

# 노-한 기계번역 시스템 연구<sup>1)</sup>

- 노어 격어미와 한국어 조사의 대응관계를  
바탕으로 한 형태소 사전 구성 방안 고찰 -

김 성 래 (한국외대)

## 1. 머리말

본 논문은 노-한 기계번역 시스템에서 번역 목표언어인 한국어가 실제 번역과정에서 어떻게 생성될 것인가를 고려하여 대상언어인 러시아어를 분석하는 것을 목표로 삼고 있다. 본 논문의 주된 연구 범위는 굴절어라는 언어적 특성에 의해 다양한 형태로 나타나는 러시아어 체언의 격어미와 한국어 조사의 대응관계를 밝히는 것이다. 현재 개발 중인 노-한 기계번역 시스템은 과학, 기술 텍스트의 번역을 목표로 하고 있으나, 본 논문은 러시아어의 기초분석을 한다는 측면에서 초급 러시아어 수준의 단문 (simple sentence)만을 분석대상으로 한다.

## 2. 격(Case) 체계

### 2.1 격(Case)의 개념

격에 대한 개념은 크게 전통적인 통사기능(syntactic function) 중심의 격체계와 격문법의 의미기능(semantic function) 중심의 격체계로 나눌 수 있다. 먼저 전통구조문법에 따른 러시아어와 한국어에서의 격에 대한 정의는 다음과 같다.

(1) а. Категория падежа — словоизменятельная

1) 본 논문은 필자의 석사학위 논문(한국외대, 1994년 2월)을 제3장을 중심으로 요약한 것이다.

морфологическая категория, выражающая отношение имени существительного к другим словам в словосочетании и предложении. (격범주는 어결합과 문장에서 다른 단어들에 대한 명사의 관계를 표현하는 단어변화의 형태적 범주이다.)<sup>2)</sup>

b. 격이란 구문상 체언류와 다른 어류와의 관계를 표시하는 형태류이며, 그 관계는 본질적으로 전체에 대한 개체의 기능이 성립된다.<sup>3)</sup>

위의 격개념은 표면구조상의 격형태와 통사기능 중심의 격개념이라 할 수 있다. 이 개념에 따라 러시아어에서는 격형태에 따른 6격 체계를 가지며, 한국어에서는 용언에 대한 체언의 통사기능을 범주화하여 주어, 목적어, 보어, 관형어, 부사어, 서술어, 독립어 등의 통사범주를 설정하여 그 성분자리에 실현된 체언과 조사류에 격개념을 부여하였다. 이러한 통사기능에 의한 격을 Nilsen(1973)은 surface case(표층격:이후부터 약어 SCase로 표기)로 표현하였는데 본고에서도 이 용어를 따른다.

다음으로 격문법에 따른 격에 대한 정의는 다음과 같다.

(2) a. ... the term case to identify the underlying syntactic-semantic relationship...<sup>4)</sup>

b. 격은 단문(simple sentence)의 내면구조상에서 서술어(predicator)를 핵으로 한 명사(구)들의 통사·의미론적 관계(syntactosemantic relations)이다.<sup>5)</sup>

위의 정의들은 심층구조상의 의미론적 격개념으로 용언에 대한 체언의 의미기능을 범주화하여 Agent, Patient, Experiencer, Beneficiary ... 등의 의미범주를 설정하여, 러시아어의 경우 의미범주로 실현된 명사류와 전치사에 격개념을 부여하고, 국어의 경우 체언과 조사류에 격개념을 부여할 수 있다. 이러한 의미기능에 따른 격을 심층격(deep case:이후부터 약어 DCase로 표기함)라고 한다.

2) Лопатин, в. в., et al.(1989), Современный русский язык: теоретический курс Москва: Наука, p. 51.

3) 김민수(1975), 「국어문법론」, 서울:일조각 p.160.

4) Fillmore(1968), "The Case for Case", p. 21.

5) 김영희(1974), "한국어 조사류어의 연구", p. 256.

위 두가지 유형의 정의 중 첫번째(1:a-b)를 따를 경우 러시아어의 격형태(Case Form: 이후부터 약어 CF로 표기)로 표현되는 통사성분을 조사로 표현하는데 효율적이고, 두번째(2:a-b)를 따를 경우 러시아어의 전치사로 표현되는 다양한 의미를 한국어 조사로 옮기기에 유용하다. 따라서 본 논문에서는 기계번역에서 효율적인 한국어 조사의 생성을 위해 의미론적 격개념인 심층격을 기본적으로 수용하되, 러시아어의 격형태가 대부분의 러시아어 문장에서 기본적인 통사기능을 수행하므로 표층격도 받아들여 이원적으로 격개념을 정의한다.

## 2.2 표층격과 심층격의 설정

### 2.2.1 표층격의 설정

표층격은 문장성분들의 표면적인 문법관계를 나타내는 것으로, 본고에서는 주어, 목적어, 간접목적어에 주격(SUBJ), 목적격(OBJ), 여격(DAT)의 표층격을 부여한다. 문법기능으로서의 주어, 목적어 등은 서술을 완성하기 위한 보충어들의 관계 내지 기능이며 논리적인 서술-논항(predicate-argument)구조와 상통하는 문법개념으로<sup>6)</sup>, 기계번역의 러시아어 구문분석 단계에서 이러한 문장성분들이 추출되어야한다. 위의 3가지 표층격에 더하여 체언 상호간의 통사기능(NP\_ NP 구조. 예: книга сестры, темнота ночи)에서는 러시아어의 격형태가 생격(gen)인 명사구(NP)를 표층격 생격(GEN)으로 설정한다. 이러한 4가지의 표층격은 심층격에 우선하여 조사를 부여한다. 위 사항을 정리하면 다음과 같다.

- (3)           Он   рубил   дрова   топором.  
           CF: ном           acc       inst  
           SCase: SUBJ                   OBJ

(4)

표층격(SCase)	주로 쓰이는 노어의 CF
주격 (SUBJ)	nom, dat
목적격(OBJ)	acc, gen
여격 (DAT)	dat

6) 장석진(1993), 「정보기반 한국어 문법」, p. 46.



### 2.2.2 심층격의 설정

심층격은 명사구가 서술어나 기타 다른 논항에 대하여 가지는 의미론적 정보를 나타내는 것으로 한국어 조사를 결정하는데 중요한 역할을 한다. 심층격은 Fillmore(1968)가 격문법(Case Grammar)에서 처음 제시한 것으로 그 이론의 개요는 다음과 같다.

#### ① 기저부 규칙

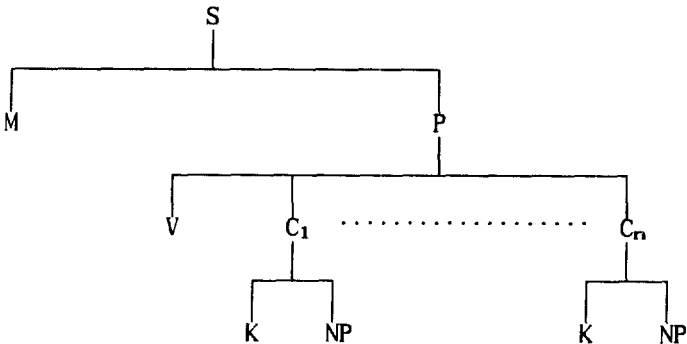
a)  $S \rightarrow M + P$  [M:modality(서법범주: 부정(negation), 시제, 상을 포함) ; P=proposition(명제)]

b)  $P \rightarrow V + C_1 + \dots + C_n$  (명제는 동사와 하나 이상의 격으로 구성된다.)

c)  $C \rightarrow K + NP$  [C:Case, K:Case marker, NP:noun phrase]

이때 명사구의 격관계(case relation)는 가장 기본적 통사현상이며 격체계(case system)는 언어 보편적(universal)이라고 한다. 위의 체계는 다음과 같이 나타낼 수 있다.

(5)



이에 따라 Fillmore(1968)는 Proposition에 6격(Agentive, Instrumental, Dative, Locative, Factitive, Objective)을 설정하였다.

(6) John opened the door with a key.

위 예문에서 John은 Agentive, the door는 Objective, with a key는 Instrumental이 된다.

② 단문의 심층구조상 격관계는 동일한 것이 한번 이상 발생할 수 없다. (one instance per clause principle)

(7) \* The key opened the door with a chisel.

이 예문은 단문 내에서 2회의 도구격이 발생하여 비문이 되었다.

③ 한 동사에는 동일한 격을 지닌 명사구만 접속될 수 있다. (principle of contrast)

(8) \* The janiter and the key opened the door.

이 예문은 open이라는 하나의 동사에 행위격과 도구격이라는 다른 격기능의 명사가 접속되어 비문이 되었다.

④ 격형태(CF)는 언어 특유의 표면 구조 현상으로 각언어마다 격어미, 전치사, 후치사, 어순방식 등의 여러 방법을 이용하여 명사구의 격관계를 보인다.

⑤ 격틀(case frame) - 동사와 함께 나타날 수 있는 서로 다른 종류의 격들이 모여 격환경(case environment)를 만드는데 이것을 격틀이라한다. (1)의 open의 격틀은 [\_\_\_\_\_ O+I+A]이 된다.

Fillmore를 계승한 Nilsen은 심층격을 15개로 정하였고<sup>7)</sup>, 최근의 전산언어학자들은 그 수를 계속 늘려가고 있다. 예를 들면 Nagao 등은 영-일 기계번역 시스템인 Mu에서 영어에 32개, 일본어에 33개의 심층격을 설정하고 있고<sup>8)</sup>, 이기용 교수(1989)는 영-한 기계번역을 위해서 심층격이란 용어 대신 의미역할

7) Nilsen(1973), The Instrumental Case in English, p.121 참조.

8) Margaret King(1989), "Simulation of Language Understanding : Semantic Models and Semantic Interpretation", p.442-443.

(semantic role)이란 용어로 56개를 설정하였다.

심층격을 세분화하는 것은 번역 대상언어나 목표언어 문장의 의미론적인 관련을 정확하게 포착하기 위한 것이나, 심층격 할당에 있어서 여러부분이 겹치고 일관성이 없는 등 문제가 발생할 수 있다. 따라서 본 연구에서는 총 36개의 심층격을 설정하여 한국어 조사생성에 이용한다. (3)의 예문에 심층격을 표시하면 다음과 같다.

(9)		Он	рубил	дрова	топором.
	CF:	nom		acc	inst
	SCase:	<del>SUBJ</del>		<del>OBJ</del>	
	DCase:	AGT		OBJ	<del>INST</del>

표시는 조사가 부여되는 격을 나타냄.

위의 예문에서 표층격은 심층격에 우선하여 조사를 부여하기 때문에 표층격 주격(SUBJ)으로 설정된 он에 주격 조사(은/는, 이/가)가 부여되고, 표층격 목적격(OBJ)으로 설정된 дрова에 목적격 조사(을, 를)가 부여된다. 심층격에서는 이미 표층격에서 격조사가 부여된 он과 дрова는 제외되고 심층격 INST(instrument)으로 설정된 топором에 조사 '로, 으로'가 부여된다.9)

지금까지 논의된 격체계를 나타내는 표지는 다음과 같이 정리한다.

(10) 격형태(CF): 러시아어에서는 격어미(case ending)로 표시되는 6가지의 형태론적 격형태를 nom, gen, dat, acc, inst, prep 로 표기한다. 한국어에서는 체언이 곡용변화(declension)를 하지 않기 때문에 격형태를 인정하지 않는다.

(11) 격표지(case marker: 약어 CM으로 표기) : 러시아어에서 격표지로 격형태와 전치사를 설정한다. 한국어에서는 모든 조사류(후치사)를 격표지로 한다.10)

9) 제 3 장 '3.6 구문분석과 의미해석' 참조.

10) 전치사와 후치사에 관련하여 Fillmore(1968:15)는 다음과 같이 언급하고 있다. "It seems to me the discussion of case could be seen in a somewhat better perspective if the assignment of case forms were viewed as exactly analogous to the rules for assigning preposition in English, or postpositions in Japanese."

### 2.3 격과 어순과의 관계

러시아어의 어순은 theme과 rheme의 구분으로써 의미론적 해석에 영향을 주지만, 격표지와는 아무런 관련이 없다. Harper(1975)에 따르면 노-한 기계번역이 목표로 삼고 있는 러시아의 과학, 기술 텍스트에서는 대부분의 어순이 S+V+O로 영어와 유사하다.<sup>11)</sup> 결국 기본어순만을 고려한다면 러시아어는 머리어 선두(head-initial)언어, 한국어는 머리어 후미(head-final)언어라고 부를 수 있다.

어순에서 한 문장의 구성성분의 배열 뿐만 아니라, 하나의 명사구나 동사구 안에서의 구성성분들이 이루는 순서도 중요하다. 여기서는 격과 관련된 명사구와 전치사구만 살펴본다.

① 전치사 없이 쓰이는 명사구에서는 러시아어 명사 어간(stem)과 격형태인 격어미(case ending)의 어순은 한국어 명사와 격표지인 조사의 어순과 동일하다.

(12)	NP		NP
Stem	Case ending	→	N P
книг	- у		책 을

② 러시아어의 전치사구(prepositional phrase)는 한국어의 후치사구(postpositional phrase)에 대응되며, 그 어순은 거울 이미지(mirror image)<sup>12)</sup>로 반사되는 형태를 갖는다.

11) Kenneth E. Harper(1975), "A Preliminary Study of Russian", pp.67-68. 그에 따르면 주어와 동사를 포함한 1,528개의 예문중 81%가 S+V의 어순이며, V+S의 어순의 경우는 대개 수동태 혹은 재귀동사구문에서 나타난다. 또한 동사와 목적어를 포함하는 1,129개의 예문중 88%가 V+O의 어순이다.

12) Kim, Jong-Hyun, Chung, Hee-Sung (1990), "Adjunction Structure Formalism for English-Korean Machine Translation", p.205 참조.





- (15) Я дочитал книгу до середины.  
 DCase : STO  
 (나는 책을 중간까지 읽었다.)

③ 러시아어의 даже, и, ни, только 등과 같은 소사(частицы)도 한국어의 후치사에 대응될 수 있다. 이러한 소사는 격과 관련이 없고, 그 대역어는 표층격과 심층격에 대응하는 조사에 우선하여 쓰이거나(16), 함께 쓰인다(17).

- (16) Это даже ребёнок понимает.  
 (그것은 어린아이조차도 이해한다.)

- (17) Я посмотрел только статью.  
 (나는 논문만을 읽었다.)

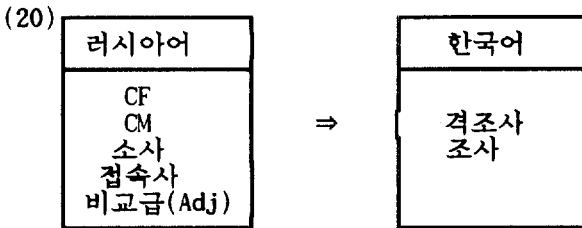
④ 러시아어의 형용사 비교급 표시도 한국어의 조사에 해당한다.

- (18) Брат старше сестры. (혹은 Брат старше, чем сестра.)  
 (형은 누나보다 나이가 많다.)

⑤ и, как, не только...но и .... 같은 러시아어 접속사도 한국어 조사에 대응된다.

- (19) Я купил книгу и альбом.  
 (나는 책과 앨범을 샀다.)

위의 분류를 요약하면 다음과 같다.

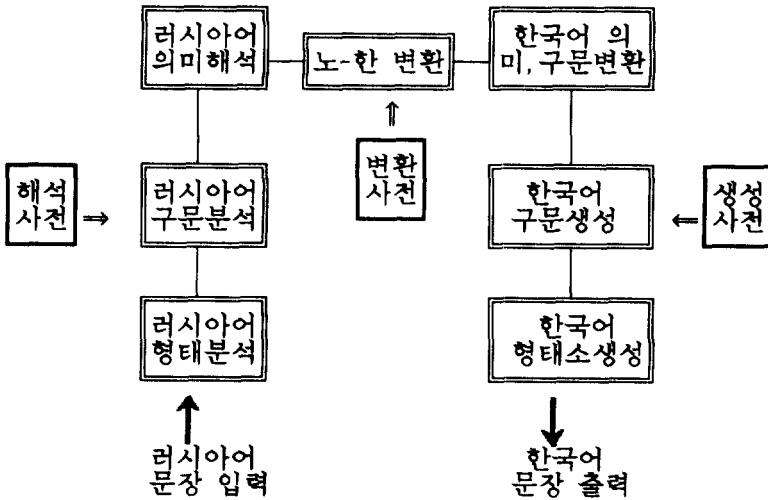


### 3. 형태소 분석과 사전 구성

#### 3.1 사건의 종류와 역할

기계번역 방식은 크게 직접번역(direct translation) 방식, 중간언어(interlingual)방식과 변환(transfer)방식으로 나누며<sup>14)</sup>, 각 방식에 따라 요구되는 사건의 종류와 구성도 달라진다. 본 논문에서는 한국과학기술원에서 노-한 기계번역에 사용하는 변환방식을 따르기로 한다. 변환방식에서의 번역과정은 3단계(대상언어 해석, 변환, 목표언어 생성)로 나누며, 각 단계에서 대상언어 해석사전, 변환사전, 목표언어 생성사전이 사용된다.<sup>15)</sup>

(21) 노-한 기계번역 시스템 구조와 사전<sup>16)</sup>



14) 김재훈, 김철호, 최기선, 김길창(1987), "PIVOT방식의 기계번역에서 한국어 격구조 설정과 중간언어로부터 조사 생성", p. 216.

15) 그외에도 사용자 사전, 전문어 사전, 학습사전 등이 이용된다.

16) 시스템공학연구소(1993), 러한 기계번역 기술 개발에 관한 연구(I)(미발간), p. 8.

러시아어 격체계 분석에 필요한 해석 사전은 분석단계에 따라 형태어 사전, 구문 사전, 의미 사전으로 나눌 수 있다.

### 3.2 형태소 해석과 형태어 사전

가장 흔히 쓰이는 단어의 형태소 처리 방법은 분리 알고리즘 (stemming or stripping algorithm)이다. 분리 알고리즘에서는 단어의 첫부분과 끝부분에서 접사들(affixes)을 반복적으로 떼어 내면서 기본 형태(base forms)를 제시한다. 제시된 기본 형태가 사전에서 발견된다면 그 단어는 분석 가능한 단어가 된다.<sup>17)</sup>

분리 알고리즘에 사용될 수 있는 사전은 단어기반(word-based) 사전과 형태소기반(morpheme-based) 사전이 있다. 사전 구성을 단어기반으로 할 것인가 혹은 형태소기반으로 할 것인가는 단어의 효율적인 탐색과 정보의 이용에 커다란 영향을 주는 중요한 사항이다. 본 논문에서는 최근 크게 확장된 컴퓨터의 기억 용량과 사전구축 작업의 용이성 등을 고려하여 단어기반 사전을 사용하기로 한다.

단어기반 사전에서는 굴절어라는 러시아어의 특성을 고려하여 모든 접두사(prefix)와 접미사(suffix)는 분리시키지 않고 체언의 굴절어미(inflexional ending)와 용언의 활용어미(conjugational ending)만을 분리시키는 것을 원칙으로 한다. 이렇게 분리되어 제시되는 기본형태(base form)는 어간(stem)이 된다.

(22) 단어기반 사전에서의 단어분석의 예  
 입력단어: междугородный

분 석		정 보
междугородн- -ый	형용사 어간 형용사 어미	(기본 형태) 형용사 남성, 주격, 단수, 장어미

이제 단어기반 사전 구축을 위한 품사의 분류와 형태소 해석과 정에서 품사의 인식과정과 품사에 들어있어야 할 정보 등을 격체계 분석에 필요한 명사, 동사, 전치사를 중심으로 알아본다. 우

17) William O'Grady and Michael Dobrovolsky(1989), *Contemporary Linguistics*, pp.420-424 참조.

선 러시아어의 품사는 다음과 같이 9개로 분류한다.

(23) 러시아어 품사의 분류 [ ( )는 약어 표시]

\* 품사: part of speech (POS)

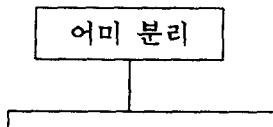
명사(N), 형용사(A), 동사(V), 부사(ADV), 수사(NUM),  
대명사(PRON), 접속사(CONJ), 전치사(PREP), 조사(PART).

### 3.3 명사의 형태소 분석과 사전구성

러시아어의 격체계 분석에서 가장 중요한 부분은 명사의 형태적 특성인 격어미(case ending)를 인식하여 그 명사의 격형태를 알아내는 것이다. 러시아어 명사는 같은 철자의 격어미로 여러가지 격형태를 나타내는 경우가 있기 때문에 명사의 격형태를 인식하는데 어려움이 있다.

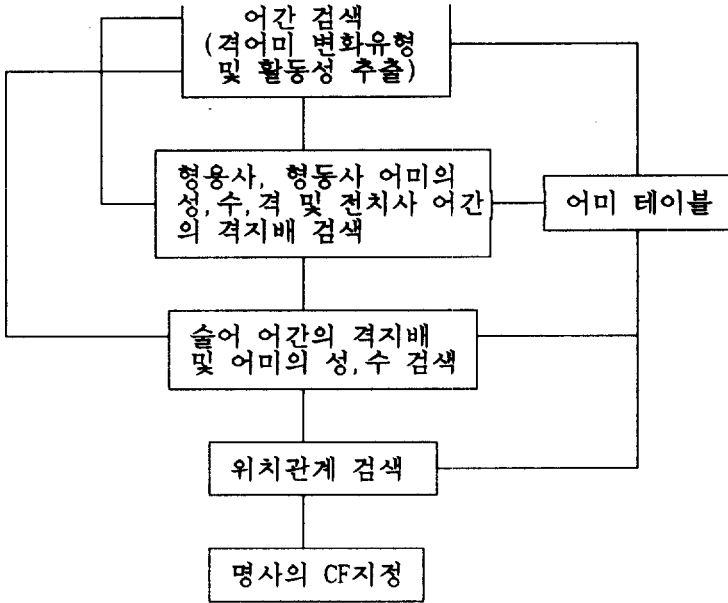
이러한 문제의 해결을 위해 우선 어간에 격어미 변화 유형과 활동성의 유무를 입력시킨다. 이 경우 여격, 조격, 전치격 형태들은 거의 인식되어질 수 있다. 이 과정에서 인식되지 않은 명사는 수식하는 형용사와 형동사 장어미의 격 일치관계와 전치사의 격 지배관계 그리고 술어의 격지배와 술어 어미의 성, 수 등을 어미 테이블에서 검색하여 격형태를 정한다.<sup>18)</sup> 그래도 격형태가 인식되지 않는 명사는 위치(position)관계를 고려한다. 위치관계는 주로 생격형태를 인식하는데 유용하게 쓰인다. Harper (1975: 71-72)에 따르면 명사+명사 유형으로 쓰인 예의 617개 중 두번째 명사가 생격 형태인 경우가 93.4%이고 조격 형태인 경우가 3.9%이다.<sup>19)</sup> 위의 명사 격어미 인식과정은 다음 (24)와 같이 나타난다.

#### (24) 명사의 격형태 인식 과정



18) Kenneth E. Harper(1975), pp. 70-71 참조.

19) 조격 형태는 분리된 격어미만으로도 인식된다.



명사 어미의 인식에 이어서 명사 어간의 인식에 대해서 알아본다. 명사 어간의 분석에서 고려해야할 사항은 출몰모음(mobile vowel)에 의한 교체어간을 어떻게 표기할 것인가이다. 본 논문에서는 Крысин(1992)의 교체어간 표기 규칙을 수정하여 사용한다. 수정된 Крысин의 교체어간 표기 규칙과 적용 예는 다음 (25)와 같다.<sup>20)</sup>

(25) 교체어간 표기 규칙

어간의 교체되는 부분은 X(A)Y, X(A;B)Y, X(A;B;C)Y 등으로 표시한다. X, Y, A, B, C 는 '!' 표시, 콤마, 괄호를 포함하지 않는 분절된 단위이며, 하나 이상의 철자로 이루어질 수 있다. 괄호안의 '!' 표시는 그것으로 구분된 단위 중 하나만을 선택하는 조건을 나타내며, 괄호안에 '!' 표시가 없이 하나의 단위만 있을 경우에 그 단위는 생략될 수 있다. 이 표시들은 다음과 같이 두 개 혹은 그 이상의 어간 세트를 나타낸다.

$$X(A)Y = \{XAY, XY\}; \quad X(A;B)Y = \{XAY, XBY\};$$

20) Крысин, Л.П. (1992) p. 41.

$X(A;B;C)Y = \{XAY, XBY, XCY\}$   
 X와 Y자리는 공란이 될 수 있으나, A, B, C는 공란이 아니다. 공란을 표시할 때는 '#' 기호를 사용한다.  
 $X(A;B;#)Y = \{XAY, XBY, XY\}$

(26) 교체어간 표시 규칙의 적용 예<sup>21)</sup>  
 сосна의 어간표시:  $\text{сос}(e)_H-$   
 незнакомец의 어간표시:  $\text{незнаком}(e)_ц-$

위의 분석으로 인식된 명사가 검색될 때 사전에 수록되어 있어야 하는 정보는 다음과 같이 분류될 수 있다.<sup>22)</sup>

① 명사 어간에 수록되는 정보 [( )는 약어 표시]

- \* 품사 (POS): 명사(N)
- \* 격어미 변화 유형 inflexion type (Infl Type): 사전구축의 편리함을 위하여 '노-한 사전'에 수록된 A에서 S까지의 유형을 그대로 사용함.
- \* 활동성 (Animacy): 활동체(anim), 불활동체(inanim)
- \* 성 (Gender): 남성(masc), 여성(fem), 중성(neut)
- \* 의미표지 semantic mark (Semant Mark)<sup>23)</sup>

② 명사 어미에 수록되는 정보 [( )는 약어 표시]

- \* 수 (Number): 단수(sg), 복수(pl)
- \* 격 형태 (CF): 주격(nom), 생격(gen), 여격(dat), 대격(acc), 조격(inst), 전치격(prep)

지금까지 논의한 형태소 분석 방법과 사전 구성 방법에 따라 명사에 대한 정보를 추출하면 다음의 예와 같다.

(27) 명사 студента의 분석 및 정보추출  
 \* 형태소 분석: студент - a

21) 동사 어간에 적용시킨 예는 (29), (30)을 보시오.

22) 사전에 수록되어야 할 정보의 목록 작성은 '영-한 기계번역'에 사용된 영어 해석 사건의 모형을 주로 참고 하였고, 러시아어에 특징적인 목록들은 Клысин(1992:38-51)과 Harper(1975:68-80)를 참고하여 작성하였다. 이후에 작성되는 동사, 전치사에 대한 정보도 마찬가지이다.

23) 과학기술처(1992a:175-179)에서 영-한 기계번역에 사용한 의미표지를 이용한다. 보다 자세한 분류는 이기용(1980:101-120) 참조.

← (형태소 분리 방향)

\* 사전정보

어간 : студент-

어미 : -а

POS	N
Infl Type	A
Animacy	anim
Gender	masc
Seman Mark	인간

CF	gen
Num	sg

### 3.4 동사의 형태소 분석과 사전구성

동사의 형태소 분석에서 고려해야할 사항은 어간에서의 자음교체의 다양성으로 인한 분석상의 어려움과 형동사와 부동사의 처리이다. 먼저 어간의 인식을 위하여 어미를 분리하여야한다.

어미가 분리된 나머지 부분이 어간이 되며, 그 어간은 사전으로 보내어져서 탐색된다. 사전에서 동사의 어간 형태 구성시에는 상(aspect), 출몰모음(mobile vowel), 자음교체(consonant alternation), 어간모음(theme vowel) 등이 고려되어야 한다.

러시아어 동사에서 상은 형태적으로 접두사(예: при-, про-), 접미사(예: -ива-, -ива-)에 의해 구분되거나 다른 어근을 가진 형태로 나타난다. 본 논문에서 사용되는 단어기반 사전은 동사의 상에 대한 정보를 나타내는 접두사, 접미사 등을 분리하지 않기 때문에 사전 구성에서 완료상 어간과 불완료상 어간을 동시에 하나의 표제어로 수록한다. 예를 들면 читала라는 동사는 어미 -ла가 분리된 어간 чита-가 사전에 보내어지며, 사전의 표제어는 다음과 같다.

(28) читала의 사전구성

чита-	прочита-
impf.	pf.

완료상과 불완료상으로 구분된 어간은 출몰모음, 자음교체, 어

간모음 등이 모두 표시될 수 있는 형태로 수록된다. 표시방법은 명사 어간 표시에 사용된 규칙을 적용한다.<sup>24)</sup>

## (29) писать의 사전구성

пи(са!ш)-	напи(са!ш)-
impf.	pf.

위의 표시 방법으로 나타내기 힘든 불규칙 형태의 동사 어간들은 모두 표제어로 표기하거나, 상과 시제에 따라 분류가 가능한 동사는 분류하여 표기한다.

(30) обойти의 사전구성<sup>25)</sup>

обо(д!ж)-	обойд-	обо(ш!ше!мед)-
impf.	pf.	pf. past

동사는 위의 방법으로 인식되고 사전으로 보내어진다. 사전에 수록되어야 할 동사에 대한 정보는 다음과 같다.

- ① 동사 어간에 수록되어야 할 정보 [( )는 약어 표시]
- \* 품사(POS): 동사(V)
  - \* 상(Aspect): 완료상(pf.), 불완료상(impf.)
  - \* 미정형 Infinitive (Inf)
  - \* 활용 유형 Conjugation Type (Conj Type)
  - \* 타동성 Transitivity (Trans): 타동사(trans), 자동사(intrans)
  - \* 의미범주 Semantic Category (Semant Cat)<sup>26)</sup>
- ② 동사 어미에 수록되어야 할 정보 [( )는 약어 표시]
- \* 시제 (Tense): 현재, 미래(non-past)<sup>27)</sup>, 과거(past)

24) 본 논문의 '(25) 교체어간 표시 규칙'을 보시오.

25) обойти동사의 모든 활용형태는 제시된 세가지 어간에 해당된다. обож-у, обож-ишь... обойд-у, обойд-ешь... обож-ел, обож-ла... обойд-енный... обожед-ший...

26) 과학기술처(1989a), p. 269-270.



- \* 인칭 (Person): 1인칭(1-per), 2인칭(2-per), 3인칭(3-per)
- \* 성 (Gender)<sup>28)</sup>: 남성(masc), 여성(fem), 중성(neut)
- \* 수 (Number): 단수(sg), 복수(pl)
- \* 진술 유형 Representation Type (Repre Type):  
 서술형(decl), 미정형(impf), 형동사(G adj), 부동사(G adv) [ 형동사 장어미는 6격형태에 관한 정보를 가짐]
- \* 태(Voice): 능동태(active), 수동태(passive)

지금까지 논의한 동사의 형태소 분석 방법과 사전구성에 의해서 동사에 관한 정보를 얻으면 다음과 같다.

(31) произносится의 분석 및 정보 추출

\* 형태소 분석: произнос - ит - ся  
 ← (형태소 분리 방향)

\* 사전 정보:

어간 : произно(с'иш)-

어미 : -ит-

POS	V
Aspect	Impf.
Inf	произноситься

Tense	non-past
Person	3-per
Number	sg
Repre Type	decl

접미사 : -ся

Voice	passive
-------	---------

### 3.5 전치사의 형태소 분석과 사전구성

전치사에는 어미가 붙지 않기 때문에 형태소 분석은 명사나 동사에 비해서 훨씬 간단하다. 전치사는 어미 분리 과정을 거치지 않고 바로 단어를 사전에 입력하여 검색한다. 전치사의 어간은 o

27) 현재와 미래시제의 구분은 어미가 아니라 어간에 수록된 상정보에 의한다. 이 방법은 Крысин(1992:40)에 따른 것이다.

28) 성에 관한 정보는 과거시제 어미와 형동사 단어미에 수록된다.

(#i6i6o), κ(o), β(o), c(o) 등으로 표시한다. 사전에 수록되는 정보는 다음과 같다.

- \* 품사(POS): 전치사(PREP)
- \* 격형태 지배 case form government (CFG): 생격(gen), 여격(dat), 대격(acc), 조격(inst), 전치격(pre)
- \* 격형태에 따라 설정될 수 있는 심층격(DCase)

(32) 전치사 β의 사전정보  
어간 : β(o)

POS	PREP	
CFG	acc	prep
DCase	DIR, DUR	LOC, TIM

### 3.6 구문분석과 의미해석

#### 3.6.1 구문분석

구문분석을 위해서는 구문유형에 관한 정보가 구문 사전의 술어로 쓰이는 품사 항목에 수록되어야 한다. 이러한 러시아어 구문에 관한 정보들은 한국어와의 대조분석을 통하여 얻어져야 하고, 한국어로의 구문변환도 고려되어 체계적으로 수록되어야 한다. 동사를 예로 들면 그 동사가 지배할 수 있는 논항(argument)에 대응할 수 있는 소성들 즉 표층격, 심층격, 구문형태(syntactic form)<sup>29)</sup>, 명사의 의미표지 등이 격틀(case frame)이나 하위범주화(subcategorization)의 형태로 구문 사전에 수록되어야 한다.

본 논문의 구문분석에서는 위에서 예를 든 동사의 구문 사전에 수록된 항목 중에서 표층격 정보만을 활용하여 격조사 생성을 시도한다. 구문 사전에 수록되어야 할 표층격에 대한 정보는 술어가 지배할 수 있는 표층격과 그 표층격의 격형태로 구성된다. 예를 들어 동사 решить의 구문 정보는 다음과 같다.

(33) 동사 решить의 구문 정보

SCase	SUBJ	OBJ
-------	------	-----

29) 논항을 구성하는 구문형태는 명사, 대명사, 동사의 미정형, 전치사구, 접속사로 이끄러는 절 ... 등으로 분류될 수 있다. 과학기술처(1989a: 113-114) 참고.

CF	nom	acc, *gen
----	-----	-----------

\*gen은 부정생격을 나타냄.

구문분석 단계에서는 위에서 예로 보인 구문 정보를 이용하여 전치사 없이 쓰인 명사구에 표층격을 설정할 수 있다. 즉 형태소 분석으로 밝혀진 명사의 격형태를 그 문장에서 사용된 술어의 구문정보에 보내면 바로 표층격이 설정된다.

설정된 표층격은 노-한 변환 단계에서 한국어 격조사 생성 규칙에 의해 조사로 옮겨진다. 변환 사전에 수록되어야 할 한국어 격조사 생성 규칙은 다음과 같다.

(34) 한국어 격조사 생성 규칙

SCase	→	한국어 조사	SCase	→	한국어 조사
SUBJ		은/는, 이/가	DAT		에게
SCase	→	한국어 조사	SCase	→	한국어 조사
OBJ		을, 를	GEN		의

위의 한국어 격조사 생성규칙을 решить가 사용된 간단한 예문에 적용시키면 다음 (35)의 ㉔와 같이 격조사가 생성된다.

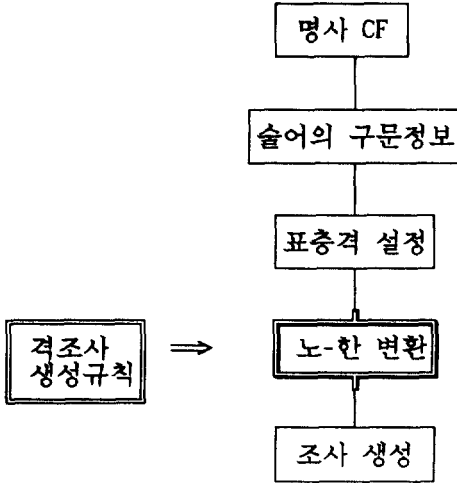
(35) 예문분석:

	Я	решил	эту задачу .
㉔ CF :	nom		acc
㉕ 구문정보:	nom→SUBJ		acc→OBJ
㉖ SCase :	SUBJ		OBJ
㉗ 격조사 :	은/는, 이/가		을/를

지금까지 논의한 구문분석 단계에서의 표층격 설정과 격조사 생

성 과정은 다음과 같이 나타낼 수 있다.

(36) 표층격 설정과 격조사 생성 과정



3.6.2 의미해석

의미해석 단계는 러시아어 분석단계에서 마지막 단계로 한국어 조사 생성을 위한 러시아어의 격체계 분석을 완결하고 그에 대한 정보들을 노-한 변환단계로 보내야한다. 격체계 분석은 심층격이 설정됨으로써 완결된다. 여기서는 심층격이 설정되는 과정을 설명하고 그 과정을 규칙화하며, 설정된 심층격에서 조사가 생성되는 과정을 보이고 그 과정을 단축시킬 수 있는 방법을 제시한다.

심층격은 원칙적으로 모든 논항(argument)에 설정되어야 한다. 따라서 사전에는 동사마다 함께 나타날 수 있는 심층격들에 대한 하위범주화 정보가 수록되어야한다. 동사 встретить를 예를 들면 다음과 같다.

(37) a. 동사 встретить의 심층격 정보

DCase	AGT	OBJ	TIM	LOC	PUR
-------	-----	-----	-----	-----	-----

b. 예문:

Я	встретил	его	в понедельник	.
---	----------	-----	---------------	---

DCase: AGT OBJ TIM

위의 예와 같이 정확한 심층격 설정을 위해서는 심층격에 동사의 하위범주화 작업이 요구되나 본 논문의 심층격 설정 과정에서는 이 작업을 생략한다. 또한 본 논문은 한국어 조사 생성을 목적으로하기 때문에 구문분석 단계에서 이미 표층격으로 설정되어 한국어 격조사가 생성된 논항의 심층격 설정에 관한 논의는 생략하고, 표층격으로 설정되지 않은 논항에 대한 심층격 설정 과정을 논의하기로 한다. 위 (37:b) 예문에 구문분석 단계에서 이미 설정된 표층격을 표시하면 다음과 같다.

(38) 예문:

	Я	встретил	его	в понедельник .
SCase:	SUBJ		OBJ	
DCase:	AGT		OBJ	TIM

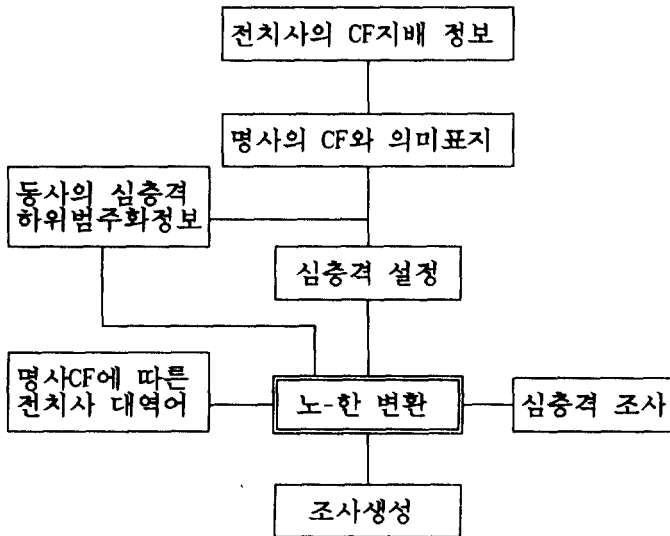
위 예문에서 я와 его는 구문분석 단계에서 각각 표층격 주격(SUBJ)과 목적격(OBJ)으로 설정되어 격조사가 결정되었기 때문에 의미해석 단계에서는 в понедельник에 어떻게 심층격을 설정하고 조사를 생성시킬 것인가를 분석해야 할 것이다.

위에서 언급한 바와 같이 의미해석 단계에서는 구문분석 단계에서 표층격으로 설정되지 않은 논항인 조격 형태의 명사구와 전치사구에 심층격을 설정한다. 심층격 설정에는 형태소 분석에서 얻어진 명사, 동사, 전치사에 대한 형태 정보를 이용한다. 가장 단순한 심층격 설정 과정은 전치사가 사용되지 않은 명사구에서 명사의 격형태와 의미표지 그리고 동사의 정보를 이용하므로써 이루어진다. 조격 형태로 쓰인 명사구가 이에 해당되나 본 논문에서는 전치사구만을 다룬다. 전치사구의 심층격 설정은 전치사의 격형태 지배 정보와 명사의 의미표지를 이용하므로써 이루어진다. 이 과정에서 설정되지 않을 경우에는 동사의 하위범주화 정보를 이용해야 하지만, 본 논문에서는 심층격에 대한 하위범주화 작업을 생략했기 때문에 이 정보의 이용은 생략한다.

설정된 심층격은 노-한 변환 단계로 보내어지며, 변환 단계에서는 전치사의 한국어 대역어와 심층격에 따른 한국어 조사를 검색하여 일치하는 것을 조사로 결정하되, 이 과정에서 동사의 정보중

의미범주를 참고로 한다. 동사의 정보를 참고하는 이유는 보다 정확한 조사의 선택을 위한 것이며, 이 부분은 한국어 생성 단계에서 할 수도 있다. 예를 들면 심층격 LOC(location)에 해당하는 한국어 조사 중 '에'와 '에서'의 구분은 상태동사일 때는 '에', 행위동사일 때는 '에서'가 선택되므로써 가능한 것이다.<sup>30)</sup> 위 과정은 다음과 같이 나타낼 수 있다.

(39) 의미해석 단계의 심층격 설정과 변환단계의 조사 생성과정



위 과정에 전치사 B가 쓰인 전치사구를 적용시켜보자. 전치사 B는 대격, 전치격 형태의 명사를 지배하며, 각 격형태에 따라 대격 형태에는 심층격 DIR(direction)과 DUR(duration)이, 전치격 형태에는 LOC(location)와 TIM(time)이 설정될 수 있다.<sup>31)</sup> 이 정보들은 형태어 사전에 수록되어있다.<sup>32)</sup> 심층격의 결정은 명사의 의미표지에 따른다. 즉 대격 형태로 쓰인 명사의 의미표지가 장소,

30) 예문:

철수는 공원에 있다.  
철수는 공원~~에서~~ 산책했다.

31) 대표적으로 설정될 수 있는 네개의 심층격만 살펴본다.

32) (32) '전치사 B의 사전정보' 참고.

위치 등을 나타내면 DIR(direction)로, 시간, 기간 등을 나타내면 DUR(duration)로 심층격이 결정되고, 전치격 형태로 쓰인 명사의 의미표지가 구체물, 장소 등을 나타내면 LOC(location)로, 시간 등을 나타내면 TIM(time)으로 심층격이 설정된다. 이러한 심층격 설정 규칙은 의미 사전에 다음과 같이 수록되어 있어야 한다.

(40) 'B+명사구(NP)'의 심층격 설정 규칙:

PREP	CF	Seman Mark	→	DCase
B	prep	장소		LOC
PREP	CF	Seman Mark	→	DCase
B	prep	시간		TIM
PREP	CF	Seman Mark	→	DCase
B	acc	장소		DIR
PREP	CF	Seman Mark	→	DCase
B	acc	시간		DUR

노-한 변환단계에서는 B의 한국어 대역어인 '에, 에서, 로, 동안' 등을 설정된 심층격 조사에서 검색하고 동사 정보를 이용하여 최종적으로 조사를 결정한다.<sup>33)</sup> 즉 심층격이 TIM으로 설정되었으면, B의 대역어 '에, 에서, 로, 동안'을 TIM에 대응하는 조사 '에, 전에, 후에, 무렵에...'에서 검색하여 일치하는 '에'를 조사로 결정한다.

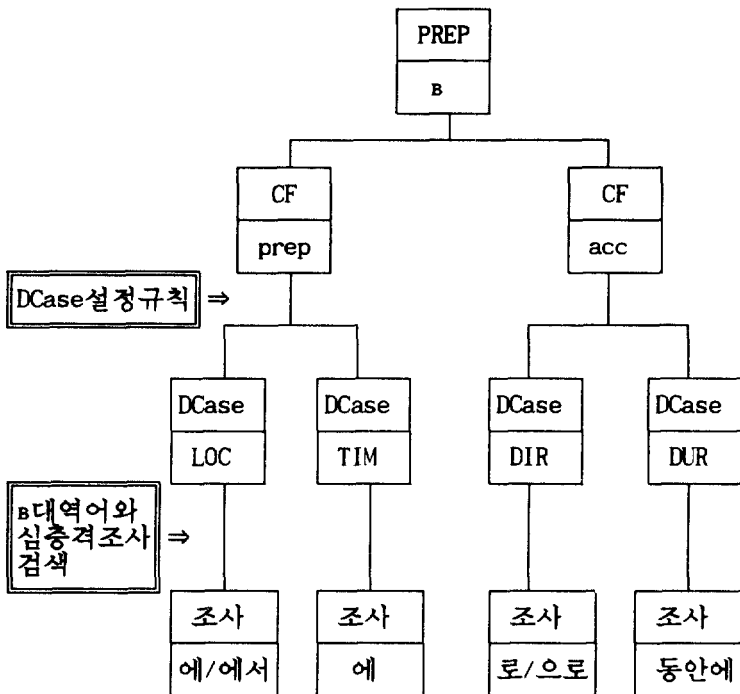
33) 'B+명사구(NP)'에서 설정될 수 있는 심층격들에 대응되는 한국어 조사들은 다음과 같다.

- DIR: 로, 으로, 쪽으로...
- DUR: 동안, 동안에...
- LOC: 에, 에서, 앞에, 뒤에, 옆에...
- TIM: 에, 전에, 후에, 무렵에...

이러한 심층격들은 다른 전치사에 의해서도 설정될 수 있으며, 그 경우에는 해당 전치사의 역할을 위의 심층격 조사에서 검색하여 일치하는 것을 조사로 결정한다.

다. 이 과정을 나머지 심층격에 모두 적용시키면 러시아어 전치사 B에 대응하는 한국어 조사는 다음과 같이 나타난다.

(41) 전치사 B에 대응하는 한국어 조사 생성

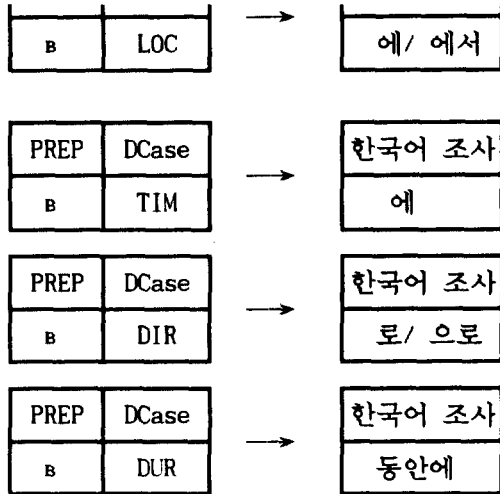


지금까지 논의한 과정에 의해서도 한국어 조사의 생성이 가능하지만, 노-한 변환 단계에서의 전치사 대역어와 심층격 조사의 검색 과정은 변환 사전에 적절한 규칙을 수록하므로써 보다 단순하고 명확해 질 수 있다. 즉 변환 사전에 전치사에 따라 설정된 심층격과 그에 대응하는 한국어 조사 생성을 규칙화하여 입력하면, 지금까지 논의한 과정보다 훨씬 간단히 조사가 생성될 수 있다. 위에서 예를 든 전치사 B에 대응하는 조사 생성 규칙은 다음과 같이 만들 수 있다.

(42) 변환 사전에 수록될 전치사 B에 대한 한국어 조사 생성 규칙







위에서 만든 심층격 설정 규칙과 한국어 조사 생성 규칙을 실제로 적용시키면 다음과 같다.

(43) 예문 분석

Мы гуляли в парке.

\* 사전정보<sup>34)</sup> :

어간 : в(о)

POS	PREP	
CFG	acc	prep
DCase	DIR, DUR	LOC, TIM

어간 : парк-

어미 : -е

POS	N
Seman Mark	장소

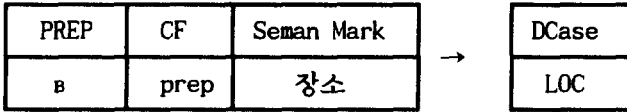
CF	prep
----	------

어간 : гуля-

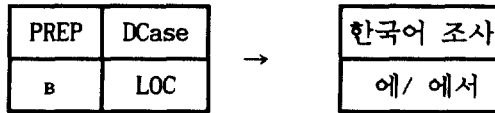
POS	V
Seman Cat	동작

34) 심층격 설정에 필요한 정보만 보임.

\* 심층격 설정 규칙:



\* 한국어 조사 생성 규칙:

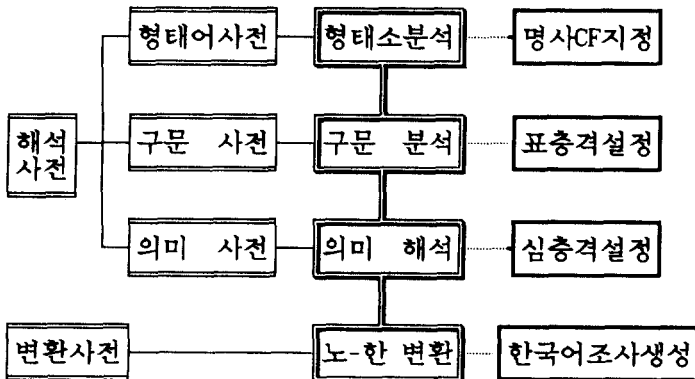


지금까지 제3장에서 논의한 내용은 러시아어 형태소 분석 방안과 사전구성 방법 그리고 표층격과 심층격 설정에 의한 한국어 조사의 생성 등이었다.

#### 4. 맺음말

본 논문은 변환(transfer)방식을 취하고 있는 노-한 기계번역 시스템의 러시아어 해석 단계에 따라 격체계를 분석하여 한국어 조사생성에 이용하였다. 즉 형태소분석에서 명사의 격형태를 지정하였고, 구문분석에서는 표층격을, 의미해석에서는 심층격을 설정하여, 최종적으로 노-한 변환단계에서 한국어 격조사를 생성시켰다. 이 과정을 전체적으로 나타내면 다음과 같다.

(44) 러시아어 격체계 분석을 통한 한국어 조사 생성 과정



본 논문에서 연구한 분석 방법은 격어미(case ending)를 갖는 러시아어 체언의 형태론적 특성을 격체계 분석에 이용하고, 어휘부(lexicon)에 최대한 많은 정보를 입력하여 러시아어 해석에 이용하므로써 통사부의 부담을 최소화 하였다. 위의 연구를 실용화 하기 위해서는 보다 많은 분량의 자료 분석이 앞으로 필요할 것이다. 특히 구문분석과 의미해석에서 정확하게 표층격과 심층격을 설정하기 위해서는 동사에 대한 하위범주화(subcategorization) 작업과 의미범주의 세밀한 분류 그리고 명사에 대한 의미표지의 세분화가 요구된다.

앞으로 노-한 기계번역을 위해서는 본 논문에서 다루지 않은 한국어와의 대조분석을 통한 러시아어 구문유형의 정리, 의존구조의 체계화 등에 관한 연구가 필요할 것이다.

### 주요 참고 문헌

- 과학기술처(1987), 「자동번역 시스템 개발기술에 관한 연구 (I)」, (1989a), 「한영·영한 기계번역시스템 개발(I)」, (1989b), 「자동번역 시스템 개발기술에 관한 연구(II)」, (1992a), 「기계번역을 위한 언어모델링 및 골격시스템 개발 (3)」, (1992b), 「영한기계번역시스템(III)」: 문법개발지원환경 및 해석문법개발.
- 민현식(1982), 「현대국어의 격에 대한 연구」, 국어연구 제49호, 서울: 국어연구회.
- 성광수(1979), 「국어조사의 연구」, 서울: 형설출판사.
- 시스템공학연구소(1993), 「러한 기계번역 기술개발에 관한 연구」, 시스템공학연구소 (미발간).
- 이기용, 박병수, 임홍빈(1989), 「영한 기계번역을 위한 어휘부 구동의 문법모형의 구축과 그 적용」, 시스템공학연구소.
- АН СССР(1982), Русская Грамматика, Москва: Наука.
- Лопатин, В. в., et al.(1989), Современный русский язык теоретический курс, Москва: Наука.
- Крысин, Л. П.(1992), Лингвистический Процессор для сложных информационных систем, Москва: Наука.
- Цыганенко Г. П.(1982), Словарь служебных морфем русского языка, Киев: <<Радянська школа>>.
- Colin Beardon et al.(1991), Natural Language and

- Computational Linguistics: An Introduction, West Sussex: Ellis Horwood Limited.
- Fillmore, C.J.(1968), "The Case for Case" , Universals in Linguistic TheoryInd, E. Bach and T. Harms, eds., New York:Holt, Rinehart and Winston Inc.
- Kim, Jong-Hyun, Chung, Hee-Sung (1990), "Adjunction Structure Formalism for English-Korean Machine Translation", Proceedings SICONLP '90, Seoul:Seoul National University.
- Kenneth E. Harper(1975), "A Preliminary Study of Russian", Machine Traslation of Languages, Westport:Greenwood Press.
- Margaret King(1989), "Simulation of Language Understanding : Semantic Models and Semantic Interpretation", Computational Linguistics Berlin : Walter de Gruyter.
- Nilsen, D.L.F.(1972), Toward a Semantic Specification of Deep Case, The Hague : Mouton.
- Somers,H.L.(1987), Valency and case in computational linguistics, Edinburgh:Edinburgh University Press
- William O'Grady and Michael Dobrovolsky(1989), Contemporary Linguistics, New York:St Martin's Press. pp.413-447.