불-한 연어 데이터베이스 구축을 위한 굴절 정보의 처리

윤 애선* 정 휘웅** 권 혁철**
부산대학교 불어불문학과*, 인지과학 협동과정**, 전자전기정보컴퓨터 공학부**
부산 광역시 금정구 장전동 산 30번지
{asyoon, hckwon}@pusan.ac.kr

Processing of Inflectional forms for the French-Korean Collocational Database

Aesun Yoon* Hwi-Woong Jeong** Hyuk-Chul Kwon***
Department of French*, Interdisciplinary Program of Cognitive Science**
School of Electrical and Computer Engineering**
Pusan National University,
San 30, Jang-jeon-dong, Geum-jeong-gu, 609-735, Pusan
{asyoon, hckwon}@pusan.ac.kr

요약

구(phrase) 단위 또는 문장(sentence) 단위의 연어(collocation) 정보는 자연언어 처리를 위한 단일어 또는 이중어 데이터베이스를 구축할 수 있는 중요한 기초 자료가 될 뿐 아니라, 외국어 학습에서도 어휘 단계를 넘어선 학습 자료를 제공할 수 있다. 불어는 굴절 언어(inflectional language)로서 기본형 대 굴절형의 비율이 약 1:9 정도로 비교적 굴절 비율이 높은 언어다. 또한 불어 표제어 중 95% 이상을 차지하는 불어의 동사, 명사, 형용사 중 상당한 비율이 암기해야 할 목록(list)이라는 특성을 갖기 때문에 검색과 학습에 있어 오류가 지속적으로 일어나는 부분이다. 표제어의 검색의 경우 불어 굴절 현상을 지원하는 전자 사전이 개발되어 있지만 아직까지 연어 정보에서 굴절형을 지원할 수 사전 또는 데이터베이스는 개발되어 있지 않다. 본 연구의 목적은 전자 사전과 형태소 분석기를 이용하여 굴절형 처리를 지원할 수 있는 불-한 연어 데이터베이스를 구축하는데 있다. 이를 위해 부산대학교 언어정보 연구실에서 개발한 불어 형태소 분석기 *Inflection*와 불-한 전자 사전 *FranCo*를 사용하였으며, 지금까지 구축된 불-한 연어 정보는 94,965 개이다. 본 고에서는 두 정보를 이용하여 불어 굴절형 정보를 분석 및 생성하는 방식 및 불-한 연어 데이터베이스 구조를 살펴 본다.

1. 들어가기

언어 정보의 자동 처리를 위한 데이터베이스 구축은, 국제 무대에서 벌어지고 있는 정보화 시대에 들어서면서 단순한 정보 획득 뿐 아니라 다양한 산업의 부가가치를 재생산하는데 필수적인 분야로 인식하고 있다. 따라서 각국은 자국어를 자동 처리할 수 있는 사전이나 데이터베이스 구축은 물론이고, 경쟁 대상이 되고 있는 외국어와 자국어간의 이중어 또는 다중어 데이터베이스를 구축하고, 연구 단계를 지나 이를 산업화하는데 주력하고 있다. 국내에서도 한국어 데이터베이스의 필요성을 인식하고 어휘 및 구문 사전을 구축하고 있고, 한국어-외국어간 이중어 병렬 데이터베이스를 개발하고 있다. 하지만 오랫동안 많은 양의 이중어 및 다중어 데이터를 구성해 온 외국의 연구진과 언어 산업계에 비해, 후발 주자인 우리나라는 이중어 데이터베이스 구축에 있어 아직도 방어적인 차원에서 연구 개발이 이루어지고 있으며, 더욱이 개발되는 언어도 영어에 치우쳐 있고 부분적으로 일본어와 중국어 정도를 포함할 뿐이다[10]. 국제 언어 시장에서 주요 언어인 불어, 독어, 스페인어, 러시아어 등 이른바 제 2 외국어에 대한 언어 자료는 실험실 단위의 연구에 그치고 있다. 연구자의 규모, 연구 환경

및 지원이 매우 열악한 제 2 외국어 분야에서는 이미 개발된 자료를 재사용하거나 효과적으로 가공할 필요성이 더욱 높다.

국내 불어 정보화의 경우, 지난 10여년에 걸쳐 본 연구진에 의해, 인간 사용자를 위한 중간 규모의 전자 사전(*FranCo*)과 형태소 처리기(*Inflection*)가 개발되었고[2, 3], 이를 바탕으로 후속 연구가 계속 진행될 예정이다. 불어의 특성 상 전자 사전과 형태소 처리기의 언어 단위는 어휘이다. 하지만 자연 언어를 처리하기 위해서는 적어도 문장 단위의 분석이 필요하므로, 후속 연구는 구구조 규칙이나 어휘간 결합 단위에 대한 연구가 이루어져야 한다[5,8,12]. 인간 사용자를 위한 전자 사전의 경우, 한 표제어에 대한 내용 정보로 구(phrase) 단위 또는 문장(sentence) 단위의 연어(collocation) 정보를 상당수 포함하고 있다. 이러한 정보는 자연언어 처리를 위한 단일어 또는 이중어 데이터베이스를 구축할 수 있는 중요한 기초 자료가 될 뿐 아니라, 외국어 학습에서도 어휘 단계를 넘어선 학습 자료를 제공한다는 점에서 매우 유용하다[4, 11].

인간 사용자를 위한 전자 사전에 수록된 예문은 사전 편찬자의 직관이나 경험 또는 대용량 말뭉치에서 추출한 상당히 사용 빈도가 높은 연어 정보를 포함하고 있다[6,8]. 하지만 예문은 문장 단위로 제시되므로, 어휘의 굴절 형태가 사용된다.[1] 기 구축된 전자 사전의 예문에서부터 연어 정보를 정제(purification)하려면 적어도 기본형 정보를 필요로 한다. 이렇게 가공된 자료는 자연언어 처리용 기초 자료로 사용될 수 있고,[2] 외국어 학습에서 생산적 기능(productive skills)을 교수-학습하는 자료로도 가공할 수 있다. 표제어 검색의 경우, 불어 굴절 현상을 지원하는 전자 사전이 개발되어 있지만 아직까지 연어나 예문 정보에서 굴절형을 지원할 수 있는 데이터베이스는 개발되어 있지 않다[1].

본 연구의 목적은 전자 사전과 형태소 분석기를 이용하여 굴절형 처리를 지원할 수 있는 불-한 연어 데이터베이스(*F-K Collocation*)를 구축하는데 있다. 이를 위해 약 42 만개 굴절 변화형을 분석하고 생성할 수 있는 불어 형태소 분석기 *Inflection*와 53,650개의 표제어를 수록한 불-한 전자 사전 *FranCo*를 사용하였으며, 지금까지 구축된 불-한 연어 정보는 94,965 개이다. 본 고에서는 두 정보를 이용하여 불어 굴절형 정보를 분석 및 생성하는 방식 및 불-한 연어 데이터베이스 구조를 살펴 본다.[3] 2장과 3장에서 각각 *Inflection*에서 사용하는 형태소 분석 방식과 *FranCo*의 정보 구조 소개한 후, 4장에서는 *F-K Collocation* 구축 방식과 구조를 살펴 본다. 5장에서는 향후 연구 방향을 알아 본다.

2. *Inflection*의 형태소 처리

자연 언어의 형태소 처리 방식은 여러 가지가 있다. 이중 어느 하나가 절대적인 효율성을 갖는 것은 아니며, 어느 방법이 인간의 인지 모형에 가까운 지에 대한 논란도 계속되고 있다.[2,5,8,9] 형태소 처리 방법은 각기 다른 장 단점을 갖고 있다. 따라서 실제 자료의 형태소 처리 시스템을 구축함에 있어 대상 언어의 특성, 사용 목적, 자료의 크기, 개발에 필요한 시간 등을 고려하여 적절한 방법을 선택해야 한다.

불어의 굴절 현상을 살펴보면[5] 다음 3가지 특성을 갖는다[2]. (1) 불어에서 기본형(base form)으로부터 생성될 수 있는 어휘 또는 표현의 생산력은 합성(compound), 굴절(inflection), 파생(derivation)의 순으로 나타난다. 이 중 '양쪽에 공백이 있는 단어'는 파생과 굴절 과정을 통해 생성되며, 대부분의 합성어는 독일어와는 달리 단어간 공백을 가지므로 반드시 파생이나 굴절과 동일한 방법으로 처리할 필요가 없다. (2) 파생과 굴절 중, 본 연구에서 처리하고자 하는 굴절의 비율이 월등히 더 높다. (3) 불어의 기본형 대비 굴절 표면형의 비율은 약 9배 정도다. 따라서 기본형 어휘부의 크기가 5만 개 - 10만 개 정도라면 굴절 표면형 어휘부의 크기는 45만 개 - 90만 개 정도이므로,

---

[1] 불어는 굴절 언어(inflectional language)로서 기본형 대 굴절형의 비율이 약 1:9 정도로 비교적 굴절 비율이 높은 언어다. 또한 굴절 현상은 한국어 화자가 불어를 학습할 때 처음 부딪히는 어려움이자, 상당한 학습 단계에서 지속적으로 오류를 범하는 부분이다. 이는 불어 표제어 중 95% 이상을 차지하는 불어의 동사, 명사, 형용사 중 상당한 비율이 암기해야 할 목록(list)이라는 특성을 갖기 때문이다.

[2] 자연 언어 처리용으로 사용하려면 언어 정보에 태그가 달려 있어야 한다. 필요한 태그는 사용 목적과 개발 환경에 따라 상세한 분류에 있어서는 큰 차이를 보이나, 기본적으로 문법 범주 정보를 필요로 한다.

[3] 부산대학교 언어정보 연구실(http://langue.fr.pusan. ac.kr)에서는 53,650개의 표제어를 수록한 불-한 전자 사전 *FranCo*의 언어 정보 구축과 병행하여 47,265개 기본형의 420,543개 굴절 변화형을 분석하고 생성할 수 있는 형태소 분석기 *Inflection*을 개발하였다. 불-한 전자 사전 *FranCo*에는 *Inflection*이 연동되어 있어, 굴절 변화형의 분석과 생성을 지원한다. *Inflection*과 *FranCo*에 대한 자세한 기술은 윤애선(2000, 2001)을 보라.

[4] 형태소 처리 방법은 (1) 분절과 접합이 일어나는지, (2) 분절된 이형태(allomorpheme)가 형태소(morpheme)로 축약(reduction)되는지에 따라, 크게 3가지 표면형(surface form) 인식 방법, 형태소 인식 방법, 이형태 인식 방법으로 분류할 수 있다.[Hausser 2000]

[5] *Inflection*의 목적은 형태소 처리를 위한 범용적 시스템을 만드는 것이 아니라, 불어라는 특정한 언어에 적용할 시스템을 구축하고 이를 전자 사전 및 교육 프로그램 등과 같은 응용 프로그램과 연동하는 데 있다.

표면형 인식 방법을 사용해도 메모리와 검색 속도에 큰 영향을 미치지 않는다. 하지만 굴절형의 생성은 상당 부분을 규칙화할 수 있으므로[7], 모든 굴절 표면형을 열거하는 단순 파생형 인식 방법은 자료를 관리 및 확장하는데 효율적이지 못하다.

위와 같은 이유에서 *Inlfection*에서는 복합 파생형 인식 방법을 이용하였으며, 이에 따라 파생 과정으로 만들어지는 소수 신조어는 기본형 어휘부에 추가하는 방식으로 확장된다. *Inflection*에서 사용한 표면형 인식 방법은 텍스트를 구성하는 실제 단어형인 미분석어(unanalyzed word) 자체를 분절하지 않은 채 어휘 검색을 하여, 어휘부에 수록된 형태에 관한 분석적 정보(analyzed morphological information)를 부여한다. 표면형 인식 방법은 단순 방식과 복합 방식으로 구분할 수 있다. 단순 방식은 인공 언어같이 어휘의 수가 적을 경우, 모든 사용되는 표면형을 동일한 단계의 어휘부에 나열할 수 있다. 하지만 자연 언어의 경우, 실제 사용되는 표면형의 수가 50-60만 개를 쉽게 넘으며, 이중 다수가 기본형으로부터 규칙에 의해 생성될 수 있으므로 표면형을 한 단계에서 나열하는 것은 어휘부를 유지하거나 확장하는데 비효율적이다. 따라서 기본형 어휘부 단계와 표면형 어휘부 단계를 분리하는 복합 방식을 사용하는데, 기본형 어휘부에 굴절이나 파생 규칙을 적용하여 모든 표면형을 만들고 이를 어휘부에 등록하여 어휘 검색에 사용한다[6].

불어에서 굴절 형태의 거의 전부를 차지하는 범주는 명사, 형용사, 동사로, 명사와 형용사는 성 수의 표지를, 동사의 단순형(simple form)은 인칭, 시제, 법의 표지를 갖는다. 형태소 분석 및 생성을 위해서 *Inflection*에는 크게 3개의 기초 어휘부가 존재한다. 하나는 어간을 위한 '어간 어휘부'이고, 다른 하나는 기본형을 저장하기 위한 '기본형 어휘부'이다. 이런 어간 어휘부와 기본형 어휘부는 기본형 및 어간이 입력되면서 구성되고 유기적으로 연결된다. 하나의 어간에 하나의 어미만이 존재하는 것은 아니다. 따라서 어간과 기본형의 연결은 다중으로 연결될 수 있으며, 이를 위해 기본형과 어간을 연결하는 '규칙 어휘부'를 설계하였다. 이 규칙 어휘부는 기본 어휘부의 각 노드에 포함되어 있어서, 기본형 어휘부에 어간 어휘부를 연결하는 경우와 어간 어휘부에 기본형 어휘부를 연결하는 경우 모두를 위해 사용된다. 이 규칙 어휘부에는 '어미 접합 규칙'[6]이 제공된다. 이를 위해, *FranCo*의 표제어와 품사 정보를 기초 자료로 사용하여, 명사 29,059개, 형용사 9,957개, 동사 8,249개를 분석하였다. 그 결과, 본 연구에서 명사 57개, 형용사 53개, 동사 123개로 어미의 접합 규칙을 분류하였다[2].

3. *FranCo*의 정보 구조

인간 사용자를 위한 전자 사전를 만드는 방법은 크게 두 가지로 구분할 수 있다. 하나는 원시 말뭉치(corpus brut)를 구축하는 것처럼 아무런 표지(marque)없이 모든 자료를 단순히 디지털화하는 것이고, 다른 하나는 디지털화된 자료에 그 내용과 형식을 구분하는 표지를 다는 방법이다. 물론 후자는 시간이 많이 걸리는 방식이지만, 이렇게 구축되면 쉽고 빠르며 효율적인 검색이 가능하다. 전자 사전이 이러한 방식으로 구현되기 위해서는, 사전의 의미있는 정보를 분류하여 논리 구조를 결정하고, 이를 바탕으로 문서구조를 규정하여야 한다.

*FranCo*의 표제어는 두 단계로 구분되는데, 하나는 상위 단계에서 정렬키 역할을 하는 정렬 표제어[7]이고, 다음 하나는 하위 단계에서 품사에 따라 구분된 하위 표제어이다.[8] 각 표제어의 하위 정보를 구성하는 미시구조(micro-structure)는 [표 1]과 같다.[9]

[표 1] *FranCo*의 미시구조

어휘기술 항목	특성
구분자	정렬을 위한 표제어 내부의 동철이의어와 같은 하위 표제어 구분
발음기호	하위 표제어의 발음 (음성기호, 음성)
품사	하위 표제어의 품사
기본형	굴절 또는 파생어 관계
동의어, 반의어, 상위어, 하위어	하위 표제어의 세부 의미가 다른 어휘와 갖는 의미적 관계

[6] 본 고에서 굴절 규칙'은 전통 문법의 용어대로 인간을 위한 굴절 정보 규칙을 지칭하고, 어미 접합 규칙'은 본 연구에서 사용하는 역검색(reverse search, right-to-left search)'을 이용한 트리밍(trimming)이 가능하도록 어미 구분을 바탕으로 만든 규칙이다.

[7] 인간 사용자를 위한 전자 사전에서는 표제어가 정렬키 역할을 하므로, 두 개의 동일한 표제어가 등재될 수 없다. 따라서 이러한 동철이의어를 구분하기 위해 전자 사전에서는 정렬을 위한 단계를 따로 두어 정렬 표제어를 설정한다.

[8] 표제어를 결정하는 거시 구조(macro-structure) 문제는, 전통적인 사전에서부터 다의 관계(polysemy)와 동철이의 관계(homonymy)를 구별하는 문제 및 사전 내의 정렬 문제와 긴밀히 연관되어 왔다. 전통사전에서는 한 표제어가 많은 뜻을 가지면 그것들의 의미적 거리와 구조를 반영하는 방법으로 분류되고 정렬된다. 동철이의어는 어원, 문법적인 분류나 의미를 기초로 결정되고 따로 떨어진 표제어들로 제시되며, 여러 표제어간 동철이의 관계는 위첨자나 숫자 등으로 표시된다.

[9] 사전의 용도, 언어적 특성, 개발 인력 및 시간 등에 따라 그 범위가 제한된다. 전통 종이 사전의 경우, 한 표제어에 대한 어휘 기술은 일반적으로 발음기호, 품사, 형태적 특성, 통사 정보, 의미 자질, 화용적 기술, 어원, 목표 언어(target language)로의 번역 예, 상투 어구, 기타 다른 용법들로 구성된다.

전문 영역	하위 표제어의 세부 의미가 사용되는 전문 영역
의미	목표 언어로의 의미 기술이며, 어휘 기술항목의 다른 층위에서 반복적으로 출현 가능
구문 정보, 연어 정보	해당 표제어가 통사적 지배소(governor)인 경우, 의존소(dependant)와의 관계 제시되며, 해당 표제어의 세부 의미가 다른 어휘와 갖는 의미적으로 긴밀한 결합 관계을 갖는 경우, 원시 언어와 목표 언어 정보가 동시에 제공됨.
예문 정보	해당 표제어의 세부 의미가 사용되는 문맥이 문장 단위로 제시되며, 원시 언어와 목표 언어 정보가 동시에 제공됨.

현재 *FranCo*에 수록된 연어 정보는 8,195개이고 예문 정보는 86,770개로, 후자의 수가 월등히 많다.[10] 따라서 예문 정보를 가공하면, 매우 정제된 연어 정보와 구문정보를 추출할 수 있을 것이다. 물론 이 정제 작업은 태그의 세분화 정도에 따라 인간이 직접 분석해야 할 수작업의 양과, 연어 정보 데이터베이스의 질이 좌우된다. 따라서 *F-K Collocation*의 구축은 방대한 시간이 걸리므로 여러 단계에 걸쳐 개발하고 각 단계의 결과를 이용하고자 한다.

4. F-K Collocation의 구축

본 고에서는 *F-K Collocation* 구축의 첫 단계로, (1) *FranCo*로부터 이중어 예문을 추출하고, (2) 각 예문을 구성하는 어휘에 범주(parts of speech) 정보 및 기본형 정보를 제공하여, 이로부터 연어 정보의 정제를 용이하게 하기 위한 것이다. 불어의 굴절 변화형은 대명사, 명사, 형용사, 동사에서 나타나는데, *FranCo*에 등재된 각 품사의 기본형과 굴절 변화형의 수 및 비율은 [표 2], [표 3]과 같다.[11]

[표 2] *FranCo*에 수록된 굴절형 품사의 비율

	표제어(기본형)의 수	전체 어휘 대비 해당 품사 비율
FranCo[12]	49,857	100.0%
일반명사	29,059	58.9%
형용사	9,957	20.0%
동사	8,249	16.5%

[표 3] 품사별 기본형 대비 굴절형 비율

	기본형의 수	굴절형의 수	기본형 대비 굴절형 비율
일반 명사	29,059	63,425	2.18
형용사	9,957	39,658	3.98
명사	8,249	317,460	38.48
계	47,265	420,543	8.90

기본형대비 굴절 변화형의 비율이 높기 때문에, 불어에는 굴절 변화형 간에는 예문 (1-a) ~ (2-e)와 같은 동철이의 관계가 높은 비율로 발생한다.

[표 4] 굴절변화형의 동철이의 현상

	예문 정보	품사	기본형
(1-a)	Je **suis** étudiant.	V	être
(1-b)	Je **suis** la même itinéaire de Paul.	V	suivre
(2-a)	Ce tableau ne trouve pas le **pendant**.	N	pendant
(2-b)	**Pendant** que tu es, ramène mes affaires à la maison.	Conj.	pendant
(2-c)	Il a su ma maladie, mais ni **pendant** ni après il n'est pas venu me voir.	Adv.	pendant
(2-d)	Le people a fait un protestation contre les Etats-Unix, en **pendant** le president Bush en effigie.	V	pendre
(2-e)	C'est un procès **pendant**.	Adj.	pendant

따라서 이러한 예문을 가공하여 연어 정보 데이터베이스를 만들기 위해서는, 우선 굴절 변화형의 동철이의 관계에서 오는 중의성(ambiguity) 문제를 해결해야 한다. 이를 위해 *FranCo*를 구성하는 정보와 *Inflection*이 제공하는 기본형 및 굴절형 정보를 이용한다.

4.1. *FranCo* 정보의 추출 및 인덱싱

*F-K Collocation*을 구축하는 기초 자료는 *FranCo*에서 제공되나, 연어 정보로 정제되는 결과 및 확장 정보를 다시 *FranCo*에 사용하기 위해서는 단순한 예문 정보 추출이 아니라 *FranCo*의 위치 값을 보유할 수 있는 형태로 이루어져야 한다. 자료의 추출은 별도의 Microsoft XML parser 3.0을 이용하였다. XML parser의 X-pointer 명령을 이용하여 추출된 예문 정보는 우선 첫

[10] *FranCo*의 예문 정보는 기존 종이 사전의 예문 뿐 아니라 일반 텍스트에서 인용 또는 추출하여 이중어 예문으로 등재된 것이 많아, 연어 정보로 환원되어 있지 않다.

[11] 현대 불어의 대명사는 그 수가 적고, 굴절형을 규칙화하기에 효율성이 적은 목록의 특성을 지니므로, 대명사의 굴절형 자체가 *FranCo*에 표제어로 등록되어 있다.

[12] *FranCo*에는 자주 사용되는 고유 명사 3,793개를 포함하고 있으므로 49,857은 이를 제외한 수이다. 이 밖의 고유명사는 독립된 다른 사전으로 등재되어 따로 관리한다.

번째 예문정보 테이블에 저장된다. 하나의 예문 정보가 저장될 때, 스페이스 문자를 기준으로 각 표제어 정보를 분리하여 일련번호를 부여하고, 표제어 테이블에 부여된 일련번호와 함께 저장한다.

[표 5]는 *FranCo*에서 추출된 각 정보의 구성을 보여준다. 추출된 자료는 분석의 편의성 및 속도를 고려하여 관계형 데이터베이스에 저장하였다. 분석 정보는 예문정보와 그 위치를 저장하는 테이블과, 각 예문 정보에서 추출된 단어 정보와 [표 6]과 같이 변화형 정보를 XML 포맷으로 저장하는 테이블로 나누었다. 테이블을 두 개로 나눈 것은 한 개의 테이블에 저장할 때 예문 정보와 분석정보를 하나의 XML 정보에 저장해야 하나, 이 경우 사용자의 가독성이 떨어지고, 표제어 검색시 각 XML정보를 분석해야 하므로, 검색 속도를 저하시키기 때문이다.

[표 5] 예문정보 테이블 구조

필드명	특성
fldID	단어 정보와 연결되기 위해 하나의 예문 정보에 부여되는 고유값
fldENTRY_ID	*FranCo*에 저장된 표제어의 고유값
fldENTRY	자료 검색의 효율성을 위해 *FranCo*에서 추출된 표제어 정보
fldEX_Pointer	사전 자료내 예문의 위치를 설명하는 값. 이 값은 표제어 정보에 따라 유동적이다.
fldEX_F	불어 예문 정보
fldEX_K	불어 예문의 한글 번역 정보

[표 6] 추출된 표제어 테이블 구조

필드명	특성
fldEX_ID	연계된 상위 예문의 ID
fldSEQ	문장 내부에서 추출된 단어의 위치를 시스템이 자동으로 부여하는 번호.
fldENTRY	실제 예문에 저장되어 있는 정보
fldINF	형태소 분석이 된 정보

4.2. *FranCo*와 *Inflection*을 이용한 기본형 및 품사 태깅

*FranCo*에서 추출된 정보는 시스템 메모리에 저장된 *Inflection* 정보를 호출하여 자신의 기본형 정보를 찾는다. 기본형이 한 개만 존재하는 경우 테이블의 구조는 두 개로 구성할 수 있으나, 품사 및 기본형이 두 개 이상 검출되는 경우에는 별도의 테이블을 구성해야 한다.[13] XML로 저장된 정보는 검색의 효율성을 위해 기본형 검색 전용 테이블을 [표 7]과 같이 구성하였다.

[표 7] 검색용 테이블

필드명	특성
fldENTID	관련된 단어 정보. 품사가 하나 이상인 경우를 대비하여 필드내 중복값을 가질 수 있다.
fldPOS	품사 정보
fldBF	기본형 정보. 인덱싱 되어 빠른 검색을 보장한다.

이 구조를 활용할 경우 비록 검색을 위한 테이블을 위해 추가 코드와 데이터베이스 저장 공간, 데이터베이스 인덱싱 시간을 요구하나, 추출된 단어 정보와 예문 정보를 역추적하여 관련된 정보를 빠르게 검색할 수 있는 장점이 있다. 이는 사용자에 대한 반응시간을 최소화하여, 이용상 편리성을 극대화할 수 있다.

4.3. 활용: 연어 및 예문 검색

F-K Collocation을 이용하면 웹 환경에서 사용자들이 자신이 원하는 형태로 사전의 표제어 및 예문을 검색할 수 있다. 기본형 정보를 바탕으로 관련된 모든 예문을 검색할 경우 다음의 순서에 따라 검색된다.

① 검색어(변화형 혹은 기본형)를 입력하고, 사용자는 자신의 검색 옵션을 선택한다. (변화형 검색 및 품사 선택 여부)
② *Inflection*에 의해 변화 가능한 형태를 검색한다. 변화형이면 기본형을 분석한 뒤, 다시 모든 굴절 변화형 목록을 추출한다.
③ 검색 테이블을 바탕으로 기본형 정보 관련 자료를 검색한다.
④ 사용자의 옵션에 따라 추출표제어 테이블을 바탕으로 변화형 정보를 검색한다.
⑤ 두 검색 결과를 바탕으로 최종 예문 정보를 검색한다.

4. 마치면서

본 연구의 목적은 전자 사전과 형태소 분석기를 이용하여 굴절형 처리를 지원할 수 있는 불-한 연어 데이터베이스의 구축 가능성을 살펴보는 것이며, 지금까지 구축된 불-한 연어 정보는 94,965 개이다. 형태소 분석기나 전자 사전의 설계 및 구현, 또 이것을 응용 시스템에 연동하는 것 자체는 결코 새로운 개념은 아니다. 국외에는 수많은 자연 언어 처리 기초 도구 및 자료가 개발되어 있으나, 이들은 대부분 경제적, 법적 때로는 기술적인 이유로 국내에서 사용하기 어렵다[1].

[13] 이렇게 추적된 품사 정보가 2개 이상일 때 중의성 제거가 필요하다. 중의성 제거는 통사 규칙와 인접 단어의 품사 정보를 이용한 품사의 자동 태깅으로 수작업을 최소한으로 줄여 효율성을 기하고 있으며, 이 부분에 대해서는 후속 논문으로 발표할 예정이다.

하지만 형태소 분석기는 전자 사전과 함께 자연언어를 자동처리하기 위해 반드시 필요하다.

F-K Collocation은 현재 품사와 기본형이라는 정보만이 태깅된 매우 기초적 단계에 머물러 있다. 앞으로 이를 사용하는 목적에 따라 상세한 태그 세트의 설정과 이를 F-K Collocation에 적용하게 될 것이다. 또한 연어 정보를 구문 정보로 가공할 수 있는 가능성의 모색이 필요하다.

참고 문헌

[1] 윤애선. 1998. 불어-한국어 정보화 환경. 불어불문학 연구 36집, pp. 483-500.

[2] 윤애선. 2000. 불어 굴절 변화형의 자동 분석-생성기. 불어불문학 연구 43집. pp. 365-392.

[3] 윤애선. 2001. 표준화된 전자사전 개발을 위한 정보구조 및 사용자 환경 설계와 그 응용. 한국프랑스학 논집 33집. pp. 41-58.

[4] 윤애선/이은정. 1997. 통신망을 이용한 외국어 학습(1): 사용자 요구 및 환경 분석을 중심으로. 불어불문학 연구 34집. pp. 607-623.

[5] 이기용. 1999. 전산 형태론. 고려대학교 출판부. 서울, 328 p.

[6] Courtois, B. 1990. Un système de dictionnaire électronique pour les mots simples du français, *Langue Française 87*, pp. 11-22.

[7] Degoud, N. ed. 1995. *Conjugaison*, Larousse. Paris, 191 p.

[8] Hausser, R. 1999, *Foundations of Computational Linguistics: Man-Machine Communication in Natural Language*, Springer. Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Pais, Singapore, Tokyo, 534 p.

[9] Karttunen, L./K.R. Beesley. 1992. *Two-Level Rule Compiler*. Xerox Palo Alto Research Center. 40 p. (ISTL-92-2)

[10] Kwon, H.Ch. 1997. Current status and trend of NLP Technologies in Korea. *Korea-France Joint Workshop on Language Industries*. pp. 7-11.

[11] Levy, M. 1997. *Computer-Assisted Language Learning: Context and Conceptualization*. Oxford University Press. Oxford.

[12] Zinglé, H. 1999. *La Modélisation des langues naturelles: Aspects théoriques et pratiques (Travaux du LILLA, numéro spécial)*, Presse Universitaire de Nice-Sophia Antipolis. Nice [France]150 p.

<형태소 분석기 데모 웹사이트>

Inflection: http://parole.fr.pusan.ac.kr/inflection
FranCo: http://parole.fr.pusan.ac.kr/franco

* 본 연구는 한국 학술진흥재단의 2000 선도연구 지원사업 (과제번호 A00359)의 지원을 받아 이루어졌음을 밝힌다.