특집

한국어 전자 사전

남 지 순
KAIST 인공지능 연구센터 한글공학연구실

I. 서 론

컴퓨터로 자연어를 처리한다는 것은, 궁극적으로 인간이 가지고 있는 모든 '언어 능력(Language Competence)'을 컴퓨터에 그대로 재현할 수 있기를 추구하는 작업이다. 제한된 언어 현상의 습득을 통해, 전혀 접하지 못했던 유형의 문장들을 생성해 낼 수 있는 인간의 언어 능력은, 기계와는 달리 그 기억 공간이나 능력에 분명 현실적인 한계가 있는 점을 감안해 볼때, 단순한 데이터의 축적만으로 이루어지는 100% 경험적인(Empirical) 방법으로만 작용할 수 없는 것은 틀림없는 사실이다. 따라서, 인간의 언어 사용은, 일정 경험으로부터 주어진 현상을 분석하고 이해하여, 언어 현상을 연역적으로 추론할 수 있는 지적 능력을 토대로 하여 비로소 가능할 것이다. 즉, 인간의 언어 사용이, 이와 같은 '내재적인(Inherent)' 하나의 메카니즘에 의해 이루어진다고 볼 때, 바로 이 메카니즘의 실체를 파악할 수만 있다면, 컴퓨터에 인간과 동일한 언어 능력을 부여해 주는 작업이 실현 불가능한 일은 아닐 지도 모른다.

실제로 이와 같은 인간의 언어 능력의 메카니즘에 대한 가설을 세우고, 그것을 밝혀내기 위한 여러가지 이론과 모델들이 1950년대를 전후하여 제시되어 왔다. 그 시기 이전의 언어 이론들은, 언어 사용에 대한 철학적인 접근의 오랜 전통의 한 흐름과 (유럽 문헌학적, 규범적 전통), 철저히 관찰에 의해 밝혀질 수 있는 현상을 기술하는 구조주의적 방법론의 한 흐름 (북미 인디언들의 언어와 같이, 그 당시 언어학자들이 기본 지식을 전혀 갖지 못한 미지의 언어에 대한 연구)으로 대표될 수 있다. Chomsky등에 의해서, 언어에 대한 이와 같은 순수 '기술적 (Descriptive)' 관점을 넘어서서, 인간의 내적 언어 능력에 대한 '가설(Hypothesis)'을 제시하는 '변형 생성 문법(Transformational Generative Grammar)'의 이론이 대두된 것은, 기계에 의한 인간 언어 처리에 새로운 가능성을 제시하는 계기가 되었다. 그러나,

과연 이와 같은 언어 능력의 메카니즘이 진정으로 어떠한 것인지, 과연 밝혀질 수 있는 것인지에 대한 질문에 있어서는 현재로서도 아직 아무런 결론을 내릴 수가 없다. 분명한 것은, 현 단계의 컴퓨터는 반드시 입력된 정보에 대해서만 언어 처리 능력을 가진다는 점이며, 인간의 뇌에서 이루어지는 유추나 추론의 과정을 밝혀내어 그것을 이식시키기 위한 어떠한 이론이나 모델로도 인간과 동일한 언어 능력을 갖춘 컴퓨터의 구현은, 현재로는 기대하기 어렵다는 점이다. 그러므로, 컴퓨터로 자연어를 처리하는 작업은, 얼마만큼 충분한 데이터를 얼마만큼 체계적으로 저장하였는가 하는 질문에 전적으로 연관되는 문제이며, 시스템의 효율성은 바로 이와 같은 문제에 밀접하게 관련된다. 이와 같이, 컴퓨터에 의한 자연어 처리가, 저장된 데이터의 질과 양에 의해 좌우되는 것이라면, 이때 어떠한 데이터들이 어떠한 원칙하에서 입력되어야 하는지를 결정하는 일은 가장 기본적이고도 핵심적인 작업중의 하나가 될 것이다.

인간에 의한 자연어 문서 처리를 위해 필요한 언어 정보들은, 일반적으로 '사전 (Dictionary)'이라는 형태로 저장된다. 사전의 '기본 단위 (Basic Unit)'는 대개 '단어 (Word)'라고 불리우며, 대부분의 경우, 품사 정보 및 그 의미 해석 정보들이 함께 수반되는데, 이와 같은 '단어'라는 개념에 대한 정의는 사실상 명시적으로 주어져 있지 않다. 전통 문법에서는 '최소의 의미 단위'를 '형태소 (Morpheme)'라 정의하고, 이러한 형태소들의 결합으로 이루어지는 것을 단어라고 정의하고 있는데, 예를 들어, '사람', '칼'과 같은 형태가 단일 형태소로 이루어진 단어들이라면, '신세대', '식칼'과 같은 형태는 하나의 의존 형태소와 하나의 자립 형태소의 결합으로 구성된 단어들이다. 그러나 이러한 관점에서 단어의 개념을 이해할 때, 기존 사전들의 모든 기본 단위를 '단어'로 간주하기는 어렵다. 가령 '죽이는'과 같은 형태는 '죽 (동사 '죽다'의 어간)', '이 (사역형 어미)', '는 (관형형 접속 어미)'의 3 개의 형태소로 구성된 하나의 '단어'로 간주되어야 함에도 불구하고, 사전에는 기본형으로 설정된 '죽다', 또는 '죽이다'와 같은 형태들만이 '표제어 (Lexical Entry)'로 입력되어 있기 때문이다.

실제로, 자연어 문서속에 나타나는 어휘 형태들은, 기존 사전들에서 기본 단위로 처리한 '원형 (Canonical Form)'의 형태로만 실현되지 않는다. 컴퓨터에 의한 자연어 처리는, 우선 입력 스트링을 인식하는 작업으로 시작되는데, 두개의 '분리 기호 (Separator)', 즉 여백 (Blank)이나 쉼표 (Comma), 마침표 (Period)등에 의해 형성된 하나의 스트링이 인식되면, 이것은 사전 정보와의 매칭 (Matching) 작업을 일일이 거쳐야 한다. 이때, 입력 스트링에 대한 올바른 언어 정보를 사전으로부터 얻어내지 못하면 이 스트링은 인식되지 못한 상태로 남게 된다. 컴퓨터용 사전에서 정보의 기본 단위가, 문서내에서 발견되는 기본 스트링의 형태와 일치하게 되면 자동 처리 프로세싱의 효율성은 매우 높아질 것이다. 그런데, 이와 같은 기본 스트링의 형태는 개별 언어마다 다르게 나타난다. 가령, 영어의 경우, 두개의 분리 기호에 의해 인식되는 스트링의 형태가 기존 사전에서 표제어로 설정한 '단어'의 개념과 그렇게 다른 형태가 아닌 반면, 중국어의 경우에는 아무런 여백도 없이 모든 문장 성분이 연이어 나타나는 표기상의 특성으로 인하여, 기본 입력 스트링이 하나의 문장에 해당할 수도 있다. 한국어의 경우에는, 기존 사전의 표제어 차원보다는 더 확대된 형태가 입력 스트링으로 인식되는 경우가 대부분이다. 즉, 문서안에서 발견되는 스트링들은 '집에', '운동을'과 같은 형태를 띠는데, 기존의 어느 사전에도 이와 같은 형태가 기본 정보 단위로 설정되어 있지는 않다.

사전의 기본 단위가 반드시 '집', '에', '운동', '을'과 같은 형태로 이루어져야 한다는 절대적인 원칙은 없다. 또한 인간 사용자를 대상으로 한 경우와 컴퓨터를 위해 고안된 경우가 서로 다르다. 인간 사용자를 위한 경우, 오히려 너무나 자명하거나 반복적인 현상에 대해서는 생략을 하는 경우가 바람직한데, 가령 동사나 형용사에 파생 접미사 '기'나 '음'이 결합되어 유도되는 명사형들과 같은 경우이다. 그러나 컴퓨터용 사전의 경우, 모든 정보는 빠짐없이 체계적으로 저장되어야 하므로, 위

와 같은 정보는 어떠한 형태로든 모든 동사, 형용사 엔트리에 일일이 결합되어야 한다. 이러한 이유로 컴퓨터용 사전은 전자화된 기존 사전들의 정보 유형을 그대로 이용할 수가 없다. 전자 사전이라는 용어가 '컴퓨터용 사전'과 '인간 사용자를 위한 전자화 사전'사이의 모호성을 가질 수 있는 점을 고려하여, '기계 사전 (Machine Dictionary : MD)', "기계 가독형 사전 (Machine-Readable Dictionary : MRD)', 컴퓨터 가독형 사전 (Computer-Readable Dictionary : CRD)', "어휘 데이터 베이스 (Lexical Data Base : LDB)'라는 표현들외에도 '기계 처리용 사전 (Machine-Tractable Dictionary : MTD)'이라는 표현등을 도입하고 있는 것 [Wal95]도 바로 이러한 이유 때문이다. 컴퓨터를 위한 전자 사전의 주요 원천은 우선은 기존 사전들의 다양한 언어 정보이며, 그것으로 처리되기 힘든 유형의 언어 현상, 가령 고유명사의 인식이라든지, 복합 명사구등의 인식을 위해서, 대형 코퍼스 (Large Corpus)로부터 '공기 (Collocation)' 정보등을 얻어내어 함께 사용한다. 분석시 발생하는 엄청난 형태의 중의성 문제를 해결하기 위하여 '빈도수 (Frequency)' 를 계산하는 확률 모델도 미등록어등의 사전 처리를 위하여 도입된다.

이 글에서는, 외국의 사례나 또는 외국의 언어 유형에 입각한 사전의 개념 및 그 이론적 모델등에 관한 논의 보다는, 실제로 한국어에서 관찰되는 개별적인 어휘 특성들에 대한 검토를 토대로 하여, 어떠한 형태의 '한국어 전자 사전 (Korean Electronic Dictionary)'이 구축되어야 하는 지에 촛점을 맞추어 논의를 전개할 것이다. 여기서 일컫는 '사전'이란, 궁극적으로 모든 언어 정보가 저장된 '어휘 정보 베이스 (Lexical Information Base)'의 형태로서, 종래의 좁은 의미의 사전 개념으로부터 보다 더 확장된 개념으로 이해되어야 한다. 그러므로, 이와 같은 형태의 사전에서 언어 정보가 저장되는 '기본 단위 (Basic Unit)'에는, 원칙적으로 어떤 절대적인 조건이 전제될 수 없다. 이 글에서는 우선, '어떤 형태'가 사전의 기본 단위로 설정될 수 있는지에 관하여 살펴 보고, 그러면 '어떠한 정보'들이 이러한 사전 엔트리에 결합되어 저장되어야 하는지에 관하여 살펴 보기로 한다.

II. 전자 사전의 기본 단위(Basic Unit)의 유형

1. 품사 단위 (POS Unit)

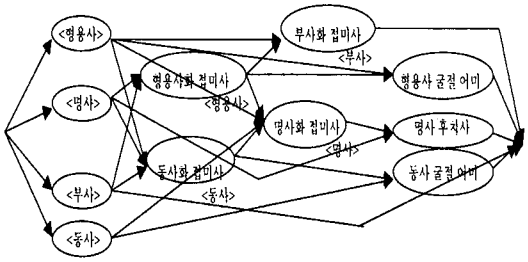
일반적으로 전자 사전이 제공하는 언어 정보의 가장 기본적인 형태는, 가능한 모든 '어휘 목록 (List of Lexical entries)'에 대한 정보라고 할 수 있다. 앞서 논의한 바와 같이, 어느 형태를 사전의 기본 어휘 엔트리로 삼을 것인가 하는 것은 개별 언어 유형에 따라, 또 사전의 사용 범위에 따라 달라질 수 있다. 만일 단일 형태소로만 이루어진 모든 품사 단위 (POS Unit)들을 기본 형태로 설정한다면, 여러 형태소들이 결합하여 구성되는 어휘 형태는 별도의 처리를 거쳐야 한다. 가령, 명사의 경우, 파생 접사 (Derivational Suffix)나 복수 표지가 결합되지 않은 단순 원형들만이 표제어로 등재되고, 동사나 형용사의 경우, 파생 접사나 굴절 접사 (Inflectional Suffix)등이 첨가되지 않은 원형 (Root) 형태만이 저장된다. 이때, 접사들도 단순어 원형과 마찬가지로 독립적으로 사전 항목을 이루게 되며, 이때 각 기본 품사 단위들간의 상호 결합 정보는 별도로 기술된다.

인간 사용자를 위하여 구축된 기존 사전들을 토대로 컴퓨터용 사전을 개발할 경우, 특히 다음과 같은 문제점들이 발생한다. 첫째, 기존 사전의 경우, 단일 형태소 품사 단위들의 결합 형태는 부분적으로만 제시되어 있어, 실제 전체 복합 형태 어휘류의 극히 일부분에 대한 정보만을 제공하고 있으며, 둘째, 이때 이와 같은 불충분한 목록을 체계적으로 완성할 수 있는 어떠한 방법도 제시되어 있지 않으며, 세째, 복합 형태들의 내부적 구조에 대한 정보가 미흡하여 신조어에 대한 예측등을 어렵게 한다. 가령, '좋다', '싫다', '즐겁다' 등의 형용사들은 파생 접미사 '어하다'와 결합하여 '좋아하다', '싫어하다', '즐거워하다' 등의 동사를 구성

하는 것으로 간주되고 있는데, '괴롭다', '안타깝다', '섭섭하다' 등으로부터 파생될 수 있는 '괴로워하다'나 '안타까와하다', '섭섭해하다' 와 같은 형태들은 기존 사전에 아무런 언급이 없다. 'Adj-어하다'의 구조는 짧은 형용사에 기초하는 경우 하나의 스트링으로 실현되기 쉬우나, 여러 음절로 이루어진 형용사에 기초한 경우, 두개의 분리된 형태로 나타나는 경우가 더 많고, 그 수는 사전에 수록된 형태들보다 훨씬 더 많다. 사전에 이와 같은 형태들을 기본 단위로 처리하지 않으려면, 몇몇의 경우만이 삽입되는 오류를 피해야만, 일정 규칙(Rule)의 형태로 전체가 처리되는 일관성(Coherence)을 유지할 수 있게 된다.

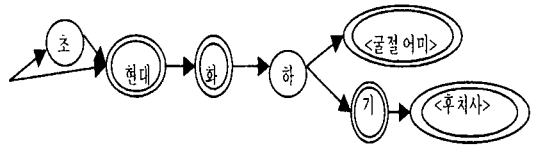
그런데, 위와 같은 파생어 형태나 굴절 변화형들을 기본형으로부터 올바르게 예측하여 자동으로 처리하는 일은 쉽지 않다. 더우기, 한국어의 경우, 하나의 동사나 형용사 원형으로부터 유도될 수 있는 형태 변화형의 구성은 영어나 기타 유럽 언어들의 경우와는 비교할 수도 없을 만큼 복잡한 양상을 띠고 있으며, 일정 파생 접사들에 의해 다른 품사 유형으로부터 동사나 형용사등이 유도된 경우, 다시 이와 같은 복잡한 굴절 변화형들이 결합되어야 하므로, 단일 형태소 유형의 품사 단위로 사전 항목을 설정할 경우, 그들 상호간의 결합 정보를 처리하는 모듈은 엄청난 부담을 안게 될 것이다. 다음 그래프는 한국어의 4 가지 기본 품사 단위들 (즉, 'N', "V', "Adj', "Adv') 간의 '품사 전이 관계'를 보여준다. 〈그림 1〉

위의 그래프에, 품사 전이를 가져오지 않는 '접



〈그림 1〉

두사'등에 의한 파생 관계들까지 고려하게 되면, 이러한 어휘 결합 정보는 더욱 복잡한 형태로 나타날 것이다. 실제로, 품사 단위들간의 결합은, 이미 이루어진 둘이상의 결합체에 다시 여러번 반복되어 일어날 수 있다. 가령, '현대'라는 명사에 접미사 '화'가 결합되어 '현대화'라는 파생어가 형성될 수 있는데, 이때, '현대'라는 명사로부터 바로 유도될 수 없었던 '하다'류 동사 (즉, *현대하다)는, '현대화'와 같은 파생어가 구성되면, '현대화하다'라는 형태로 실현될 수 있다. 마찬가지로, 접두사 '초'가 '현대화'에 다시 결합되면, '초현대화'와 같은 파생어를 생성해 내고, 이것은 다시 '초현대화하다'의 동사 형태를 취하게 되는데, 이와 같이, 'N-하다'류 형태에 대한 정보는, 여러 단계의 파생 관계를 고려하지 않고서는 체계적으로 구성되기가 어렵다. 위의 '초현대화하다'에 파생이 여전히 계속되면, 가령 명사화 접미사 '기'가 여기에 다시 결합될 때, '초현대화하기'와 같은 형태가 형성된다. 이때 '하다' 동사 형태는, 명사화 접미사 '기'에 의해서 명사 후치사 (즉 '조사')들이 뒤따르는 명사류로 전환되었기 때문에, 예를 들어 '초현대화하기를', '초현대화하기에는'등과 같은 다양한 형태로 실현된다. 다음 전이 그래프는, '현대'라는 명사로부터 유도될 수 있는 위와 같은 파생어 형태들 사이의 관계를 나타내는 '유한 오토마타 (Finite State Automaton)'를 보여준다〈그림 2〉.



〈그림 2〉

품사 단위로 사전 정보가 저장될 때, 검토되어야 할 기본적인 문제중의 하나가, 바로 '품사 분류의 원칙'에 관한 문제이다. 가령, '온 누리', '전 세계' 등에서 실현된 '온' 과 '전'은 뒤에 나타난 명사에 들러붙어 하나의 단일 형태를 구성하기도 하는데,

이들은 기존의 사전들을 보면, 때로는 '관형사 (Determiner)'로, 때로는 '접두사 (Prefix)'로 분류된다. 품사 분류의 원칙이 체계적이고 명시적인 방법에 의하여 구성되지 않으면, 품사를 아무리 세분화해서 분류하여도, 그 품사 범주에 어느 어휘 요소를 배당해야 하는지가 모호해 진다. '연구회'의 '회'와 '송별회'의 '회'는, 하나는 명사로, 하나는 접미사로 분류가 되고 있는 것을 발견할 수 있는데, 이 경우 역시, 품사 분류가 개념적으로만 나누어져 있을 뿐, 실제적으로 적용될 수 있는 '형식적 기준(Formal Criterion)'이 주어져 있지 않은 데에서 발생하는 문제점이다. 체계적이지 않은 품사 분류에 의거하여 하나의 품사가 어휘 요소들에 배당되었다면, 이와 같은 단일 형태소 품사 단위들 사이의 여러 결합 관계들에 관한 정보를 기술하는 작업은 결코 기대된 결과를 가져오지 못하게 될 것이다. 가령, '합리적', '논리적', '과학적'과 같은 어휘들은 명사, 또는 관형사로 기존 사전에 분류되어 있는 것을 관찰할 수 있는데, 이들은 일반 명사들이 실현될 수 있는 주어나 목적어 위치등에 결코 나타나지 못하며 (*과학적이 인간을 돕는다 / *그는 논리적을 좋아한다), 의미적으로도 하나의 형용사 술어와 더 가깝다 (그는 언제나 이성적이다 / 냉정하다 / 침착하다). 즉, 이 형태들은 오직, '으로'를 동반하여 부사 형태로 쓰이거나 (즉, '합리적으로', '과학적으로' 등), '이다'를 동반하여 형용사적으로 쓰이기 때문에 (즉, '합리적이다', '과학적이다' 등), 이들을 명사로 처리하게 되면, 문장내에서 격표지를 갖고 하나의 논항으로 쓰일 수 있는 일반 명사들과 동시에 같은 통사 규칙을 적용받아야 하는 어려움에 부딪히게 된다. 마찬가지로, 5 000 여개의 한국어 형용사중, '하다'를 동반하는 형태가 60 % 정도에 이르고 있는데 (Cf. DECOS-AS/V01 [Nam96a]), 이때 상당수의 어간 형태들은 어휘적 독립성을 전혀 갖고 있지 못한 것들로서 (즉, '예민하다'의 '예민', 또는 '영리하다'의 '영리'와 같은 경우, 문장내에서 주어나 목적어 위치등에 홀로 사용될 수 없다), 하나의 '명사'라는 품사의 배당을 받아서는 안된다. 그러나, 기존의 많은 사전들에서는, 아직도 이러한 한자어들을 어원적, 의미적 특성을 고려하여, 하나의 '명사' 엔트리로 취급하고 있다. 이와 같은 형태들을 명사로 취급하는 한, 명사 부류에 일반적으로 나타나는 통사적 특성들에 있어서도 예외의 경우들을 고려해야 하게 되며, 그렇지 않으면, 제시된 규칙들의 높은 오류율을 감수해야 하게 된다. 이것은 품사의 배당이 어원적 한자등에 기초하여 이루어져서, 실제 한국어 구문에서 고려되어야 하는 '구문적 단위 (Syntactic Unit)'를 중심으로 설정되지 못한 데에서 오는 문제점이다.

2. 합성어 단위 (Complex Unit)

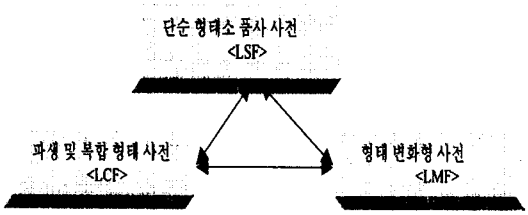
단일 품사 단위들 사이의 결합 관계를 별도의 규칙으로 기술하는 작업은, 경우에 따라서는 결합된 형태들을 바로 사전 엔트리로 저장하는 것보다 훨씬 더 복잡하고 비효율적인 프로세싱을 가져올 수 있다. 단일 형태소 품사 단위들이 여러개 결합하여 구성될 수 있는 '합성어 (Complex Unit) 유형'을 분류해 보면, 다음과 같다〈그림 3〉.

유형 <1>, <2>와 같이 '어휘소'들만이 결합하여 이루어진 형태들은 '파생 형태론 (Derivational Morphology)'으로 다루어지고, <3>, <4>, <5> 등과 같이 문법소가 함께 나타난 형태들은 '굴절 형태론 (Inflectional Morphology)'으로 다루어지는 것이 일반적이다. 한국어나 영어, 기타 유럽 언어들의 기존 사전들에서, <1>, <2>와 같은 형태들은, 단순 어휘 형태들과 함께 부분적으로 사전 엔트리류로 다루어지고 있고, <3>, <4>의 형태들은, 별도의 형태 변화 정보가 추가되어 유도될 수 있도록 구성되어 있다. <5>의 경우는 한국어를 비롯한 일정 언어에서 발견되는 특별한 현상으로, 가령 영어와 같은 언어권에서는 사전 구축시 우선적으로 제기되는 문제가 아니다. <5>의 형태는 다음 장에서 다시 논의될 것이므로, 여기서는 앞의 <1>, <2>의 파생 형태들과 <3>, <4>의 굴절 형태들에 관해서만 살펴보기로 한다.

위의 복합 형태들이 모두 사전의 기본 정보 단위로 저장된다는 것은, 다시 말해서 다음과 같이 세 가지 유형의 사전이 구축되어, 그들 사이의 관

	Type	Sub-Class	Example/Structure	
1	'어휘소(Lexeme)'들의 결합	복합어 (Compound Form) : 두개 이상의 '자립 형태소 (AL)'의 결합	'나이테'	AL-AL
2		파생어 (Derived Form) : 하나 이상의 '의존 형태소 (DL)'와 하나의 '자립 형태소 (AL)'의 결합	'신역사학파'	DL-AL-DL-DL
3	'문법소(Mor-pheme)'들의 결합	두개 이상의 '문법소 (DM)'들의 결합	'시겠지만은'	DM-DM-DM-DM-DM
4	'어휘소'와 '문법소'들의 결합	용언 어절 (Predicative String) 유형 : 동사, 또는 형용사 어간에 굴절 후치사가 결합되어 구성	'작았지만'/ '비좁아서'	AL-DM-DM/ DL-AL-DM
5		명사 어절 (Noun String) 유형 : 명사에 문법 표지 후치사가 결합되어 구성	'연구소에'/ '학교에도'	AL-DL-DM/ AL-DM-DM

〈그림 3〉



〈그림 4〉

계가 FST (Finite State Transducer)등과 같은 데이터 구조에 의해 구현된다는 것을 뜻한다 [Sil93]〈그림 4〉.

1) 형태 변화형 사전

'형태 변화형 사전(Lexicon of Morphological Forms : LMF)'은, 위의 <그림 3>에서 <3>과 <4>와 같은 굴절 형태들을 기본 단위로 갖는 사전 형태로, 이와 같은 사전을 구축하기 위해서는 모든 '문법소', 즉 '용언 활용 어미'들간의 결합 관계를 낱낱히 기술하여야 하며, 또한 기본 원형과 결합할 때 발생하는 어형 변화들이 빠짐없이 조사되어야 한다. 영어와 같은 경우, 각 '동사 (Verb)'에 대한 활용 형태의 수는 그렇게 많지 않기 때문에 이와 같은 사전의 구축은 그 구현이 비교적 쉽다. 유럽어의 경우, 예를 들어 프랑스어의 경우, 한 동사는 50 여개의 활용 형태까지 가질 수 있으므로 영어의 경우보다는 그 규모와 복잡도가 더 심각해 진다. '형용사 (Adjective)'의 경우에도, 영어와는 달리 성과 수의 변화를 갖게 되므로 4 가지 유형의 변화형을 고려해야 한다. 한국어의 경우에는 문제가 훨씬 더 심각해 진다. 하나의 동사뿐 아니라 형용사에도 여러 유형의 활용 어미 (가령, 시제, 양태, 시상, 문장 유형등에 대한 정보)들이 연달아 들러붙어 실현될 수 있기 때문에, 주어진 동사, 또는 형용사가 몇 개의 활용 형태를 가질 수 있는지를 밝혀내기 위해서는, 여러 유형의 활용 어미들 사이의 그 결합 관계를 완벽하게 기술해야 한다. 가령, '떠나시겠다고'와 같은 형태는 기본형 어간 '떠나'에 '시', '겠', '다', '고'의 4 개의 어미가 결합되어 나타난 형태이다. 어미들의 조합 순서 및 그 제약 조건들은 몇 가지 규칙의 형태로 자동 추출될 수 없다. 다시 말해서, '*떠나-겠-시-다-고'라든지, '*떠나-시-겠-고-다'와 같은 유형은 불가능한데, 이와 같은 조합 가능성은 굴절 어미들을 일일이 조합해 보지 않고서는 기술될 수 없기 때문이다.

2) 파생 및 복합 형태 사전

'파생 및 복합 형태 사전(Lexicon of Complex Forms : LCF)'은, 위 <그림 3>에서 나타난 <1

>과 <2>와 같은 형태들을 기본 단위로 갖는 사전 형태로, 이러한 사전 정보가 일정 규칙의 형태로 구현되기에는 그 형성 관계가 너무나 복잡하다. 이것은 위에서 살핀 '형태 변화형 사전'의 경우보다 훨씬 복합적인 양상을 띠는데, 다음과 같은 이유로 설명될 수 있다.

(1) 파생 연산의 반복

첫째, 앞서 논의한 바와 같이, 단순 형태소 명사에 접사(Affix)가 결합하는 파생(Derivation)이 여러 번 반복될 수 있다. 가령, '탈옥하다'는 명사 '옥'에 접두사 '탈'이 결합하여 다시 '하다'가 결합된 것으로, 이와 같은 파생 형태들은 각 어휘 요소들에 대한 개별적인 검토없이 선험적인 규칙으로서 생성해 낼 수 없다. 여기서는, 서술성을 갖지 않은 명사 '옥'이 서술형 접두사 '탈'에 의해 '하다' 동사류 구성이 가능해진 것인데(즉, '*옥하다' vs. "탈옥하다'), 이때, 서술성이 없는 '감'과 같은 접사가 나타나면, 'PF-N-V'구성은 불가능해 진다(즉, '*감옥하다'). 여기서, 파생 명사의 목록이 원칙상 무한하게 확장될 것으로 생각되어질 수 있다. 그러나 자연어에서 무한하게 반복되는 연산(Operation)은 결코 존재하지 않는다. Chomsky의 '언어 형식 모델'에서, 무한히 반복될 수 있는 것으로 가정된 '관계절(Relative Clause)'의 연산도, 실제 문장에서는 최대로 설정해도 10번 이상 결코 이와 같은 현상이 반복되어 나타나지 않는다는 사실은, 형식적인 언어 모델이 구체적인 현상에 대한 검토없이 수학적 가정에 의해서만 제시되었을 때 갖게 되는 문제점을 보여주는 것과 같다.

(2) 복합 형태들의 '비분리성(Non-Compositionality)' 현상

합성어 유형중에서 '자립 형태소'들이 연이어 결합하여 이루어진 '복합 명사(Compound Noun)'들과 같은 경우, 단순 품사형으로부터 자동으로 그 목록이 구성되기는 매우 어렵다. 가령 '밤낮'과 같은, 두 명사의 결합으로 이루어진 복합어의 경우, 그 의미는 두 단어의 산술적인 합이 아닌 전혀 새로운 뜻으로, '늘', '항상'과 같은 시간 부사들과 동의 관계의 의미를 가진다. '집사람'과 같은 복합어의 경우에도, 그 뜻이 '집'과 '사람'의 두 단어의 결합으로 유추될 수 있는 것이 아니어서, '아내', '부인'등의 어휘와 대응관계를 가지며, 어휘적으로도 자유롭지 못하기 때문에,

*아파트사람, *건물사람, *방사람, *집여자

등과 같은 표현으로 실현될 수 없다. 이와 같은 복합어들은, 개별적으로 그 어휘 결합 정보가 기술되어야 하므로, 규칙화하기 어려운 경우들이 대부분이므로, 복합 구성 형태가 바로 사전 엔트리로 등재되는 것이 더 바람직하다.

그런데, 복합어류에는 이와 같이 굳어진 정도가 아주 강한 것들이 있는가 하면, 비교적 자유롭게 구성될 수 있는 것으로 보이는 '가죽 장갑', '털 장갑' 등과 같은 구체 명사류와, '국민 의식', '신용사회'등과 같은 추상 명사류들이 존재한다. '가죽 장갑', '털 장갑'과 같은 경우, 재질을 나타내는 명사와 그와 같은 재료로 만들어진 구체적인 대상물의 명사가 함께 결합되어 만들어지는데, 가령 '가죽 잠바', '가죽 가방', '면 바지', '털 조끼', '비닐 덮개' 등과 같이, 조합 가능 형태들이 대단히 생산적인 것으로 보인다. 그런데 이러한 형태들중에는, '고무 장갑', '비닐 봉지' 등과 같이 그 의미의 결속력이 위의 형태들보다는 각별해 보이는 것들이 있다. 추상 명사들의 경우에도 마찬가지여서, '국민 투표', '신용 대출'과 같은 형태들은 하나의 단위로 인식될 수 있는 의미적 응집력이 비교적 높다. 그러나 실제로 존재하는 복합 형태들의 유형을 살펴보면, 굳어진 정도를 '이분화 (Binary Division)'하여 기술하기는 어렵다. 언어 이론적으로는, '굳어진 형태 (Frozen Expression)'를 '복합 명사(Compound Noun)'라 정의하고, 자유로운 구성을 '명사구 (Noun Phrase)'로 나누어 분류하는데, 이 두 유형 사이의 경계는 현실적으로 불분명하다.

자연어의 실제 데이터들을 관찰해 보면, 사실상 '완전히 자유로운 명사구'란 존재하지 않는다. 두 어휘 사이에는 반드시 어떠한 형식으로든지 일정 '의미적 제약 (Semantic Restriction)'이 존재하기 때문이다. 여기서, 과연 이와 같은 모든 명사구들을 빠짐없이 기술해낼 수 있을까 하는 의문과, 또한 가능하다 해도 이것들을 모두 사전에 등재시킬 필요가 있는가 하는 의문이 제기될 수 있다. 우선,

두번째 질문은 사전을 어떠한 개념으로 이해하느냐 하는 입장과 무관하지 않은데, 사전이 궁극적으로 모든 언어 정보를 제공하는 대규모 언어 베이스의 형태로 이해된다면, 이러한 작업 자체의 정당성은 분명히 입증된다. 두 어휘 요소들 사이의 결합 가능성은 단순한 의미적 유추만으로 예측될 수 없기 때문에, 가령 '치마단'과 '원피스단'에서는 유사한 의미를 갖는 것처럼 보이는 '치마'와 '원피스'가, '치마바람'과 같은 구성에서는 '*원피스바람'과 같은 형태로 치환이 이루어지지 않는 사실은, 일반 규칙의 설정으로는 제어하기 어렵다. '과연 모든 명사구들을 기술해낼 수 있을까' 하는 첫번째 질문에 대해서는, 경험적으로밖에는 답할 수 없다. 두 개의 단순 명사 형태가 결합할 수 있는 가능성에 대해서만 고려해 보아도, 이론적 산술에 의하면, n 개의 단순 명사에 대해서, 그 복합체의 수는 최대 n(n-1)개에 이를 것이다. 따라서, 15 000 개에 이르는 한국어 단순 명사의 경우 (Cf. DECOS-NS/V01 [Nam94]), 2 개의 명사로 이루어진 복합체의 수는 2억 2천만 (15 000×14 999) 여개에 이를 것으로 계산된다. 그러나, 앞서도 언급한 바와 같이, 실제로 자연어 데이터들을 조사해 보면, 그 이와 같은 n(n-1) 개의 조합은 결코 이루어지지 않는다. 여기서 보다 현실적인 문제의 접근은, 그러면 과연 '어떠한 방식으로 이러한 정보들을 구축해 나갈 것인가'하는 질문일 것이다. 모든 명사구 형태들을 한꺼번에 다 고려할 수 없기 때문에, 아주 굳어진 표현들로부터 몇 단계의 유형 분류를 거쳐, 한 부류씩 기술해 나가야 한다. 우선, 명사구가 'Noun-Noun'의 결합 형태인지, 'Adjective-Noun'의 형태인지 등과 같은 분류가 이루어져야 하고, 가령, 'N-N'의 경우, 추상 명사들인지, 구체 명사들인지, 또는 그 내부 논항 구조가 '목적어-술어'인지 (즉, '대통령 선출'), '주어-술어'인지(즉, '대통령 연설') 등과 같은 하위 분류가 이루어 져야 한다. 굳어진 정도가 아주 심한 경우를 제외하고는 복합 명사들의 경우, 두 개 이상으로 분리된 형태로도 실현되므로 이와 같은 정보들도 아울러 고려되어야 한다.

(3) 형태 정보와 구문 정보사이의 관계

'서술성 (predicativity)'을 갖는 '연설', '공격'등과 같은 명사들은, 접미사 '하다'와 결합하여, '연설하다', '공격하다'등과 같은 하나의 파생 동사를 구성한다. 이때, 이들을 사전에 기본 엔트리의 형태로 일일이 등재시키는 대신, 서술성을 가진 것으로 판단되는 명사들로부터 자동적으로 생성되는 파생 형태들로 규칙화하는 방법을 가정해 볼 수 있다. 그러나, 이와 같은 '서술성'의 자질은, 우선 명시적으로 정의하기 어려울 뿐 아니라, 설령 그러한 의미값이 주어진다 하더라도, 이것은 '충분 조건'이 아닌 하나의 '필요 조건'에 불과하다. 가령, '임종'과 '죽음'은 모두 서술성을 갖는 명사로 여겨지는데, 이때, 다음과 같이,

임종하다 / *죽음하다

'임종하다'가 가능하다면, '*죽음하다'와 같은 형태의 구성은 불가능하다. '죽음'이 한자어가 아니라서 발생된 문제인 것인가 하는 입장에서, '우정'과 '사랑'의 쌍을 비교해 보면, 상황이 반대가 되는 것을 볼 수 있다. 즉, 한자어 '우정'은 '하다' 구성을 허용하지 않는 반면('*우정하다'), 순우리말 '사랑'은 '사랑하다'와 같은 동사 구성을 허용한다. 이와 같은 현상은, 주어진 명사의 의미 자질등에 대한 검토만으로, 파생 가능한 어휘 성분들을 추측할 수 없음을 보여 주는데, 그러므로 'N-하다'류 구성을 허용하는 명사들은 개별적으로 조사되어 사전에 수록되어야 한다.

그런데 여기서, 'N-하다'류 파생 동사에 대한 처리는, 앞서 살펴 본 다른 파생어들의 경우와는 좀 다르게 이루어져야 할 것으로 보인다. 15 000 개의 단순 명사로부터, 절반에 가까운 6 500 여개의 명사들이 이러한 'N-하다'의 구성을 허용하는 것으로 관찰되는데, 다른 파생어들의 경우와는 달리, 이들은 80 여개의 경우를 제외하고는, 모두 다음의 (1b)와 같은 하나의 술어구 (Predicative Phrase)와 대응 관계를 보인다.

(1a) N-하다

(1b)= N-를 하다

가령, '공격하다'는 '공격을 하다'와 같은 형태로 실현될 수 있는데, 이때, '공격을 하다'는, '공격하다'와 같은 하나의 파생 어휘 요소와는 그 통사적

속성이 다르다. '공격을 하다'의 경우, '빨리', '먼저' 등과 같은 부사어의 삽입이 가능하고 ('공격을 먼저 하다'), '공격'을 수식하는 관형형의 삽입이 허용된다 ('예상치 못한 공격을 하다'). 이러한 특성은 하나의 단일어인 '공격하다'에서는 관찰할 수 없는 현상이다. 이와 같은 하나의 '파생어'와 하나의 '통사적 구문'사이에 아주 규칙적으로 대응 관계가 존재한다는 사실은, 'N-하다'류 파생 동사를 사전에 바로 등재시키는 작업을 주저하게 만든다. 이러한 파생 동사들과 평행적으로 존재하는 6 500 여개의 명사들이 사전에 다시 등재되고, 이와 같은 구문적 특성들이 각 명사를 중심으로 반복되어 기술되며, 또한 이러한 파생 동사들과의 동의 관계를 다시 한번 재언급하는 부담을 안게 되기 때문이다. 따라서 이때, (1a)과 (1b)에서 나타난 두 형태가 일정 변형 관계에 의해 서로 연관되어 있다고 가정하고, 하나의 형태만을 사전 엔트리로 수록하게 되면, 동일 어휘 요소에 대한 이중 처리의 부담을 줄일 수 있다. 다음과 같은 형태들을 포함한 80 개의 '하다'류 동사들<그림 5>을 제외하고는, 'N-를 하다' 의 구성과 대응되지 못하는 'N-하다' 동사는 없으므로 (그 역은 반드시 참이라고 하기 어렵다. 가령, '올해는 고추를 했다 (= 재배했다)'에서 '고추하다'와 같은 결합형은 다소 부자연스러운 듯이 보이기 때문이다), '명사' 엔트리를 중심으로 'N-를 하다'와 'N-하다'의 구문 정보를 저장하는 것이 바람직하다.

가하다	거하다	과하다	금하다	논하다
간과하다	격하다	구하다	기하다	다하다
갈구하다	고무하다	굴하다	꾀하다	달하다
감하다	고하다	권하다	노하다	etc...

〈그림 5〉

위의 '하다'류 동사들은, 3 개만을 제외하고는 모두 '단음절 (Mono-Syllabic) 한자어' 형태에 '하다'가 결합된 유형이다. 이들은 반드시 결합형으로만 나타나는 동사들이므로 (즉, '가하다' vs. "*가를 하다'), '가', '거', '과'등이 하나의 명사로 등재되어서는 안된다. 즉, 이 동사들은 파생 동사가 아닌, 하나의 단일 동사들이므로, 사전에 그대로 등재되어야 한다.

'N-하다' 동사형을 사전 엔트리로 등재하지 않고, 해당 명사들에 대한 일종의 구문 정보로서 저장하게 되면, 다음과 같이, 기존 사전에서 고려되지 않은, 'Nc (구체 명사 : Concrete Noun)-하다' 유형의 동사 형태들에 대한 정보를 동일한 방법으로 처리할 수 있게 된다. 가령, 다음의 쌍들을 비교해 보면,

(2a) 공격하다, 비판하다,

(2b) 저녁하다, 목걸이하다

사전에는 전혀 표시되지 않은 (2b)와 같은 형태의 스트링을 구성하는 '구체 명사'들이 상당수 존재한다는 것을 알 수 있다. 이러한 'N-하다'의 유형은, '하다'가 일종의 '대동사 (Pro-Verb)'적으로 사용된 경우가 많아, 예를 들어 살펴보면 다음과 같다 <그림 6>.

유 형	대응되는 동사 형태	예 문
Type 1	N-을 하다=N-을 경영하다	탁아소를 하다(=경영하다)
Type 2	N-을 하다=N-을 마련하다	혼수감을 하다(=마련하다)
Type 3	N-을 하다=N-을 착용하다	목걸이를 하다(=착용하다)
Type 4	N-을 하다=N-을 요리하다	갈비찜을 하다(=요리하다)

〈그림 6〉

'하다'를 중심으로 하나의 파생어 형태 (즉 'N-하다')와 통사적 구문 (즉 'N-를 하다')사이에 나타난 일정 대응 관계는, '되다', '시키다', '당하다', '주다', '받다'등과 같은 일종의 '기능 동사 (Light Verb)'들이 '서술성 명사 (Predicative Noun)'들을 동반할 때에도 관찰된다. 즉, 다음과 같이 두 구조 사이의 동의 관계가 관찰된다.

N-(되다+시키다+당하다+주다+받다)
=N-Post(되다+시키다+당하다+주다+받다)

예를 들면 다음과 같다.

비판되다 = 비판이 되다
벌받다 = 벌을 받다
공격당하다 = 공격을 당하다

공부시키다 = 공부를 시키다

이러한 동사 형태들은, 한국어에 있어서, '형태론'의 범위에서 다루어져야 할 부분과 '구문론'의 범위에서 다루어져야 할 부분 사이에 구별이 명확하지 않은, 하나의 예를 보인다. 'N-하다' 형태를 동사 엔트리 유형으로 처리하지 않고, 그 해당 명사들에 이와 같은 정보를 결합시킨다는 입장은, 바꿔 말하면, 이러한 현상을 '형태 정보'의 한 유형으로 간주하지 않고, 명사 엔트리에 덧붙여진 '구문 정보'의 한 형태로 간주한다는 것을 의미한다.

3. 어절 단위(Typographical Unit)

'어절 단위 사전'이란 위의 <그림 3>에서 본 바와 같이, 형태 변화형 및 합성어 유형뿐 아니라 명사에 후치사들이 동반되어 나타난 형태들을 모두 포함하는 사전 형태이다. 즉, 한국어 문서내에서 두 개 이상의 분리 기호 (Separator)에 의해 형성되는 모든 유형의 스트링을 가르킨다. 가령, 다음 문장은

대도시에서는 남자들에 비해서 여자들의 비율이 더 높다

모두 7 개의 어절로 구성되어 있다. 이때, 하나의 어절 '대도시에는'은 2 개의 합성어 ('대도시'와 '에는')로 이루어져 있고, 이 2 개의 합성어는, 다시 각각 2 개의 단일 형태소 품사들의 결합 (즉, 2 개의 어휘소 '대'와 '도시'의 결합인 파생어 '대도시'와, 2 개의 문법소 '에'와 '는'의 결합인 복합 후치사 '에는')으로 구성된다. 기존의 어느 사전에도 '대도시에서는'과 같은 형태는 기본 엔트리 유형으로 저장되어 있지 않은데, 그것은 마치 용언의 활용형들이 사전에는 나타나지 않는 것과 같다. 한국어의 경우, 하나의 명사는 대부분 하나 이상의 후치사가 동반되어 실현되므로, 자연어 문서 처리 프로세싱에 있어서 이와 같은 결합 형태들을 인식하지 못하게 되면, 사전 매칭은 실패로 돌아가게 된다. 마치 한 동사의 기본형만을 알고 있을 때, 그 형태 변화형들이 이루어지기 위한 결합 정보들을 갖지 못하면, 이와 같은 형태들의 인식이 불가능한 것과 같다.

실제로, '명사-후치사'의 스트링은 '동사어간-굴절어미'의 경우와 마찬가지로 하나의 응집된 형태로 실현되므로, 문서 처리의 첫 단계에서 이와 같은 결합에 대한 정보가 주어져야 비로소 프로세싱이 진행될 수 있다. 컴퓨터용 사전에 이러한 형태가 기본 정보 단위로 입력된다면, 주어진 입력 스트링들에 대한 인식은 매우 효율적으로 이루어 질 것인데, 이때 이와 같은 사전 형태의 구현이 쉽게 구상되지 않은 첫째 이유는, 동사의 활용형의 경우와는 달리, 명사와 후치사가 접목되는 부분에는 어떠한 형태 변화도 일어나지 않기 때문이다. 가령, '줍다'와 같은 동사는 활용 어미 '어'가 동반되면, '주워'와 같은 형태로 바뀌게 되는데, 명사 '길'과 후치사 '에서'가 결합하게 되면, 이들 사이의 결합은 아무런 형태상의 변이를 수반하지 않는다. 이때, 후치사 '에서'를 '서'의 변이형 (Variant)으로 간주할 수도 있다. 그러나, 주어진 명사에는 아무런 형태 변화가 일어나지 않으며, 후치사 '에서'와 '서'의 선택은 명사의 끝 음소가 '자음 (Consonant)'인가 '모음 (Vowel)'인가에 따라 예외없이 예측될 수 있다. 둘째 이유는, 명사의 경우, 앞서 본 바와 같이, 용언류와는 달리 합성어등의 구성이 훨씬 풍부하고 복잡하기 때문에, 이러한 형태들에 결합 가능한 후치사 형태까지 일일이 고려하여, 그와 같은 어절들을 모두 사전에 등재한다는 것이 과연 가능할 것인가 하는 의문점 때문이다. 가령, 접두사 '여'가 결합되어 파생 명사를 구성할 수 있는 명사는, 다음에서 보듯이 높은 생산성 (Productivity)을 가질 것으로 보이는데,

여선생, 여기자, 여사장, 여학생, 여배우, ...

이때 이러한 파생 명사들에 결합 가능한 후치사 형태들까지 모두 고려한다면, 그와 같은 어절의 수는 기하급수적으로 증가될 것이다.

그러나, 다음과 같은 경우를 보자. 이것은, 어절 형태에 대한 정보가 사전에 바로 저장되지 않고, 별도의 규칙으로 첨가되어 있을 때 발생하는 문제점을 보여 준다.

정치가

위 스트링은 두 가지 분석의 중의성을 갖는다. 즉, 명사 '정치'에 사람을 가르키는 접미사 '가'가 동반하여 이루어진 하나의 '파생 명사'로 해석되거

나, 또는 명사 '정치'에 후치사 '가'가 결합된 하나의 '주어 어절'로 해석되거나 하는 경우이다. 사전에 입력된 기본 단위가 합성어 수준까지만 되어 있어, 어절 단위에 대한 정보가 바로 얻어질 수 없을 때, 다음과 같은 문맥에서

한국 정치가 많이 달라졌다

'정치가'가 주어 어절로 올바르게 분석될 수 있기 위해서는, 주어진 스트링을 무조건 사전에 등재된 형태와만 매칭시켜서는 안된다. 따라서 주어진 스트링은 '분절 (Segmentation)'의 과정을 반드시 거쳐야 한다. 즉, 명사로 인식될 수 있는 단위가 판별되면, 그것에 결합된 형태가 후치사의 일종으로 인식될 수 있는지의 검증을 거쳐야 하는데, 가령, '정치가'라는 스트링으로부터 가능한 모든 명사 후보를 사전으로부터 찾아내고 난 뒤, 후치사 유형을 검색해야 한다. 그런데, 이와 같이 '분절'의 프로세싱을 거쳐야 하는 것은, 컴퓨터 처리에 있어서 매우 심각한 결과를 가져온다. 가령, '뽕나무누에'와 같은 명사 형태가 합성어 사전에서 매칭되었다 하더라도, 이것이 '명사+후치사'의 결합 형태인지 아닌지에 대한 확인 (가령, '뽕나무누(명사) + 에(후치사)'의 구조인지)을 위해서 분절의 과정을 반드시 거쳐야 하므로, 오른쪽 우선 분석 (Right-to-Left), 또는 왼쪽 우선 분석 (Left-to-Right) 등의 어느 알고리즘을 사용하든지 간에, 가능한 모든 명사 후보들과 모든 후치사 후보들을 찾아낼 때까지, 분절을 통한 사전 매칭은 여러번 이루어져야 한다.

실제로, 한국어 처리를 위한 현행 형태소 분석기 시스템들의 대부분이 이와 같이 분절의 과정을 전제해야 하는데, 바로 분절의 과정을 전제해야 하는 데에서 나타나는 피할 수 없는 문제가, 수없이 많은 사전 검색의 횟수와 더불어, 엄청난 수의 오분석의 발생이다. 이와 같이 발생하는 중의성 (Ambiguity)을 해소하기 위해서 여러 가지 방법론 및 모델, 알고리즘들이 제시되고 있으나, 어절 단위 사전을 이용하지 않고는 어떠한 방법으로도, 이와 같이 발생할 수 있는 중의성들을 해결하기 어렵다. 이때, 여기서 말하는 '중의성'은, 구문상 발생할 수 있는 중의성 그 이전의 모든 유형을 가르킨다. 가령, '정치가'가 하나의 주어 어절일지, 하나의 파생 명사일지에 대해서는 통사적 문맥이 주어져야 비로소 판단될 수 있는 것으로, 사전에 주어진 어절 정보만으로는 그와 같은 선택은 불가능하다. 반면, '남자가'와 같은 형태는 명사 '남자'에 접사 '가'가 결합하여 유도된 파생어 형태가 사전에 존재하지 않고, 하나의 어절 단위 (즉, '남자(명사) + 가(후치사)')로서만 구성이 가능하므로, 자연히 이와 같은 오분석의 가능성은, 이 단계에서 사전 매칭을 통하여 미리 제거될 수 있는 것이다.

4. 구, 또는 절, 문장 단위

사전의 기본 단위가 어절의 형태로부터 더 확장되면, '명사구 (Noun Phrase)', "동사구 (Verbal Phrase)', "형용사구 (Adjectival Phrase)', 또는 '부사구 (Adverbial Phrase)' 등과 같은 '구' 단위로 사전 엔트리 유형이 설정될 수 있다. 가령,

아양을 떨다

주책을 부리다

와 같은 형태로 나타난 동사구의 경우, 그 결합 관계는 아주 구속적이어서, '아양', '주책'등과 같은 명사들은, 일반 명사들과는 달리 술부를 벗어나 주어나 다른 보어의 위치에 쉽게 나타나기 어렵다. 마찬가지로, '떨다', '부리다'와 같은 동사들의 경우에 있어서도, 함께 실현될 수 있는 명사들의 목록은 아주 제한적이다. 따라서, 이와 같은 형태들이 사전에 하나의 기본 단위로 등재되면, 명사류들에 대한 통사적인 특성을 기술할 때, 별도로 특별한 장치를 고안해야 할 필요가 없다. 시간을 나타내는 부사구와 같은 경우에도, 구 형태 전체가 일종의 '부분 문법(Local Grammar)'의 형태로 사전에 바로 입력되면, 구문 분석 이상의 층위에서 매우 효율적인 처리를 가져올 수 있다. 가령, 다음 두 문장에서,

회의는 <u>두 시 십 분전에</u> 시작된다

회의는 <u>한 시 오십 분에</u> 시작된다

'두 시 십 분 전에'와 '한 시 오십 분에'는 같은 시각을 나타내는 표현으로, 이들이 그대로 사전 입력 단위가 되면, 이 두 표현 사이의 동의 관계를 기술하는 방법이 훨씬 용이해 질 것이다. 다음과 같은

절이나 문장의 경우에는,

어제 도착한 편지가 온 데 간 데 없다

가지 많은 나무에 바람 잘 날 없다

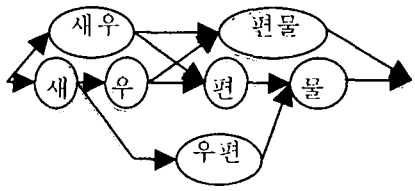
굳어진 정도가 매우 심해, 가령 '온 데 간 데 없다'의 경우, '*간 데 온 데 없다', '*간 곳 온 곳 없다' 등과 같이 유사한 의미의 표현들과 서로 '치환(Permutation)'될 수 없다. 이러한 표현들이 그대로 사전에 입력될 때, '구문 분석(Syntactic Analysis)'이나 '기계 번역(Machine Translation)' 등의 분야에서 매우 유용한 정보를 갖게 될 것이다.

5. 기본 단위 유형들 사이의 비교

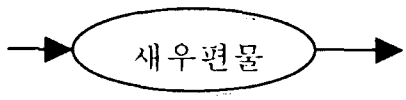
앞서 살핀 네 가지 유형의 사전 기본 단위(Basic Unit) 들을 비교하기 위해, 다음과 같은 문장을 살펴 보자.

새우편물 특별 요금 제도가 이번달부터 실시된다

위에서 '새우편물'이라는 스트링은, 단일 형태소 품사 단위로 사전을 구성하게 되면, 사전과의 매칭을 위하여 반드시 분절의 과정을 거쳐야 하는데, 이때 다음과 같은 7 가지의 사전 엔트리 유형들과 매칭이 이루어지게 된다. 따라서, 이 스트링은 모두 5 가지 유형의 분절 결과를 보이게 된다 <그림 7>.



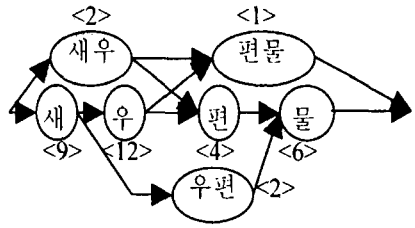
〈그림 7〉



〈그림 8〉

반면, 합성어 단위로 사전 엔트리 유형을 구성하게 되면, 위와 같은 분절에 의한 중의성은 사라지고, 다음과 같이 단 하나의 올바른 사전 후보와 매칭된다 <그림 8>.

이때, 위의 7 가지의 단일 품사 엔트리들은 여러 가지의 어휘 의미적 중의성을 가지고 있으므로, 그것들을 계산하면 다음과 같이 모두 2 832 개의 분석 가능성이 나타난다 <그림 9>.



〈그림 9〉

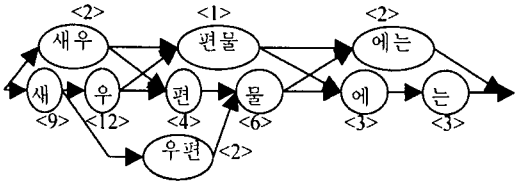
여기서 볼 수 있는 것과 같이 합성어 중심으로 기본 단위가 설정되지 않으면, 사전 정보와의 매칭을 위해서 분절을 거쳐야 하고, 따라서 가능한 모든 조합의 가능성으로부터 단 하나의 올바른 형태를 찾아내기 위해서는, 저장해야 하는 어휘 형성 결합 규칙의 수가 엄청나게 확장된다. 더구나, 단일 형태소들의 경우, 여러 가지의 의미를 가질 수 있는 '어휘적 중의성(Lexical Ambiguity)'이 매우 복잡하게 나타나는 반면, 합성어 형태로 그 단위가 확대되면, 중의적 해석의 가능성은 매우 줄어 들어, '새우편물'과 같은 형태는 한 가지 해석만이 가능하게 된다.

그런데, 다음과 같은 문장을 보자.

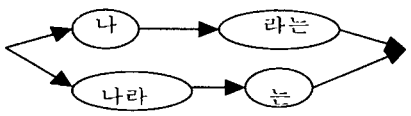
새우편물에는 우체국의 도장이 생략된다

이 경우, '새우편물에는'과 같은 스트링의 인식을 위해서 사전 매칭을 하게 되면, 단일 품사 엔트리의 경우, 위에서 보다 그 수가 기하 급수적으로 증가하게 되어, 다음에서 보이는 바와 같이, 2 백만개가 넘는 (즉, 2 597 511 개) 분석 가능성을 갖게 된다 <그림 10>.

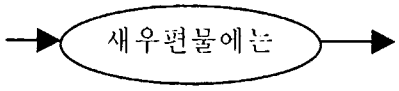
이와 같은 스트링을 합성어 단위의 사전과 매칭을 하게 되면, 앞의 '새우편물'의 경우와는 달리, 사전 엔트리로 저장되어 있지 않은 형태이므로, '분절'의 프로세싱을 거쳐야 분석이 이루어진다. 즉, 주어진 스트링에서 명사 후보가 하나 나타났다



〈그림 10〉



〈그림 11〉



〈그림 12〉

고 해서, 그것으로써 모든 분절의 가능성을 무시해 버릴 수는 없기 때문이다. 가령, '나라는'이라는 스트링에, 대명사 '나'가 매칭되고, 그 다음에 후치사 '라고'가 나타날 수 있다고, 분석을 여기서 완료해 버리게 되면, '나라'라는 명사에 후치사 '는'이 결합된 가능성을 파악할 수 없게 된다. 그러므로, 다음과 같이 두 가지 경로의 가능성이 동시에 설정되어야 하는 것과 같다 <그림 11>.

따라서, <그림 10>에서와 같은 어절을 합성어 단위의 사전과 매칭시키게 되면, 결국 단일 품사 사전을 사용할 때와 마찬가지로, 모든 분절의 가능성을 찾기 위하여 여러 차례에 걸친 사전 검색을 거치게 될 것이다. 여기서, 어절 단위 사전의 필요성이 확인된다. 어절 단위 사전에는 모든 유형의 입력 스트링들이 바로 엔트리 형식으로 구성될 것이므로, 위와 같은 스트링은 사전과의 매칭에 의해서 단 한번에 찾아진다 <그림 12>.

여기서, 사전에 입력되는 단위가 확대되면 될수록, 자연어 문서 처리시 발생하는 중의적 해석의 가능성은 현격하게 줄어든다는 중요한 원리가 다시 확인된다. 자연어 문서의 중의성들은 문맥을 확대해서 고려해야 해결될 수 있는 경우들이 대부분이다. 즉, 위에서 '새우'라는 명사로의 분석이 가능한지의 여부는 뒤에 나타난 '편물'이라는 명사를 보아야 비로소 결정될 수 있는데, 즉 명사 '새우'가 결합하여 이루어질 수 있는 합성어의 형태들, 바꿔 말해서 명사 '새우'가 나타날 수 있는 모든 어휘적 문맥이 밝혀졌을 때, 비로소 중의성의 해소가 가능해 진다. 합성어 사전이란 바로 이와 같은 문맥 정보를 사전에 빠짐없이 수록한 것으로, 이러한 사전류를 사용하게 되면 중의성의 해결에 큰 도움을 얻을 수가 있다. 그런데, 대부분의 명사들이 '후치사'들을 동반하여 실현된다는 점을 감안할 때, 어절 단위 사전이 없이, 합성어 사전만으로는 자연어 처리를 하게 되면, '분절'의 과정을 필연적으로 거쳐야 하므로, 결국 단일 품사 사전의 경우와 마찬가지 횟수의 사전 검색을 통해서 모든 명사 후보들을 찾아야 한다. 그러므로, 어절 단위 사전에 기초한 경우에 비해서 그 효율성이 엄청나게 저하된다. 그러나, 사전에 어떠한 형태를 기본 단위로 설정할 것인가 하는 점은 선험적으로 미리 결정지을 수 있는 문제가 아니다. 어떠한 유형의 단위를 기본 엔트리로 할 때, 그 엔트리의 수가 얼마에 이를지는 개별 언어마다 차이가 있고, 따라서 그러한 정보를 사전 엔트리 형태로 저장할 것인지 아니면 별도의 정보 형태로 결합시킬 것인지는 실제 프로세싱의 효율성의 정도에 따라 결정되어야 한다. 그러나 무엇보다도 중요한 문제는, 모든 가능한 어휘 결합 형태에 대한 완벽한 기술이 우선 이루어 져야 한다는 점이다.

III. 관련 정보의 유형

1. 음운 정보

사전에 등재된 기본 단위들의 '음운 정보(Phonic Information)'는 응용 분야에 따라 매우 중요한 역할을 한다. 자동 음성 인식 시스템의 개발을 위한 음성 데이터 베이스의 구축시, 음운 정보를 가진 사전 엔트리들과의 매칭은, 데이터에서 발견되는 수많은 '동음 중의성(Homophonic

Ambiguity)'의 해결을 돕는다. 가령, '꼰망울'과 같은 음성 데이터가 인식되었을 때, '꼰'의 실제 표기는 다음과 같이 3 가지 형태로 나타날 수 있다.

꼰, 꽂, 꽃

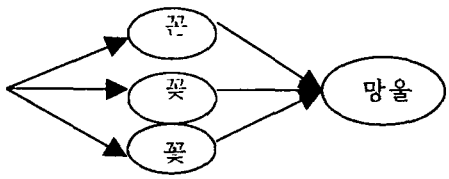
즉, 다음에서 보듯이,

꼰/꼰/ 새끼줄을...

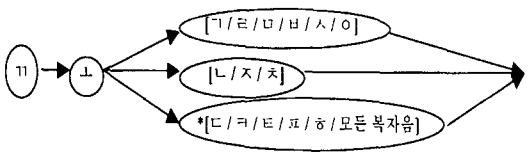
연필을 꽂/꼰/는 ...

꽃/꼰/망울

세 가지 다른 표기 형태가, 일정 문맥 조건안에 실현될 때, 동일한 음가를 가지고 실현되는 것을 알 수 있다. 이때 '꽃망울'과 같은 합성어가 사전에 저장되어 있고, 그와 같은 합성어가 어떠한 음운 속성을 갖게 되는지에 대한 정보가 제공되면, 다음과 같은 세가지 후보로 부터 올바른 매칭이 이루어질 수 있다 <그림 13>.



〈그림 13〉



〈그림 14〉

위와 같이 3 가지 후보가 가능하다는 사실은, 실제로 일반적인 한국어 음운 정보를 이용해서 이미 선택한 결과에서 얻어진 것이다. 가령, 다음 '음절 구성 오토마타'에서 보여지듯이, 'ㄲ'와 'ㄴ'는, 'ㄷ'이나 'ㅋ', 'ㅌ', 'ㅍ', 'ㅎ' 등의 자음들과 복자음등이 종성 위치에 나타나서, 하나의 음절을 구성하는 것을 허용하지 않는다(한두 가지 방언의 경우를 제외하고는). 즉, 이러한 정보가 주어졌기 때문에, 위의 자음들을 종성으로 하는 음절들은 미리 그 가능성의 후보 목록에서 제외된 것이다 <그림 14>.

위의 오토마타에서, 처음 셋트는 'ㄴ'으로 음가가 바뀔 수 없는 자음들을 가르키고, 둘째 셋트는 제한된 일정 상황속에서 'ㄴ'으로 바뀌어 나타날 수 있는 자음들을 나타낸다. 그리고 세째 셋트는 'ㄲ+ㄴ'와 함께 하나의 음절을 구성하지 못하는 종성 자음들을 가르킨다.

그런데, 위와 같은 데이타의 결과는 이대로 일반화되어질 수 없다. 우선 첫째로, '꽃'이 '망울'과 같은 어휘와 함께 나타나게 되면, '꼰'으로 실현되나, '밭'과 같은 어휘와 나타나면, 그 음가는 더이상 '꼬+ㄴ'이 아닌 '꼬+ㄷ'이 되는데, 이때 뒤에 실현되는 자음의 음가가 무엇이냐 ('ㅁ' vs. "ㅂ') 에 따라 그 제약 조건이 기술된다는 일반적인 원칙을 가정해 볼 수 있을 것이다. 그러나, 다음을 보면, 이와 같은 음가의 결정은, 단순히 음소의 형태와 대응되는 자동 메카니즘에 의한 것이 아닌 것을 알 수 있다.

꽃이슬 /꼰니슬/

꽃이 /꼬치/

위에서 '꽃'은 두 경우 모두, 'ㅇ'로 시작하는 동일 음절 '이' 앞에 나타났다. 그러나 실현된 음가는 전혀 다르게 나타났다. 문법적 구성 조건등이, 주어진 형태의 음가 결정에 개입되는 것으로 보여지는데, 이때, '꽃이'와 같은 형태의 음가를 알기 위해서는 '어절' 단위로 엔트리 정보를 가지고 있는 사전과의 매칭이 필요하다. 따라서, '음소' 단위로 실제 음가에 대한 제약 조건을 기술하는 것은 부적당하다. 주어진 문맥의 성격은, 하나의 단순 형태소, 합성어의 차원을 넘어서서, 하나의 어절 형태로 확대되어 기술되어야 비로소 올바르게 밝혀질 수 있다.

위에서 살핀, '꼰망울'의 예를 단순히 일반화시킬 수 없다는 두번째 근거는, 위와 같은 데이타로 부터, 'ㄴ / ㅈ / ㅊ'과 같은 형태들은 'ㅁ'과 결합할 때 'ㄴ'의 음가를 갖게 되며 ('꽃망울'처럼), 'ㄱ / ㄹ / ㅁ / ㅂ / ㅅ / ㅇ' 와 같은 형태들은 그러한 음가의 실현이 불가능하다는 규칙을 유도하게 되면, 그것은 오류이기 때문이다. 가령, 다음을 보면,

꽃가루 CN, <C : NN> + E, /꼬ㄷ까루/	꽃밭 CN, <C : NN> + E, /꼬ㄷ빠ㄷ/
꽃과 NP, <C : N> + <P : C>, /꼬ㄷ꽈/	꽃밭에 NP, <C : NN> + <P : L>, /꼬ㄷ빠테/
꽃과는 NP, <C : N> + <P : CM>, /꼬ㄷ꽈는/	꽃병 CN, <C : NN> + E, /꼬ㄷ뼝/
꽃나무 CN, <C : NN> + E, /꼰나무/	꽃신 CN, <C : NN> + E, /꼬ㄷ씬/
꽃놀이 CN, <C : NV> + E, /꼰노리/	꽃씨 CN, <C : NN> + E, /꼬ㄷ씨/
꽃눈 CN, <C : NN> + E, /꼰눈/	꽃은 NP, <C : N> + <P : M>, /꼬츤/
꽃눈이 NP, <C : NN> + <P : S,P>, /꼰누니/	꽃이슬 CN, <C : NN> + E, /꼰니슬/
꽃다발 CN, <C : NN> + E, /꼬ㄷ따발/	꽃잎 CN, <C : NN> + E, /꼰닙/
꽃마다 NP, <C : N> + <P : M>, /꼬ㄷ마다/	꽃잎이 NP, <C : NN> + <P : S,P>, /꼰니피/
꽃망울 CN, <C : NN> + E, /꼰망울/	꽃이 NP, <C : N> + <P : S,P>, /꼬치/
꽃말 CN, <C : NN> + E, /꼰말/	꽃처럼 NP, <C : N> + <P : M>, /꼬ㄷ처럼/

〈그림 15〉

콧물 /콘물/
촛물 /촌물/

과 같이, 'ㅅ'이 'ㅁ'앞에 실현될 때, 그 음가는 'ㄴ'으로 나타났다. 실제로 'ㅅ'은 'ㅊ'과 유사한 음운 제약 조건을 따르는 경우가 많다. 가령, '붓고'와 '쫓고'에서 'ㅅ + ㄱ'의 결합과 'ㅊ + ㄱ'의 결합은 두 경우 모두 'ㄷ + ㄲ'의 음가로 실현된다. 그러한 현상이, '꼿'의 경우에 나타나는 'ㅅ'에는 적용되지 않은 것인데, 이것은 실제로, 음절 '꼿'이 '꼿꼿이', '꼿꼿하다'의 두 가지 형태속에서만 실현되기 때문이다. 즉, '꼿'은 'ㄲ', 'ㅇ', 'ㅎ' 외의 다른 문맥에서는 발견되지 않기 때문에, 'ㄲ + ㅗ'를 뒤따른 'ㅅ'의 경우, 위의 오토마타에서 보았듯이 'ㄴ'의 음가로 바뀌는 것으로 분석될 수 없는 것이다. 이와 같은 현상이, 다른 음절속에 실현된 종성 'ㅅ'에 적용될 수는 없으므로, 위와 같은 데이타로부터 'ㄴ / ㅈ / ㅊ'의 3 가지 형태만이 'ㅁ'과 결합할 때 'ㄴ'의 음가를 갖는다고 규칙화할 수는 없다.

어절 단위의 언어 정보를 제공하는 사전에서 나타날 음운 정보는, 하나의 예를 들면, 다음과 같은 형태를 갖는다 <그림 15>.

여기서 볼 수 있듯이, 음운 정보가 '단일 품사', 또는 '합성어'의 수준에 머물 때 제공하지 못하는 음가의 변화 현상들이 빈번하다. 가령, 위에서 '꽃잎'은 '/꼰닙/'의 음가를 갖는 것으로 나타나는데, 이때, 후치사 '이'가 뒤따르게 되면, '/꼰니피/'와 같이 종성 'ㅍ'의 음가가 다시 살아나게 된다. 이와 같은 음가 결정에 어절 단위는 매우 중요한 문맥 형태로 나타난다. 다음에서,

(3a) 꽃잎이 바람에 날린다
(3b) 꽃잎 이십 장이 떨어졌다

'꽃잎'의 뒤에 나타난 '이'의 음가는, (3a)에서와 같이 하나의 어절을 이룰 때는, '피'로 실현되나, (3b)의 경우와 같이 '꽃잎'과 별도의 어절을 형성할 때는, '이'의 음가를 그대로 유지하게 된다. 여기서, 음운 정보의 처리를 위해서도, '어절 단위'의 사전 형식이, 보다 효율적인 결과를 가져오게 한다는 것을 알 수 있다.

2. 형태 정보

다음과 같이 실현된 형태들이,

죽겠거든, 죽어서, 죽었다, 죽으면, 죽으려고

모두, 동사 '죽다'에 관련된 변화형들이라는 정보는, 단일 형태소 품사 사전에 규칙의 형태로 첨부되든지, 합성어 사전에 기본 엔트리로서 등재되든지 간에, 어떠한 형태로든 저장이 되어야 한다. 이와 마찬가지로, 파생어나 복합어등의 경우에도, 주어진 기본 형태에 대한 모든 형태 변화형들이 함께 기술되어야 하는데, 예를 들어, '솟다'에 접두사 '치'가 결합된 파생 동사 '치솟다'의 경우, '솟다'처럼 '치솟아서', '치솟는다', '치솟고' 등과 같은 변

화형들에 대한 정보가 제공되어야 한다.

이러한 정보의 기술 (Description)은 우선, 변화형 유형에 따른 '기본 동사'들의 하위 분류가 이루어진 후에, 같은 변화 유형을 갖는 파생 동사들이 함께 처리되는 방식을 통해야 한다. '기본 동사'들의 모든 '형태 변화형 (Morphological Variation)'들을 찾아내기 위해서는 결합 가능한 굴절 어미들의 목록을 작성하고, 그 다음, 그들 사이의 결합 조건들을 조사해야 한다. 이때, 어미들 사이의 결합 제약을 규칙으로 추정하는 것보다는 '언어학적 조합' 방식에 의하여 결합 가능한 모든 어미 셋트를 구성하는 것이 훨씬 더 효율적이다. 그것은 가령, 이와 같은 어미들 사이의 결합 관계를 규칙으로 표현하려면, 다음과 같이 그들을 지시할 수 있는 메타 언어적 문법 표지들을 이용해야 하는데,

선어말 어미 <NTS> --- 시, 겠, 었, ..

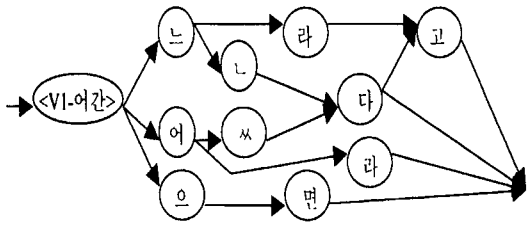
어말 어미 <TS> --- 다, 까, 니,...

이때, 모든 선어말 어미, 어말 어미들은 동일한 결합 제약을 따르지 않는다. 또한 그들 상호간에도 일정 결합 조건이 존재하는데, 이때, 이러한 관계를 규칙으로 표현하려면, 실제 어휘 요소의 수만큼이나 새로운 '메타 언어'들 (또는 '규칙'들)이 요구되는 최악의 상황에 놓일 수도 있다. 더구나, 이들 굴절 어미들의 내적 결합 관계는, 함께 실현되는 용언 어간과는 별도로 구성될 수 있으므로, '부분 문법 (Local Grammar)'의 형태로 기술될 수 있는데, 이와 같이 어미들 사이의 내적 구조들이 유한 오토마타 형식으로 구현되면, 각 동사나 형용사의 기본형들은 이러한 내적 구조들 유형중에서 올바른 셋트와 결합될 수 있다. 가령, 위의 '<NTS>', "<TS>' 등과 같은 메타 언어를 이용한 규칙의 나열을 피하고, 다음과 같은 방법으로 굴절 어미들 사이의 내적 구조들을 표현하는 유한 오토마타를 구축한 후 <그림 16>,

이러한 어미 셋트를 취하는 동사 'V1'에는 어떠한 것들이 있는지 조사하여 그 정보를 제공하는 방법이다. 다음은 위와 같은 어미 셋트를 취하는 몇 가지 동사들의 예이다.

먹다, 섞다, 넘다, 묵다

가령, 어간의 끝음절 종성이 'ㅂ'으로 나타난 동



〈그림 16〉

사나 형용사들을 살펴 보면, 모두 동일한 어미 셋트를 취하는 것이 아니라는 사실을 알 수 있다. 즉, '노엽다'와 같은 형용사의 경우, 변화형 셋트는 소위 전통 문법에서 말하는 'ㅂ불규칙 형태'를 취하는데 ('노여워', '노여우면' 등), '비좁다'와 같은 형용사는 규칙 활용을 한다 ('비좁아', '비좁으면' 등). 이때 'ㅂ'으로 끝나는 용언중 어떠한 것들이 위와 같은 변화 형태형을 취할 것인지는 음운적, 의미적, 그 어떠한 다른 정보로도 예측될 수 없다. 이것은 철저히 '어휘적'인 특성인데, 동사의 경우, 가령 'ㅂ불규칙 변화형'을 취하는 동사들이 전체 23 개의 'ㅂ종성 동사류'중 7 개에 불과하다면, 형용사의 경우에는 정반대로, 'ㅂ불규칙 변화형'을 취하는 경우가 99 %에 이른다. (즉, 519 개의 'ㅂ종성 형용사류' 중 514 개). 다음은, 기본형과의 관련 정보를 갖춘 '형태 변화형 사전'의 한 예를 보인다[Nam96a] <그림 17>.

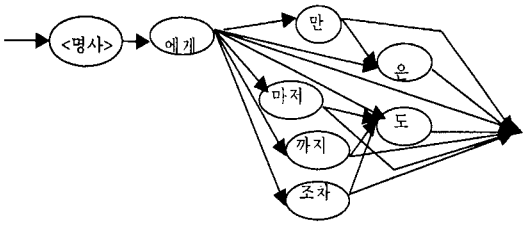
불행해서 /PA5. Conj	불행해서만은 /PA5. Conj
불행해서가 /PA5. Conj	불행해서만이 /PA5. Conj
불행해서나마 /PA5. Conj	불행해서만이라도 /PA5. Conj
불행해서는 /PA5. Conj	불행해서뿐만아니라 /PA5. Conj
불행해서도 /PA5. Conj	불행해서뿐아니라 /PA5. Conj
불행해서라도 /PA5. Conj	불행해서야 /PA5. Conj
불행해서만 /PA5. Conj	불행해야 /PA5. Conj
불행해서만도 PA5. Conj	불행해야만 /PA5. Conj

〈그림 17〉

한국어에 있어서, 하나의 동사 또는 형용사는 최대 무려 6 000 여개에 이르는 복합 어미 셋트와 결합 가능한 것으로 평가된다(Cf. DECO-POST/

V01). 따라서, 이러한 형태들을 기본 엔트리로 하는 사전 유형의 경우, 사전 구조 및 사전 탐색 알고리즘등의 효율성을 최대로 하기 위한 노력이 매우 중요한 열쇠가 된다.

마찬가지로, 명사구 어절 형태로 실현되는 모든 단일 품사 및 합성어들의 경우에 있어서도, 그들이 수반하는 후치사들의 셋트가 어떠한 것들인가에 대한 정보가 체계적으로 기술되어야 하는데, 가령 다음과 같이 <그림 18>,



〈그림 18〉

후치사들의 조합 가능성은 규칙으로 예측되기에는 매우 복잡한 양상을 띤다. 이 경우에도 일종의 부분 문법으로 실현되는 것이 바람직한데, 즉, 후치사들끼리의 내부 결합 제약은, 동반되는 '명사'유형에 간섭받지 않으며, 일단 그와 같은 순서 및 결합 가능성들이 기술되고 나면, 이때 어떠한 복합 후치사 셋트가 어떠한 명사 유형에 결합될 수 있는지가 선택될 수 있다. 위에서 살핀 후치사 결합 셋트는, 예를 들어 반드시 '인물성(Human Feature)'을 갖는 명사들에만 결합하는 것으로, 가령 다음과 같은 명사들과 결합될 수 있다.

친구, 어머니, 학생, 장관

접사가 결합되어 생성된 파생어들의 경우, 가령 '인물성' 표지를 갖는 '원', '인', '자', '가', '사' 등과 같은 접미사가 결합되어 구성된 다음과 같은 파생 명사들은 모두, 이러한 접미사들로 인하여 '인물성'의 표지를 갖게 되어 위와 같은 복합 후치사 셋트와 결합할 수 있게 된다.

간호원, 언론인, 사회자, 음악가, 운전사

복합 명사들의 경우에도 이러한 정보가 표시되면, 그 후치사 셋트에 대한 목록이 구성될 수 있는데, 이때 '소꼽 친구'처럼 그 구조의 '머리(Head)'가 되는 어휘 (여기서는 '친구')가 '인물성'의 표지를 갖고 있으면, 이러한 복합어들은 위의 경우와 동일한 '후치사 셋트'를 취하게 된다.

다음은, 결합 가능한 후치사 형태가 고려되어 '어절'의 형태로 사전 엔트리가 구성된, 어절 사전의 한 예를 보인다 <그림 19>.

친구에게 PN1,/Dat.Hum	친구에게도 PN1, /Dat.Hum
친구에게는 PN1,/Dat.Hum	친구에게까지 PN1, /Dat.Hum
친구에게만 PN1,/Dat.Hum	친구에게까지도 PN1, /Dat.Hum
친구에게만은 PN1,/Dat.Hum	친구에게조차 PN1, /Dat.Hum
친구에게마저 PN1,/Dat.Hum	친구에게조차도 PN1, /Dat.Hum
친구에게마저도 PN1,/Dat.Hum	친구에게서 PN1, /Dat.Hum

〈그림 19〉

하나의 명사에 결합될 수 있는 복합 후치사 셋트는 최대 1 500 여개 정도에 이르는 것으로 분석된다. [ef.DECO－POST/vo1] 위에서 살핀 용언의 경우와 마찬가지로, 이와 같은 어절 형태를 엔트리로 하는 사전의 경우, 그 구조와 검색 엔진의 개발에 많은 주의를 기울여야 한다.

3. 구문 정보

'구문 정보 (Syntactic Information)'는 일반적으로, '형태소' 또는 '단어' 차원을 넘어서서, 일정 언어 형태가 하나의 문장안에 실현될 때, 다른 '문장 성분들 (Constituant)' 과의 통사적인 관계를 나타내는 것으로 이해된다. 따라서, 명사등에 격표지가 주어져 하나의 문장 성분으로서의 기능이 표시되거나, 또는 부사어등이 술어를 수식하는 성분으로 분석되기 위하여 필요한 정보등을 가르킨다. 영어의 경우, '형태 정보'와 '구문 정보'는 그 경계가 비교적 분명하나, 한국어의 경우, 소위 '형태소 분석 (Morphological Analysis)' 단계라 일컫는 컴퓨터 처리 모듈은, 이미 상당 부분의 '구문 정보'를 포함해야 한다. 일정 구문 정보들이 제공되지 않으면, 경우에 따라서는 올바른 형태소 후보들이

제시될 수 없기 때문이다. 가령, 위의 예에서 본 바와 같이, '에게'를 내포한 복합 후치사 형태들이 결합될 수 있는 명사 유형들을 알기 위해서는, 바로 어떠한 명사들에 이러한 후치사류가 허용되는지 '구문상의 제약' 조건이 기술되어야 알 수 있다. 마치, 영어에서 전치사 'in'이나 'on'과 같은 형태들이 '장소 (Location)'를 나타내는 명사류와 함께 나타날 수 있다는 정보가 제공되어야, 이와 같은 '전치사구' 구문에 대한 분석이 가능한 것과 같은 이치이다.

전치사 유형의 분류에 의해 어떤 명사들이 함께 나타날 수 있는지에 대한 구문 정보가 주어지고, 또한, 동사 유형에 의해 어떤 전치사들이 실현될 수 있는지에 대한 정보가 제공될 수 있는 것처럼, 한국어의 경우에도 '서술어(Predicative Term)'의 유형을 보면, 어떠한 어절의 형태(즉, 어떠한 '명사구'의 형태)가 함께 실현되어야 하는지에 대한 구문 정보가 얻어진다. 가령, 영어에서 동사 'put'은, 원칙적으로 3 개의 '필수 논항 (Inherent Argument)'을 가진 문장 구조속에 실현될 수 있는 술어 성분으로서, 이때 하나는 '주어'의 형태로, 하나는 '직접 목적 보어'의 형태로, 그리고 또 하나는 전치사 'on'을 동반한 '간접 목적 보어'의 형태로 실현되는 구문적 특성을 갖는다. 마찬가지로, 한국어에서, '비교하다'와 같은 동사는 원칙적으로 3 개의 필수 논항을 요구하며, 그 세 가지 논항의 구문적 특성은 다음과 같다.

N0-가 N1-를 N2-와 비교하다

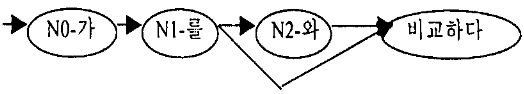
이때, 위의 세 논항은, 술어 '비교하다'의 필수 성분들로서, 후치사들의 형태는 술어 성분에 의해 내재적으로 미리 결정되어 있다. 따라서, 논항 자리에 나타날 수 있는 명사 성분과는 관계없이, 다음과 같은 형태의 후치사 형태들은 허용되지 않는다.

*N0-가 N1-에게 N2-를 비교하다

*N0-가 N1-를 N2-가 비교하다

또한, 위의 세 논항은 술어에 의하여 필수적으로 요구된 형태들이기 때문에, 다음과 같은 논항 삭제는 불가능하게 보인다 <그림 20>.

예를 들면,



〈그림 20〉

대의원단은 그 문제를 이 문제와 비교했다

* 대의원단은 그 문제를 비교했다

에서 보이는 바와 같다. 이러한 구문적 특성은, 위와 동일한 형식 구조속에 실현될 수 있는 '결정하다'의 경우와 대조를 이룬다.

대의원단은 그 문제를 인근 주민들과 결정했다

위의 '결정하다'는 '비교하다'의 경우처럼 ' N0-가 N1-를 N2-와 Verb'의 통사 구조속에 실현되었다. 그러나, 이 경우, 'N2-와'는 술어에 의해 요구된 필수 논항이 아니기 때문에, 다음과 같이 생략이 가능하다.

대의원단은 그 문제를 결정했다

이와 같은 현상에 입각하여, 모든 술어 형태를 중심으로 문장의 구문 정보를 기술할 수 있다. '어휘-문법(Lexicon-Grammar)' [Gro75]은 바로 이러한 정보를 '매트릭스(Matrix)'의 형식으로 저장하는 '어휘 기반 통사 기술 문법'으로, 다음은 한국어 형용사의 한 부류(여기서는 'AWS'로 정의된 '대칭 형용사' 부류)에 대한 매트릭스의 예 [Nam96a]를 보인다 <그림 21>.

N0				N1													Nsym						
N0 =: Nhum	N0 =: N-hum	N0 =: Qu P		N1 =: Nhum	N1 =: N-hum	N1 =: Qu P	N0 N1-wa Adj = N1 N0-wa Adj	Insertion de sôlo	[N0 et N1] Adj	N0 Adj et N1 Adj	(Npl)0 sôlo Adj	(Npl)0 Adj	N1-wa = N1-e	N0 N1-e Adj = N1 N0-e Adj	[N0-ii N1-wa-ii Nsym]0 Adj	[N0-wa N1-ii Nsym]0 Adj	Nsym =: kwankye	Nsym =: sai	Nsym =: kôli	N0 N1-wa Adj-Sd Nsym-ita	N0 N1-wa Adj-Sd Nsym-e is'ta	N0 N1-wa N2(= Nrestructuré)-i Adj	[N0-ii N2][N1-ii N2]-wa Adj
+	+	+	나란하다	+	+	+	+	+	+	-	+	-	+	+	-	-	-	-	-	-	-	+	+
+	-	-	내밀하다	+	-	-	+	+	+	-	+	-	-	-	+	+	+	+	-	+	+	-	-
+	+	+	다르다	+	+	+	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	+	+
+	+	+	다름없다	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	-	-	-	+	+
+	-	-	다정스럽다	+	-	-	+	+	+	-	+	+	-	-	+	+	+	+	-	+	+	+	-
+	-	-	다정하다	+	-	-	+	+	+	-	+	+	-	-	+	+	+	+	-	+	+	+	-
+	+	+	대등하다	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	-	-	-	+	+
+	+	+	대동하다	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+
+	+	+	대립적이다	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+
+	+	+	대조적이다	+	+	+	+	+	+	-	+	+	+	+	-	-	-	-	-	-	-	+	+
+	+	-	대칭적이다	+	+	-	+	+	+	-	+	+	+	+	-	-	-	-	-	-	-	+	+

〈그림 21〉

한국어에서 술어 구조는 다음과 같이 세 가지 유형으로 분류될 수 있는데,

동사 <Verb>

형용사 <Adjective>

서술 명사 + 기능 동사 <Predicative Noun + Light Verb>

각각의 예를 들어 보면, 다음과 같다.

<V> A-가 B-를 돕다

<ADJ> A-가 B-에게 알맞다

<PN-LV> A-가 B-에게 도움을 주다

동사의 경우, 'N-하다'의 형태로 구성되는 6 500 여개를 제외하고 나면, 단순 동사의 수는 7 000 여개(Cf. DECOS-VS/V01 [Nam96b])이며, 단순 형용사의 경우는 5 000 여개에 이른다. 서술 명사에 기능 동사가 결합하여 하나의 술어 성분으로 쓰일 수 있는 경우는, 대표적인 '하다', '되다' 류에 국한시키면, 단순 서술 명사를 기초로 한 경우, 위에서 언급한 바와 같이 6 500 여개의 '하다' 구성과, '2,000' 여개의 '되다' 구성을 관찰할 수 있다. 그러나 실제로, 기능 동사로 분류할 수 있는 통사적 정의가 명시적으로 주어져 있지 않으므로, 고유한 논항 구조를 요구하는 모든 '술어 형태'들을 제시한다면, 그 수는 훨씬 증가하게 될 것이다.

구문 정보가 저장된 사전의 형태는 예를 들면 다음과 같다. 즉, 위에서 제시한 바와 같은 '구문 정보 매트릭스'와 연관된 정보가 개별 사전 엔트리를 중심으로 주어지는데, 가령 다음은 형용사 부류들을 엔트리로 갖는 사전의 일부 형태이다 [Nam96a] <그림 22>.

가공적이다 AQ	가느스레하다 AN
가깝다 AWS / ARR	가느스름하다 AN
가깝디가깝다 AWS / ARR	가늘다 AH
가깝하다 AEP	가늘디가늘다 AH
가난하다 AH	가능하다 AER / AP
가냘프다 AH	가닥가닥하다 AN
가느다랗다 AH	가당찮다 AER

〈그림 22〉

4. 의미 정보

인간 사용자들을 위한 기존의 언어 사전들의 경우, 사전의 가장 핵심적인 내용은 기본 단위들의 '의미 (Meaning)'에 대한 정보이다. 주어진 엔트리에 대한 '의미' 정보는, 구체 명사와 같은 경우, '내재적인 속성(Intrinsic Feature)'들의 나열로 이루어지거나, '외연적인 구체물(Extrinsic Feature)'들의 목록으로 주어지곤 한다. 가령, '사람' 이라는 어휘의 의미는, '눈', '코', '입', '팔', '다리'등과 같은 내재적인 속성들과, '남자', '여자', '어린이' 등과 같은 외연적인 관계들을 통해서 기술될 수 있다. 그러나, 추상 명사나, 명사외의 다른 품사류의 경우에는 이러한 원칙을 적용하기 어렵고, 또한 실제로 이러한 방법을 이용한다 하여도, 컴퓨터에 의한 처리 과정에서 이와 같이 기술된 의미 정보를 어떻게 이용할 수 있을지 구체화하기 어렵다.

의미 정보의 형식화를 위한 여러가지 논리 모델들이 제시되고, 이렇게 기술될 수 있는 의미 정보를 컴퓨터로 구현하기 위해서, 적절한 프로그래밍 언어들이 고안되어 왔음에도 불구하고, 의미 정보의 체계적인 기술은 여전히 저조한 상태에 있다. 의미 기술은 주관적인 특성을 벗어나기 어려워, 객관적 적용이 쉽게 이루어 질 수 없기 때문에, 가령 '시소러스 (Thesaurus)' 나 '단어망 (Word Net)'과 같은 의미 정보 시스템은, 그 필요성의 증가에도 불구하고, 모듈화하여 어느 누구나 그 시스템 구축에 참여할 수 있는 형식적 기준이 쉽게 마련되지 못하고 있는 실정이다. 하나의 예로써, 명사로 사용된 '물'은, 다음과 같이 4 가지의 의미를 갖는 것으로 분석될 수 있는데 <그림 23>,

어휘	예	문
물1	산소1과 수소2의 화합물	그가 물을 마신다
물2	빛깔	옷에 파란 물이 들었다.
물3	영향력	그 아이는 이미 나쁜 물이 들었다.
물4	채소등이 나오는 차례	포도가 이제 끝물이다

〈그림 23〉

그러나, 여기서, 2와 3의 '물'이 과연 두 가지의 의미로 나뉘어져 설명되어야 하는지, 아니면 하나의 원 뜻에 비유적인 정도가 가미되어 3과 같은 의미로 사용된 것인지는, 판단하는 개인 주체에 따라 달라질 수 있다(기존 사전에는 나뉘어져 있지 않다). 또한 '쓰다'와 같은 동사의 경우, 이것의 의미는 다음과 같이, 3 가지에서 10 가지 이상으로 계속 분류될 수 있다. 이때 이와 같은 의미들 사이의 차이를 과연 어떻게 형식화시킬 수 있을 것인가는 무척 의문스럽다 <그림 24>.

어 휘	의 미		예 문
쓰다1	그리다	1	편지를 쓰다
쓰다2	걸치다	2	모자를 쓰다
		3	먼지를 쓰다
		4	누명을 쓰다
쓰다3	사용하다	5	바른말을 쓰다
		6	돈을 쓰다
		7	힘을 쓰다
		8	억지를 쓰다
		9	뇌를 쓰다(산소
		10	말을 쓰다(윳놀이)

〈그림 24〉

위와 같은 근본적인 의문을 제기할 때, 발견하게 되는 중요한 점중의 하나가 바로, 위에서 방금 '쓰다'의 여러 의미를 설명하기 위하여 우리가 사용한 방법에 대한 검토이다. '쓰다'가 여러 의미를 갖는다는 사실을, 우리는 주어진 어휘 요소가 나타날 수 있는 '문맥 (Context)'을 제시함으로써 밝혀내었다. 즉, 동일한 'N-를 쓰다' 구조 속에서도, '쓰다'가 '편지'라는 명사와 실현될 때와 '모자'라는 명사와 실현될 때, 그 의미가 비로소 달라지는 것이다. 바꿔 말하면, '공기 (Collocation)'하는 명사 목록의 제시가, 바로 이와 같은 '쓰다'의 의미를 보여줄 수 있는 유일한 객관적 방법이라 할 수 있다. 여기서 언급한 '공기'의 개념은, 단순히 바로 연접하는 어절들에 대한 개념이 아닌, 구문적으로 '쓰다'의 직접 목적 보어 성분으로 실현될 수 있는 성분들을 가르킨다. 따라서, '편지를 쓰다'에 대응되는, 다음과 같이 다양한 유형의 문장들속에서 찾아질 수 있는 문법적 '보어-술어' 구조들이다.

그는 친구들에게 편지를 썼다

그가 친구들에게 쓴 편지

편지를 매일 하루에 한 장씩 쓴 그가, ..

보어 성분으로 어떠한 명사가 나타났는가 하는 점을 살핌으로써, 바로 이와 같은 의미의 '쓰다'를 정의해줄 수 있는 것이라면, 중요한 점은, 이때의 동사 '쓰다'의 의미가, 몇 가지 대표적 보어 명사들의 목록에 의해서 제시되어서는 안된다는 점이다. 앞서 언급한 바와 같이, 컴퓨터에는 전혀 추론의 능력이 없기 때문에, 모든 사용 가능한 명사들의 목록이 완성되지 않는 한, 몇 개의 보어 예를 통해 제시된 한 동사의 의미 정보는 실제적인 유용성을 갖지 못한다. 가령, '모자를 쓰다'에서 사용된 의미의 '쓰다'는, '바가지를 쓰다'나 '가면을 쓰다'에서 나타날 수 있는 '쓰다'의 또 다른 의미와는 구별되어야 하며, '장갑'이나 '양말'과 같이, 유사한 종류의 의복을 지시하는 개념들임에도 불구하고, 다음과 같이,

*장갑을 쓰다, *양말을 쓰다

동사 '쓰다'와 함께 사용될 수 없는 명사류들도 개별적 검토없이 예측될 수 없기 때문이다.

의미 표현은 때로는, '동의어 (Synonyme)'의 사용에 의해서 이루어지기도 한다. 그러나, 이 경우에도 마찬가지로, 동의어의 설정이 주관적이기 쉬우며, 또한 동의어로 판단할 수 있는 어휘가 쉽게 찾아지지 않는 경우도 있다. 가령, 위에서 '모자'와 '족도리'는 동의 관계인가, 또한 '장갑'의 동의어로는 무엇을 가정할 수 있는가 하는 것과 같은 문제들이다. 더구나, 동사의 경우, '농약을 쓰다'와 '표준말을 쓰다'에서 사용된 '쓰다'는 모두 '사용하다'의 동의어로 설정될 수 있을 듯 하나, 이 두 구조속에서 '쓰다'는 실제로는 각각 다른 동의어군을 형성하고 있다.

농약을(쓰다+사용하다+살포하다+뿌리다+*구사하다)

표준말을(쓰다+사용하다+*살포하다+*뿌리다+구사하다)

즉 이것은, '쓰다'의 동의어 유형은, 반드시 이러한 술어가 사용된 '문맥'속에서, 다시 말해, '통사 구문 (Syntactic Structure)'내에서 설정되어야 함을 보여준다. 문맥을 벗어나 '단어' 또는 '형태소' 차원에서 그 의미 기술이 이루어지는 것은 그러므로 큰 의미가 없다. 의미 정보가, 어휘 형태, 문장 구조 형태, 더 나아가 그러한 문장 성분들 사이에 나타나는 제약 조건들에 대한 기술들과는 별도로, 최종의 단계에서 다루어져야 하는 것이라는 인위적인 설정은, 이런 의미에서 볼때, 분명히 수정되어야 한다.

다음은, 구문 형태와 논항 명사들에 대한 정보를 통해 제시할 수 있는 의미 정보만을 기술한 사전의 한 예를 보인다. 여러 의미의 '쓰다'는 예를 들어 다음과 같은 형태로 저장된다. 이때, N1, N2의 목록은 완성된 형태로 제공되어야 한다.<그림 25>

쓰다1, ADJ/N0-가 쓰다
쓰다2, VT/N0-가 N1-를 쓰다 <N1 = Nhum, N2 = 소설, 편지, 글>
쓰다3, VT/N0-가 N1-를 쓰다 <N1 = Nhum, N2 = 모자, 마스크, 탈>
쓰다4, VT/N0-가 N1-를 쓰다 <N1 = Nhum, N2 = 돈, 표준말, 힘>

〈그림 25〉

의미 정보는, '기계 번역(Machine Translation)' 용 대역어 사전등에서, 비교적 효율적으로 기술될 수 있다. 그러나 이 경우, 대상 언어를 고려하지 않고, 모든 언어들에 적용시킬수 있는, 일반적 '의미 정보 사전'을 구축하려고 하면, 문제는 다시 원점으로 돌아가게 된다. 가령, 한국어에서 동사 '찾다'는, 영어로 'search'와 'find' 모두에 해당하는 어휘 형태이다. 영어와의 번역을 위한 대역어 사전을 구축하기 위해서는, 이러한 '찾다'가, 위와 같은 2 가지 의미로 분류되어야 하나, 한국어처럼 하나의 어휘 형태로 이러한 두 가지 의미가 나타나는 언어와의 번역을 위한 의미 정보 사전에서는, 단일 항목만으로도 사전이 구성될 수 있다. 영어의 동사 'wear'는, 영-한 사전을 위한 의미 정보를 기술할 때, '쓰다', 입다', '신다' 등과 같은 한국어의 여러 어휘 유형과 연결되기 위해, 여러 의미를 갖는 엔트리로 쪼개어져 기술되어야 할 것이나, 'porter'라는 하나의 어휘 형태에 대응되는 불어와 같은 언어로의 번역을 위한 의미 사전에서는, 위와 같은 엔트리 하위 분류가 불필요할 것이다.

IV. 결 론

컴퓨터에 의한 자연어 처리의 효율성은, 결국 얼마만큼 정확한 데이터를, 얼마만큼 체계적으로 시스템내에 저장하였는가에 하는 문제에 궁극적으로 좌우된다. 이것은, 전혀 미지의 한 시퀀스 abcd 가 컴퓨터에 입력되었을 때, 이와 같은 시퀀스를 올바르게 분석해 가는 과정이, 마치 완벽한 '변수 (Variant)'의 형태로 주어진 하나의 입력 형태를 완벽한 '상수 (Coefficient)'의 형태로 치환해 가는 프로세싱으로 비유될 수 있는 것과 같다. 이와 같은 프로세싱은, 주어진 언어 정보 데이터들과의 반복적인 매칭을 통해서 이루어지는데, 이러한 처리 과정에서 많은 중의성 (Ambiguity)이 발생하게 된다. 모든 언어 정보들을 저장하는 '어휘 정보 시스템 (Lexical Information System)'이, 여기서 정의된 '사전 (Machine-Readable Dictionary)'의 개념으로 이해되어야 하는데, 만일 주어진 시퀀스 abcd 가, 그 형태 그대로 사전에 등재되어 있다면, 단 한번의 사전 검색으로 이 시퀀스는 완벽한 상수 형태로 치환되어, 컴퓨터 처리는 종료될 것이다. 입력 시퀀스는, 하나의 단어나 어절의 형태만으로 이루어질 수도 있고, 또는 하나의 문장 (Sentence)으로 구성될 수도 있다. 그러나 현실적으로는, 두 개 이상의 문장으로 구성된, '문서 (Text)' 형태들이 입력 시퀀스로 주어지게 되므로, 이와 같이 문서 자체를 엔트리로 갖는 사전의 구현은 불가능하다. 여기서 분명한 것은, 사전 검색을 위한 기본 단위가 가능한한 최대로 확장되면, '상수' 형태로의 수렴을 위한 과정에서 발생하는 '오분석'의 비율도 줄어들 뿐 아니라, 처리 프로세싱의 효율성도 향상될 것이라는 점이다. 가령, 둘 이상의 분리 기호에 의해 형성되는 스트링 형태들

이 사전에 바로 저장되어 있으면, 입력 스트링에 대한 '분절 (Segmentation)' 과정없이 사전 정보와의 매칭을 바로 시행할 수 있다. 한국어의 경우, 이러한 입력 스트링은, '형태소'나 '단어'보다 더 확대된 단위인 '어절'의 형태로 나타난다.

그러나 반대로, 사전을 구성하는 기본 엔트리 형태가 축소되면 될수록, 사전 구축 작업은 더 용이해 진다. 이때, 사전의 엔트리 차원에서 제시되지 못하는 정보들은 별도의 처리를 통해서 제공된다. 실제로, '어절 중심 사전'을 구현함으로써, 보다 확장된 정보를 사전에 바로 저장하는 방법이 더 바람직한지의 여부는 이론적으로 평가되기 어렵다. 그러나 한국어의 경우, 어휘 형성 제약 조건들의 '복잡성(Complexity)'과 '개별성(Idiosyncrasy)' 등을 고려할 때, 규칙으로 그와 같은 언어 정보들을 저장한다는 것은 현실적으로 구현이 매우 어렵다. 따라서 만일 이와 같은 '어절'들을 기본 엔트리로 하는 사전을 구축하는 경우, 어떠한 형태로 사전을 구축해야 할지 '사전 구조(Dictionary Data Structure)'에 대한 연구가 병행되어야 하며, 빠른 탐색을 위한 알고리즘(Searching Algorithm)등이 개발되어야 한다. 대용량의 데이터를 저장할 수 있는 메모리등의 하드웨어적인 개발도, 이와 같은 선택에 큰 몫을 할 것이다. 여기서 다시 언급하지만, 언어 현상의 모든 정보를 사전 엔트리의 형태로 저장하여서, 그 악세스 시간(Acces Time)이 줄어들고 사전 탐색만으로 분석이 이루어지는, '사전 기반 시스템(Lexicon-Based System)'을 구현할 것이가, 아니면 '규칙(Rule)'의 설정과 알고리즘의 모듈은 매우 복잡하지만, 사전 구축 비용이 적고, 사전 사이즈가 적어지는 '규칙 기반 시스템(Rule-Based System)'을 구현할 것인가 하는 문제는, 사실은 부차적인 문제이다. 이러한 논의들에 선행하는, 보다 중요한 문제는 모든 유형의 언어 정보들을 우선 빠짐없이 기술해 내야 한다는 기본 원칙이다. 이런 관점에서 볼때, '대형 코퍼스(Large-Scale Corpus)'의 사용은 '어휘 정보 시스템'의 구축을 위한 하나의 중요한 '도구 (Tool)'인 것이지, 그것 자체로서, 이와 같이 제공되어야 하는 언어 정보가 자동적으로 주어질 수 있는 것은 아니라는 점을 고려해야 한다. '학습(Learning)'에 의해 '언어 정보의 세공(Refinement)'을 도모하는 '확률 기반 방법론(Statistic Method)'도, 완벽하게 모든 언어 정보를 밝혀내는 것을 궁극적인 목표로 한다면, 이 글에서 다루는 '사전 기반 시스템'과 결국 그 근본에 있어서는 크게 다르지 않을 것이다.

참 고 문 헌

〔1〕 최현배, 우리 말본, 서울 :정음 문화사, 941p, 1929 (1989 재출간)

〔2〕 Gross, Maurice, Methodes en Syntaxe, Paris : Hermann, 414p, 1975.

〔3〕 Nam, Jee-Sun, Dictionnaire des Noms Simples du Coreen, Rapport Technique N-46, Laboratoire d'Automatique Documentaire et Linguistique, University Paris 7, 82p + 140p, 1994.

〔4〕 Nam, Jee-Sun, Classification Syntaxique des Constructions Adjectivales en Coreen, John Benjamins Publishing Company : Amsterdam / Philadelphia, 352p + 186p, 1996a.

〔5〕 Nam, Jee-Sun, Dictionary of Korean Simple Verbs, Rapport Technique N-49, Laboratoire d'Automatique Documentaire et Linguistique, University Paris 7, 71p + 130p, 1996b.

〔6〕 Silberztein, Max, Dictionnaires Electroniques et Analyse Automatique de Textes - The System INTEX, Paris : Masson, 233p, 1993.

〔7〕 Walker, Donald E.;Antonio Zampolli; Nicoletta Calzolari, Automating The Lexicon, Clarendon Press Oxford, 413p, 1995.

저 자 소 개

南 芝 順

1962年 2月 19日生
1994年 6月 프랑스 파리 제7대학 형식언어학 박사
1987年 2月 연세대학교 언어학 석사
1985年 2月 연세대학교 언어학 학사

1997年 7月～현재 KAIST 한글공학 연구실
1994年 10月～1997年 6月 프랑스 마른느-라-발레 대학교 전산과 IGM 연구소
1989年 10月～1994年 6月 프랑스 파리 제7대학 LADL 연구소

주관심 분야 : 전자사전, 구문분석연구