# 대화체 기계번역을 위한 사전의 통사, 의미 정보*

강 범 모**

〈차 례〉



## 0. 서 론

본 연구는 궁극적으로 대화체 한국어의 기계번역을 위해 구축해야 할 사전에 필요한 정보에 관한 기초적 조사이다. 물론, 음운, 통사, 의미 등 모든 정보가 중요하고, 특히 대화체 언어가 대상인 만큼, 음운적 정보를 어떻게 다룰 것인가가 아주 중요한 것이지만, 여기서는 이 문제를 다른 연구에 미루고, 주로 통사, 의미 정보의 종류 및 표시에 관하여 고찰하고자 한다.

이 연구가 기계번역이라는 궁극적 용도를 염두에 두고 실행되지만, 여기서 채택하는 기본적 태도는, 기계번역뿐 아니라, 다른 어떤 용도에도 쉽게 쓰일 수 있도록 가능한 한 포괄적이고 많은 정보를 사전에 수록하는 것이 중요하며, 이러한 태도가 결국 기계번역을 위한 사전을 만드는 정도인 동시에 첩경이라는 것이다. 특별히, 대화체의 경우 문어와는 달리 생략 및 축약이 빈번하여, 문서의 기계번역의 경우보다도 더 많은 통사, 의미 정보가 필요할 것이다. 생략된 만큼의 정보를 추출해 내려 할 때, 이미 언어지식으로 가지고 있는 언어적 정보가 기반이 되어야 하기 때문이다. 음운의 인식이라는 아주 기본적인 층위에서 고찰하더라도, 더 많은 통사, 의미 정보의 필요성은 확실해진다. 예를 들어, 우리말의 모음 ㅐ와 ㅔ는 음운, 철자상 구분이 되지만 빠른 대화에서 그 음성적 차이는 없거나 아주 미미하다. 이 경우 철자상에서 발생하지 않는 중의성의 문제가 생길 수 있으며, 그 중의성 해결의 중요한 요건은 통사, 의미 정보이다. 따라서, 자세한 통사, 의미 정보의 표시가 우리의 궁극적 목적인 대화체 기계번역을 위한 사전에 필수적임을 알 수 있다.

* 이 논문의 연구는 한국전기통신공사 연구개발단에서 지원하는 「한국어특질 및 대화체 기계번역에 관한 연구」의 일부로 수행되었음.

** 고려대학교 언어학과 조교수

사전 연구는 항목의 선택 및 배열에 관한 거시구조(macro-structure) 연구와, 사전의 각 어휘항목에 필요한 정보의 선택, 표시 및 배열에 관한 미시구조(micro-structure) 연구로 나눌 수 있다. 본 연구는 순전히 후자에 속한다. 실제 기계번역의 과정에서 거시구조 또한 중요한 요인이 될 것이나, 이것은 이 연구의 대상에서 제외한다.

## 1. 대화체 한국어에서의 중의성

앞서 언급했던, 대화체에서의 빠른 발음으로 인한 단어 또는 구절의 중의성에 대하여 구체적인 예들을 통하여 좀더 고찰해 보자.

어절의 중의성의 문제는 문서에서도 많이 나타난다. 김경서 외(1991)에 따르면, 보통의 형태소 분석기로 일반적인 문장의 어절을 분석할 때, 13% 정도의 어절에 대하여 원형이 두 개 이상으로 분석되는, 사전 표제어 중의성 현상이 나타난다. 예를 들어, '물어'(1. 묻다 ; 2. 물다)와 같은 용언류와 용언류 사이의 중의성, '한'(1. 하다 ; 2. 한)과 같은 용언류와 체언류 사이의 중의성, 그리고 '강의'(1. 강의 ; 2. 강)과 같은 체언류와 체언류 사이의 중의성이 있다. 여기서 용언류와 체언류의 문법적 성질이 다르므로, 이 둘 사이의 중의성은 형태적, 문법적 용례를 제한함으로써 많이 해결할 수 있다(김경서 외 1991). 문제는 체언류와 체언류, 용언류와 용언류 사이의 중의성인데, 이것들은 앞으로 제시할 통사, 의미 정보를 어휘항목에 많이 공급함으로써 가능한 한 국부적으로 해결할 수 있을 것이다.

그런데, 앞의 13%라는 수치는 문서상에 나타나는 언어의 표제어적 중의성이다. 소리로 이루어진 대화체 텍스트에는 이보다 훨씬 심한 정도의 중의성이 나타날 것이다. 그 중 몇 가지를 생각해 보자.

한국어에서 긴 소리와 짧은 소리가 변별적 기능을 할 때가 있으나, 실제 발음에서 구분이 안 될 수 있다. 이 경우는 문서 텍스트에서도 중의적이다.

(1) 눈, 말, 밤, 주사, 소망

모음 중, ㅐ와 ㅔ는 일부 방언에서 전혀 구별이 되지 않을 뿐 아니라, 이것이 구분되는 방언에서조차 그 구분이 어려울 적이 많다. 이로 인하여 많은 어절들이 문서 텍스트에서와는 달리 중의성을 갖는다.

(2) 개/게, 축재/축제, 분재/분제, 재동/제동, 재수/제수, 재독/제독, 재물/제물, 재소/제소, 재정/제정, 새우/세우, 새길/세길, 새차/세차, 재삼/제삼
ㅐ 또는 ㅔ는 때로 ㅖ와도 구별이 잘 안된다.

(3) 가게/가제, 개간/계간, 개시/계시, 개집/계집

모음 ㅗ와 ㅜ도 사람에 따라서는 큰 구별 없이 발음하는 수가 있다.

자음의 경우, ㄱ과 ㅋ, ㄷ과 ㅌ, ㅂ과 ㅍ, ㅈ과 ㅊ, 등 보통 소리와 거친 소리가 어두에서 잘 구별되지 않을 수가 있는데 예를 들어 '달/탈', '발/팔', '자다/차다' 등의 발음이 혼동될 수 있다.

사실상 이러한 대화에서의 중의성은 언어적, 비언어적 맥락 속에서 자연히 해결되는 수가 많은데, 사전에 이러한 맥락적 정보를 가능한 한 많이 기재해 놓음으로써, 사전을 이용해 (발음된) 문장을 분석할 때 이용할 수 있다. 예를 들어,

(4) ㄱ. 개가/게가 짖는다.
ㄴ. 개가/게가 집게로 물었다.

전자의 경우, '짖다'라는 표현이 개를 포함한 짐승에만 쓰인다는 통사, 의미적 정보가 사전에 기재되어 있고 활용 가능할 때, 개/게의 중의성 문제는 발생하지 않는다. 후자는, 게만이 집게를 가지고 있다는 백과사전적 정보가 관계하여 중의성이 해결된다. 따라서, 우리는 통사, 의미적 정보뿐 아니라 때로는 어느 정도의 백과사전적 지식이 사전 정보의 일부로 등재되어 있어야 효과적인 기계번역이 가능함을 알 수 있다.

이제, 구체적으로 어떤 정보들이 필요한지 살펴 보고, 이러한 정보의 표시의 문제를 생각해 보자.

## 2. 통사론적 정보

전통적으로 사전항목의 가장 중요한 통사적 정보로서 품사의 분류와 형태론적인 정보가 인식되어 왔다.

후자는 명사는 격, 수, 성, 그리고 동사의 시제, 인칭, 수에 따른 형태의 변화를 말하는데, 영어 등 굴절어의 경우 각각의 어휘항목을 완전히 굴절된 형태로 등재하고 이러한 형태론적 정보를 제공하는 것이 가능하다. 특히 기계번역을 위한 사전의 구성에서는 이 방법이 많이 쓰인다(이 기용 1989 등). 그러나 첨가어의 하나인 한국어의 사전 구성을 이런식으로 할 수는 없으며, 설사 가능하다고 할지라도 비경제적이다. 첨가어에서는 어간과 어미의 구분이 명확하고, 각 어미가 개별적인 문법적, 비문법적 의미를 갖기 때문이다. 어미의 숫자도 방대하여, '-게, -고, -고자, -구먼, -기로니, -나, -냐, -노라고, -는군요, -는다마는, -는단다, -니까, -니만큼, ...' 등의 어미와 결합된 모든 동사의 목록을 사전에 올릴 수는 없는 노릇이다. 따라서 한국어 사전의 항목은 형태소 분석의 가정 하에 결정되어야 할 것이며, 이것을 바탕으로 구성되는 영한, 한영 등의 변환 사전도 이러한 점을 반영해야 할 것이다.

품사의 지정은 한 언어 표현이 그 언어의 어절, 또는 문장에서 다른 표현과

갖는 대강의 결합관계(syntagmatic relation) 및 대치관계(paradigmatic relation)를 압축적으로 표시해 놓은 것이다. 그러나 사실상 전통적인 품사의 구분은 표현의 문법적 쓰임에 대한 정보만을 제공할 뿐이다. 따라서, 외국어를 배우는 사람이 한 표현의 품사만을 알았다고 해서 그 표현을 문장에서 정확히 사용할 수 있는 가능성은 희박하다. 예를 들어, '때리다'가 동사임을 알고 이것이 여러 가지 동사 어미를 취할 수 있음을 안다고 하더라도 '영수가 영희에게 순자를 때렸다'와 같은 문장의 비문법성은 알아차릴 수 없다. 이것을 알기 위해서는 '때리다'가 동사 중에서도 주어와 목적어만을 취하는 동사라는 정보가 필요한 것이다. 이러한 정보가 바로 하위범주화(subcategorization) 정보로서, 기존의 사전은 자동사/타동사의 구분 등 극히 일부의 하위범주화 정보만을 제공할 뿐이다.

영어 사전 중에서 빈약한 하위범주화 정보의 예외로서, Oxford Advanced Learner's Dictionary(OALD) 와 최근의 Longman Dictionary of Contemporary English (LDOCE: 1978, 1987), 그리고 Colins COBUILD English Language Dictionary (1987) 등이 있다. 이러한 사전들은 고유한 문법코드(grammar code)를 사용하여 동사 및 명사의 자세한 하위범주화를 시도한다. Akkerman (1989)는 앞의 두 사전, OALD와 LDOCE의 문법코드 체계를 비교 검토하여, LDOCE의 체계가 우월함을 보이고 있는데, 그 주요 특징을 살펴 보자.

LDOCE는 명사, 형용사, 동사를 각각 품사에 따른 기준에 의해 자세히 분류한다. 이것은 주로 A, F, I, T, C, U 등 대문자에 의해 코드화된다.[1] 형용사는 수식하는 명사의 앞에만 올 수 있는지(예: main), 명사 뒤에만 올 수 있는지(예: elect), 서술적으로만 쓰일 수 있는지(예: asleep), 또는 위치에 구애를 받지 않는지에 따라 구분된다. 명사는 가산성과 수의 일치 등과 관련하여, 가산 명사, 비가산 명사, 집단 가산 명사(예: committee), 집단 비가산 명사(예: admirality), 단수로만 쓰이는 명사, 복수로만 쓰이는 명사 등으로 구분된다. 동사는 자동사, 타동사, 계사(copula) 등으로 크게 구분되며, 초판에서는 타동사를 취하는 목적어의 갯수에 따라 다시 분류하였으나 수정판에서는 타동사 내에서의 세분화는 이제 곧 언급할 논항의 갯수, 종류 구분과 중복되므로 피하였다.

명사, 형용사, 특히 동사는 그 앞과 뒤에 어떠한 표현이 올 수 있는가에 따라 더욱 하위범주화된다. 즉, 어떤 표현도 따라나올 수 없는 것, 하나 또는 그 이상의 명사나 대명사가 따라나올 수 있는 것, to-부정사가 따라나오는 것, to 없는 부정사가 따라 나오는 것, -ing 형태가 따라나오는 것, that-절이 따라나오는 것, wh-단어가 따라나오는 것, 등으로 구분이 된다.

여기서 주목할 점은 사전에서 말하는 하위범주화와 문법이론에서 말하는 하위범주화가 약간의 차이를 가진다는 것이다. 일반적으로 문법이론에서의

1) LDOCE의 초판과 개정판의 문법코드는 좀 다르나, 기본적인 하위범주화의 체계면에서는 근본적으로 동일하다.

하위범주화는 어떤 어휘항목이 취하는 보어의 순서 및 종류에 관한 것이다. 이것은 앞의 LDOCE의 하위범주화 기준 중 후자의 것만을 의미한다. 물론 동사에 관하여 말하자면, 보어의 순서와 종류가 동사 구분의 모든 것일 수 있으며, 문법이론에서의 하위범주화가 주로 동사의 분석과 관련되어 온 사실을 생각해 볼 때, GB, HPSG(Pollard and Sag 1987, 1991) 등 문법 이론에서 이것만을 하위범주화로 파악한 것이 이상한 것은 아니다. 그러나 앞에서 언급한 바와 같이 명사 및 형용사의 경우, 어떠한 보어를 취하느냐와 더불어 다른 분류 기준이 필요하므로, 하위범주화의 표시는 이런 점까지를 고려해야 하는 것이다.

한국어의 경우에도 자세한 하위범주화의 표시가 명사 및 동사에 필요할 것이다. 형용사의 경우, 명사의 앞뒤 위치에 따른 하위범주화가 필요하지 않은데, 이는 명사 앞에만 오는 수식어가 관형사라는 고유한 품사로 인정되어 있기 때문이다. 동사의 하위범주화는 일찌기 그 필요성이 인정되어 문장 유형과 관련하여 연구되어 왔으나, 명사의 경우 많은 연구가 없었던 듯이 보이므로 이 점에 대하여 좀더 깊이 살펴 보자.

한국어의 특수한 명사로서, 불완전 명사가 있다. 이는 완전 명사와 달리 홀로 쓰일 수 없고 반드시 다른 표현과 같이 연결되어 쓰일 수 있는 명사이다. 불완전 명사의 정보는 기존의 분류체계를 따르면 되겠으나(남 기심, 고 영근 1985 등) 아직 완전하지는 않은 것 같다.

먼저 완전 명사의 수(number)에 대하여 알아 보자.

한국어의 수는 인구어의 경우와 달리 단수, 복수의 구분이 그리 명확하지 않은 것이 사실이다. 또한 개념적으로 가산, 물질(비가산) 명사의 구조가 명사+수사+분류사(classifier)등의 동일한 구조로 표현된다.

(5) ㄱ. 개 세 마리
ㄴ. 물 세 잔

그러나, 한국어에서도 가산, 비가산 명사의 구분이 필요하다는 것은 다음과 같은 면에서 두 종류의 명사가 다르게 행동한다는 점에서 분명하다. 우선, 복수성을 나타내는 '-들'은 가산명사에만 붙을 수 있다.

(6) ㄱ. 개들
ㄴ. *물들

다음, 명사 앞에 올 수 있는 한정어가 제한될 경우가 있는데, 예를 들어 '각'은 가산명사 앞에만 올 수 있다.

(7) ㄱ. 각 소년
ㄴ. *각 우유

명사 뒤에 오는 일부 특수 조사의 경우도 마찬가지다.

(8) ㄱ. 소년마다
ㄴ. *진흙마다

이상에서 우리는 한국어에서도 가산, 비가산명사의 구분이 중요한 통사적 정보임을 알 수 있고, 이 점을 사전의 작성에서 고려해야 할 것이다. 집단성도 어느 정도 필요한 정보라고 여겨진다. '모이다'와 같은 집단적 주어를 요구하는 동사가 존재하기 때문이다.

(9) ㄱ. *그 소년이 모였다/만났다.
ㄴ. 그 위원회가 모였다/만났다.

이제, 명사가 취하는 보어의 종류에 따른 하위범주화에 대하여 알아보자. 영어에서와 마찬가지로 보어를 취할 수 있는 명사는 보어 없이도 쓰일 수 있다는 점에서 명사의 보어들은 수의적이라는 가정에서 출발하기로 하자.

첫째, 관계적 의미를 갖는 명사는 다른 명사구 보어를 가질 수 있다. '아버지, 어머니, 할아버지, 아들' 등의 친족관계 명사는 '김씨의 아들' 등 속격의 보어를 취할 수 있다. '왕, 대통령, 사장' 등 직함을 나타내는 명사도 국가, 기관명 등의 속격 보어를 취할 수 있다. 좀더 특수한 예로서, '위, 옆, 아래, 뒤, 앞' 등의 명사는 다른 명사(일반적으로 격조사가 생략된 속격)과 결합하여, 영어 등에서 전치사가 나타내는 의미를 표현할 수 있다.

(10) 책상(의) 위, 탁자(의) 아래, 그 군인(의) 앞

둘째, 보문을 취하는 명사들이 있다. 이러한 명사가 취하는 보문의 형태는 어미가 '-고 하는'으로 끝나는 것과 '-는/은/을'로 끝나는 것의 두가지로 크게 구분할 수 있는데 후자는 다시 '-는, -은, -을' 중 어느 것들을 취할 수 있는가에 따라 세분할 수 있다(강 범모 1983 등).

(11) ㄱ. 그 야구팀이 우승했다고 하는 사실
ㄴ. 그가 춤을 추는 모습

특히, 많은 불완전 명사가 이 부류에 속하는데, 앞에 올 수 있는 보문 형태의 제한이 강하다.

세째, '연구, 수행' 등 동사적 성격을 가지고 있는 추상명사들로서, 이 명사들은 주로 '-하다'가 붙어 완전한 동사가 될 수 있다.

네째, 불완전명사 중 분류사 등의 세분류가 필요하다. 남 기심, 고 영근(1985)에는 관형어와 조사와의 통합에 있어 큰 제약을 받지 않는 '것' 등의 보편성 불완전명사(의존명사), '있다, 없다'와만 결합하여 주어로만 쓰이는 '수, 리, 나위' 등 주어성 불완전명사, '따름, 뿐, 터, 때문' 등 서술성 불완전명사, '줄, 채, 김, 만큼' 등 부사성 불완전명사, '벌, 마리' 등 분류사의 성격을 띠고 있는 단위성 불완전명사 등으로 분류하고 있다.

## 3. 의미론적 정보

기본적으로 필요한 의미 정보는 하위범주화와 관련하여, 동사 및 명사 의미의 논항구조(argument structure)에 관한 것이다. 예를 들어, 동사 '먹다'의 경우 이 동사는 통사적으로 명사구 목적어와 명사구 주어가 필요한 동시에, 의미적으로 이 동사를 주 술어로 포함한 문장이 진술을 위해 성공적으로 쓰이기 위해서는 먹히는 것과, 먹는 사람 또는 동물이라는 의미적 대상이 '먹다'가 표현하는 의미와 결합하여야 하는 것이다. 즉, '먹다'가 의미하는 바는 두 개의 적절한 논항을 필요로 한다. 따라서, 논항구조는 하위범주화의 통사 정보와 밀접한 관계가 있으나, 논항구조와 하위범주화 정보가 일치하거나 잉여적인 것은 아니다. 상황의미론의 관점에서 볼 때(Barwise 1989, Devlin 1991 등), 동사가 표현한다고 생각되는 관계 중 대부분이 시간적, 공간적 장소의 논항을 가지고 있으나, 동사 자체는 장소 표현을 하위범주화하지 않는 수가 있다. 시간, 장소는 일반적으로 부가어로 표현되기 때문이다. 물론, 하위범주화되지 않는 표현의 논항의 종류와 숫자을 정하는 것이 쉽지 않은 문제이지만, 적어도 어떤 동사 및 명사 표현과 자주 자연스럽게 결합되는 부가어와 관련되는 (의미적) 논항은 사전에 표시하는 것이 좋을 것이다.

논항의 종류와 숫자보다 더 긴요한 의미 정보는, 하위범주화된 표현과 대응하는 논항에 대하여 그 논항의 자리를 차지할 수 있는 의미적 대상의 종류에 대한 제약이다. '먹다'의 경우 (그 기본 의미에 있어) 사람이나 동물만이 그 행위를 할 수 있으며, 먹히는 것도 액체나 고형체의 물질적 대상이어야만 하는 것이다. 따라서, 책상은 전쟁과 먹는다는 관계에 있을 수가 없다. 이러한 선택제약에 관한 정보는, 앞에서 언급했듯이 특히 대화체 언어의 분석에서 아주 유용할 것이다. 비슷한 발음의 단어들의 혼동을 피할 수 있으며, 또한 적극적으로 적절한 뜻의 논항을 예상케하여 분석을 용이하게 할 수도 있겠다.

선택제약에 관한 논항의 의미 정보를 세분하는 방법 중의 하나가 성분분석(componential analysis)이다. 의미 자질을 이용하여 전반적 어휘 체계 또는 비슷한 뜻의 단어들의 뜻을 체계적으로 세분하는 이 방법은 실제 사전의 뜻풀이를 위한 기초로서 많이 이용되어 왔었다(Ayto 1983, 정 순기, 리 기원 1984). Ayto는 한 예로써, seat, chair, bench, form, stool, sofa, settee, couch, settle 등의 앉기 위해 쓰이는 물체들을 가리키는 영어 단어들의 성분분석을 제시하고 있는데, 비슷한 방법으로 한국어의 '방석, 좌석, 의자, 걸상, 안락의자, 회전의자, 자리' 그리고 외래어인 '벤치, 소파'들을 분석하면 다음과 같다.

이것들은 모두 앉는데 쓰이는 것들인데, '좌석, 자리'는 가장 포괄적인 의

미를 지니고 있어, 앞으로 제시할 모든 기준에 구애받지 않는다. 먼저, 일인용인가 다인용인가에 따라 '방석, 의자, 안락의자, 회전의자' 등 일인용으로만 쓰이는 것, '벤치, 소파' 등 다인용으로만 쓰이는 것, '걸상' 등 양쪽으로 다 쓰이는 것이 있다. '방석, 의자, 걸상, 안락의자, 회전의자, 소파'는 움직일 수 있으며, '벤치'는 움직이거나 고정되어 있을 수 있다. '방석'에는 등받이가 없으며, '안락의자, 회전의자'에는 등받이가 있고, 나머지는 등받이가 있거나 없다. '회전의자, 소파'는 실내용으로만 쓰이고 '벤치'는 실외용으로만 쓰이나, 나머지는 양쪽으로 다 쓰인다. 기능적인 면에서, '방석, 안락의자, 소파'는 안락함을 목적으로 하나, 나머지는 그렇지 않다. 이상의 내용을 자질을 사용하여 표시하자면, [일인용], [다인용], [움직임], [고정됨], [등받이 있음], [등받이 없음], [실내용], [실외용], [안락용] 등으로 표시할 수 있을 것이다. 예를 들어, '방석'은 [+일인용, −다인용, +움직임, −고정됨, −등받이 있음, +등받이 없음, +실내용, −실외용, +안락용]으로, '벤치'는 [−일인용, −다인용, −움직임, −고정됨, +등받이 있음, −등받이 없음, −실내용, +실외용, −안락용]으로 표시할 수 있겠다.

궁극적으로 성분분석이 뜻풀이와 언어의 기계분석에 유용하게 쓰일 것이므로, 명사, 동사, 형용사, 부사 등 실질어의 전반적인 분석이 필요할 것이다. 필요한 의미자질의 목록을 어떻게 확정하는가 하는 것이 이 방법의 궁극적인 어려움이고, 이 목록이 결국 분석의 과정에서 정해지겠지만, 이 기용 외(1989)에서 제시한 의미유형의 목록에서 출발하는 것도 한 방법일 것이다.

의미 분석에서 많이 쓰이는 또하나의 방법은 어휘적, 개념적 관계를 이용하는 관계 모형(relational mode)이다(Evens 1988). 이 모형은 의미영역 내의 구조적 조직을 명시적으로 보여 주기 위해, 영역 내의 항목들 간의 관계를 이용하는 것이다. 이 관계는 어휘적(lexical)이거나 개념적(conceptual)이다. 전자는 동의어, 반의어 등 단어들 간의 관계이고, 후자는 개념들 간의 관계로서 주로 심리학적 심리 모형의 구축에 이용되어 왔다(ISA 관계 등). 사전, 또는 어휘 데이터베이스의 구성에는 어휘 관계가 좀더 중요하며(Mel'-čuk & Zholkovsky 1988, Calzolari 1988, Grimes 1988 등), 관계 모형이 제공하는 정보를 앞에 언급한 다른 정보와 함께 사용할 수 있을 것이다. 이러한 관계 모형의 이용에 대하여서는 6절에서 자세히 다루기로 한다.

한편, 이제까지 통사 정보와 의미 정보를 각각 독립적으로 기술하여야 한다는 입장에서 논의하여 왔는데, 이와 반대로 두 가지 정보가 상호 의존적, 나아가 통사 정보는 의미 정보로부터 도출될 수 있다고 하는 이론이 있으므로 이에 대하여 고찰해 보자. 예를 들어, Levin (1985)에 따른면, 동사의 하위범주화 정보는 그 동사의 의미로부터 예측가능하다고 주장한다. 예를 들어 보어를 두 개 취하는 동사 중 give는 NP 목적어와 PP 보어를 취하거나 (give an apple to John) 또는 두 개의 NP 목적어를 취하지만(give John an apple), explain은 전자의 방법으로만 쓰일 수 있다. 이것을 일반적으로 소유

의 변화 등 주제(theme)의 이동이 간여할 경우 후자의 방법으로 표현이 된다고 가정하여 설명할 수 있다고 한다. 그러나 사실상, 전자의 표현방식만을 된다고 허용하는 donate의 예에서 보다시피, 그러한 일반화가 성립하지 않는 듯하다. 더우기 Boguraev and Briscoe (1989)는 실제로 Longman 사전에 나온 이중 보어를 취하는 동사 131 개를 조사하여, 위와 같은 일반화가 성립하지 않음을 계량적으로 보여 준다. 따라서, 우리는 Gazdar, et al.(1985)에서 가정했듯이 의미적으로 예측 불가능한 하위범주화의 통사 정보가 필요함을 인정할 수밖에 없으며, 사전에도 독립된 정보로 올려야 할 것이다.

## 4. 화용론적 정보

단어의 통사, 의미 정보만을 가지고는 그것을 적절한 상황에서 올바로 쓸 수 없다. 어떤 표현이 쓰이는 비언어적 맥락(context)이 중요하다는 것이다. 우리말과 같이 경어법이 중요한 언어의 사용에서, 아들이 아버지에게 말할 때 자기 친구에게 말하듯이 한다면 이상할 것이다. 바꾸어 생각하면, 언어 사용의 맥락에 대한 화용론적 지식이 있음으로써 언어의 분석 및 이해가 용이하며, 번역의 경우 올바른 그 언어의 번역문이 나오는 것이다.

Hartman (1983)에 따르면, 맥락에 따른 언어 변이는 개인적 차이, 기능적 차이, 상황적 차이, 형식성의 차이, 지역적 차이, 사회 계층적 차이, 시대적 차이, 규범적 차이, 언어 간의 교류에 의한 차이 등의 여러 기준에 의해 구분될 수 있으나, 우리의 목적에 가장 필요한 구분들 중의 하나는 형식성의 차이에 의한 문어, 구어의 차이일 것이다. 같은 구어 중에서도 그 형식성의 정도에 차이가 있을 수 있다. 앞에서도 언급했지만, 화자, 청자 간의 관계에 따른 형식성의 차이가 직접 문법에 반영되는 한국어의 경우 이러한 정보가 필수적이다.[2] 궁극적으로는 지역 방언 및 사회 방언의 표시도 이루어져야 할 것이다. 은어 및 속어의 명시도 문장의 분석 및 번역문 생성에 중요한 역할을 할 것이다. 어느 정도의 백과사전적 정보도 효과적인 사전의 이용을 위해서 필요함은 앞에서 지적한 대로이다.

## 5. 문법이론과 사전 정보

앞에서 살펴 보았던 통사, 의미, 화용 정보는 특정한 언어 이론에 상관없이 어휘부 또는 사전의 항목이 갖추어야 할 것들이다. 그러나 궁극적으로 언어의 분석을 수행할 문법 이론이 이러한 사전 정보를 충분히 이용할 수 있어야 효과적인 기계 번역이 가능할 것이다.

2) 장석진(1990)은 Pollard and Sag (1987, 1991)의 Head-Dreven Phrase Structure Grammar(HPSG)를 이용하여 이 문제를 다루었다.

하위범주화의 통사적 정보를 효과적으로 표시하고 이용하는 문법은 범주 문법이다. 기본 범주를 바탕으로 파생 범주들을 순환적으로 정의하는 범주 문법은, 그 범주의 이름 자체가 하위범주화 정보를 표시한다. 예를 들어,

(12) ㄱ. S, NP, CN 이 기본 범주이다.
ㄴ. A, B가 범주이면, A/B 도 범주이다.
ㄷ. 그 이외의 것은 범주가 아니다.

이와 같이 정의하면, 자동사는 S/NP, 타동사는 S/NP/NP, 한정사는 NP/CN으로 정의할 수 있다. 이것은 범주 문법에서 A/B 범주의 표현이 B 범주의 표현과 결합한다는 한다는 함수적용(functional application)의 기본적 결합 규칙이 있기 때문이다.

NP에 '철희, 영희' 등의 단어를 기본 표현으로, CN에 '사람, 개, 사과' 등을 기본 표현으로 지정할 수 있고, '자다, 걷다' 등은 S/NP, '먹다, 때리다' 등은 S/NP/NP, '모든, 한' 등은 NP/CN에 속할 것이다. 타동사가 포함된 문장의 예를 들면 다음과 같다.

(13) 영희가 모든 사과를 먹었다.
NP NP/CN CN S/NP/NP

NP

S/NP

S

위의 예에서 보다시피, 타동사 '먹다'가 두 개의 명사구를 취하는 것으로 하위범주화된 것이 범주 이름 자체에 표시되어 있는 것을 알 수 있다.

근래의 범주 문법은 함수적용뿐 아니라, 함수합성, 인상 등을 인정하여 형태론, 통사론의 여러 문제를 다루고 있으나(Kang 1988, Jacobson 1990, Steedman 1991 등), 하위범주화 이외의 정보를 효과적으로 표시하기 위해서는 부족한 점이 있다. 우선, 앞의 예에서 보다시피, 한정사와 명사의 결합은 한정사가 함수로서 기능하여 전체 명사구에서 주도적 기능을 하는 것으로 나타나 있으나, 실제적으로 명사구의 중심 요소는 명사이다. 또한 일반적으로 명사의 하위범주화는 많이 다루어지지 않았으므로 이에 대한 표시방법의 제시가 필요하다. 어순과 관련하여서는 일반적으로 두 개의 인접한 표현만이 결합할 수 있다는 인접성 조건(Steedman 1987)의 계약이 지켜질 경우, 한국어와 같은 어순이 비교적 자유로운 언어의 기술이 다소 복잡해진다. 더우기 대화의 경우 어순은 문어보다도 훨씬 자유롭다는 것을 감안한다면, 엄밀한 인접성 조건의 준수는 비실용적일 듯이 보인다. 그리고 무엇보다도, 하위범주화 이외의 정보를 표시하기 위해서는 앞에서 언급한 범주의 이름만을 가지고는 그것을 나타낼 수 없다. 따라서 다른 여러가지 통사, 의미 자질이 필요하다.

이러한 점들은 범주문법가들에 의해 이미 잘 인식이 되어 왔던 바이며, 통사자질 등의 첨가를 통한 확장된 문법이 묵시적으로 전제되어 왔었다. 이러한 요구를 명시적으로 보여 주고 있는 것이 Pollard and Sag(1987, 1991)의 핵어중심 문법(HPSG)이다. 특히 이 문법 모형은 통사, 의미 정보뿐만 아니라, 화용론적 정보를 CONTEXT라는 자질을 통하여 언어 기호 체계 내에 표시할 수 있는 길을 열어 놓음으로써, 앞에서 제시하였던, 사전 항목의 다양한 정보를 쉽게 이용할 수 있는 길을 터 놓은 것으로 생각된다.

이상, 문법 이론과 사전 정보의 관련성에 대하여 간단히 고찰하였으나, 본 연구의 기본 입장은 역시 기계번역 사전의 기초가 될수 있는 사전에 필요한 가능한 한 많은 정보를 체계적으로 저장하는 것이 좋다는 것이다. 이러한 점에서 앞에서 언급했던, Mel'čuk 등의 어휘적 관계에 대한 정보가 비록 HPSG 등의 문법이론에서 이용되고 있지는 않으나 궁극적으로 요긴한 정보가 될 것이므로 이 점에 대하여 살펴 보기로 하자.

## 6. 어휘적 관계

동의어, 반의어 등 어휘적 관계는 전통적인 사전들에도 많이 제시, 이용되어 왔었다. 그러나 어휘적 관계를 가장 중요한 어휘 정보의 하나로 인식하여 사전 정보의 대부분을 차지하도록 구성한 사전은 Mel'čuk & Zholkovsky의 설명적 결합 사전(Explanatory Combinatorial Dictionary: ECD)이다. ECD는 모국어 사용자가 (특히 청취자가 아닌 발화자의 관점에서) 그 언어의 단어에 대하여 알고 있는 모든 것을 아주 자세하게 기술하여, 그 언어를 배우는 외국인이 그 사전 정보만을 가지고 정확한 문장을 사용할 수 있도록 목표로 하는 사전이다. ECD는 단어의 모든 형태와 사용에 대해 설명하므로 '설명적'이고, 어떤 단어가 다른 단어들과 어떻게 결합하는지에 대한 정보를 주므로 '결합적'이다. ECD에 수록된 단어의 정의, 지배패턴(government pattern) 등은 앞에서 논의한 하위범주화를 포함한 통사, 의미 정보와 중복되는 면이 많으나, ECD에서 이용하는 어휘관계는 그 수효(약 50개) 및 기술의 명시성에 있어서 혁신적이라고 할 수 있다.

기본적으로 어휘관계는 언어에서의 기초적 관계인 대치관계(paradigmatic relation)과 결합관계(syntagmatic relation)를 세부적, 명시적으로 보여주는 것이라고 할 수 있다. 따라서, 단어들 간의 이러한 관계들에 관한 정보는 해당 언어 자체에 대한 지식의 많은 부분을 차지하고 있다. 이러한 면에서, Mel'čuk이 스스로 주장하듯이, ECD는 모국어 화자의 언어 지식을 충실히 기술하려는 현대 이론언어학의 목표를 수행하는 하나의 방법인 셈이다. 동시에 기계번역 등 언어의 자동 처리가, 될 수 있는 한의 자세한 어휘 정보를 필요로 함을 상기할 때, ECD에서 제시하는 어휘적 정보가 우리 목적에 맞

는 사전 구성의 하나의 요소가 될 필요가 있겠다.[3] 어휘관계가 정확한 언어 사용의 필요조건임을 Mel'čuk 관계를 예로 들어 살펴 보자. Magn은 정도의 심함을 의미하는 관계이다. 영어에서 temperature의 정도가 심함은 high라는 수식어로써 표현하나, rain의 경우는 heavy가 적절한 표현이다. 이것을 Magn(temperature)=high, Magn(rain)=heavy 라고 나타낼 수 있다. 한국어의 경우, '기온'은 '높은 기온'이라고 할 수 있으므로 영어의 high와 '높은'의 대응에 무리가 없으나, '비'의 경우 '무거운 비'라고 할 수 없고 '심한 비' 또는 '억수같은 비'라고 해야 하므로 heavy와 '무거운'의 단순한 대응은 성립할 수 없다. 한국어의 사전에 Magn('기온')= '높-', Magn('비')= '심하-, 억수같-' 이라는 어휘적 정보가 표시되어 있을 때, 이러한 문제는 발생하지 않을 것이다.

이와 같이 어휘적 관계는 무엇보다도 자연스러운 연어(collocation)를 위해서 아주 유용하게 활용될 수 있을 것으로 생각되는데, 이렇게 결합관계(syntagmatic relation)를 보여 주는 어휘관계의 예를 앞의 Magn 이외에 몇 가지 더 들어 보자.

Liqu는 주어진 항목의 제거를 나타내는데, 예를 들어 Liqu('위원회')= '해체하-', Liqu('범죄')= '일소하-, 뿌리뽑-'. 즉 '범죄를 뿌리뽑다' 라고 하는 연결형이 가능함을 보여 준다. Ver는 올바름 또는 적절함을 나타내며,

3) 이제 논의하는 Mel'čuk의 어휘관계 표시이외에도, 주로 단어 간의 의미적 관계를 중심으로 포괄적으로 논의한 업적으로 Cruse (1986)가 있다. 단어 간의 관계를 논하는데 있어 Cruse는 우선 그 관계의 주체가 되는 어휘단위(lexical unit)를 "(상대적으로) 안정적이고 단절적인(discrete) 의미 특성을 갖는 형태-의미의 복합체"(p. 49)로 정의하고, 사전에 기재되어 있는 항목으로 정의할 수 있는 어휘소(lexeme)와 구별하는 데에서부터 그 논의를 시작한다. 따라서, 중의어는 물론이고 다의어의 경우에도 우리가 어휘관계를 논할 수 있는 것은 세부적인 의미 하나 하나라고 볼 수 있으며, 이러한 태도는 우리의 어휘관계 논의에도 암묵적으로 받아들여 지고 있는 셈이다. 구체적인 어휘관계의 논의에 있어, Cruse는 주로 대립관계(paradigmatic relation)들을 그 특성에 따라 구별하여 논하는데, 그 중에는 다음과 같은 것들이 있다. 첫째, animal-dog-spaniel 등의 예에서 보는 것과 같은 분류적 관계 (taxonomy); 둘째, body-arm-finger-nail 등에서 보는 것과 같은 부분관계 (meronomy); 세째, sentence-clause-phrase-word-morpheme 등의 예에서와 같은 가지치지 않는 층위관계 (nonbranching hierarchy). 이상은 모두 층위관계 (hierarchy)의 일종들이다. 네째, dead-alive 등 상보적 반의관계와 left-right 등 방향적 반의관계를 포함하는 반의관계; 다섯째, unblemished-spotless-flawless-immaculate-impeccable 등에서 보는 것같은 유사어 또는 동의어관계 (synonymy). 이러한 대립관계들은 앞서 예시했던 성분분석의 방법을 통하여 좀더 예리하게 분석되어 실제 기계번역 사전의 작성에 이용될 수 있을 것이다. Cruse의 논의에 있어 아쉬운 점은 연결관계 (syntagmatic relation)에 관한 논의가 극히 미약하다는 것이다(총 300여 페이지 중 약 1페이지). 문장의 적절성(appropriateness)과 자연스러운 연어(collocation) 현상을 포착하기 위하여 어떠한 종류의 연결관계가 필요한가를 알아내고 기술하는 것이 번역사전을 위하여는 필수불가결한 것임은 말할 나위가 없다. Mel'čuk가 사용한 어휘관계들 중 많은 부분이 이러한 연결관계를 나타내고 있다는 점이 바로 우리가 Mel'čuk의 이론을 검토하여 이용해 보고자 하는 주요 이유이다.

Ver('재판')= '공정하-', Ver('자')= '정확하-' 이며, 이것은'옳은 재판'이 아니라 '공정한 재판'이 자연스러운 표현임을 보여 준다. Bon은 주어진 항목에 대한 일반적인 칭찬의 말인데, Bon('신체')= '건강하-', Bon('소파')= '안락하-' 등이 있다. $Oper_i$ 는 주어진 항목이 의미하는 사건의 i번째 참여자가 주어이고 그 항목을 목적어로 취하는 동사이다. 예를 들어 $Oper_1$('변화') = '가져오-', $Oper_2$('변화')= '받-'로서, 김이 박을 변화시킨 상황에서 '김이 변화를 가져왔다', '박이 변화를 받았다' 등의 표현이 가능함을 보인다. 반대로 $Func_i$는 i 번째 참여자를 목적어로, 주어진 항목을 주어로 취하는 동사이다. $Func_1$('변화')= '-에 기인하-', $Func_2$('변화')= '-에 생기-'로서 '변화가 김에 기인한다', '변화가 박에 생겼다' 등의 표현이 가능하다. Sing은 단위를 표시하는 것으로서 분류사가 많이 쓰이는 한국어에서 특히 중요하다. Sing('집')= '채', Sing('종이')= '장' 등. 이 밖에도, $A_i$, Centr, Excess, Figur, Germ, $Labor_{ij}$, $LabReal_{ij}$, $Loc_{in, ad, ab}$, Manif, Mult, Nocer, Obstr, Plus, $Pos_i$, Pred, Propt, Prox, $Qual_i$, $Real_i$, Son 등이 결합관계와 관련이 있는 어휘관계로 파악된다.[4] 이상에서 ECD의 어휘관계 중 몇 가지를 살펴보았다. 연어(collocation)의 표시를 위해서는 유용한 방법으로 생각되나, 문제는 ECD의 관계들 모두를 전부 사용할 것인지 또는 필요한 것만을 추려서 사용할 것인지를 결정하는 일이다. 더욱 큰 일은, 만일 어휘관계를 실제로 사용하려고 한다면, 단어 하나하나의 분석이 방대한 작업이 될 것이다. 따라서, 그 유용성에 대한 합의가 먼저 이루어져야 할 것이다. 당장 생각할 수 있는 어휘관계의 효용성은, 앞에서도 지적했듯이, 단어 간의 연어를 다루어 자연스러운 문장으로의 번역을 가능하게 한다는 것이다. 다만, 숙어 또는 관용어 사전에서 다룰 완전한 숙어와의 구분, 그 표시방법의 일치 등 결정해야 할 사항이 많이 남아 있다.

## 7. 요 약

궁극적으로 대화체 기계번역을 위한 한국어 사전, 또는 그 기초가 되는 어휘 데이터베이스에 필요한 정보에 관하여 살펴 보았다. 될 수 있는 한의 자세한 통사, 의미 정보의 표시가 필요한 이유 중 하나는, 텍스트의 경우와 달리 실제 대화에서 음운정보를 신뢰할 수 있을 정도로 파악해 내기가 힘들다는 것이었다. 물론 자세한 통사, 의미 정보는 텍스트의 번역도 더 쉽게 만들 수는 있을 것이다.

필요한 통사 정보 중 엄밀하고도 자세한 하위범주화 정보가 필수적임을 논하였는데, 주로 이제까지 많은 논의가 없었던 한국어 명사의 하위범주화의

4) 어휘관계 목록과 러시아어 및 영어의 예를 Mel'čuk & Zholkovsky (1988), 또는 Mel'čuk, et al. (1984)를 인용한 Frawley (1988) 에서 찾을 수 있다.

필요성과 그 대강의 체계를 언급했다. 특히, 가산, 비가산 명사의 구분이 한국어에도 긴요함을 밝혔는데, 실제 사전의 작성에 꼭 명시해야 할 것으로 생각된다. 의미 정보는, 선택제약을 다룰 수 있도록 성분분석적 방법을 채택해야 할 것으로 보는데, 물론 이 방법의 근본적인 문제점인, 의미원소(semantic primitive)의 설정이라는 문제는 남아 있을 것이다.

통사, 의미 정보의 또 하나의 표시 방법은 Mel'čuk 등의 설명적 결합 사전(ECM)의 어휘관계를 이용하는 것이다. 특히 단어 간의 연결관계를 명시적으로 보여주는 이 방법은 자연스러운 연어(collocation)를 생산해 내기 위해서 아주 긴요하게 쓰일 것으로 생각되는데, 다른 통사, 의미 정보의 표시와 더불어, 이러한 어휘관계를 부분적으로라도 이용할 필요가 있다고 본다. 다만, 단어 하나하나의 분석이 방대한 작업일 것이므로 작업의 적정한 범위를 제한하는 것이 미리 필요할 것이다.

## 참 고 문 헌


강 범모 (1983) "한국어 보문명사 구문의 의미자질", 어학연구 19, 53-73.

김 경서, 김 대철, 정 강석, 송 만석 (1991) "말뭉치를 이용한 형태소 분석 단계에서의 중의성 해결에 관한 연구", 제 3회 한글 및 한국어 정보처리 학술대회 논문집, 36-43.

남 기심, 고 영근 (1985) 표준 국어문법론, 서울 : 탑출판사.

이 기용, 박 병수, 임 홍빈 (1989) "영한 기계번역을 위한 어휘부 구동의 문법 모형의 구축과 그 적용," 시스템 공학센터.

장 석진 (1990) "화용과 문법-자연언어 처리를 위한 화맥 연구-," 언어 15, 499-538.

정 순기, 리 기원 (1984) 사전편찬리론 연구, 사회과학 출판사.

Ayto, J.R. (1983) "On specifying meaning," in R. Hartman (ed.), 89-98.

Akkerman, E. (1989) "An independent analysis of the LDOCE grammar coding system," in B. Boguraev & T. Briscoe (eds.), 65-83.

Barwise, J. (1989) *The Situation in Logic*, Stanford: CSLI.

Boguraev, B. and T. Briscoe (1989) "Utilizing the LDOCE grammar codes," in B. Boguraev and T. Briscoe (eds.), 85-116.

Boguraev, B. and T. Briscoe (eds.)(1989) *Computational Lexicography for Natural Language Processing*, London: Longman.

Calzolari, N. (1988) "The dictionary and the thesaurus can be combined," in M. Evens (ed.), 75-96.

Cruse, D.A. (1986) *Lexical Semantics*, Cambridge: Cambridge University Press.

Devlin, K. (1990) *Logic and Information*, ms., to be published by Cambridge University Press.

Evens, M. (ed.)(1988) *Relational Theory of the Lexicon: Representing Knowledge in Semantic Networks*, Cambridge: Cambridge University Press.

Frawley, W. (1988) "Relational models and metascience," in M. Evens (ed.), 335-372.

Gazdar, G., E. Klein, G. Pullum, and I. Sag (1985) *Generalized Phrase Structuce Grammar*, Cambridge: Harvard University Press.

Hartman, R.R.K. (ed.)(1983) *Lexicography: Principles and Practice*, London: Academic Press.

Jacobson, P. (1990) "Raising as function composition," in *Linguistics and Philosophy* 13, 423−75.

Levin, B. (1985) "Lexical semantics in review: an introduction," in B. Levin (ed.) *MIT Lexicon Working Papers* 1, 1−62.

Mel'čuk, I. and A. Zholkovsky (1988) "The explanatory combinatorial dictionary," in M. Evens (ed.) 41−74.

Mel'čuk, I., N. Arbatchewky-Jumarie, L. Elnitsky, L. Iordanskaja, and A. Lessard (1984) *Dictionaire Explicatif et Combinatoire du Français Comtemporain*, Montreal: University of Montreal Press.

Kang, B. (1988) *Functional Inheritance, Anaphora, and Semantic Interpretation in a Generalized Categorial Grammar*, Ph.D. dissertation, Brown University.

Pollard, C. and I. Sag (1987) *An Information-Based Approach to Syntax and Sematnics, Volume I*, Stanford: CSLI.

Pollard, C. and I. Sag (1991) *Head−Driven Phrase Structure Grammar*, Ms., to be published by Stanford: CSLI.

Steedman, M. (1987) "Combinatory grammars and parasitic gaps," in *Natural Language and Linguistic Theory* 5, 403−39.

Steedman, M. (1991) "Structure and intonation," in *Language* 67, 260−96.

〈Abstract〉

# Syntactic and Semantic Information in the Lexicon for Spoken Language Machine Translation

Beom-mo Kang

This paper aims to examine kinds of syntactic and semantic information to be encoded in lexical entries of the lexicon which should ulticately be used for machine translation of spoken languages. The reason why we need as detailed information as possible is that in the real situation of speech, unlike the case of texts, it is difficult to get reliable phonological information. Among needed syntactic information, detailed subcategorization informaion is necessary. For Korean nouns, it is unavoidable to have the distinction of mass and count nouns, as in English. One kind of useful semantic information can be encoded with semantic features, by means of which selectional restriction may be handled well. Another way of representing syntactic and semantic information is using Mel'čuk's lexical relations used in his *Explanatory Combinatorial Dictionary*. Lexical relations give a promising method to handle collocation, which should be of prime concern in any study of (machine) translation.