(제 10회 한글 및 한국어 정보처리 학술대회)

일한 문형사전을 위한 구문연구

송영빈 최기선
한국과학기술원 전산학과 인공지능연구센터
대전시 유성구 구성동 373-1
우:305-701
ybsong@world.kaist.ac.kr kschoi@world.kaist.ac.kr

Study of Japanese Korean Syntactic Dictionary Construction

Young-Bin SONG Key-Sun CHOI
CAIR, Department of Computer Science,
KAIST

요약

이 논문은 한국어와 일본어의 대역문형사전 구축 시에 동사의 대역어 선정을 위한 구체적인 방법을 실증적으로 제시하는데 목적이 있다. 현재 자연언어 처리에서의 동사의 의미기술은 동사의 중의성(重義性)을 해소하려는데 초점이 맞추어져 있다. 그 주된 방법론은 동사와 결합하는 명사 의미속성의 기술에 의해 중의성을 해소하려는 것이다. 이 논문에서는 개별언어에 있어서의 명사의 의미속성의 분류가 다국어를 대상으로 하는 경우 어떻게 다루어져야 하는가에 대해 언어학적인 조명을 하는데 목적이 있다. 아울러 현재까지 일본에서 구축된 동사의 의미사전 들을 비교, 명사 의미속성 분류의 기준이 어떻게 설정되어야 하는가에 대해 검증한다. 특히 외국어와의 대조는 동사와 명사의 공기관계가 각 언어마다 독특한 결합관계를 갖으며 문법체계의 차이로 인해 개별언어를 대상으로 했을 때보다 의미기술의 양상이 상당히 달라짐을 보여줄 것이다[1].

1. 중의성이란

1.1 중의성의 단계

단어가 갖는 중의성의 규명은 인간의 언어지식을 망라하는 중요한 문제로 기계번역을 시작으로 정보검색 등의 분야에서 중요한 의미를 갖으며 거의 모든 언어에서 공통적으로 관찰되는 현상이다.

중의성을 유발하는 단계는 여러 가지가 있을 수 있는데 Cruse(1986)는 언어의 중의성에 대해 다음 네 가지로 분류하고 있다[2].

1. 순수히 통사적인 중의성(pure syntactic ambiguity)
   ex) old men and women→(old men) and women
   →old (men and women)
2. 준 통사적인 중의성(quasi-syntactic ambiguity)
   ex) a red pencil →a pencil painted red
   →a pencil which writes red
3. 어휘 통사적 중의성(lexico-syntactic ambiguity)
   ex) I saw the door open→문이 열리는 것을 보다
   →열린 문을 보다
4. 순수히 어휘적인 중의성(pure lexical ambiguity)
   ex) He reached the bank→은행
   →제방

본고에서 대상으로 하는 것은 3번의 어휘 통사적 중의성이다. 특히 대상언어의 한 동사가 여러 개의 의미 혹은 단일한 의미를 갖더라도 목표언어에서의 대역어는 자동사 혹은 타동사와 같은 상이한 문법적 카테고리를 갖을 수 있고 명사와 동사와의 공기관계가 대상언어의 그것과 다름으로 인해 여러 개의 다른 대역어로 표현될 수 있

1) 본 연구는 정보통신부 [지능형 멀티미디어 통합 정보베이스]와 과학기술부 [대용량 국어정보 심층처리 및 품질관리 기술개발]의 지원을 받아 수행된 연구의 일환으로 이루어졌다.

2) Cruse D.A. Lexical Semantics : Cambridge University Press

다. 이와 같은 문제를 해결하기 위한 전제로는 다음 두 가지를 들 수 있다.

(1)대량의 코퍼스에서 명사와 동사의 공기관계를 추출할 것.
(2)명사 의미속성 기술은 양 언어의 의미 대응 관계가 충실히 반영될 수 있도록 정도(精度)를 설정할 것.

1.2 중의성 해소를 위한 기존의 연구들

동사의 중의성 해소에 관한 연구는 명사의 의미기술과 밀접히 연관되어 있다. 따라서 중의성 해소를 위한 연구는 명사의 의미기술에 관해 활발히 전개되어왔다. 명사의 의미기술에 관한 현재까지의 연구는 대략 다음과 같이 세가지로 나눌 수 있다[3].

(1) 의미속성에 의한 의미구별
(2) 개념식별자에 의한 의미구별
(3) 유의어의 집합을 이용한 의미구별

의미속성(*semantic feature*)에 의한 명사의 의미구별은 단어의 변별적인 의미속성을 이용해서 단어의 의미를 구분하는 방법이다. 단어의 변별적인 의미속성, 예를 들어 [목]의 경우 [피부]나 [내부기관]과 같은 변별적인 속성에 착안해서 이를 명사에 부여함으로써 중의성을 갖는 동사 [타다]의 의미를 구별하는 것이다.

[목(피부)]이+타다→首が日に焼ける(그을리다)
[목(내부기관)]이+타다→喉が乾く(마르다)

위의 경우는 공기(共起)하는 명사의 의미속성이 다름으로 인해 동사의 의미가 달라진 예이며 일본어의 동사 대역어도 각각 [焼ける]와 [乾く]가 된다. 이 경우 일본어는 의미속성 [피부]를 변별적 특성으로 갖는 명사는 [首]가 되고 [내부기관]을 변별적 특성으로 갖는 명사는 [喉]가 된다. 이와 같이 의미의 변별적 특성을 이용해서 동사의 의미를 구별한 대표적 사전으로는 일본의 [정보처리진흥사업협회기술센터]에서 개발한 [계산기용 일본어 기본동사 사서 IPAL(1987)](이하 이를 [IPAL사전]이라 부른다)과 [NTT커뮤니케이션 과학연구소]에 의해 작성된 [일본어어휘대계](1997)(이하 [NTT사전]이라 부른다)가 있다.

개념식별자(concept identifier)에 의한 명사의 의미분류는 앞서 살펴 본 의미속성에 의한 방법과 비슷하나 개개의 단어의 의미에 대해 이를 구별할 수 있는 개념식별자를 부여함으로써 단어의 의미를 구별하는 방법이다. 예를 들어 [목]에 대해서 다음과 같은 개념식별자를 부여한다.

2bc262 신체의 한 부분
2cf8b1 신체의 순환기계의 한 기관

이와 같은 방법을 써서 의미를 구별하는 사전으로는 [일본전자화사서연구소]에서 나온 [EDR전자화사서](1993)(이하 이를 [EDR사전]이라 부른다)가 있다.

개념식별자를 쓰는 경우 각각의 단어가 갖는 개별적인 의미에 대해 개념식별자를 일일이 부여함으로써 개념의 숫자가 대단히 늘어나서 의미기술의 효율이 떨어지게 된다. 실제로 IPAL동사사전의 경우 58개의 의미속성을 써서 명사를 분류하고 있는데 비해 EDR사전에서는 약 40만개의 개념을 정의하고 있다[4].

유의어 집합에 의해 단어의 의미를 구별하는 방법은 일반적으로 thesaurus라고 불리는 사전에서 취하는 방법으로 동일한 혹은 유사한 의미속성을 공유하는 유의어들을 엮어 놓은 것이라고 할 수 있다.

이의 대표적 사전으로 Roget의 thesaurus(1992)가 있으며 31개의 대분류, 1073의 소분류로 25만개의 단어를 분류하고 있다. 일본어에서는 [「분류어휘표」 형식에 의한 어휘분류표](1996)가 대표적인 것으로 5개의 대분류, 798개의 소분류로 약 8만7천개의 단어를 분류하고 있다.

thesaurus에서의 단어의 의미구별은 의미속성이나 개념식별자를 쓰는 경우에 비해 일반적으로 의미구분이 섬세하지 못한 특징을 갖고 있다. 이는 thesaurus가 모국어 화자를 대상으로 작문 등과 같은 실용적인 목적에서 사용되는 것을 전제로 만들어지는 경우가 많기 때문에 동일한 노드에 속하는 단어들 사이의 섬세한 의미구별이 필요하지 않다는 것이 크게 작용하고 있다. 따라서

3) 松本裕治, 影山太郎, 齊藤洋典, 德永健伸, 單語と辭書, 岩波書店, PP. 172-175, 1997

4) 日本電子化辭書研究所, EDR電子化辭書仕樣說明書, EDR, 1993

섬세한 명사의 의미구별에 의해 대역어를 선정해야 하는 대역구문사전의 구축이라는 관점에서 보면 일반적인 thesaurus에 의한 동사의 중의성 해결은 그다지 효과가 없다고 할 수 있다. 실제로 일영번역을 전제로 작성된 NTT사전의 경우 약 3,000개의 의미속성을 써서 40만개의 단어의 의미를 구별하고 있는 것과 비교를 하면 기존의 thesaurus가 외국어와의 번역이나 모국어의 자연언어처리를 위한 의미분석을 위한 사전으로는 의미속성의 수가 충분치 못하다는 것을 알 수 있다.

1.3 대표적 사전

동사의 중의성 해소를 위한 가장 효과적인 방법은 통사구조와 의미구조를 동시에 처리하는 방법이다. 이와 같은 관점에서 편찬된 대표적 사전으로는 1.2에서 소개한 일본의 IPAL사전, EDR사전, NTT사전, 미국의 뉴욕대학에서 개발된 COMLEX(Grishman.et.al. 1994)등이 있다. 이들 중에서 현재 가장 상세한 통사적 정보와 의미정보를 담고 있는 사전은 NTT사전이다.

한국어와 일본어는 한자어를 공유하는 경우가 많고 비교적 문법체계가 비슷한 경우가 많다는 점에서 일본어의 의미해석용 사전은 한국어의 의미해석용 사전의 구축에도 많은 참고가 될 수 있다.

2. NTT 사전

2.1 NTT사전의 구성

NTT사전은 NTT일영 기계번역 시스템을 위해 만들어진 기계번역용 사전을 인간용으로 바꾼 것이다. NTT사전은 일영 기계번역 시스템에서의 활용을 전제로 일본어의 의미해석을 위해 개발된 것으로 [구문체계], [의미체계], [단어체계]의 세 부분으로 구성되어 있다.

[구문체계]는 일본어의 용언 6,000개를 대상으로 가능한 한 모든 의미용법을 분류 이를 약 16,000개의 문형으로 정리하고 있다. 각각의 문형은 용언의 의미를 나타내고 있고 이에 대응하는 영어의 문형을 기술해 놓고 있다.

[의미체계]는 [구문체계]의 구문에서 각각의 논항을 구성하는 명사를 약 3,000개의 의미속성을 써서 분류하고 있다.

[단어체계]는 [의미체계]의 명사들을 일본어의 자모순으로 열거하여 동일한 의미속성을 가진 명사를 검색하기 쉽게 만든 일종의 thesaurus라고 할 수 있다.

2.2 NTT사전의 문형

명사와 용언과의 의미관계를 기술하는 대표적인 문법으로는 격문법과 결합가문법이 있다. 격문법은 비교적 소수의 [심층격]을 써서 명사의 의미를 기술하기 때문에 용언과 결합하는 다양한 명사들 사이에서 나타나는 의미의 중의성을 수용하는데 한계가 있다. 이에 비해 결합가문법은 [표층]에서 용언과 임의의 격의 결합관계를 기술하기 때문에 기술능력에 있어서의 유연성이 확보될 수 있다. 다만 [표층]에 나타나는 명사와 동사와의 관계를 기술하기 때문에 많은 수의 의미속성을 준비할 필요가 있고 다국어를 대상으로 할 경우 대상으로 하는 언어에 따라 동사와 명사의 공기관계가 독특한 관계를 갖기 때문에 의미기술이 복잡해진다는 문제가 있다. 그러나 [표층]을 대상으로 하고 있다는 점에서 격문법에서는 기술이 어려웠던 관용표현까지도 기술할 수 있다는 장점을 갖고 있다. 따라서 개별언어는 물론 다국어 문형사전에서도 섬세한 의미의 기술이 가능하기 때문에 NTT사전은 결합가문법을 기반으로 하고 있다.

NTT사전의 문형은 의미속성에 의해 명사를 표현할 수 있는 일반표현문형과

문　　형 : N1ga N2wo N3ni ageru
단어 의미속성 : N1(4인간) N2(533구체물) N3("리스트/위 875지붕")
영어대역문형 : N1 put N2 on N3

한 개의 단어만이 논항의 자리에 올 수 있는 관용표현문형으로 나뉘어진다.

문　　형 : N1이 N2을 올리다
단어 의미속성 : N1(4인간) N2("약")

위의 일반표현문형의 예는 일본어 동사 [ageru (올리다)]의 여러 의미 중에 하나를 예시한 것이다. 단어 의미속성에는 문형 N1,N2,N3에 올 수 있는 명사의 의미속성이 의미코드와 함께 기술되

어있다. 이 코드를 근거로 NTT사전의 [단어체계]의 [의미속성별 단어표]에서 구체적인 단어를 검색할 수 있다. 이와 같은 구조를 통해 문형이 용언을 중심으로 한나의 의미 단위로 기술되기 때문에 구문 해석상의 중의성을 해소할 수 있으며 대역문형을 동시에 달아줌으로써 대상언어의 [의미해석]을 종료한 시점에서 목표언어의 표현구조로의 대응도 자동적으로 끝나게 되어 재차 목표언어의 용언에 대한 변환과정을 밟지 않아도 된다는 특징이 있다.

NTT사전에서는 명사 의미속성의 분류에 대해 일영번역의 경우 약 3,000개의 의미속성이면 적절한 영어의 대역어가 선정될 수 있다고 주장하고 있다. 이는 thesaurus에 있어서의 의미분류를 번역이라는 사용목적에 맞게 필요최소한도의 정밀성을 갖으면 된다는 실용적인 입장을 표명한 것으로 thesaurus와 같이 의미분류의 정도(精度)가 정밀할 수록 thesaurus로서의 가치가 높다는 입장과는 다른 입장을 보이고 있다.

3. IPAL사전과 NTT사전의 비교

IPAL사전과 NTT사전의 의미 기술능력을 비교하기 위해 일본어 동사 「あたる(ataru)」를 예로 들어 분석을 하면, NTT사전에서는 38개의 문형으로 의미를 구분하고 있는데 비해 IPAL사전의 경우 12개의 문형으로 의미를 구분하고 있다.

이는 IPAL사전이 일본어만을 대상으로 하고 58개라는 소수의 의미속성 만을 갖고 의미를 분류하고 있는데 비해 NTT사전은 영어로의 번역을 전제로 약 3,000개의 의미속성을 갖고 의미를 분류하고 있다는 점에서 의미분류 수의 차이가 난다고 볼 수 있다. 또한 NTT사전은 관용표현까지도 적극적으로 처리하고 있다는 점에서 문형의 숫자가 늘어났다고 볼 수 있다.

한편 사전 구축의 방법론이란 입장에서 보면 IPAL사전이 사전기술자의 내성에 의존해서 작성된 데 비해 NTT사전은 신문 Corpus 등 실제적인 문장들을 많이 참조로 해서 작성되었다는 점에서 비교적 일본어의 포괄적이고 다양한 의미를 충실히 반영한 것이라 할 수 있다.

위의 두 가지 사전의 의미기술의 비교를 위해 IPAL사전과 NTT사전에서 [ataru]의 의미분류를 한국어로 번역한 것을 각각 표1.1과 표1.2에 제시한다.

표1.1 IPAL동사사전의 의미구분

(1)N1ga[CON/PHE] N2ni[CON]
물건이나 현상이 무언가에 강하게 접촉하다
(2)N1ga[HUM] N2ni[PRO/PHE]
물체에 접근 혹은 접촉해서 그 작용을 받는다
(3)N1ga[HUM/ORG] N2ni/to[HUM/ORG]
누군가에 대항하다
(4)N1ga[HUM](N2wo[LIN/ABS])
N3에[HUM/PRO/LIN]
모르는 것에 대해 무엇인가에서 조사하거나 누군가에게 묻다
(5)N1ga[HUM] N2ni[DIV]
음식물, 독, 열기, 한기 등에 의해 몸이 피해를 입다
(6)N1ga [DIV] N2ni[DIV]
어떤 사람이 무언가에 해당하다
(7)N1ga[ABS] N2ni[ACT]
어떤 행위의 근거로 타당하다
(8)N1ga[ACT/MEN]
미리 예상한 것이 사실과 일치하다
(9)N1ga[ABS]
장사나 흥행이 성공하다
(10)N1ga/ni[HUM] N2de[ABS] N3ga[CON/ABS]
선발되어 무언가를 받다
(11)N1ga[HUM/ORG] N2ni[ACT]
어떤 일이나 역할을 부여받다
(12)N1ga[HUM] N2ni[HUM/ANI]
불평 불만을 주위에 터트리다

[약어표]

CON=CONCRETE PHE=PHENOMENA
HUM=HUMAN PRO=PRODUCT
ABS=ABSTRACT DIV=DIVERSE
ACT=ACTIVITY MEN=MENTAL
ANI=ANIMATE ORG=ORGANIZATION
LIN=LINGUISTIC PRODUCT

이와 같은 12개의 의미분류는 의미분류 항목수에서 매우 적은 것이라 할 수 있다. 일본의 일반적인 학습용 사전에서의 [ataru]의 의미분류 항목 수가 15개에서 20개 정도인 것과 비교를 해도 적은 편이라고 할 수 있다. 한편 NTT사전은 [ataru]에 대해 38개의 문형으로 의미를 세밀하게 분류하고 있다.

표1.2 [NTT사전]의 [ataru]

자연현상
(1)N1ga N2ni ataru N1 fill N2
[N1(2373 바람)N2("돛")]
(2)N1ni N2ga ataru N2 shine into N1
[N1(388 장소)N2(2345 빛)]
(3)N1ga N2ni ataru N1 pound on N2
[N1(2364 비)N2(533 구체물)]

행동
(4)N1ga N2ni ataru N1 work on N2
[N1(3 주체)N2(1236 인간활동)]
소유적 이동
(5)N1ga N2ni N3de ataru N2 win N1 in N3
[N1("경품/상품/특등" 533구체물)N2(4 사람) N3(1857 포상)]
속성변화
(6)N1ga N2ni ataru N1 be exposed to N2
[N1(3주체 533구체물)N2(2373 바람 2364 비)]
신체변화
(7)N1ga N2ni ataru N1 be poisoned by N2
[N1(4 사람)N2(542 어패류 838 식료)]
(8)N1ga N2ni ataru N1 be affected by N2
[N1(4 사람)N2(2363 따스함)]
(9)N1ga N2ni ataru N2 disagree with N1
[N1(3 주체)N2(748 물)]
결과
(10)N1ga N2ni ataru N1 hit N2
[N1(-2670 시간 533구체물 2422 추상적 관계) N2(-2670 시간 533 구체물 4 사람 389 시설 2422 추상적 관계)]
(11)N1ga ataru N1 prove correct
[N1(1252 꿈 1433 추측 등 1551 통지 1240 느낌)]
(12)N1ga N2ni ataru N1 win a prize in N2
[N1(3 주체)N2(933 제비 1417 선택)]
(13)N1ga N2ni ataru N2 hit N1
[N1(4 사람 535 동물)N2(-2670 시간 533 구체물 2422 추상적 관계)]
(14)N1ga ataru N1 be a hit
[N1(1037 창작물)]
(15)N1ni N2 N3de ataru N1 win N2 in N3
[N1(3 주체)N2(2590 값/금액)N3(933 제비)]
(16)N1ga ataru N1 be a success
[N1(1036 안<案>)]
(17)N1ga ataru N1 do well
[N1(677 작품)]
(18)N1ga N2de ataru N1 be called on in N2
[N1(4 사람)N2("수업")]
(19)N1ga N2ni ataru N1 face N2
[N1(3 주체)N2("강호")]
신체동작
(20)N1ga N2ni ataru N1 warm N1-self over N2
[N1(3주체)N2("불/화로/스토브/모닥불")]
(21)N1ga N2ni ataru N1 fight against N2
[N1(3 주체)N2("적" 123 적/자기편)]
감정동작
(22)N1ga N2ni ataru N1 treat N2 badly
[N1(3 주체)N2(3 주체)]
사고동작
(23)N1ga N2wo/ni tuite N3ni ataru
N1 consult N3 for N2
[N1(4 사람)N2(*)N3("사전" 1119 책(내용))]
(24)N1ga N2ni ataru N1 take charge of N2
[N1(3 주체)N2(1167 의무 1936 일)]
(25)N1ga N2wo ataru N1 look into N2
[N1(3 주체)N2("점작 가는 곳")]
존재
(26)N1ga N2ni ataru N1 lie to N2
[N1(388 장소 2610 장소)N2(2652 방향)]
속성
(27)N1ga N2ni ataru N1 constitute N2
[N1(1000 추상)N2("불이행")]
(28)N1ga N2ni ataru N1 be in N2
[N1(534 생물)N2(2694 기간)]
(29)N1ga N2ni *siturei* ni ataru
N1 be impolite to N2
[N1(1236 인간활동)N2(3-주체)]
(30)N1ni *tenbatu* ga ataru it serve N1 right
[N1(3 주체)]
(31)N1ni *tugi* ga ataru N1 be patched
[N1(813 옷)]
(32)N1ni tuite *hinan* ga ataru N1 be to blame
[N1 (3 주체)]
(33)N1ga ataru N1 be right
[N1(1236 인간활동)]
상대관계
(34)N1ga N2ni ataru N1 correspond to N2
[N 1 (*)N2(*)]
협의관계
(35)N1ga N2ni ataru N1 fall on N2
[N1(2671 역일)N2(2671 역일)]
상대관계
(36)N1ga N2ni ataru N1 be equivalent to N2
[N1(2587 양)N2(2587 양)]
결과
(37)N1ga *勝*ni ataru N1 work well
[N1(1001 추상물1236 인간활동)]
속성
(38)N1ni *hi* ga ataru N1 be sunny
[N1(388 장소 863 건축물)]

IPAL사전의 (1)번 문형 N1ga[CON/PHE] N2ni[CON]에 대해 NTT사전은 표1.2에서 보는 바와 같이 (1)(2)(3)(6)(10)(13)의 6개의 문형으로 의미를 세분화하고 있다. NTT사전의 명사의 의미속성과 영어의 대역문형을 예시하면 다음과 같다.

(1)[N1(2373 바람)N2("돛")] N1 fill N2

(2)[N1(388 장소)N2(2345 빛)] N2 shine into N1
(3)[N1(2364 비)N2(533 구체물)] N1 pound on N2
(6)[N1(3주체 533 구체물)N2(2373 바람 2364 비)]
N1 be exposed to N2
(10)[N1(-2670 시간 533 구체물2422 추상적 관계)
N2(-2670 시간 533 구체물 4 사람 389 시설
2422 추상적 관계)] N1 hit N2
(13)[N1(4 사람 535 동물)N2(-2670 시간 533 구체
물 2422 추상적 관계)] N2 hit N1

영어의 적절한 대역문형을 대응시키기 위해 IPAL사전에서 N1[CON/PHE]과 같이 [구체물]과 [자연현상]을 동시에 문형으로 표현하고 있는데 비해 NTT사전에서는 [자연현상]과 [구체물]을 따로 분리시키고 더 나아가 [자연현상]을 더욱 세분화해서 다음과 같이 분류를 하고 있다.

[2373바람] N1 fill N2
[2364비] N1 pound on N2

IPAL사전의 경우 일영번역에 대응하기에는 명사의 의미속성이 덜 세분화 되었다는 것을 볼 수 있다. 실제로 IPAL동사사전을 일영번역을 전제로 문형을 구축할 경우 약 60% 정도의 문형을 더 구축을 해야만 일영번역을 위한 일본어의 의미분석을 어느정도 만족시키는 문형이 얻어질 수 있다고 한다.[5)]

IPAL사전의 (5)번 문형의 경우 N2의 의미속성이 [DIV] 즉 어휘적 제약이 비교적 자유롭다는 속성을 부여하고 실제로 다음과 같은 단어들을 제시하고 있다.

[복어 음식물 더위 비 추위 독]

위에서 예로 나와 있는 명사들의 의미속성은 매우 성격이 달라서 [복어 음식물 독]과 [더위 추위], [비]는 영어의 대역어를 찾을 때 각각 다른 대역어가 선정될 수 있다.

한편 NTT사전에서는 (5)번의 문형을 다음과 같이 두 개의 문형으로 구분하고 있다.

(7)N1ga N2ni ataru N1 be poisoned by N2
[N1(4 사람)N2(542 어패류 838 식료)]

(8)N1ga N2ni ataru N1 be affected by N2
[N1(4 사람)N2(2363 따스함)]

[음식]에 관련되는 어휘와 [온도]에 관련되는 어휘를 구별하고 있는 것이다. 단 NTT사전의 경우 (8)번 문형에서 N2의 의미속성을 [2363 따스함]으로만 규정을 함으로써 IPAL사전에서 제시된 [추위]라는 의미속성이 빠져 있다. 실제로 [寒氣にあたって体調を崩す(한기에 노출되어 병이 들다)]와 같은 표현이 일본어에서 가능하기 때문에 NTT사전의 의미속성도 누락된 경우가 있음을 알 수 있다.

4. 한일 대역사전의 구축

4.1 동사의 문법적 카테고리

NTT사전을 기반으로 한국어와 일본어의 대역문형사전을 만들 경우 우선 고려되어야 할 사항으로 동사의 문법적 카테고리의 차이가 있다. 일본어 [ataru]는 아래의 예와 같이 한국어의 [부딪치다], [맞다], [때리다]등과 대응하고 있다. 이때 [부딪치다], [맞다]는 자타양용동사이고 [때리다]는 타동사이다.

(1) ボールが學生にあたる
(1)'공이 학생을 때리다
(2) 彼はしばらく冷たい空氣にあたってから…
(2)'그는 잠시 차가운 공기를 맞고 나서…
(3) 風が窓にあたる
(3)'바람이 창에 부딪치다

(1)과 (2)는 일본어의 자동사 구문이 한국어에서는 타동사 구문에 대응이 되는 예이다. (1)'의 문장에 나타난 [때리다]를 아래의 (1)"처럼 자동사 구문에 대입하면 어색한 표현이 된다.

(1)"공이 학생에 때리다

이 경우는 [때리다]보다는 아래의 (1)"'과 같이 [맞다]를 선택하면 자연스러운 표현이 된다.

(1)"'학생이 공에 맞다

이를 토대로 한일 대역문형을 만들어보면 다음과

5) 白井諭, 井上浩子, 小井ひとみ, 井田倉紀子, 横尾昭男, IPAL 動詞辭書の用例文に基づく日英結合価パターン對の收集,情報處理學會第53回全國大會, 4L-4, pp.2-59-60, 1996

같이 된다.

N1ga N2ni ataru [N1 구체물 N2 사람]
N1이 N2를 때리다 [N1 구체물 N2 사람]
N1이 N2에 맞다 [N1 사람 N2 구체물]

일본어의 자동사가 한국어에서 자타양용동사이거나 타동사에 해당되는 경우 대역어의 선정에 있어서 N1과 N2의 명사의 어순에 대해서도 세심한 배려가 필요하다. 한국어는 N1의 자리에 [인간]이 올 수 있는 경우와 올 수 없는 경우가 생기는데 일본어의 [ataru]의 경우는 아래의 (1)""과 같이 명사의 어순에 제약이 따르지 않는다.

(1)""學生がボールにあたる

이와 같이 문법적 카테고리의 차이는 대역어 선정에 깊이 관여한다.

(3)의 경우 논항의 명사의 배열이 다음과 같을 경우 각각의 구문에 어울리는 한국어는 다음과 같이 된다.

(4) N1=바람 N2=유리창→(자동사문형)부딪치다
(타동사문형)때리다/치다
(5) N1=바람 N2=돛 →(자동사문형)부딪치다
(타동사문형)때리다/치다

한국어의 대역어가 자동사 구문에서 실현될 경우에는 [ataru]의 대역어로 [부딪치다]가 되지만 타동사구문에서 실현될 경우에는 [때리다]와 [치다]의 두 개의 대역어가 선택될 수 있다. 이 경우, 한국어는 자동사구문이냐 타동사 구문이냐는 동작의 [강약]에 의해 선택되게 된다.

한국어에서 만일 [강약]이라는 속성을 표현하려면 [부딪치다]가 타동사구문에서 실현될 경우는 대부분이 N1의 의미속성이 다음과 같이 [인간]인 경우이다[6].

(6)철수가 책상에 머리를 부딪치다.
(7)명예회장이 카메라에 이마를 부딪치다.

예외적으로 다음과 같이 N1의 의미속성이 [비인간]인 경우가 있다.

(8)차가 노량대교의 난간을 부딪치고 밖으로…
(9)작은 낚싯배들이 서로 몸을 부딪치며…
(10)먹이가 목구멍의 내벽에 몸을 부딪치며…

이런 경우 대부분은 의인화된 표현인 경우이다.

이와 같이 자동사 구문이냐 타동사 구문이냐는 대역문형의 구축에 있어서 대역어 선정에 영향을 준다. 또한 한국어와 일본어는 서로 문법적 카테고리가 다른 동사가 많이 존재하기 때문에 대역문형 구축 시에 세심한 배려가 필요하다.

4.2 의미속성의 통합과 명사·리스트의 정비

[NTT사전]은 일영기계번역을 전제로 만들어진 것이기 때문에 이를 토대로 한국어과 일본어의 대역문형을 작성하게 될 경우 불필요한 문형의 세분화가 있을 수 있다. 일본어를 영어로 번역할 경우에는 38개의 문형이 필요하지만 일본어 문형을 한국어로 번역할 때에는 앞서 3.에서 살펴본 것처럼 [바람]과 [비]에 의한 대역어의 차이가 영어에서 처럼 발생하지 않기 때문에 NTT사전에서는 명사의 의미속성에 따른 두 개의 문형이 분리될 필요가 있었으나 한일에서는 문형을 다음과 같이 하나로 통합할 수 있다.

N1이 N2에 부딪치다 N1ga N2ni ataru
[N1(2373바람 2364비)N2("돛" 533구체물)]

마찬가지로 결과를 나타내는 NTT사전의 문형 (11)과 (16)처럼 N1의 의미속성에 따라

[1252꿈 1433예상 1551통지 1240느낌] →N1 prove correct
[1036안(案)] → N1 be success

와 같은 의미의 구별이 한국어와 일본어에서는 존재하지 않기 때문에 아래와 같은 한 개의 문형으로 충분하다.

N1이 들어맞다
[N1(1036안(案) 1252꿈 1433추량 등 1551통지 1240느낌)]

그러나 [1036 안(案)]의 경우 논항의 의미속성 별 명사의 리스트를 수록하고 있는 NTT사전의 [의미체계]에 의하면 다음과 같은 명사가 나와 있다.

6 KAIST Corpus에서 [부딪치-]의 용례 1,316개 중 N2가 목적격 [을/를]을 취하는 것은 78개이며 그 중 6개가 N1의 의미속성이 [비인간]으로 나타났다.

[의도 기획 우민정책 작전 책략 해답…]

이와 같은 경우 한국어의 대역어 [들어맞다]가 대역어로 선정될 수 있는데 같은 의미속성에 속하는 다음과 같은 명사들은 한국어 명사로써 거의 쓰이지 않거나 일본어 자체로도 거의 쓰이지 않는 것들이기 때문에 단어 리스트에서 삭제시켜도 무방한 것들이라고 할 수 있다.

[意企 一案 英案 僞計 下策 神算 選對…]

이와 같은 사실은 NTT사전이 철저한 코퍼스에 기반을 두고 작성된 사전이 아니라 여러 기존의 사전과 신문 코퍼스를 결합시켜 만들어 진 것이라는 데서 온 결과라고 할 수 있다.

4.3 의미속성의 분할

한국어와 일본어의 경우 일본어에서는 동일한 의미속성으로 분류되는 명사라 할지라도 한국어에서는 다른 의미속성을 부여해야 되는 경우가 있다.

(23)32 사고 동작
N1ga N2wo/ni tuite N3ni ataru
N1 consult N3 for N2
[N1(4 사람)N2(*)N3("사전" 1119 책(내용))]

N3의 [1119책]에 해당되는 명사를 NTT사전의 [의미체계]에서 찾아보면

[원저 원서 원본 대본 판본 성서…]

등과 같은 단어들이 나와 있다. 한국어에서는 [원본/대본/대장]의 경우 [대조하다]라는 표현이 가능하기 때문에 N3의 의미속성을 [일반적인 책]과 [기준이 될 수 있는 책]으로 세분화시킴으로써 대역어 선정을 정밀화 할 수 있다. 이 경우 문형의 격표시가 [N3을]에서 [N3와]로 바뀌게 됨으로 명사의 의미속성에 따라 문형의 격표시도 바꾸어줄 필요가 있다. 다음과 같은 경우에도 의미속성의 세분화 및 문형의 세분화가 필요하다.

N1ga N2ni ataru N1 poisoned by N2
[N1(4 사람) N2(542 어패류) 838 식료)]

N2의 [542어패류]의 경우 [복어/피조개/광어/독버섯] 등 독이 있는 생선이나 음식 또는 회로 먹는 것들이 올 경우에는 [중독되다]가 대역어로 쓰일 수 있으나 [게/뱀장어/동태/대구] 등과 같이 일반적으로 익혀서 먹는 생선일 경우 [체하다]를 대역어로 선정할 수 있다.

의미속성이 대역어 선정에 영향을 미치는 것은 위의 문형에서는 [542어패류][838식료]가 아니라 독이 있느냐 없느냐는 기준에 의해 분류가 이루어져야 한다. 이는 대역문형 구축 뿐만 아니라 일본어의 정확한 의미분석을 위해서도 필요한 작업이라고 할 수 있다.

4.4 누락된 의미의 보충

NTT사전은 현재 일본에서 나온 사전 중에 가장 상세한 의미기술을 하고 있는 사전이다. 그러나 NTT사전의 서문에도 나와 있듯이 개발 중에 있는 사전을 연구의 차원에서 공개했다는 점에서 많은 개선점이 발견된다.

다음과 같은 관용적 표현이거나 동사의 관형형에 대해서는 NTT사전에서는 기술이 안되어 있다. 이는 현재까지 구축된 문형이 주격과 목적격이라고 하는 필수격을 대상으로 하고 있다는데서 그 원인을 찾을 수 있다. 다음은 관용표현에 해당하는 문형의 예이다.

(1)N1ga N2ni ataru [N1"신발"N2 "발"]
N1이 N2에 꼭 끼다
(2)N1ni ataru [N1"더위"]
N1을 먹다

명사의 의미속성에 [신발]과 [발], [더위]라는 명사 만이 올 수 있는 경우로 한국어의 대역어는 각각 [꼭 끼다],[먹다]가 된다. 이 외에도

(3)N1ni atari [N1 "개관" "결혼"]
N1에 즈음해서

와 같이 N1의 의미속성이 [행사]와 관련있는 명사가 올 경우 한국어의 대역어는 [즈음해서]가 된다.

[ataru]가 체언을 수식하는 경우 중의성의 문제가 발생하게 되는데 이를 문형을 통해 작성해 줄 필요가 있다. 이 경우 [ataru]는 [atattuta]로 활용을 한다.

(3)atattuta N1 [N1 854 과일]

썩은 N1

이와 같이 문장 내에서의 동사의 위치도 의미와 관계하는 경우가 있기 때문에 다양한 현실의 문장에 대응할 수 있는 다양한 형태의 문형의 구축이 필요하다.

5. 결론

대역문형사전의 구축은 코퍼스에 기반을 둔 개별언어에 대한 충분한 의미분석을 토대로 이루어져야 한다. 이는 현실적으로 존재하는 문장의 의미를 해석하기 위한 첫 단계라고 할 수 있으며 대역문형 작성에 있어서도 매우 중요한 의미를 갖는다. 개별언어를 대상으로 하는 문형사전의 구축이 코퍼스에 기반을 두었을 때 앞서 4.2에서 살펴 본 바와 같이 불필요한 명사가 명사 리스트에 등재되는 것을 피할 수 있으며 이를 통해 현실적으로 동사와 결합하는 명사의 분포가 정확히 추출될 수 있을 것이다.

한국어와 일본어는 자동사,타동사라고 하는 문법적 카테고리에 의한 대역어 선택의 양상이 대역문형사전 구축 시에 중요한 의미를 갖기 때문에 명사의 의미기술과 병행해서 문법적 카테고리의 충분한 검토가 필요하다. 또한 관용표현을 포함한 어순에 의한 의미의 파생문제 등에 대해서도 적극적으로 기술을 할 필요가 있다.

이와 같은 작업들을 통해 한국어와 일본어뿐만 아니라 다국어 대역문형사전의 구축에도 유용한 의미속성의 분류와 구문의 확립을 위한 많은 지식이 축적되기를 바란다.

참고문헌

[1] Cruse D.A. *Lexical Semantics* : Cambridge University Press, 66 p. 1986

[2] 松本裕治, 影山太朗, 齊藤洋典, 德永健伸, 單語と辭書, 岩波書店, pp.172-175, 1997

[3] NTTコミュニケーション科學研究所監修, 日本語語彙大系, 岩波書店, 1997

[4] 白井論, 井上浩子, 小出ひとみ, 井田倉紀子, 横尾昭男, IPAL動詞辭書の用例文に基づく日英結合価パターン對の收集, 情報處理學會第53回全國大會, 4L-4, pp. 59-60, 1996

[5] 情報處理振興事業協會技術センター, 計算機用日本語基本動詞辭書IPAL(Basic Verbs), 情報處理振興事業協會技術センター, 1987

[6] 日本電子化辭書研究所, EDR電子化辭書仕様說明書, EDR, 1993

[7] 中野洋, 「分類語彙表」形式による語彙分類表(增補版), 國立國語研究所, 1996

[8] 長尾眞, 自然言語處理, 岩波書店, 1996