다고 볼 수 있다.

장기적으로 언어사용과 심리학 변인 간의 관계성을 증명하는 현상 발견의 차원을 넘어서 축적된 본인의 연구와 영어권 자료를 바탕으로 주요 심리학 변인과 언어변인이 대응되는 심리 언어학적 모형(Psycho-Linguistic Model)을 제시할 계획이다. 이를 통하여 언어사용 방식이 어떤 심리적 기제에 의하여 작동되는지를 밝힐 수 있을 뿐만 아니라 언어사용을

심리학 연구의 종속변인으로 사용할 수 있는 범언어적인 이론적 토대를 제공할 것으로 기대된다. 또한, 심리학 분야를 벗어나서 글이나 말을 분석대상으로 하여 쟁점이 되는 연구문제를 해결하거나 실용적인 정보를 제공할 수 있는 신문 방송학, 정치학, 사회학, 문헌정보학 등 다양한 사회과학 분야의 연구자들과의 교류를 통하여 공동연구를 추진하고 국어교육이나 영어를 포함한 이중언어 연구자들과의 *KLIWC*나 *LIWC*를 활용한 교육프로그램 개발을 위한 공동연구를 할 수 있는 가교 역할을 한국어 분석프로그램은 제공할 수 있을 것이다.

참고문헌

고진숙 (1987), 조선리론문법: 품사론, 과학백과사전출판사: 평양.

권혁철 (1990), 자연언어 처리에 있어서의 형태소 처리 및 어휘 사전, 어학연구, 26(1), 253～259.

권혁철, 김민정 (1992), 한국어 특성을 이용한 자동 색인 기법, 1992 가을 학술발표논문집, 19(2), 1005～1008.

김기찬 (1998), 영어와 한국어의 소절구문의 비교분석, 언어과학연구, 15, 53～74.

김민정 (1997), 규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거, 부산대학교 전자계산학과 박사학위논문.

김영택 외 (2001), 자연언어처리, 생능출판사: 서울, 144～174

김영화 (1997), 품사개조와 문장구조, 언어, 22(3), 413～432.

김용전 (1988), 영어의 품사에 관하여, 어문학논총, 7, 179～190.

김재훈, 이공주 (2003), 사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정, 한국정보처리학회 정보처리학회 논문지B, 10(1), 47～56.

김진형 (1996), 품사전환에서의 의미변화, 언어, 21(4), 1009～1023.

남기심 (1982), 국어의 공시적 기술과 형태소의 분석, 배달말학회 배달말, 7, 1～10.

박민경, 김민정, 권혁철 (1997), 부분 파싱을 이용한 한국어 명사구, 술어구와 접사의 색인 기법, 한국 정보과학회 '97 봄학술발표 논문집, 24(1), 491～494.

서정수 (1996), 국어문법, 한양대학교 출판원: 서울.

소길자, 권혁철 (2001), 어휘적 중의성 제거 규칙과 부분 문장 분석을 이용한 한국어 문법 검사기, 한국정보과학회 정보과학회 논문지: 소프트웨어 및 응용, 28(3), 305～315.

안신호 · 이승혜 · 권오식 (1993), 정서의 구조: 한국어 정서단어 분석, 한국심리학회지: 사회, 7(1), 107～123.

연세대학교 빈도사전 (1989), 한국어사전 편찬실.

윤애선 (2001), 표준화된 전자사전 개발을 위한 정보구조 및 사용자 환경 설계와 그 응용, 한국 프랑스학 논집, 33, 41～58.

윤애선, 권혁철 (2001), 효율적인 표제어 검색을 위한 굴절어 전자사전의 구조, 사전편찬학, 115～126.

이광정 (1997), 학교 문법에서의 품사분류, 국어교육, 94, 41～76.

이기용, 이종민, 홍정하 (2002), 언어정보처리

를 위한 데이터베이스 의미론 체계의 구축, 한국언어학회(언어) 언어, 27(3), 417~438.

이동영 (2002), 상황정보에 기반한 한국어대화의 전산적 처리와 표상구조의 구축, 한국정보처리학회 정보처리학회논문지 B, 9(6), 817~826.

이상복 (1989), 국어사전 편찬과 문법형태소의 처리: 조사와 연결어미의 기술을 중심으로, 연세대학교 언어정보개발원 사전편찬학연구, 2, 63~91.

정달영 (1991), 국어학(國語學) 및 한어학(漢語學): 국어 품사론(品詞論) 고(考), 국제어문, 12, 19~48.

정영자 (1991), 박승민의 품사분류체계에 대한 소고, 청람어문학, 5, 92~129.

조세형 (2001), 자율 학습에 의한 실질 형태소와 형식 형태소의 분리, 한국정보처리학회 정보처리학회논문지B, 8(6), 675~684.

채영숙, 최성필, 서정현 (2002), 자동 색인을 위한 한국어 형태소 분석기의 실제적인 구현 및 적용, 한국정보처리학회 정보처리학회논문지 B, 9(5), 689~700.

최덕수, 배해수, 김광해 (1998), 한국학 정보화를 위한 언어자원과 문서표준 및 기반기술의 연구: 한국학 정보 처리를 위한 학술용 시소러스 연구, 한국어전산학회 한국어전산학, 2, 171~194.

한인석 (1982), 국어어휘의 품사처리에 대한 관견, 국어교육, 42, 203~234.

허윤영, 권혁철 (1994), 의미적 한 단어 유형분석 및 형태소 분석기법, 94 한글 및 한국어 정보처리 학술발표논문집, 128~131.

유석훈 역 (1999), 언어와 컴퓨터, 고려대학교 출판부: 서울, 48~53.

Booth, R. J., Petrie, K. J., & Pennebaker, J. W. (1997), Changes in circulating lymphocyte numbers following emotional disclosure: Evidence of buffering?, *Stress Medicine*. 13, 23~29.

Buller, D. B., Burgoon, J. K., Busling, A., & Roiger, J. (1996), Testing Interpersonal Deception Theory: the language of interpersonal deception, *Communication Theory*, 6, 268~289.

Choi, Wankyu 외(2001), Begining PHP4, *Wrox Press*: Birmingham, 511~518.

Chung, Y. J., Park, J. H. Kwon, H. Ch., & Yoon, A. S. (2000), An Improved Korean Morphological Analyzer, *Artificial Intelligence and Soft Computing, Banff, Alberta, Canada*, 466~471

Freud, S. (1901), *Psychopathology of everyday life*, New York: Basic Books.

Goffman, E. (1959). *The presentation of self in everyday life*, Garden City, NY: Doubleday.

Hart, R. P. (2001). Redeveloping DICTION: theoretical considerations, *See West 2001*, 43~60.

Jeong, H. W., & Yoon, A. S. (2001), Linguistic Database Handling using XML in Web Environment, *Proceedings of 2001 IEEE International Conference on Industrial Electronics*, 833~838.

John. O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 4b*, Berkeley: Institute for Personality and Social Research, University of California.

Kim, M. J., & Kwon, H. Ch. (1997), A Korean Phrasal Indexing using a Greedy Parsing, *IRAL*, 88~94.

Klein, K., & Boals, A.(2001). Expressive writing can increase working memory capacity. F. *Journal of Experimental Psychology: General*, 130, 520~533.

Knapp, M. L., Hart, R. P., & Dennis, H. S. (1974), An exploration of deception as a communication construct, *Human Communication Research*, 1, 15~29.

Kwon, H. Ch., & Yoon, A. S. (1991), Unification-Based Dependency Parsing of Governor-Final Languages, *Second International Workshop on Parsing Technology*, 182~192.

Kwon, H. Ch., Park, Y. U., Yoon, A. S. (1990), A Korean Parser with Unification-Based Dependency Grammar, *'90 Seoul International Conference on Natural Language Processing*.

Lepore, S. J. (1997), Expressive writing moderates the relation between intrusive thoughts and depressive symptoms, *Journal of Personality and Social*

Psychology, 73, 1030~1037.

Mehl, R. M., & Pennebaker, J. W. (2003), The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversation, *Journal of Personality and Social Psychology*, 84, 857~870.

Mehl, R. M., & Pennebaker, J. W. (2003), The social Dynamics of a Cultural Upheaval: Everyday Social Life in the Aftermath of the September 11 Attack on America, *Psychological Science*,14(6) 579-585.

Mergenthaler, E. (1996), Emotion-abstraction patterns in verbatim protocols: a new way of describing psychotherapeutic processes, *Journal Consulting and Clinical Psychology*, 64, 1306~1315.

Morgan, C., & Murray, H. A. (1935), A method for investigating fantasies: The Thematic Apperception Test, *Archives of Neurology and Psychiatry*, 34, 289~306.

Murray, E. J., & Segal, D. L.(1994), Emotional processing in vocal and written expression of feelings about traumatic experiences, *Journal of Traumatic Stress*, 7, 391~405.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003), Lying words: predicting deception from linguistic styles, *Personality and Social Psychology Bulletin*, 29, 665~675.

Pennebaker, J. W. (1997), Writing about emotional experience as a therapeutic process, *Psychological Science*, 8, 162~166.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: disclosure, personality, and social integration, *Current Directions in Psychological Science*, 10, 90~93.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference, *Journal of Personality and Social psychology*, 77, 1296~1312.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001), Linguistic Inquiry and Word Count (LIWC): LIWC 2001. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997), Linguistic predictors of adaptive

bereavement, *Journal of Personality and Social Psychology*, 72, 863～871.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves, *Annual Review of Psychology*, 54, 547～577.

Petrie, K. P., Booth R. J., & Pennebaker J. W. (1999). The immunological effects of thought suppression, *Journal of Personality and Social Psychology*, 75, 1246～1272.

Petrie, K. P., Booth, R. J., Pennebaker, J. W., Davison, K. P., & Thomas, M. G. (1995), Disclosure of trauma and immune response to a hepatitis B vaccination program, *Journal of Consulting and Clinical Psycology*, 63, 787～792.

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (Submitted). *Language Use of Depressd and Depression-Vulnerable College Students.*

Stirman, S. W., & Pennebaker, J. W. (2001), Word use in the poetry of suicidal and non-suicidal poets, *Psychosomatic Medicine*, 63, 517～522.

Stone, L. D., & Pennebaker, J. W. (2002), Trauma in real time: Talking and avoiding online conversations about the death of princess Diana, *Basic and Applied Social Psychology*, 24, 172～182.

Stone, P. J., Durphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966), *The General Inquirer: A Computer Approach to Content Analysis*, Cambridge, MA: MIT Press.

Yoon, A. S., & Kwon, H. Ch. (1997), Rule-Based Morphological Disambiguation using Statistics in an Agglutinative Language, *The Proceedings of the 16th International Congress of Linguists*, (published in CD-ROM, ISBN 0-08-043- 438X).

Yoon, A. S., & Kwon, H. Ch. (2000), Building Reusable Contents for Electronic Dictionaries, *Proceedings of the SICOL 2000*, 320～330.

Yoon, A. S. (1989), *Les attitudes interlocutives: Les formes de politesse en français et en coréen*(대화자 태도: 불어와 한국어의 경어법(in French)), Thèse de Doctorat, Université de Paris-Sorbonne.

부록 1. LIWC의 72개 변인과 관련 단어 정보

차원	약자	단어 예	해당단어수
I. STANDARD LINGUISTIC DIMENSION			
Word Count	WC		
Words per sentence	WPC		
Sentences ending with?	Qmarks		
Unique words(type/token ratio)	Unique		
% words captured, dictionary words	Dic		
% words longer than 6 letters	Sixltr		
Total pronouns	Pronun	I, our, they, you're	70
1 st person singular	I	I, my, me	9
1 st person plural	We	we, our, us	11
Total first person	Self	I, we, us	20
Total second person	You	you, you'll	14
Total third person	Other	she, their, them	22
Negation	Negate	no, never, not	31
Assents	Assent	yes, OK, *mmhmm*	18
Articles	Article	a, an, the	3
Prepositions	Preps	on, to, from	43
Numbers	Number	one, thirty, million	29
II. PSYCHOLOGICAL PROCESSES			
Affective Emotional Processes	Affect	happy, ugly, bitter	615
Position Emotions	Posemo	happy, pretty, good	261
Positive feeling	Posfeel	happy, joy, love	43
Optimism and energy	Optim	certainty, pride, win	69
Negative Emotions	Negemo	hate, worthless, enemy	345
Anxiety or fear	Anx	nervous, afraid, tense	62
Anger	Anger	hate, kill, pissed	121
Sadness or depression	Sad	grife, cry, sad	72
Cognitive Processes	Cogmech	cause, know, ought	312
Causation	Cause	because, effect, hence	49
Insight	Insight	think, know, consider	116
Discrepancy	Discrep	should, would, could	32

Inhibition	Inhib	block, constrain	64
Tentative	Tentat	maybe, perhaps, guess	79
Certainty	Certain	always, never	30
Sensory and Perceptual Processes	Senses	see, touch, listen	111
Seeing	See	view, saw, look	31
Hearing	Hear	heard, listen, sound	36
Feeling	Feel	touch, hold, felt	30
Social Processes	Social	talk, us, friend	314
Communication	Comm	talk, share, converse	124
Other reference to people	Otherf	1st, pl, 2nd, 3rd, per prns	54
Friends	Friends	pal, buddy, coworker	28
Family	Family	mom,brother, cousion	43
Humans	Humans	boy, woman, group	43
Ⅲ. RELATIVITY			
Time	Time	hour,day, oclock	113
Past tense verb	Past	walked, were, had	144
Present tense verb	Present	walk, might, shall	256
Future tense verb	Future	will, might, shall	14
Space	Space	around, over, up	71
Up	Up	up, above, over	12
Down	Down	down, below, under	7
Inclusive	Incl	with, and, include	16
Exclusive	Excl	but, except, without	19
Motion	Motin	walk, move, go	73
Ⅳ. PERSONAL CONCERNS			
Occupation	Occup	work, class, boss	213
School	School	class, student, college	100
Job or work	Job	employ, boss, career	62
Achievement	Achieve	try, goal, win	60
Leisure activity	Leisure	house, TV, music	102
Home	Home	house, kitchen, lawn	26
Sports	Sports	football, game, play	28
Television and movies	TV	TV, sitcom, cinema	19
Music	Music	tunes, song, cd	31

Money and financial issues	Money	cash, taxes, income	75
Metaphysical issues	Metaph	God, heaven, coffin	85
Religion	Relig	God, church, rabbi	56
Deth and dying	Death	dead, burial, coffin	29
Physical states and functions	Physcal	ache,breast, sleep	285
Body states , symptoms	Body	ache, heart, cough	200
Sex and sexual	Sexual	lust, penis, fuck	49
Eating, drinking, dieting	Eating	eat, swallow, taste	52
Sleeping, deaming	Sleep	aslep, bed, dream	21
Grooming	Groom	wash, bath, clean	15
APPENDIX: EXPERMENTAL DIMENSIONS			
Swear words	Swear	damn, fuck, piss	29
Noufluencies	Noufl	uh, rr*	6
Filers	Filers	youknow, Imean	6

부록 2. KLIWC의 시험판에 적용한 단계별 한국어 품사 분류의 예 (일부 발췌)

구분	대분류	중분류	소분류
체언(n)	명사(nn)	일반 명사(nng)	
체언(n)	명사(nn)	고유 명사(nnp)	
체언(n)	명사(nn)	의존 명사(nnb)	일반 의존명사(nnbg)
체언(n)	명사(nn)	의존 명사(nnb)	단위 의존명사(nnbc)
체언(n)	대명사(np)	인칭 대명사(npp)	
체언(n)	대명사(np)	지시 대명사(npd)	
체언(n)	수사(nr)	양수사(nrc)	
체언(n)	수사(nr)	서수사(nro)	
용언(v)	동사(vv)	일반 동사(vvg)	자동사(vvgi)
용언(v)	동사(vv)	일반 동사(vvg)	타동사(vvgt)
용언(v)	동사(vv)	보조 동사(vvx)	
용언(v)	형용사(va)	일반 형용사(vag)	
용언(v)	형용사(va)	존재 형용사(vat)	
용언(v)	형용사(va)	지정 형용사(vac)	긍정 지정 형용사(vacp)
용언(v)	형용사(va)	지정 형용사(vac)	부정 지정 형용사(vacn)
용언(v)	형용사(va)	보조 형용사(vax)	보조 형용사(vax)
용언(v)	관형사(md)	지시 관형사(mdd)	지시 관형사(mdd)
용언(v)	관형사(md)	수 관형사(mdq)	수 관형사(mdq)
용언(v)	관형사(md)	성상 관형사(mds)	
용언(v)	부사(mb)	일반 부사(mbg)	
용언(v)	부사(mb)	일반 부사(mbg)	
용언(v)	부사(mb)	접속 부사(mbc)	
용언(v)	감탄사(ic)		

부록 3. LIWC의 72개 변인과 관련 단어 정보

언어학적 변인

1. 형태소 태깅				
문장 당 어절 수				
문장 당 형태소 수				
체언				
	명사	일반명사		
		고유명사		
		의존명사		
		일반-고유명사		
	대명사	인칭대명사		
		지시대명사		
용언				
	동사	동사		
		일반동사		
		자동사		
	형용사	형용사		
수식언				
	관형사	수관형사		
		관형사		
	부사	일반부사		
		접속부사		
독립언				
	감탄사	감탄사		
수사		양수사		
		서수사		
접사		접두사		
		접미사		
기타		속담구		
		관용구		
		한자어		
		명사추정 외래어		
		명사추정 범주		
		외래어		
		경구		

심리학적 변인

차원	단어	전체 단어수
2. 심리학적 과정		
감정 또는 정서적 과정	행복한, 사랑, 즐거움, 두려움	
긍정적 정서	행복한, 예쁜	
긍정적인 느낌	즐거움, 사랑, 행복함	
낙천성 또는 활동성	승리	
부정적 정서	싫음, 두려움	
불안	두려움	
화	미움	
슬픔 또는 우울	울음, 슬픔	
인지적인 과정	원인, 알다, 앎	
원인	영향, 가정	
사고	분석, 의식	
기대	그러나, 기대하다	
제한	억제, 의무	
추측	어디선지	
확신	신임, 신뢰	
감각 & 지각적 과정	나타나다, 눈, 맛	
사회적 과정	말하다, 우리	
체면	체면치레, 차리다	
의사소통	말하다, 주장	
타인참조	어느 누구, 그, 그녀	
또래(친구)	파트너, 룸메이트	
가족	아버지, 어머니	
인간	소녀, 아줌마	

3. 자기영역		
자기활동	작업, 실패	
학교	학생, 반	
직장 & 일	작업, 사업	
성취	목표, 시도하다	
여가활동	야구, 클래식	
집	집, 부엌	
운동	야구, 축구	
TV 영화	TV, 코미디	
음악	클래식, 팝	
돈 & 재정적 이슈	은행, 수표	
신체적 상태와 기능	화, 자다	
몸 상태와 증상	화, 심장	
성 & 상징	섹스	
식사 & 음주 & 다이어트	마시다, 먹다	
수면 & 꿈	자다, 잠	
형이상학적 이슈		
죽음	죽다, 돌아가시다	
종교	기독교, 불교	
4. 실험적 차원		
속어		
맹세 단어		
어눌한 말		
Filler		

부록 4. KLIWC 주요 변인들의 신뢰도 계수치

변인	전체 M	전체 SD	변인-전체 상관				
			의식적 글쓰기	감정적 글쓰기	3월 글쓰기	5월 글쓰기	전체
문장당 어절	39.43	10.56	.71	.50	.62	.75	.69
문장당 형태소	71.87	20.72	.66	.40	.54	.71	.63
명사	77.43	12.65	.88	.79	.75	.87	.86
조사	57.41	10.33	.83	.78	.72	.82	.80
동사	42.60	7.02	.84	.76	.63	.80	.81
형용사	20.26	4.78	.57	.58	.43	.58	.61
어미	88.37	13.82	.90	.88	.79	.90	.89
일반명사	62.07	10.54	.85	.75	.70	.83	.83
일반부사	19.94	4.68	.65	.59	.32	.60	.61
선어말어미	8.70	2.83	.38	.38	.30	.39	.46
감정 혹은 정서	19.86	4.77	.56	.58	.32	.59	.54
인지적 과정	24.57	5.91	.65	.68	.51	.69	.69
감각과 지각	13.30	3.27	.58	.52	.43	.48	.56
사회적 과정	15.45	4.42	.45	.61	.37	.52	.54
자기활동	13.25	3.64	.44	.42	.30	.54	.50
형이상학적이슈	1.92	.94	.30	.32	.12	.32	.35

부록 5. KLIWC 2판에서 추가될 기능

1. 한국어 분석 시 발생하는 중의성 문제

교착어인 한국어는 1개의 용언이 가지는 활용형의 형태가 2천 개 정도이기 때문에 활용형 간 동형이의어의 비율이 매우 높다. 예를 들어 다음 표에서처럼 '가는'이라는 어절은 3개의 품사 분석, 4개의 기본형 분석, 7개의 의미 분석으로 구분되는 동형어의 어절이다(소길자, 권혁철, 2001; Kwon & Yoon, 1991; Kwon, Park & Yoon, 1990)

이러한 중의성을 해결하려면 논항의 구조, 통사 분석, 의미/문맥 분석이 필요하나, 아직 현 단계에서 한국어의 통사 및 의미 분석 기술은 만족할 만하지 않다. 따라서 구(phrase) 단위로 부분 분석 방식을 일반적으로 사용하는데, 본 연구에서도 후자를 이용하여 가능한 한 중의성의 수를 낮추고자 한다.

2. 유의미한 분석 수위 및 언어적 변인 결정

1차년도에 선행 연구를 이용한 문헌 연구 및 파일럿 테스트를 통해 도출한 결과를 이용하여 KLIWC에서 중요한 변인으로 작용하는 언어적 구성요소를 결정한다. 또한 과잉분석으로 인한 중의성 증가 문제를 해결하기 위해 적절한 분석 수위를 결정한다. 즉, 특정한 어휘를 제외하고 품사, 기본형, 의미 단계 중에서 어느 단계로 분석할 때 유의미한 결과를 낳는지 검토하고 이를 실용화된 KLIWC 프로그램 개발에 반영한다.

3. 실용화된 프로그램 개발

KLIWC 프로그램은 일과성의 연구에 의해 완성된 버전으로 개발하는 것으로 끝날 수 있는 것이 아니다. 오히려 다양한 분야에 응용되면서 새로운 변인의 추가, 기존 변인의 삭제, 변인 간 논리 구조와 상호 관련성의 변경 등이 지속적으로 요구된다. 따라서 언어분석만을 하는 프로그램이 아니라 갱신된 내용을 언제나 반영할 수 있는 전문 개발자용 관리(Manager) 기능이 필요하다. 이를 위해 1차년도에 개발한 Thesaurus Builder, 사전, 규칙 관리 시스템 등을 연동하여 자료 갱신이 가능한 KLIWC Manager을 개발한다. KLIWC Manager의 주요 기능은 다음과 같다.

(1) 계정 관리 및 권한 부여

1) 최고 관리자(administrator)

① KLIWC manager 운영에 필요한 모든 권한 소유

② 일반 사용자 그룹에서 Power User 지정

2) High Level Power User

① Thesaurus Builder 운영에 필요한 모든 권한 소유

② Post-Processing System 운영에 필요한 모든 권한 소유

3) Low Level Power User

① Thesaurus Builder의 단어 관리(3단계) 권한 부여

② Post-Processing System 사용 권한 소유

4) 일반 사용자(Users)

① Thesaurus 검색 권한 소유

② Post-Processing System 사용 권한 소유

(2) Thesaurus Builder

1) 1단계 - 차원(Dimension) 관리

① 등록 - 개별적인 차원 단위로 등록하며 텍스트 파일을 통한 다수 차원의 일괄 입력 기능을 지원하지 않는다.

② 차원에 관련된 모든 작업은 하위 hierarchy의 사전구조에 영향을 미친다.

2) 2단계 - 변인(Variable)의 관리

① 등록 - 상위 레벨인 차원의 중의성을 갖지 않으며 개별 입력은 물론 텍스트 파일을 통한 다수 변인의 일괄 입력 기능을 지원한다.

② 변인과 변인 간에 hierarchy를 가질 수 있고 변인에 관련된 모든 작업은 상위 hierarchy의 사전구조에 어떠한 영향도 미치지 않으나 하위 hierarchy의 사전구조에는 영향을 미친다.

3) 3단계 - 어휘(Thesaurus)의 관리

① 등록 - 상위 레벨인 변인의 중의성을 가질 수 있으며 개별 입력은 물론 텍스트 파일을 통한 다수 어휘의 일괄 입력 기능을 지원한다.

② 어휘에 관련된 모든 작업은 상위 hierarchy의 사전구조에 어떠한 영향도 미치지 않는다.

4) 검색 및 출력

① 차원 검색 - 해당 차원의 하위 레벨에 속하는 변인 및 어휘를 검색과 출력한다.

② 변인 검색 - 상위 레벨의 차원과 하위 레벨의 어휘를 검색 및 출력한다.

③ 어휘 검색 - 상위 레벨의 차원과 변인을 검색 및 출력한다.

(3) Post-Processing System

1) tagging system에 의해 분석된 string의 오류 수정

① 고유명사: tagging 결과 고유명사로 인한 오류가 발생할 때는 고유명사 등록을 통하여 문제를 해결한다.

② 기타 오류: 맞춤법, 수사, 약어, 발음, 한자어, 복합어 등의 오류들은 오류 수정 모드에서 해당 오류 사항을 지정한 후 직접 수정한다.

2) 중의성 감소를 통한 string의 정확한 어휘 정보 추출

① 중의성을 갖는 string의 경우 관련된 모든 기본형을 체크리스트로 제시하여 선택할 수 있도록 구성한다.

1 차원고접수: 2004. 11. 30
2 차원고접수: 2005. 3. 2
최종게재승인: 2005. 6. 5