일본의 韓日 機械飜譯 시스템에 있어서의 誤譯과 그 언어환경

강 용 희
東京대학 대학원 總合文化연구과 언어정보과학전공

Errors and Their Circumstances in Korean Japanese M/T Systems in Japan

KANG YONG HEE
THE UNIVERSITY OF TOKYO LANGUAGE AND INFOMATION SCIENCES
kang@tooyoo.l.u-tokyo.ac.jp

일본의 韓日 機械飜譯 시스템을 평가한 결과 각기 다른 번역 시스템임에도 불구하고 誤譯의 패턴에 있어서는 類似한 점이 많았다. 이는 辭典의 입력 단위와 構文분석의 해석단계에서 誤譯의 언어환경에 대비하지 못한 점을 지적할 수 있다. 본 연구는 誤譯의 TYPE을 언어적 환경과 기계적 환경으로 구분하여 그 영향관계를 밝혀서 誤譯의 환경에 대비한 辭典과 모듈의 작성을 목적으로 삼는다.

1. 서론

현재 일본에서 시판되고 있는 韓日機械飜譯 소프트에는 日立(HITACHI 以下 HITACHI)의 韓日 機械飜譯 시스템(HICOM／MT)과 高電社(KODENSYA 以下 KODENSYA) 의 I-seoul K/J이 있다. 본 연구는 한국이 장래에 TV의 NEWS를 문자 방송으로도 할 경우에 대비해서, 일본의 NHK측이 일본 국내에서 시판되고 있는 韓日 機械飜譯를 자체평가 한 결과의 일부분이다.

NHK 는 아직 독자적인 한일 기계번역 시스템을 갖고 있지 않다. NHK측은 주로 뉴스를 대상으로 선정했기에 필자는 CORPUS로써 조선일보를 선택했다. 조선일보(RTF 형식) CD-ROM의 RTF 형식의 파일을 텍스트파일로 바꾸어, 그 중에서 400개의 뉴스 기사(2,000개의 문장)를 각각의 번역 시스템에 입력하여 그 결과문을 평가했다. 그 결과 각기 다른 시스템임에도 불구하고, 誤譯의 패턴에 있어서는 유사한 점이 많다는 것을 발견했다. 이는 韓國語의 CORPUS의 언어환경과 誤譯의 상관관계가 辭典의 단위와 구문 분석의 해석단계에서 誤譯의 언어환경에 대비하지 못했다는 점을 감안해서 誤譯의 TYPE를 먼저 계산하고 그와 같은 환경에 대비가능한 기계번역 시스템을 고려해야 한다.

2. 본론

2. 1. 韓國語의 語節 환경

한국어의 표기는 한글 및 한자와 기호, 알파벳 등으로 이루어져 있으며 동일한 내용이라도 표기수단을 한자나 한글 등으로 교체할 수 있으며 표기방법에 있어서도 專用과 混用이 가능하다는 점에서 사전등록에는 2중의 부담이 된다. 그러나 내용과는 별도로 띄어쓰기의 기준이 되는 양식은 어절 단위이며 출현환경은 제한되어 있다고 해도 과언은 아니다.

韓國語의 語節구성은 다음과 같이 구분할 수 있다.

單獨型 (명사, 대명사, 부사, 관형사, 용언연체형(용언의 관형形)
수사, 불완전 명사, 감탄사, 접속사)

結語型 (조사, 용언의 어간, 용언의 어미, 접속사형 어미,
접두사, 접요사, 접미사, 助動詞 (보조어간))

이는 機械飜譯에 있어서 構文해석 단계에서 해석가능한 어절의 조합은 계산이 가능하다는 것을 의미한다. 또한 結語型중의 조사, 어미, 접미사, 助動詞等은 語節 중 선두위치에 올 수 없으므로 辭典검색의 단계에서 동음이의어의 환경에서 위치에 의한 선택적 制約을 할 수 있다. 또한 어절중에 있을 수 있는 同音異意語의 환경(명사 + 조사 : 용언의 연체형等)에서 어떤 식으로 변별자질인 공기정보(co-occurrence)를 辭典중에 記述해야 총체적으로 대응할 수 있을까도 알 수 있다. 例를 들면 어절인 [물을]을 조선일보의 6,000개의 paragraph로 부터 검색해보면 다음과 같다. [물+을]은 17건 : 책임을 물을(묻다)은 2건이다. 이는 [물을]의 변별자질이 빈도가 높은 [물+을]에 있다고 판단하기 어렵다. 빈도가 적은 [책임을 물을]에 있으며 [책임을 물을]以外의 [물을]은 자동적으로 [水+を]이 된다. 아울러 韓國語에는 자음과 모음에 의한 조사의 분포가 언어 환경에 따라 선택적인 변별성을 가지게 되는 경우가 있다.

2. 2 韓國語 조사의 상보분포 (Complementary Distribution)

한국어의 어절중 체언과 조사의 분포는 다음과 같다.

체언 (V) + 조사 (C) 가,를,로
체언 (C) + 조사 (V) 이 ,을 ,으로
체언 (V) + 조사 (V) 와
체언 (C) + 조사 (C) 과
체언 (ㄹ) + 조사 (C) 로

위의 상보분포가 助詞와 體言과의 관계에서 相關的이나, 조사를 우선적으로 인식한 후에 체언의 명사를 검색한 결과 오역이 발생하기도 했다.(KJ:KODENSYA의 번역, HJ:HITACHI의 번역)

예)000씨 (동양석판 전기과 근무)
HJ1> * 東洋席版電氣と勤務
KJ1> * 東洋石版電氣KWA勤務

HITACHI는 구문해석 단계에서 어절의 뒷부분에 위치하고 있는 조사를 우선적으로 처리하며, 辭典에서 조사를 제외한 부분이 테이블 및 최장일치법등의 조건에 맞으면 번역어를 선택한다. 그로 인해 韓國語에 있는 音韻의 결합규칙에 어긋나는 결합형태를 형성하게 되었다. HJ1>의 [電氣と]라는 번역어가 되기 위해서는 원문이 [전기와]이어야만 된다. 한편 KODENSYA는 접미사의 [과]를 미등록어 처리를 하고 있다. 여기서 우리는 KODENSYA가 접미사의 [과]를 포함하는 語節 [전기과]가 [전기]와 조사 [과]와의 결합형의 가능성을 배제했다는 것에 주목할 필요가 있다.

2. 3 기계번역의 평가 작업과 오역의 분석

다음의 표는 번역의 평가 기준과 오역의 환경과 오역의 발생건수이다.
(대상은 100개의 뉴스기사이며 문장의 수는 500개이다)

표1	HITACHI	KODENSYA
1. 형태분석 ERROR	927	208
2. 同音異意語	216	135
3. 다의어	89	70
4. 生成형	26	24
5. 생략형	93	35
6. 관용표현	83	55

7. 부정의 표현	14	24
8. 수동,사역	5	24

표2	HITACHI	KODENSYA
1. 명사	425	142
2. 대명사	32	3
3. 고유명사	357	105
4. 형식명사	16	4
5. 動詞	141	100
6. 복합動詞	17	15
7. 형용사	8	12
8. 관형사	28	11
9. 부사	23	18
10. 조사		
은／는	16	4
이／가	7	8
을／를	13	8
으로／로	64	21
에서	10	2
에	15	3
도	1	0
와／과	2	0
의	11	11
에게	0	1
이나／이랑	0	1
부터	0	1
그 밖의 조사	6	1
11. 어미	50	32
12. 접속사	19	12
13. 활용형	91	29
14. 접미사	203	36
15. 접두사	37	26
16. -하다 -되다 動詞	47	10
17. TENSE , ASPACT	25	18
18. 수사	5	22

위의 결과는 誤譯의 환경과 품사별로 나누어 표시했다. 즉 표1는 환경이고 표2는 구체적인 품사 및 文法기능소이다. 감탄사 및 일부 文法 기능소가 평가 기준에 빠져있는 이유는 CORPUS를 신문기사로 삼았기 때문에 발생빈도가 적은 품사 및 문법소는 취급하지 않았다.

위의 數値중 현저하게 差가 있는 형태분석 ERROR(표1의 1)와 고유명사(표2의 3)의 경우는 두 시스템이 취하고 있는 번역방식에서 起因한 것이다. HITACHI는 DIRECT 構文방식를 채택하고 있기 때문에 사전에 등록되어 있지 않는 고유명사, 관용어 등의 환경에서도 형태소를 단음절로 분리해서 검색하는 등의 경향이 많았고 무리하게 일본어로 생성하는 TYPE의 오역이 대부분을 차지했다. 그것에 비해 KODENSYA의 I-seoul K/J은 BOTTOM UP방식를 채택하고 있어 主語부와 述語부로 나누고 한 단계의 분석이 끝나면 다시 句와 語節로 나누어 분석하는 多段階 방식으로 사전에 없는 단어는 미등록어 처리하고 있으며 동음이의어와 다의어는 번역 가능한 候補

단어를 同時併記하는 방식을 채택하고 있다. 필자는 HITACHI의 DIRECT 構文방식을 적극적 방식이라 보고 있으며 KODENSYA의 I-seoul K/J의 BOTTOM UP방식은 소극적 방식이라 보고 있다. 본 연구에서 주목한 POINT는 다른 품사에 비해 接頭詞(표2의 15)와 接尾詞(표2의 14)의 誤譯의 數値가 높다는 것이다. 이는 번역방식에 差를 넘어서 최종적인 단계의 語節에서 어떤 식으로 처리하느냐의 문제인 것이다. 접미사 오역의 비율이 접두사 오역의 비율에 비해 압도적으로 높은 이유는 언어적 환경에 기인했다고 보기 어렵다. 그렇지만 韓國語의 접두사와 접미사의 환경(단일어, 복합어, 파생어, 합성어)이 다르므로 誤譯의 比率에 差가 있어도 相關관계는 희박하다고 추정할지 모르나 이는 語節의 분석단계에서 語節의 뒷부분부터 끊어오면서 辭典단어를 검색하는 [최장일치법]을 채택하는 구조적 체계에서 발생하는 數値이다. 또한 [으로]/[로]의 오역이 다른 조사의 오역보다 數値가 높은 점도 주목해야 한다. 참고로 강(1997)에서는 [으로]/[로]는 多義語型 조사로 분류하고 있다.

재료+ [으로]/[로]	→ で,から	수단+ [으로]/[로]	→ で
방법+ [으로]/[로]	→ で	근거+ [으로]/[로]	→ で,と
도구+ [으로]/[로]	→ で	원인+ [으로]/[로]	→ で,から
이유+ [으로]/[로]	→ で,から	방향+ [으로]/[로]	→ に,へ
장소+ [으로]/[로]	→ で,に	변화+ [으로]/[로]	→ に
신분+ [으로]/[로]	→ に,で	자격+ [으로]/[로]	→ として

위의 범주를 알고 있어도 語句의 지배를 기술하는 방법이 간단하지 않다.
HITACHI는 Filmore의 Case이론에 근거하여 동음이의어 및 다의어에 대응하고 있다. 동음이의어의 환경에서 Case이론에 의한 기술방법(강,1996)이 어느 정도 효과가 있는 것은 사실이나 다의어의 환경에는 무리가 있다. 강 (1996)을 인용해서 비교해보면 다음과 같다.

볼 | 축구공 | 공-을 | 를 + * (차다)
물-이 + * (차다)

結合價理論(Valenzhteorie)은 동사와 격조사의 기술에 그 선택지를 동사에 두고 동사에 따라 요구 되는 문형을 기술하고자 하는 이론이다. (1, 2, 3)은 結合價이다.

N_0(물)이 ADJ(차다)	
N_0(물)이 V(차다)	1>가/이 차다(1)
N_0(사람)이 N_1(공,볼)을 V(차다)	1>가 2>을/를 차다(2)
N_0(사람)이 N_2(허리)에 N_1(칼)을 V(차다)	1>가 2>에 3>을/를 차다(3)
N_0(사람)이 C_1(숨)이 V(차다)	1>가 숨(이) 차다 (1)

2. 4. 辭典의 입력단위와 모듈처리

2. 4. 1 입력단위

HITACHI의 HICOM／MT와 KODENSYA의 I-seoul K/J의 辭典은 그 대부분이 학교文法의 품사 단위의 형태와 유사하며 사용자가 일반사전의 대용으로도 사용할 수 있게 되어 있다.

KODENSYA:품사 21종: 명사(3종), 부사, 동사, 형용사(イ形容詞), 형용동사(ナ形容詞), 姓, 名, 人名, 數詞접두사, 數詞, 名詞접두사, 名詞접미사, 助詞,대명사, 이름뒤에 붙는 接尾辭, 관형사, 감탄사, 조사접속句, 그 밖의 句.

HITACHI:28종의 품사와 125개의 등록형으로 사전이 이루어져 있다.
품사: 명사, -하다型어간, 부사, 관형사, 접속사, 인사말, 형용사, 동사, 대명사, 접두사, 접미사, 數詞用접미사, 數詞선두어, 英字, 숫자, 제목, 고유명사,

인명(姓), 인명(名), 連用수식어, 連體수식어, 지명, 법인명, 終조사, 조사,
接續조사, 조동사, 기호(4종) .
등록형; 형용사 7형, 동사8형, 변칙동사5형, 조사 6형, 접속조사 7형, 조동사 8형,
조동사(10종)
(등록단위중 활용형과 동음이의어의 환경에 대비한 2중 등록형을 포함하고 있다)

辭典의 용량이 10메가를 超過하고 있으며 접두사, 접미사, 복합명사등의 단위로 이루어져 있고, 경우에 따라 같은 형태소라도 어휘의 일부분(복합명사) 혹은 어휘의 기능적인 요소단위(접두사, 접미사)로 중복되어 등록되어 있는 例가 많다.

예술 : 藝術 (명사), 가:家 (접미사) ,예술가 : 藝術家 (명사)

이는 語節단위로 이루어지는 우리말의 문자단위에서는 誤譯이 발생할 확률이 높고 형태분석의 ERROR가 발생하므로 機械飜譯에 있어서 치명적인 誤譯이 발생한다. 한국의 韓日機械飜譯 시스템과 일본의 韓日 機械飜譯 시스템이 형태분석의 단계에서는 [최장일치법]을 공통으로 採擇하고 있지만 語節의 飜譯方式에 있어서 반드시 가장 긴 音節의 一致가 정확한 분석의 척도가 된다고 보기 어렵다.

例) 통신은 중앙선거위원회를 인용
KJ2>: * 通信は中央線ガチョウ元會を引用(容認) [[중앙선][거위][원][회]]
HJ2>: * 通信はセンターラインがちょうウォンを引用 [[중앙선][거위][원][회]]

[중앙선거위원회]의 語節단위는 [[중앙][선거][위원][회]]로 이루어 졌다고 가정하는데 타당하겠지만, 각기 다른 소프트의 辭典에서 [최장일치법]이 적용되어 같은 음절 경계단위로 형태분석을 시행하고 있다. 또한 재미있는 점은 같은 경계단위로 분석하고도 사전의 등록단어의 차이에 의한 다른 번역어와 표기를 산출했다는 것은 같은 경계단위의 內部에도 다시 오역의 환경이 만들어짐을 反證하는 좋은 例이다. 선행연구의 강(1996,1997)에서는 위의 類型을 誤譯의 환경과 영향에서 경계단위 분석 ERROR와 연쇄 ERROR로 규정하고 있다. 원인으로는 사전에 [중앙선거위원회]라는 형태로 입력되어 있지 않고 이를 분석하는 과정에서 辭典의 등록된 단어중 語節의 조합단위를 선별하고,적합한 판단인가를 정할 때 長短으로 선택하는 방법은 초보적인 辭典에서는 유용할지 모르나 辭典의 규모가 확대되면 될수록 등록단어의 중복으로 因한 辭典內의 간섭(Interference)이 심하게 될 것이다.
參考로 조선일보의 6,000개의 paragraph중 [중앙]을 포함하는 語節은 81개 였다.
이들중 오역을 음절과 환경별로 비교하면 다음과 같다.

경계 요소 분석 ERROR: [중앙][당] ,[중앙][대] ,[중앙][협], [중앙][회]
생략형 ERROR : [중앙][아] 5개 국가들은, [중앙][아] 5개국 새연방 계획
cf: [중앙 아시아]의 형태에서는 오역이 발생하지 않았다.

또한 다음의 例는 복합명사의 성분 요소중 접두사, 접미사와 명사, 고유명사의 축약형의 분석이 어렵고 그로 인한 誤譯은 구성 요소 분석 ERROR가 된다.
例) 미영화연구소 : [[미][영화][[연구][소]]] KJ3>: *[美英花宴舊ソ] ; [美][英][花][宴][舊ソ]
HJ3>: *[未映畵研究所] ; [未][映畵][研究所]
재판장: [[재판][장]] ; [[裁判][長]] HJ4>: * [裁判][場]
재판정: [재[판정]] ; [再][判定] HJ5>: * [裁判][亭]

[재판장]은 구성의 경계까지는 정확히 판단했지만 [장]의 동음이의어의 인식과정에서 실패 했고, [재판정]은 구조적으로 1음절을 뒤에서 앞으로 끊어오는 과정에서 단어의 구성 요소의 경계분석의 ERROR가 된 경우이다.

2. 4. 2. 테이블의 문제

韓國語에는 어느 특정의 단어 혹은 文法소가 존재할 수 있는 文法상의 制約 또는 전후의 품사위치 관계등의 이유로 연속형의 制約이 있다. HITACHI의 품사table(강,1996의 보충자료)은 각각의 품사의 전,후에 올 수 있는 품사들의 group를 전, 후의 위치별로 나누지 않고, 一體형으로 다루고 있다. 즉 統辭적 위치 制約에 있어 HITACHI는 語節중 조합이 가능한 품사의 선택을 전,후에 올 수 있는 넓은 범위에 머물고 있으며 적극적으로 전,후의 위치별로 통사규칙을 적용할 수 없는 弱點이 있다.

2. 4. 3. 모듈처리와 辭典처리

韓日 기계 시스템중 辭典에 의지할 수 있는 부분과 모듈처리를 해야만 하는 요소는 무엇일까? HITACHI는 어간과 어미 그리고 助動詞를 辭典에 전부 등록하는 방식을 채택하고 있으며, 그 처리 방법은 번역어의 활용형을 정하는 부분과 어미 및 助動詞의 의미번역을 첨가하는 방식을 채택하고 있다. HITACHI는 부정의 [안],[못]은 모듈처리를 하고 있고, 可能의 표현(수 있다)과 부정의 [지 않다]는 사전처리를 하고 있다. 助動詞(보조어간)와 어미 그리고 조사등이 연속적으로 나타나는 예문의 경우에 誤譯이 많이 나타난다.

그는 밥을 먹고 있다.	彼はご飯を食べている.	○
그는 밥을 먹고 있지 않다.	彼はご飯を食べていない.	○
그는 술을 못 먹는다.	彼はお酒が飲めない.	HJ6> * 彼は酒をない

2. 5. 統辭규칙의 差異

2. 5. 1. 自動詞와 他動詞의 문제

강(1996)에서는 HITACHI의 韓日 機械飜譯 소프트는 自動詞와 他動詞의 구별이 없고 文중의 조사에 의지하는 번역 방식을 사용하고 있었지만 본인의 지적에 의해 현재 自動詞와 他動詞의 구별을 일본어의 환경에서 구별하도록 조처하고 있다. 부연하면 SORCE 언어인 韓國語에 自動詞와 他動詞의 변별자질을 두지 않고, TARGET 언어인 日本語에 그 統辭적 규칙을 적용하는 방식이다. 日本語에서는 自動詞와 他動詞의 구별에 의한 varient가 있는 [ている],[てある]의 처리문제(강,1996)가 있고 補助動詞와 本動詞의 용법에 의한 [-される],[-できる],[-になる],[-する] 의 선택문제가 있다.

한자어 + 되다 → 1> 自,他동사의 구별 = [-される] 개최되다. 정리되다
→ 2> [-になる] (本動詞) 선생님이 되다, 얼음이 물이 되다.
→ 3> [-できる] (本動詞,補助動詞) 정리되다.
→ 4> [-する] 모순되다,완성되다 = 矛盾する.完成する.

이 문제는 統辭적 요소와 어휘적 요소가 동음이의어의 환경이 되는 경우로 文法 기능소의 [되다]와 어휘소의 [되다]의 처리의 문제가 대두된다. HITACHI는 1996년에는 自動詞와 他動詞의 구별을 하지 않았으나 강(1996)의 지적 및 대처 방안에 의해 현재는 1>의 변별성이 있다 그러나 I-seoul은 本動詞 2> [되다]의 [-になる]로 전부 처리하고 있다.

예)윤활유 원료 2백ℓ들이 5드럼 1천ℓ가 바다에 유출돼 이 일대 해안이 오염되고 있다.

KJ4>: * 潤滑油原油2百リットルが5ドラム1千リットルが(に)海に流出[になり]
この一代(一帶)海岸が汚染[になっている].

2. 5. 2. 형용어의 문제

韓國語에는 관형사와 형용사가 있고 日本語에는韓國語와 달리 2가지 형용사가 있다.또한 어용론의 관점에서 보면 대응관계가 부분적으로 맞지 않는 경우가 있다.

	크다(大)	큽니다(大)	큰 문제(大きな問題)	큰딸을(長女)
HJ7>:	大きい	大きいです	*大きい問題	* 大きいむすめを
KJ5>:	大きい	大きいです	大きな問題	*大きな注ぐ

HITACHI는 연체형을 辭典에 등록하는 구조를 취하고 있기 때문에[*大きい問題]라는 번역을 하게 된다. 즉 HITACHI는 韓國語의 動詞와 형용사의 연체형을 辭典에 등록시켜 자동적으로 日本語의 연체형의 번역어를 얻고 있다. 또한 일반적 관형어와 명사의 대응관계에 있어서는 별도의 어휘를 필요로 하는 例도 이다. 또한 KODENSYA는 활용형의 처리에 문제가 있다 즉 [따르다]의 활용형으로 [딸을]을 인식하고 있다. 이는 활용형의 설정의 ERROR와 형용사(연체형)와 動詞가 연속되는 ERROR를 범하고 있다.

2. 5. 3. 合成語의 문제

[꿀]과 [물]은 韓日 兩言語에 있는 어휘이지만 韓國語에는 合成語인 [꿀물]이 어휘로써 존재하지만 日本語에는 존재하지 않는다. 機械飜譯의 구조상 표제어와 번역어가 각기 존재할 경우에는 그 합성어가TABLE상에서 명사+명사의 구조이므로 결합가능하다고 판단할 것이다. 理想론이 될지도 모르나 필자는 韓日 2言語대응 시스템 사전과 일본어 CURPUS사전 이 별도로 필요하다고 믿는다. 일본어 CURPUS사전은 단어 및 품사의 정보만으로 구성되어 있어 생선된 일본어 단어 ,어미 및 조동사의 존재여부를 확인할 수 있다. 그러나 합성어의 성질상 간단하게 대응관계가 이루어지지 않는 경우가 많다. 복합動詞도 예외는 아니다 선행연구에서는 어순이 다른 복합動詞는 韓國語에서 미리 어순을 바꾸어 日本語에서 합성하는 방식을 채택하고 있지만 어순뿐만 활용형에도 선책사항이 있으므로 동사의 합성법을 유출하기가 어렵다고 본다.

꿀 + 물 = 꿀물 : はちみつ + みず = *はちみつみず

넘어지다	넘다+지다	ころぶ	복합동사(×)
넘어서다	넘다+서다	こえる	복합동사(×)
끌려가다	끌다+가다	引いて行く	[て]활용형
고쳐쓰다	고치다+쓰다	書きなおす	逆順,연용활용형
써보다	쓰다+보다	書いてみる	使かってみる [て]활용형
갈아타다	갈다+타다	乘り換える	逆順,연용활용형

2. 6. 生成형 誤譯

2. 6. 1. 二重助詞의 문제

ㄱ.>그는 꽃을 사진을 찍었다. HJ8>: *彼は花を寫眞をとった.
(彼は花の寫眞をとった./彼は花を寫眞にとった.)

ㄴ.>내가 좋아하는 빵을 그가 먹었다. HJ9>: *私が好むパンを彼が食った.
(私の好きなパンを彼が食った.)

ㄷ.>그는 사과를 담을 바구니를 가지고 있지 않다.
(彼はリンゴを入れる箱を持っていない.)

ㄹ.>그는 사과를 담을 바구니를 찾고 있는 어른을 보았다.
(彼はリンゴを入れる箱を探している大人を見た.)

[사진(을) 찍다], [숙제(를)하다]등의 동사는 [을/를]생략할 수 있으며, 일본어에서도 격조사가 일반적으로 생략이 가능하다는 점에는 닮았다. 日本語에서는 모든 문장에서 조사[を]가 두번 이상 사용될 수 없는 것은 아니지만(例:ㄷ,ㄹ), 한 문장에서 조사 [を]나 [が]를 두번 以上 사용할 수 없는 경우(例:ㄱ,ㄴ)가 있다. 그러나 우리말에는 자유스럽게 조사 [을/를]이나 [이/가]사용가능한 경우가 있다.

2. 6. 2. 사이시옷의 처리

윗물이 맑아야 한다	HJ10>: * 上水が清かろねばならない
아랫물이 맑다	HJ11>: * ああ # 水が清い
빨랫줄	HJ12>: * 洗う # 竝
빨래줄	HJ13>: * 洗濯物竝

HITACHI는 辭典에 단어를 입력하고 合成語는 語節을 최장일치법에 의해 처리하고 있지만, 사이시옷이 語中挿入(epenthesis)되는 환경에서는 ERROR처리 된다. [윗물]은 관용어 취급하여 등록되어 있지만 [아랫물]은 일반화된 명사도 아니고, [아][랫][물]로 분석했다. [빨랫줄]의 경우도 1음절씩 나누어 [빨][랫][줄]로 분석하고 있다. 표기법과는 다르지만 [빨래줄]이라고 입력했을 경우에는 [빨래]까지는 정확히 분석했으나 [줄]은 구성 요소 분석 ERROR (강,1996,1997)로써 [줄]을 서다의 [줄]로 인식하고 있다. 한국어의 단어중 사이시옷(198字)을 분리해서 분석할 경우 모든 사이시옷의 단어를 입력할 필요는 없다(빨랫방망이,빨랫비누). 그러나 일반화된 단어(돗자리)와 일본어의 환경에서는 합성명사가 아닌 명사는 입력해야 한다. 부연설명하면 [-ㅅ]의 환경에서는 [-ㅅ]를 분리해서 (코드가 변한다) 단어검색을 한후 日本語로 生成할 때 속격인[の]를 추가하는 모듈로 처리하지 않으면 안된다. 현재 HITACHI와 KODENSYA는 사이시옷에 대한 모듈처리를 하지 않고 있으며, HITACHI는 필자의 지적에 따라 차후에 모듈처리할 예정이며, KODENSYA는 辭典처리를 하고 있으며 필자의 지적에 대해 차후의 과제라고 표명했다.

2. 6. 3 생략형(축약형)의 문제

지명,국가명의 혹은 일반화된 생략형 단어(일본어의 환경에는 일반성이 없다)의 환경과 접속어미의 생략된 환경에서 오역이 많았다. 한국어의 추상명사의 복수형이 일본어에서는 허용되지 않는 경우가 있다. 복수형의 접미사(-들)가 [ら],[達]의 variant이 있고 추상명사에 따라 복수형의 접미사를 분리해서 번역할 필요가 있다.

3. 결론

1. 최장일치법과 어절중의 출현 빈도를 우선적으로 고려하는 방식을 병용할 필요가 있다.
2. 동음이의어와 다의어의 고별적이고 구체적인 공기정보의 추출이 필요하다.
3. 오역의 유형을 고려한 프로그램의 개발을 추진할 필요가 있다.
4. 兩언어의 통사적 제약을 강화하고 兩國언어의 CORPUS 사전을 활용해야 한다

[참고문헌]

幸田香,1997 [日本語とドイツ語の動詞結合價の對照-母語干涉論の觀點から-] [外國語としての日本語]教育システム確立のための基礎的研究,pp56-64) (東京大學)
福井玲,1997 [韓國語研究におけるCD-ROMの利用について] [外國語としての日本語]教育システム確立のための基礎的研究, pp.75-79) (東京大學)
崔杞鮮,金泰完,1996 [日韓機械飜譯システムの現況および分析],言語處理學會 第2回年次大會 發表論文集,pp.433-443
姜龍熙,1996 [한.일 기계번역에 있어서의 오역 및 고찰],제8회 한글 및 한국어 정보처리 학술대회,pp351-366
姜龍熙,1997 [韓日機械飜譯における助詞の誤譯の問題],言語處理學會 第3回年次大會 發表論文集, pp.47-50