

컴퓨터 통신어의 어휘 유형 체계

김 윤식

세명대학교 컴퓨터수리정보학과

나 동열

연세대학교 정보기술학부

한 경호

나사렛대학교 정보과학부

Lexical analysis for computer chatting language

Yun-Sik Kim

Dept. of Computer Aided Mathematical Information Science, SeMyong University

Dong-Yul Ra

Division of Information Technology, Yonsei University

Kyoung-Ho Han

Division of Information Science, Korea Nazarene University

요약

가상공간에서의 정보교류가 늘어나면서 컴퓨터 통신언어도 자연어처리의 새로운 연구 분야로 등장하였다. 가장 일상화된 컴퓨터 통신어이며 인터넷상의 실시간 대화어인 채팅언어는 신속한 정보 전달 매체의 기능과 문자언어로서 정보교류와 대화 상황을 함께 내포해야 하는 문어와 구어의 양면성을 동시에 가지는 언어이다. 이러한 특수성으로 컴퓨터 통신언어의 경우 발화자의 개성에 따라 한 의미의 단어도 표기법이 다양하여 어휘수가 매우 많다. 본 연구에서는 컴퓨터 통신에서 사용되는 다양한 어휘를 컴퓨터로 처리하기 위한 어휘의 유형을 체계화하는 방안을 제시한다.

1. 서론

통신 교류와 인터넷 대화방을 통한 교류는 점점 늘어나고 있어 이제 컴퓨터 통신언어도 문어, 구어의 틀이 아닌 제3의 언어로서 연구할 필요성이 있으며, 새로운 생활언어로서의 컴퓨터 통신언어를 정리할 필요가 있다.

본 논문에서는 일상생활에서도 사용되는 컴퓨터 통신언어 말투나 어휘를 올바르게 활용 정착시키고, 컴퓨터에서 효율적인 어휘 처리 방안을 제시하고자 한다.

2. 인터넷 언어

전자 언어(electronic language)란 용어는 Ong에서 처음 언급되었으며, 국내에서는 김수업에서 언급되었다(임규홍, 2000).

권연진(2000)에 의하면 전자 언어의 부분집합인 컴퓨터 통신언어(computer mediated communication language)는 통신상에서 이루어지는 모든 의사소통 과정에서 사용되는 음성·문자를 총칭한다. 컴퓨터 통신언어는 채팅(chatting), 머드게임(MUD: multi user dungeon), 전자게시판(BBS: Bulletin Board System), 전자우편(e-mail) 등 컴퓨터상에서 사용되는 언어와 이모티콘(emoticon) 뿐만 아니라 휴대폰(cellular), 문자 메시지(SMS)까지 포함하는 광범위한 개념으로 정의된다.

통신언어는 일반 언어와 특이한 양상을 보이는데, 시간상의 문제로 인한 축약어의 사용, 매체의 특성상 가질 수 있는 신조어의 사용, 시각에만 의존하는 대화의 극복 방안으로서의 기호의 사용, 발신자의 개인적 성향에 따른 특이한 어투의 사용 등을 들 수 있다(권순희, 2001). 채팅언어는 실제 대화가 사람의 음성을 통하지 않고 컴

퓨터를 키보드를 이용한 문자를 통해 의사 전달이 이루어진다는 점에서 문어적 성격을 띠어야 하나, 비록 대면 대화는 아니어도 실시간으로 이루어지는 대화이므로 구어체에 가까운 문자적 표현을 쓴다.

국내 이 분야의 연구는 1998년부터 시작되었으며, 주요 연구주제는 통신언어의 언어학적 특징(임규홍, 2000), 컴퓨터 통신언어의 특성(권순희, 2001), 통신언어의 유형별 실태(권연진, 2000), 채팅 표현(노형남, 2000), 채팅 언어에 대한 연구(전병철, 2002), 컴퓨터 대화방 어휘의 계량적 연구(임철성, 2000), 채팅이 언어생활에 미치는 영향(권오성, 2001) 등 대부분의 연구가 언어학적, 교육학적, 사회학적 관점에서의 연구였다. 전산언어학적 관점의 연구로는 채팅어 사전연구(김윤식&나동열, 2003), 음운 변형된 채팅어의 복원 알고리즘 연구(김윤식&나동열, 2004) 등이 있다.

본 논문에서는 컴퓨터 통신언어 중에서도 인터넷상에서 사용되는 채팅 언어에 초점을 맞출 것이며, 컴퓨터 자료처리를 위한 컴퓨터 통신어휘에 대한 전산언어학적인 접근을 시도하려고 한다.

3. 채팅언어 분석

일상 언어에 사용되는 문장이 평균 6.5어절인데 비하여 채팅에 사용되는 문장은 평균 2.5어절로 구성된다(임철성, 2000). 일상 언어는 3음절이나 4음절로 구성된 어휘가 많은데 비하여 채팅언어는 2음절과 3음절로 구성된 어절이 80%에 이른다(김윤식&나동열, 2004). 조사도 앞 음절에 축약된 경우가 많고, 서술어의 어간도 시제나 피동, 사동 등 변화보다는 발화자는 개성에 따른 변화가 주류를 이룬다.

표 1. 채팅문장 한 어절의 음절수(200개 조사)

음절 수	1음절	2음절	3음절	4음절	5음절	6개 이상
원시데이터	15	69	71	25	11	9
공백 수정 후	21	84	70	19	6	0
비율	11%	42%	35%	10%	2%	0%

채팅 문장을 파싱하는 경우에는 일반적으로 문장구성과 어휘사이의 의미나 형태적 중의성이 많아 효과적이지 못하며, 채팅 언어의 문장 분석 여건은 관련연구 자료의 부족으로 매우 어렵다. 다만 채팅 언어는 실시간 대화언어이고 빠른 문답 형식으로 진행되므로 문장이 간단하고 어절수가 비교적 적으므로 문장을 분석하기보다는 어휘를 분석하는 것이 더 중요하다. 사용빈도까지 고려하면 채팅 언어는 일반 언어에 비하여 수식어, 지시어, 의성어, 의태어, 감탄사가 매우 많고 반면에 개념어, 서술어는 현저하게 적다(임철성, 2000).

표 2. 언어의 유형별 사용 비율

어절	개념어	지시어	수사	수식어	서술어	독립어
음성 언어	43%	8%	1%	9%	37%	2%
문자 언어	43%	5%	3%	7%	41%	1%
통신 언어	38%	10%	3%	13%	28%	8%

개념어의 많은 부분은 은어, 외래어, 합성어가 차지하고 있다. 이와 같은 개념어와 이모티콘 등은 채팅 사전에 의하여 표준 어휘와 일대일 대응을 시킬 수 있으므로 본 논문에서는 조사를 포함하는 연음, 자모음 축약, 자모음 추가, 자모음 치환을 중심으로 채팅 언어의 생성 규칙을 분석하기로 한다.

하나의 한국어 어절은 음절의 결합으로 나타나고, 하나의 음절은 초성(onset), 중성(peak), 종성(coda)으로 조합되므로 어절을 구성하는 음절을 w 라 할 때 임의의 어절 W 는 다음과 같이 표현할 수 있다.

$$W = \prod_{i=1}^m w_i = \prod_{i=1}^m (opc)_i \quad (1)$$

여기서 o 는 초성 자음(onset consonant), c 는 종성 자음(coda consonant)이며, p 는 모음(vowel)이다.

하나의 채팅 문장이 존재할 때 불필요한 공백을 제거한 채팅 문장을 구성하는 각 어절을 W_i 라 하면 채팅 문장 S 는 다음과 같이 표현된다.

$$S = \sum_{i=1}^n W_i \quad n \leq 4 \quad (2)$$

표준 어휘를 W' 이라 하고 표준 어휘를 채팅 어휘로 변환시키는 연산을 f 라 하면

표준 어휘와 채팅 어휘의 매핑은 다음과 같이 표현할 수 있다.

$$W = fW' = f \prod_{i=1}^m (opc)_i \quad m \leq 3 \quad (3)$$

서로 연결된 음절과 음절사이의 자음 연산을 다음과 같이 정의한다.

$$f(opc)_i (opc)_j = \alpha_{ij} (opc'_i) ; (o'pc)_j + \beta_{ij} (o'pc)_i ; (opc'_j) \quad (4)$$

여기서 α_{ij} , β_{ij} 는 초성과 종성 자음의 변환 정도를 나타내는 자음변환계수이다.

또한 서로 연결된 음절과 음절사이의 모음 연산을 다음과 같이 정의한다.

$$f(opc)_i ; (opc)_j = \gamma_{ij} (o ; p_{ij}' c) + \delta_{ij} (op'c)_i ; (op'c)_j \quad (5)$$

여기서 γ_{ij} , δ_{ij} 는 모음의 변환 정도를 나타내는 모음변환계수이다.

서로 연결된 음절과 음절사이의 자모음 연산은 다음과 같이 정의한다.

$$f(opc)_i(opc)_j = \eta_{ij}(o_i p_{ij}' c'_i) \quad (6)$$

여기서 η_{ij} 는 자음과 모음의 변환 정도를 나타내는 자모음변환계수이다.

받침이 없는 음절의 종성을 $c=n$, 첫 소리가 목음인 음절의 초성을 $o=n$ 으로 표시하면, 두 음절이 결합하는 자음종류에 따라 자음 연산 (4)식은 다음과 같이 고려할 수 있다.

1) $\beta_{ij}=0$ 인 경우

가) $c_i \neq o, o_j = o$ 이면 $c_i n = n c_i$
 맞아→마자, 밀으로→미트로, 인간이→인가니, 싫어→시퍼, 놀아→노라, 물어→무러, 있어요→이썬요, 알아→아라→아러, 분위기→부늬기, 멋있다→머싯다, 좋아→조아, 많이→마니, 싫어→시러

나) $c_i \neq o, o_j = \emptyset$ 이면 $c_i n = n c'_i$
 축하→추카, 편하게→편케, 착하게→차카게, 열심히→열씨미

다) $c_i = \emptyset, o_j \neq o$ 이면 $c_i o_j = n o'_j$
 좋다→조타, 그렇지→그러치, 그렇죠→그러쵸, 어떻게→어떠케, 많다→만타, 싫다→실타,

라) $c_i \neq o, o_j \neq o$ 이면 $c_i o_j = c'_i o'_j$
 반가와→방가와, 친구→칭구, 얼른→얼릉→언능, 쫓나→졸라, 있다→있다, 했습니다→헛습니다,

2) $\alpha_{ij}=0$ 인 경우

가) $o_i = o'_i$

감사→캄싸, 줌→쫘, 사랑해→따랑해, 사랑→싸랑, 사부→싸부, 컴퓨터→컴뿌터, 했어→헛쨌, 알지→알쨌, 맛있어→맛있쨌,

나) $c_j = n$ 이면 $c_j = o/\wedge/\square/\flat$

아니다→아니당, 네→넵, 나아→나얏→놔얏, 예쁘다→예브당, 하이→하잇, 외롭다→외롭당, 되세요→되세엌, 예→얏, 굿바이→굿바잇, 맛있냐→맛있남→마싯남, 없다→없당, 했어요→했어엌, 빨리→빨랑, 안녕하세요→안나세엌→안나세엌, 어디 살아요→어디사썬

두 음절이 결합하는 자음의 결합에 따라 모음 연산 (5)식은 다음과 같이 고려된다.

1) $\delta_{ij}=0$ 인 경우

가) $c_i = n, o_j = n$ 이면 $c_i = c_j$

메일→멜, 애인→앤, 게임→젼, 제일→젤, 토요일→토을, 서울→설, 어서와→어와, 마음→맘, 다음→담, 처음→침, 너를→널, 나는→난, 그럼→금, 했습니다→헛습니다, 미안합니다→미안함다, 짜증나→짱나, 지금→짐, 그냥→강,

나) $c_i = n, o_j = n, p_j = \uparrow/\downarrow$ 이면 $p'_j = \uparrow/\downarrow$

언니야→언냐, 어디야→어다, 자기야→자갸, 미안→만, 오디오→오도, 라디오→라도, 비디오→비도

다) $c_i = n, o_j = n, p_{ij} = p_i + p_j$

수업→썬, 고마워→고맬, 선생님→썬, 들어와서→들와서, 이야기→야그, 알았지→알지, 어서 오세요→어샵샤, 모범생→범생, 잠시만→잠만, 반갑습니

다→방갑따, 이야기→야기, 습니다→ 습다, 홈페이지→홈페이지

2) $\gamma_{ij}=0$ 인 경우

가) ㄴ→ㄷ, ㄹ→ㄺ, -/|→-
 놀려고→놀려구, 기쁘고→기쁘구, 고
 등→고딩, 감사→감사, 이런→이른, 그
 럼→구름, 뭇일→뭇일, 뭐야→모야, 죄
 송→지송, 그래→구래, 저요→저여, 전
 데요→전데여, 안녕→안녕,
 나) |→기, ㅏ→나
 알았지→알았취, 아니지→아니취, 했
 지→했취, 들었지→들었취, 그래야지
 →그래야취, 그런데→그런데여, 그럼
 →그러엄, 감사→감수아

두 음절의 자음과 모음 결합에 따라 자
 모음 연산 (6)식은 다음과 같이 고려할 수
 있다.

가) $c_i=n, c_j=ㅇ/ㄴ/ㅇ/ㅁ/ㅂ$ 이면 $c_i=ㅇ$;
 너무→넙, 아무튼→암튼, 어느→언, 우
 리→울, 그리구→글구, 재미→잼, 여러
 분→열분, 아무거나→암꺼나, 그러면
 →글면

나) $c_i=n, o_j=ㅎ, p_j=ㅈ/ㅊ$ 이 면
 $p'_{ij}=ㅈ/ㅊ$

시험→섬, 무지하게→무자게, 그렇게
 →글게, 미안해→마내

다) $c_i=n, c_j=ㄴ$ 이 면 $c_i=ㅇ$;
 때문→땨, 하면→함, 싫으면→실음, 하
 여튼→하튼, 때문에→땨에→땨메, 그
 런데→근데, 그렇군요→글کنون, 그러
 면→글면→금

4. 결론

문자 언어 입자에서 볼 때 채팅 언어는
 보다 풍부한 감정이나 시각적인 느낌을
 표현하는 효과를 부여하며, 음성 언어 입
 장에서 볼 때 채팅 언어의 기록과 저장을
 통하여 문자 언어와의 격차를 좁힌다는
 효과를 기대할 수 있다. 채팅 언어가 일종
 의 문자 언어와 유사하게 기록된다는 점
 에서 컴퓨터를 이용한 자료처리 문제가
 대두되며, 우리는 본 논문을 통하여 컴퓨
 터에서 처리하기 위한 전 단계로서 보편
 적으로 사용되는 채팅 언어의 어휘 분석
 과정을 제안하였다. 이를 바탕으로 채팅
 언어에 대한 표준 문장으로의 변환 과정
 을 체계화할 수 있으리라고 생각된다.

[참고문헌]

[1] 권순희 (2001). "컴퓨터 통신 대화의 언어
 적 특성 고찰", 국어교육, 제105권, 143-169.
 [2] 권연진 (2000). 컴퓨터 통신 언어의 유형별
 실태 및 바람직한 방향, 언어과학, 제7권 2호,
 5-27.
 [3] 권오성, 김윤덕, 김민기 (2001). ICT활용을
 통한 교과교육발전방향 : 컴퓨터 채팅이 아동
 들의 언어생활에 미치는 영향, 한국정보교육학
 회논문집, 제6권 2호, 227-233.
 [4] 김윤식, 나동열(2004), 음운 변형된 채팅어
 의 복원 알고리즘 연구, Journal of The
 Korean Data Analysis Society(한국자료분석학
 회지), 제6권 1호. 256-265.
 [5] 노형남 (2000). 한국어 채팅 표현에 관한
 연구, 사회언어학, 제8권 2호, 107-137.
 [6] 임규홍 (2000). 컴퓨터 통신언어에 대하여,
 배달말, 제27권, 23-59.
 [7] 임철성 (2000). 컴퓨터 대화방 글말의 어
 휘에 대한 계량적 고찰, 국어교육학연구, 제10
 권, 385-406.
 [8] 전병철(2002). 채팅언어에 대한 연구, 개신
 어문연구, 제17권, 115-137.