日本語學硏究 第26輯 (2009. 12. 20. pp.133~146)

NARA ver. 2.0 : 세종 한일 병렬 말뭉치 검색 시스템*

宋相憲**

〈 요 지 〉

병렬 말뭉치는 비교 언어학, 언어 교수법, 자연어 처리 등의 분야에 유용한 자원으로서 그 중요성이 날로 증대하고 있는 상황이다. 그러나 대부분의 연구자들이 병렬 말뭉치를 자신의 연구에 곧바로 적용하기는 어려운 측면이 있다. 이는 무엇보다 병렬 말뭉치를 손쉽게 사용할 수 있는 도구의 부재에 기인한다. 본고의 목적은 2007년까지 구축된 자신의 연구에 곧바로 적(형태소 주석 버전)를 언어 연구자, 교수자, 그리고 학습자들이 간편하게 용하할 수 있는 온라인 는 어을 소개하고, 그 개발 과정의 고려 사항을 되짚어 보는 것이다. NARA라고 명명된 이 용하기환경은 일반적인 인터넷 용하기페이지넷 유사한 인터페이스를 지니고 있으며, 데이터베이스로 구현되어 온라인에서 서비스되으며괬계로 '언제', '어디서나', '누구나' 사용할 수 있구현되특징을 지닌다. 아울러, 연구자의 요구에 맞게 그 용하기방법을 다양화할 수 있구현되것은 큰 장점이 되며, 검색의 결과를 별도로 저장할 수 있는 기능을 두어 연구자의 편의를 도모하였다.

논문분야 : 코퍼스 기반 일본어학 연구
키 워 드 : NARA, 세종 한일 병렬 말뭉치, 코퍼스, 검색 시스템

1. 서론

Abeillé (2003)는 언어 연구에 있어 언어자원이 가지는 역할에 대하여 두 가지 측면에서 살피고 있다. 하나는 '어떻게 구축할 것인가 (how to build)'이고, 다른 하나는 '어떻게 사용할 것인가 (how to use)'의 문제이다. 이는 실제 언어자원에 관한 연구에서 '구축'과 동시에 '사용'의 측면이 동시에 고려되어야 함을 의미한다. 언어자원이란 실제 연구에서 활용되는 참고자료 또는 판별기준이며, 동시에 어떠한 시스템 등을 만드는 데 있어 그 근간이 되는 대상이 된다. 따라서 구축 단계에 있거나 또는 구축이 완료된 언어자원을 어떻게 사용할 것인가 하는 문제는 자료기반 언어연구 전반에 있어 늘 고려의 대상이 되어야 한다. 다른 한편으로

* NARA라는 명칭은 한국어와 일본어의 두 측면을 비교적으로 살펴볼 수 있는 가교 역할을 수행한다는 점을 상징적으로 드러낸다. 본 논문은 제20회 한국일본어학회 학술발표회에서 발표한 것을 수정 및 가필 한 것이다. 본 논문의 많은 부분은 저자가 BK21 글로벌 인턴쉽 수혜 대상자로 선정되어 일본 NiCT Language Infrastructure Group (情報通信硏究機構 言語基盤グループ)에서 진행한 연구에 힘입었다. Francis Bond, Kiyotaka Uchimoto, Kentaro Torisawa, Takayaki Kuribayashi, Miwa Shiga 및 여러 NiCT 관계자분들께 감사드린다. 본 연구의 기획 및 발전 단계에서 고려대 언어학과 최재웅 교수님과 고려대 일어일문학과 이한섭 교수님으로부터 많은 도움을 받아왔다. NARA ver. 1.0은 2009년 4월 14일 교토대학의 한국어 회화 클럽에서 구두 발표되었다. 교토대학의 Yukinori Takubo 교수님, TIDA Syuntaroo 박사님, 그리고 Yoshihiko Asao로부터 여러 조언을 얻어 ver. 2.0으로 발전시키는 기회를 얻었다. 아울러 본 연구의 순수 기술적인 부분은 2009년 8월 6일 싱가포르 Suntec에서 진행된 ACL 2009 Linguistic Annotation Workshop에서 포스터 발표된 바 있다. 본 논문을 작성하는 단계에서 대진대학교 교양학부의 Hasegawa Yumi 선생님과 고려대 언어학과의 Kuroyanagi Shigeo로부터도 도움을 받았음을 밝힌다.

** Dept. of Linguistics, Univ. of Washington. 박사과정. 전산언어학 (한일/한영 기계번역) 전공

Abeillé (2003)에서는 '주석'과 '도구'의 문제를 지적한다. '주석'은 그 해당 언어 자원에 어떠한 정보를 부가할 것인가에 대한 사항이고, '도구'는 그 언어자원을 어떠한 환경에서 구축하거나 사용할 것인가의 문제이다. 이는 '구축' 뿐만 아니라 '사용'의 각 영역에 동시에 적용되는 사안이라 할 것이다. 즉, 언어자원을 실제로 사용하는 데 있어서 어떠한 주석 정보를 검토할 것이며, 그 정보를 어떠한 방식으로 살펴 볼 것인지가 결정되어야 한다.

바로 이러한 점이 본고의 주된 대상이다. 본고에서는 21세기 세종계획의 일환으로 구축되어 배포된 '세종 한일 병렬 품사 주석 말뭉치'를 대상으로 하여, '사용'의 입장을 논의하고자 한다. '주석'에 해당하는 부분은 실제 말뭉치에 부착된 각 정보 (품사, 형태소, 어절 등)를 대상으로 할 것이며, '도구'에 해당하는 부분은 인터넷을 이용한 온라인 검색 시스템을 기반으로 할 것이다.

2. 연구의 내용

본고에서 제안하는 시스템의 기획의도에 대해 우선 살펴보기 위해 Abeillé (2003)에서 언급된 사항을 정리하는 아래 표를 참조하도록 하자.

〈표1〉 말뭉치 연구의 여러 단계

	구축	사용
주석	A	B
도구	C	D

대부분의 말뭉치 연구는 위 표1에서 A, 다시 말해 말뭉치에 어떠한 주석을 부착할 것인가의 측면에 큰 초점을 두고 진행하여 온 바가 크다. 이는 말뭉치 연구가 그 정보를 풍부하게 기술하여 구축되어야만 여타의 추가 연구가 가능하다는 점에서 타당하지만, 실제로 말뭉치를 사용하는 입장에서 보면 D의 부분이 매우 중요하다. 말뭉치를 올바르게 그리고 손쉽게 사용할 수 있는 도구가 뒷받침되어야만 그 구축된 말뭉치의 진가가 발휘될 수 있기 때문이다. 본고에서 제안하는 온라인 검색 인터페이스는 주로 위 D에 해당하는 연구이다. 또한, 그 도구를 사용하기 전에 어떠한 주석 정보를 연구를 위해 사용할 것인가가 (B) 결정되어야 하기 때문에 그 관련 역시 살펴볼 것이다. 더하여 말뭉치가 구축될 당시 어떠한 도구가 사용되었는지 (C) 역시 고려의 대상이 된다. 아래 표2는 위 표1을 본고의 내용에 맞게 재편한 것이다. 아래에서 '세종 한일 병렬 말뭉치' 부분은 지난 기간 때안 21세기 세종계획의 일환으로 구축된 결과이며, 'NARA'는 본 연구의 결과이다.[1)]

〈표2〉 말뭉치 및 NARA ver. 2.0 구축 및 사용

	구축		사용
주석	세종 한일 병렬 말뭉치	단락 및 문장 단위 정렬	형태 분석
	NARA	구 단위 및 단어 단위 정렬	
도구	세종 한일 병렬 말뭉치	세종 품사부착 도구 (한국어) ChaSen (일본어)	온라인 검색 인터페이스
	NARA	GIZA++/ Moses / BLEU	

1) 필자는 '세종 한일 병렬 말뭉치' 구축의 과정에는 참여하지 않았음을 밝힌다.

3. 병렬 말뭉치

초기 본래의 텍스트를 그대로 활용하던 원시 말뭉치 연구에서 벗어나 말뭉치에 다양한 주석을 삽입하게 되면서, 그 목적과 필요성에 따라 다종의 말뭉치가 출현하게 되었다. 대표적인 것이 각 단어의 품사 정보를 기술한 품사 부착 (POS-tagged) 말뭉치, 문장의 구문 정보를 구조적으로 기술한 트리뱅크 등이 있다. 이 연장 선상에서 비교적 근래에 출현하여 각광을 받고 있는 주석 말뭉치가 바로 병렬 말뭉치이다.[2)]

병렬 말뭉치는 말 그대로 두 개 이상의 언어 문장 또는 텍스트를 병렬적으로 구성하여 각 문장이 다른 언어의 문장과 어떻게 연결되는지에 대한 사항을 주석 처리한 것이다. 다시 말해, 병렬 말뭉치는 원시 언어(Source Language)의 각 언어 단위가 대상 언어 (Target Language)의 어떠한 부분과 동등한지를 기술한 언어자원을 말한다. 여기서 서론에서 밝힌 바와 같이 이러한 기술을 어떠한 준거에서 수행하는가에 관련된 '주석'의 문제가 제기되는데, 본고의 대상이 되는 세종 한일 병렬 말뭉치는 XML에 기반한 인덱싱 방식을 택하고 있다. 아래는 그 실례이다.

(1)〈linkGrp domains="1.1.p625 ; 1.1.p623"〉
〈link xtargets="1.1.p625.s1;1.1.p623.s1"〉
〈link xtargets="1.1.p625.s2;1.1.p623.s2"〉
〈/linkGrp〉
(2) 〈s id=1.1.p623.s2〉
나는 나/NP+는/JX
그냥 그냥/MAG
웃었다. 웃/VV+었/EP+다/EF+./SF
〈/s〉
(3) 〈s id=1.1.p625.s2〉
私 私/NNPG
は は/PRE
笑っ 笑う/VIN
た た/AU
。 。/SYF
〈/s〉

병렬 말뭉치에서 핵심이 되는 요소는 양 언어의 언어 요소를 어떠한 층위에서 상호 연결할 것인가에 대한 결정이다. 자연언어에는 형태소, 단어, 구, 절, 문장, 및 단락 등과 같은 다양한 층위가 존재하는 데, 어떠한 층위에서 병렬 구성을 할 것인가에 대해서는 각 연구의 목적 마다 상이하다. 통상의 병렬 말뭉치는 문장 단위 정렬을 기본으로 하는데, 무엇보다 문장 이하 단위에서 병렬 말뭉치를 구성하는 것의 난이도가 상당히 높기 때문이다 (Abeillé 2003). 본고의 대상이 되는 '세종 한일 병렬 말뭉치'의 경우에도 위 (1-3)의 각 인덱스

2) 병렬 말뭉치 (Bilingual Corpus)는 경우에 따라 다국어 말뭉치 (Multilingual Corpus) 또는 평행 말뭉치 (Parallel Corpus) 등으로 불리기도 한다.

정보가 나타내는 바와 같이 문장 단위 정렬을 골자로 하고 있다.

3.1 세종 병렬 말뭉치[3)]

이후 논의의 편의를 위해 우선 세종 병렬 말뭉치의 특성에 대해 알아보도록 하자. 세종 병렬 말뭉치가 여타의 병렬 말뭉치[4)]와 차별점을 보이는 주요 특성은 아래와 같다.[5)]

- 대표성 : 세종 말뭉치 구축의 가장 근본적인 목표는 한국어에 대한 균형 말뭉치 (balanced corpus)를 만드는 것이었다. 이는 병렬 말뭉치 구축에 대해서도 정확하게 적용된다. 대표성을 지니는 말뭉치를 구축하기 위해 가장 긴요한 점은 다양한 장르를 포괄하여야 한다는 점과 실제 언중이 사용하는 텍스트를 대상으로 하여야 한다는 점이다. Europarl (Koehn 2005) 과 같은 병렬 말뭉치가 특정 장르 (예: 신문기사)만을 대상으로 하고 있다는 점에서 차이가 있으며, Tanaka 말뭉치 (Bond et al. 2008) 등이 기술하는 당사자가 문장을 임의로 만들거나 번역한다는 점을 감안하면 세종 병렬 말뭉치의 1차적인 특성이 여기에 있다고 하겠다. 다양한 장르를 다루는 점은 한국어의 특성을 올곧게 반영한다는 점에서 분명한 장점이 되는 것은 분명한 사실이나 반대로 그 분포적 특성이 넓고 얕아진다는 점에서는 단점이 될 수도 있다. 실제 텍스트를 사용한다는 측면도 마찬가지이다. 장점은 언어자원의 실제성을 높여 보다 견고한 처리를 가능하게 한다는 것이지만, 동시에 문장 대 문장 단위의 정렬이 어려워져 일부 처리 효율을 저하시키기도 한다.
- XML : 앞서 살핀 (1-3)에서 보는 바와 같이 각 말뭉치는 3개의 파일로 구성된다. ‘al’ 확장자가 붙은 정렬 XML 파일과 원시 언어 XML 파일, 그리고 대상 언어 XML 파일이다. 병렬 말뭉치가 통상 문장 대 문장을 순차적으로 기술하는 것과는 사뭇 다른 부분이다. 이러한 구성은 병렬 말뭉치의 구조를 데이터베이스로 전환하기에 용이한 환경을 제공한다.
- 의역체 : 이는 위에서 말한 바와 같이 실제 텍스트에 기반하였기에 나타나는 현상이다. 즉, 구축대상을 선정하는 과정에서 대상 언어에서 의미가 그 언어 직관에 비추어 자연스럽도록 의역체에 가까운 문장 대응을 선택하였기 때문이다. 이 역시 매끄러운 번역이라는 점에서는 분명한 장점이지만, 동시에 효율성의 측면에서는 때로 부정적인 영향을 결과에 미치기도 한다.
- 어절/분절 : 한국어의 형태소 주석은 세종 형태 주석 지침을 따르고 있으며, 일본어의 경우에는 ChaSen을 기반으로 구축되었다. 따라서, 한국어 분석의 기본 단위는 어절이며 반면 일본어 분석의 기

3) ‘세종 한일 병렬 말뭉치’ 자체가 본고의 논의 대상은 아닌 관계로, 여기에서는 이후의 논의의 과정에서 참조해야할 주요 사항만을 간단히 정리하고자 한다. 병렬 말뭉치에 대한 보다 자세한 정보는 ‘한일 병렬 말뭉치 개발 말뭉치 관련 지침’ (국립국어원 2007)을 참조하기 바란다.

4) 다른 일본어 관련 병렬 말뭉치 가운데 대표적인 것을 보자면, 우선 여행자용 예문을 모아 다국어로 번역한 BTEC (Business Travel Expression Corpus)을 들 수 있다. 이는 전형적인 다국어 말뭉치로서 한-일 병렬 이외에 영어, 중국어 등의 다양한 언어를 대상으로 하고 있다. 또한 Tanaka 말뭉치도 활용도가 높은 일본어 기반 병렬 말뭉치이다. 이는 영어 및 일본어의 각 문장을 그 대응 언어로 번역한 것을 모아둔 것으로서 온라인을 통해 무료로 입수할 수 있다는 점에서 큰 장점을 지닌다. 보다 자세한 사항은 Bond et al. (2008)을 참조하기 바란다.

5) 말뭉치는 그 구축 방법과 목적에 따라 상호 구별되는 특성을 지니기에 마련이다. 따라서 장점만을 지닌 말뭉치도 존재하지 않으며, 반대로 단점만을 지닌 말뭉치 역시 존재하지 않는다. 결국 어떠한 말뭉치를 선택할 것인가는 본질적으로 연구자의 몫이라 하겠다. 여기에서 언급된 ‘세종 한일 병렬 말뭉치’의 특성 또한 같은 맥락에서 이해해 주길 바란다.

본 단위는 분절이 된다. 이는 각 언어 이론의 기본적 관점을 수용한 측면에서는 바람직하지만, 동시에 양자가 동등한 수준에서 연결되지 않는다는 점에서 때때로 처리의 어려움을 낳는다.

4. NARA 시스템[6)]

NARA로 명명된 본 시스템의 특성에 대해 압축적으로 설명하자면, 인터넷을 통해 세종 한일 병렬 말뭉치를 검색할 수 있도록 구성된 시스템이라고 할 수 있다. 서론에서 밝힌 바와 같이, 언어자원은 구축하는 것만큼이나 그것을 능동적으로 사용하는 것이 중요한데, 병렬 말뭉치는 비단 구축뿐만 아니라 그 사용이 쉽지 않은 자원이다. 이는 내부적으로는 복수의 이질적인 언어가 인덱싱(indexing)의 형태로 정렬된 구조로 제공되기 때문에, 이를 상호 참조할 수 있는 도구를 개발하는 일이 쉽지 않기 때문이다. 아울러 외부적으로는 병렬 말뭉치를 연구자 자신의 분야에 곧바로 적용하는 연구를 계획하기가 쉽지 않다는 점도 원인이 된다. 본 NARA 시스템은 위와 같은 연구자의 어려움을 극복하는 데에 그 기본적인 목적이 있다. 다시 말해, 우선 한국어-일본어 교차 언어 정보를 손쉽게 얻을 수 있는 도구를 제공하는 것이며, 동시에 한국어-일본어 기계번역을 위한 기초 연구 환경을 마련하는 것이다. 2009년 9월 현재 사이트의 주소는 아래와 같다.

- http://corpus.mireene.com/nara.php

이어지는 각 소절에서는 앞서 살핀 표1과 표2의 각 내용을 순차적으로 살펴볼 것이다. 4.1에서는 주석 정보의 내용과 도구를 통한 그 확장 방법론에 대해 개괄할 것이다. 이는 주로 위 그림1에서 A와 C에 초점을 둔 것이다. 반면에 4.2는 B와 D에 초점을 두어, 본고의 최종적 지향점인 온라인 검색 인터페이스에 대한 전반적인 소개를 할 것이다. 먼저 각 주석 정보를 검색의 과정에서 어떻게 사용할 것인가를 살펴볼 것이며, 다음 단계로 NARA 시스템을 실제로 어떻게 활용할 수 있는지 언급할 것이다.

4.1 '주석' 정보의 확장 및 부착 도구

세종 한일 병렬 (형태 주석) 말뭉치는 기본적으로 문장 대 문장의 정렬 방식을 택하고 있다. 또한 각 문장이 포함된 단락과 단락의 연결에 대한 정보 역시 포괄하고 있다. 또한 실제 텍스트 자료를 이용하여 구축된 말뭉치의 특성을 지니기 때문에 각 문장이나 단락은 해당하는 문헌 정보를 가지고 있다. 이는 여타의 병렬 말뭉치가 문장 단위로만 이루어진 것과 달리 나름의 큰 장점이 된다. 따라서 이러한 장점을 최대한 살리는 것이 필요하다. 즉, 검색의 결과를 보여주는 방식을 택할 때, 문장 대 문장의 정렬을 기본으로 하되, 경우에 따라서 각 문장이 속한 단락 정보는 물론 각 문헌의 출처 역시 제시될 수 있어야 한다.

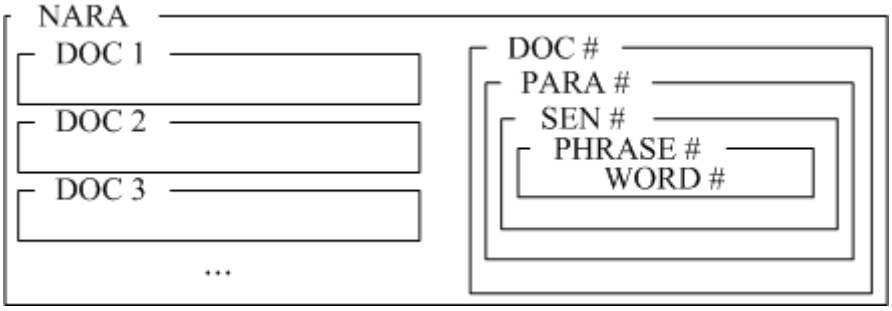
---

6) NARA의 개발 과정의 이력에 대해 간단히 설명하면 다음과 같다. 2008년 하반기 필자가 일본 NiCT로의 인턴쉽 연수가 확정된 이후, 한국어-일본어 병렬 말뭉치를 적극적으로 활용해야할 필요성에 따라 2008년 12월경 본 시스템의 초기 형태가 최초로 개발되었다. 이후 고려대학교 이한섭 교수님의 조언을 발판으로 삼아 2009년 1월 NARA ver. 1.0이 제한적으로 오픈되었다. 이후 일본 연수 과정을 통해 NiCT 관계자의 여러 건의사항과 일본 교토대학 발표회 당시의 조언을 종합한 후, 2009년 3월 병렬 말뭉치에 대한 통계적 처리 노하우를 습득하여 NARA ver. 2.0으로 발전시킬 수 있었다. 일본 연수가 종료된 이후, 다시 여러 연구자분들의 세심한 배려를 통해 지금의 골격을 지니게 되었다.

다음으로 병렬 말뭉치의 활용도를 보다 높이기 위해서는 단어 혹은 구 단위의 정렬이 필요함을 지적하고자 한다. 이를 위해서는 우선 배포판 말뭉치의 정보를 보다 확장하는 작업이 요구된다. 이를 위해서는 언어 자원의 정렬을 위한 도구인 GIZA++, Moses, BLEU 등의 소프트웨어가 사용되었다 (Och and Ney 2003, Koehn et al. 2007, Papineni et al. 2002). 그 결과 단어 및 구 단위의 한국어-일본어 정렬 정보를 추가로 획득하였으며, 보다 풍부한 주석을 기초로 하여 시스템을 구성할 수 있었다.[7)]

결론적으로 말해 NARA 시스템에 포괄된 각 주석 정보는 가장 큰 문헌 정보에서부터 시작하여 가장 작은 단어 단위까지 망라하고 있으며, 이는 언어학 연구의 거의 전반을 망라하는 것이다. 단, 형태소 단위의 정렬은 이루어지지 않는데, 이는 무엇보다 한국어와 일본어의 형태소에 대한 개념이 상호 이질적이기 때문이다. 전체적인 그림은 아래와 같다.

〈그림1〉 주석 정보의 기본 구조



위 그림1에 따라 현재 시스템에서 분석의 대상이 될 수 있는 주석의 형태는 아래와 같다.

- 문헌 정보 : 한국어 및 일본어의 문헌 출처와 그 속성을 중심으로 하여 검색을 할 수 있다. '세종 형태분석 한일 말뭉치'는 총 50개의 문헌을 대상으로 하고 있으며, 그 가운데 12개가 원전이 일본어이며 나머지는 한국어가 원전이다. 단, 어절 규모로 비교해 보면, 한국어와 일본어가 거의 동등한 수준을 유지하고 있다. 실제 논문을 기술하는 단계에서는 각 문장의 출처가 어디에서 비롯된 것인가를 병기해야 할 필요성이 있으므로, 문장 단위 검색 등을 시도한 다음 [출처정보] 버튼을 통해 이 정보를 입수할 수 있다.
- 단락 정보 : 각 문장이 포함된 단락을 단위로 살펴 볼 수 있다. 예컨대, 문장으로 검색을 하고 난 뒤, 각 문장의 [단락보기] 버튼을 클릭하면 그 문장이 속한 단락 정보의 전체를 한국어-일본어, 또는 일본어-한국어의 순으로 살필 수 있다. 이는 각 문장의 화용적 맥락을 살피는 연구에 상당히 유용할 것으로 기대한다. NARA에서 다루는 한국어 및 일본어 단락은 각각 1,713개와 1,702개다.
- 문장 정보 : 가장 기준 틀이 되는 정렬의 집합이다. 대부분의 연구 및 교수안 확보는 이 층위에서의 검색을 통해 획득할 수 있을 것이다. NARA가 포괄하는 한국어 문장의 개수는 4,030이며, 일본어 문장은 총 4,038개이다.
- 구 정보 : 한국어의 구 단위와 일본어의 구 단위 경계가 제시된다. 기계적인 처리 과정에 의해 얻어진 이 정보는 약 72%의 정확성을 보인다 (Song and Bond 2009). 비교 언어학의 관점에서는 이 층위의

7) Song and Bond (2009) 참조

정보가 매우 유용할 수 있는데, 무엇보다 통사적 구성 단계에서 양 언어가 얼마나 유사하고 얼마나 다른지를 살필 수 있기 때문이다. 예컨대, 일본어의 경우 속격 표지 'の'의 쓰임이 매우 생산적인데 비해, 한국어의 '-의'는 그렇지 않다. 따라서 일본어 단어 'の (助詞-連体化)'가 포함된 구조를 추출하여 이것에 대응하는 한국어의 대응쌍에서 얼마나 '-의'가 빈번히 출현하는지를 관찰하면 양자의 차이를 수량적으로 검토할 수 있다.[8)]

- 단어 정보 : 한국어의 각 단어와 그에 대역되는 일본어 또는 그 역방향의 검색이 가능하다. 이는 특히 한국어 및 일본어 교수자 및 학습자에게 유용한 정보로서 각 단어가 대응 언어의 어떠한 단어로 번역 가능한지를 살피는 용도로 활용될 수 있다. 또한, 이 정보를 일괄되게 추출하면 한국어-일본어의 대역어 사전을 자동으로 구축하는 일 또한 가능하다.
- 형태소 정보 : 본고에서 제안하는 NARA 시스템은 한국어-일본어의 언어적 정보를 가능한 한 정확히 활용할 수 있도록 사람에 의해 반자동 구축된 '형태 주석' 말뭉치를 사용한다. 따라서 한국어와 일본어의 각 형태소를 기준으로 할 수 있기 때문에, 동사 등의 원형을 기준으로 검색할 수 있다. 예컨대, '食べる'로 검색을 하면, '食べる', '食べ' 등의 활용형을 모두 검색의 대상으로 한다. 단, 앞서 설명한 바와 같이 형태소 이하 단위의 한국어-일본어 정렬은 이루어지지 않았기 때문에 이를 파악하는 것은 연구자, 교수자 및 학습자의 역할이 된다.
- 품사 정보 : 각 품사를 기준점으로 선택할 수 있다. 또한 동음이의어 등을 고려하여 형태소와 품사 정보 등을 결합한 검색도 가능하다. 한국어 품사 표지는 세종 형태 분석 말뭉치 처리 지침에 따른 52개이며, 일본어의 경우에는 ChaSen에 근거한 74개 표지를 사용한다. 이들 표지는 구축된 병렬 말뭉치 정보를 그대로 사용하였다. 각 표지에 대한 자세한 설명은 NARA 시스템 상단의 메뉴 바를 통해 팝업창에서 확인할 수 있다.

4.2 온라인 검색 '도구'

NARA의 기획의도는 '누구나', '언제나', '어디서나'로 요약될 수 있을 것이다.

- 누구나 (anyone) : 대부분의 연구자들이 말뭉치 사용에 어려움을 겪는 이유가 그 사용법에 익숙하지가 않기 때문이다. 따라서 연구자의 배경이나 컴퓨터 환경에 대한 친숙도에 큰 상관없이 말뭉치를 곧바로 사용할 수 있도록 지원하는 일이 긴요하다. 예컨대, 사용자가 자신의 컴퓨터에 말뭉치 검색 소프트웨어를 별도로 설치하고 여기에 자신의 말뭉치 파일들을 연계하는 일조차 때로는 쉽지 않을 수 있다. 현재 사용자에게 일반적으로 가장 친숙한 인터페이스는 웹이다. 구글과 같은 검색 엔진을 일상적으로 사용하지 않은 연구자는 거의 없기 때문이다. 현재 거의 모든 컴퓨터에는 기본적으로 웹 브라우저가 기본으로 설치되어 있으며, 자료는 중앙의 데이터베이스 시스템에 보관되어 있기 때문에 사용자는 단지 그냥 사용하기만 ('just use')하면 된다. 즉, 누구라도 '세종 한일 병렬 말뭉치'를 간편하게 사용할 수 있도록 하는 것이 NARA의 첫 번째 목적이다.
- 언제나 (anytime) : 2009년 현재 21세기 세종계획의 결과물은 대체로 DVD를 통해 배포되며, 일부 파일들은 국립국어원 홈페이지상에서 다운로드 받을 수 있다. 그러나 모든 연구자들이 이들 자원을 쉽게 입수할 수 있는 것은 아니며, 특히 해외의 연구자들의 경우에는 이 과정이 쉽지 않을 것이다. 또한

8) 실제로 이 관련 연구는 현재 필자에 의해 진행 중이며, NARA를 활용한 코퍼스 기반 대조 언어학 연구의 방식을 택할 것이다. 이 부분에 대한 자세한 언급은 본고의 목적에서는 사뭇 벗어나는 관계로 별도의 논문으로 발표할 계획이다.

DVD 자료를 늘 소지하고 있지 않다면, 세종 말뭉치의 결과물을 확인할 수 없다. 그에 비해 온라인에 구축된 시스템은 사용자가 인터넷에 연결된 컴퓨터에만 접근할 수 있으면 바로 세종 말뭉치의 자원을 쓸 수 있다는 장점을 지닌다.

- 어디서나 (anywhere) : 윈도우 환경이 주종을 이루는 한국의 경우에는 말뭉치를 검색하는 기본 환경이 윈도우와 EUC-KR 인코딩으로 거의 고정되어 있다고 해도 과언이 아니다. 그러나 세계적으로 보면 이것이 반드시 일반적이라고는 할 수 없다. 실제로 해외에 나가 있는 연구자들이 자신의 로컬 컴퓨터 상에서 한글을 입력할 수 없는 경우도 많으며, 특히 '흔글' 등과 같이 특정 프로그램에 의존하여 구축된 말뭉치의 경우에는 사용이 쉽지 않다. 반면 데이터베이스로 구현되어 온라인으로 전송되는 자원은 환경에서 자유롭다. 다시 말해, 지구 반대편이라 할지라도 '세종 병렬 말뭉치'를 아무런 장소의 제약이 없이 사용할 수 있도록 구성되었다.

이러한 기획의도에 따라 본 NARA 시스템은 전술한 바와 같이 인터넷 상에서 검색을 지원하는 방식을 택하고 있다. 이는 아래와 같이 세 가지 측면에서 독창성을 지닌다.

첫째, 사용자의 환경에 독립적인 시스템을 제공할 수 있다. 즉, EUC-KR을 기본 인코딩으로 하는 한글 윈도우 사용자든지, Shift-JS를 기본으로 하는 일본어 윈도우 사용자든지, UTF-8 인코딩의 리눅스 사용자든지, 본 NARA 시스템을 사용하는 데 아무런 문제가 없으며 누구나 동일한 결과를 볼 수 있다. 이는 말뭉치의 '접근 가능도'를 높이는 일이다.

둘째, 인터넷 상의 일반적인 검색 엔진과 같은 형태의 인터페이스를 사용하고 있기 때문에 사용자가 어렵지 않게 시스템을 쓸 수 있다. 실제로 본 시스템은 인터넷 상에서 전자우편을 주고받을 수 있을 정도의 컴퓨터 지식만 있으면 누구나 사용가능할 수 있도록 구성되었다. 이는 말뭉치의 '사용자 친숙성'을 높이는 일이다.

끝으로, 말뭉치의 자료를 중앙의 데이터베이스로 일원화할 수 있다. 이는 사용자에 따라 약간씩 다른 버전의 말뭉치를 대상으로 자신의 연구를 진행하는 문제점을 미연에 방지할 수 있다. 이는 말뭉치의 '객관성'를 높이는 일이다.

4.2.1 검색 방식

검색의 방식은 다섯 가지 차원에서 나누어 살필 수 있다.

첫째로 검색의 방향을 결정하는 것이다. 말뭉치가 한국어-일본어의 병렬 자료이기 때문에 당연히 그 방향은 '한국어→일본어'와 '일본어→한국어'의 양방향을 지원한다.

둘째로 검색의 대상이다. 검색의 기본 단위를 말하는 데, 본 시스템에서는 '품사', '형태소', '어절'의 세 층위에서 검색을 할 수 있다. 또한, 검색의 편의를 위하여 '전방위 일치', '부분 일치', '완전 일치'의 옵션을 제공한다.

셋째로 검색 결과의 선택이다. 검색 결과는 단어 단위, 문장 단위로도 볼 수 있으며, 찾고자 하는 단어를 중심어로 하여 그 앞뒤 몇 단어를 지정하여 살피는 문맥 검색도 지원한다.

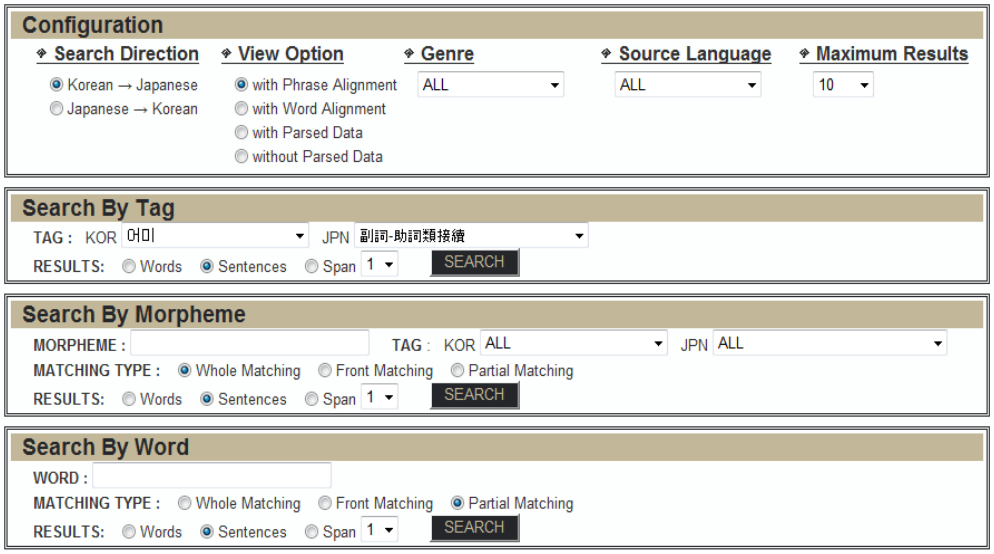
넷째로 검색의 옵션을 세분화할 수 있다. 예컨대, 말뭉치의 장르를 한정시켜서 볼 수도 있으며, 원시 언어가 한국어인지 또는 일본어인지를 선택할 수도 있다.

끝으로 결과의 형태를 이야기할 수 있는데, 우선 기본적으로 검색의 결과는 웹 브라우저 상에 제시되지만, 그 검색 결과를 엑셀 파일로 저장할 수 있는 기능 또한 지원한다. 이는 연구자가 실제 논문을 작성하거나 교

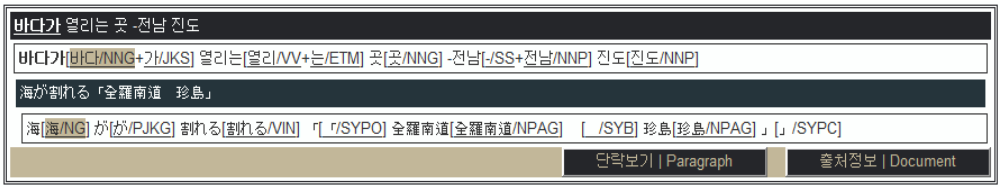
안을 만드는 등의 추가 작업을 수행하는 것을 편리하게끔 만드는 장치이다.

아래 그림은 기본 검색 화면과 그 결과 화면을 스크린 캡쳐한 것이다.

〈그림2〉 검색 인터페이스



〈그림3〉 결과 화면



여기서 한 가지 사용자를 위해 중요한 기능은 병렬의 문장이 나열되는 것 이외에, 구 또는 단어 단위가 어떻게 상호 연결되고 있는지 역시 NARA에서 확인할 수 있다는 점이다. 위 그림3에서 밑줄이 쳐진 단어 구는 대응하는 언어와 연결 정보를 지닌 것이라는 의미이다. 이때, 그 한국어 혹은 일본어 단어 위에 마우스 포인터를 올리면 위 그림3의 '[바다/NNG]-[海/NG]'와 같이 진한 배경색으로 상호 대응됨이 표시된다.

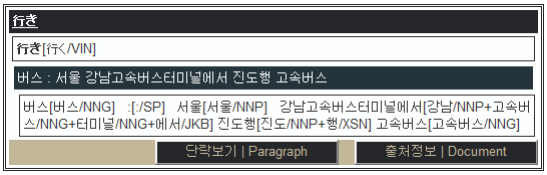
다음으로 앞서 말한 다섯 가지 사항이 NARA에서 어떻게 적용되는지 구체적으로 살펴보기로 하자.

- 검색 방향 : 이를 결정하기 위해서는 검색 인터페이스 상의 'Search Direction' 옵션에서 'Korean → Japanese' 또는 'Japanese → Korean'의 라디오 버튼을 체크해 주면 된다. 예컨대, 아래 그림과 같이 '한국어 → 일본어' 방향의 검색을 하거나 '일본어 → 한국어' 방향의 검색을 하여도 같은 결과를 얻을 수 있다.

〈그림4〉 한국과 일본의 원형 / 韓国と日本の原形

- 검색 대상 : 예컨대, '나'라는 단어를 가지고 검색을 한다고 가정을 해 보자. 앞의 (2)의 두 번째 행 '나는 나/NP+는/JX'에서 제시된 바와 같이, 이 형태는 세 가지 정보를 기본으로 지니고 있다. 우선, 이 단어가 1인칭 대명사라고 했을 때, 이것의 품사 표지는 'NP'이다. 또한 형태소 단계로서는 '나/NP'와 같은 형태로 주석이 되어 있을 것이다. 끝으로 '나는'과 같은 형태로 다른 형태소 단위와 결합하여 어절을 이룬다. NARA는 기본적으로 이 세 정보를 모두 기반으로 하여 검색할 수 있도록 하였다. 우선 표지로 검색을 하고자 한다면, 인터페이스 상에서 'Search by Tag'를 택하여 하단 드롭다운 메뉴에서 '대명사'를 선택할 수 있다. 또한 형태소 단위에서 검색하고자 한다면 'Search by Morpheme'에서 입력창에 '나'를 입력하고 표지를 '대명사'로 설정할 수 있다. 실제로 '나-'는 '새싹이 나다'와 같이 동사로서의 쓰임이 있는 동음이의어이기 때문에 이러한 과정을 통해 사용자가 원하는 형태만을 추려서 검색할 수도 있다. 다음으로 단어 즉 어절(분절) 단위 검색을 하고자 한다면 'Search by Word'를 택하면 된다. 중요한 점은 일본어의 경우 형태소와 어절(분절)의 경계가 한국어와 다르다는 것이다. 즉, 일본어의 'Search by Word'는 표면형을 중심으로 한 검색 방식이 될 것이며, 'Search by Word'는 그 형태소를 기준으로 한다. 따라서 'Search by Word'에서 '行き'를 검색하면 '行き[行く/VIN]'와 같은 형태를 살필 수 있지만, 'Search by Morpheme'에서는 기본형인 '行く'를 가지고 검색을 해야 한다.

〈그림5〉 行き[行く/VIN]



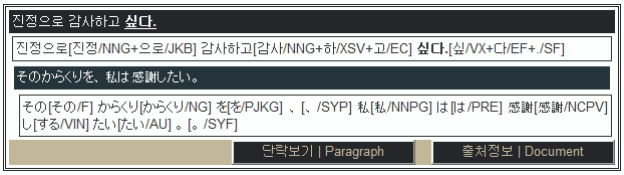
형태소와 어절 검색의 경우 주어진 검색 문자열의 일치의 정도를 결정할 수도 있다. 예컨대 '집'과 같은 형태의 경우 '완전 일치'는 '집' 뒤에 어떠한 형태도 붙지 않는 경우만을 찾아낸다. '전방 일치'는 앞에서부터 일치하는 것만을 찾아내는 것으로서 '집안에'등을 찾아내지만 '친척집'등을 찾을 수 없다. '부분 일치'는 '집'이 일부라도 포함되어 있는 것을 찾아내는 것으로서 당연히 '집안에'와 '친척집' 및 '친척집에' 등을 모두 찾을 수 있다.[9)]

- 검색 결과 : 검색의 결과는 기본 옵션인 '문장' 단위 이외에 '어절' 또는 '문맥'을 기준으로 할 수도 있다. 예컨대, 위 그림5는 어절을 선택한 것에 해당한다. '문맥' 검색은 해당 검색열을 중심으로 하여 앞의 뒤 몇 개의 형태소 또는 어절 까지를 보여줄 것인가이다. 이는 형태-통사론적 연구를 위해서 아주 유용하게 쓰일 수 있다. 예컨대, 복합명사구나 '본동사+보조 용언' 구성을 추출해 내고 이것이 일본어

9) 발표장에서 지적된 사항과 마찬가지로 '면서'를 가지고 어절 검색을 하면 '면서기'와 같은 형태까지도 결과에 포함된다. 이는 검색의 기본 옵션이 '부분 일치'이기 때문이다. 따라서 연결어미 '-면서'만을 따로 뽑고자 할 경우 어절 검색 보다는 형태소 검색을 이용하는 것이 바람직하다.

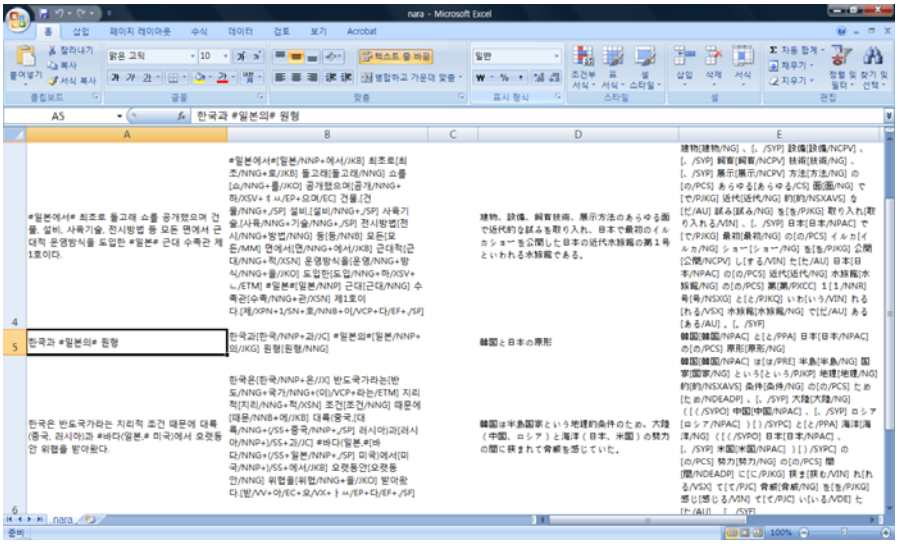
표현에서는 그 양상이 어떠한지를 추정할 수 있게끔 한다.

〈그림6〉 보조용언 '-싶'



- 검색 옵션 : 현재 NARA ver. 2.0에서는 네 가지 검색 옵션을 제공한다. 첫 번째 'View Option'는 검색 결과를 보여주는 방식으로서 구 단위 또는 단어 단위 정렬를 지원해서 보여줄 것인지 단순히 분석 결과만을 함께 보여줄 것인지 또는 아무런 분석 결과를 보여주지 않을 것인지를 결정한다. 두 번째 'Genre'는 원전의 텍스트 성격을 규정한다. 현재 '세종 한일 병렬 말뭉치'에는 '잡지', '신문', '비소설', '소설', '논문' 등의 다섯 개 장르가 망라되어 있다. 세 번째로 'Source Language'는 원전이 한국어인지 일본어인지를 결정할 수 있다. 번역의 과정에 따른 차이점을 비교하고자 하는 연구를 진행할 경우 이 옵션을 유용하게 사용할 수 있다. 끝으로 'Maximum Results'는 검색 결과의 개수를 정할 수 있다. 연구의 성격에 따라 약간의 자료만 보아도 충분한 연구가 있는 반면 대단위의 자료에 기반하여야만 의미를 지니는 연구가 있을 수 있으므로 그 때에는 이 옵션을 유용하게 활용할 수 있다.
- 결과 형태 : 검색 결과는 기본적으로 웹 브라우저 상에 표시된다. 그러나 실제 연구에 있어서는 이 결과를 별도의 파일로 저장하여 사용하는 것이 보다 유용하다. 따라서 이 부분에 대한 지원도 별도로 제공된다. 이른바 '검색 결과의 저장'에 해당한다. 검색 결과는 사용자에게 비교적 친숙한 MS Excel 형태로 제공된다. 검색 결과 화면에서 상단 또는 최하단의 'Extract Current Results to Excel' 버튼을 클릭할 경우, 현재 검색을 완료한 결과를 엑셀 형태로 다운로드 받을 수 있다. 이때, 사용자의 검색열은 # 기호로 둘러싸여 제시되며 한국어 문장과 일본어 문장의 경계는 빈 셀 하나를 사용한다.

〈그림6〉 Excel 결과 추출 (검색열 '일본')



4.2.2 활용 분야[10)]

본 시스템은 크게 네 부류의 사용자를 기본 전제로 하여 구축되었다.

첫째는 한국어-일본어 대조 언어학 연구자이다. 현재 시스템은 한국어와 일본어의 언어정보 대응양상에 대한 구조적 제시를 하고 있기 때문에 각 연구자는 자신의 연구 주제에 따라 말뭉치의 정보를 응용할 수 있다.

둘째는 한국어 교수자 또는 일본어 교수자이다. 예컨대, 가르치고자 하는 어휘 등을 선택하여 그 용례 등을 추출하여 교안 등을 손쉽게 구성할 수 있다.

셋째는 한국어 학습자 및 일본어 학습자이다. 자신이 공부하고 있는 표현의 일본어나 한국어 대역어가 어떻게 구성되는지를 실제의 문장을 가지고 확인할 수 있다.

끝으로, 본 시스템은 한국어-일본어 기계 번역 시스템을 연구하는 자연어처리 전문가에게도 연구의 각 단계마다 활용하고 참조할 수 있는 일종의 워크벤치로 역할을 할 수 있다.

5. 향후 계획

향후의 발전 계획은 두 가지 측면에 집중되어 있다.

하나는 포괄하는 말뭉치의 크기를 보다 크게 하는 것이다. 현재 약 4천 문장을 망라하는 자원은 그 크기가 적절하다고 할 수 없으므로 이후의 지속적인 말뭉치 확장 과정을 통해서 수년 계획으로 최소 몇 만 단위 규모의 문장 정렬을 할 수 있도록 계획을 잡아나갈 생각이다. 이 과정에서 여러 일본어학 연구자 및 일본어 교수자의 적극적인 도움이 절실한 것은 당연한 사실이다.

다음으로 현재의 버전은 말뭉치를 검색하는 과정 즉, 위 표1에서 C영역(사용+도구)에만 초점을 두고 있는데, 다음 버전이 되는 NARA ver. 3.0은 본 시스템을 구축의 도구(표1에서 B영역)로까지 확장해서 사용할 수 있는 기능을 추가적으로 부착할 것이다. 즉, 향후의 NARA는 한일 병렬 말뭉치를 구축하고 사용할 수 있는 종합 도구로서 그 성격을 키워나갈 것이다.

10) 본 NARA 시스템을 사용하면서 가지게 된 의문점이나 건의사항 등은 언제든 필자의 전자우편을 통해서 전달해 주길 독자 여러분에게 부탁드리는 바이다. 다양한 사용자의 다양한 의견을 통해 본 시스템의 환경이 더더욱 발전할 수 있음은 주지의 사실이다.

【參考文獻】

국립국어원 (2007) 『한일 병렬 말뭉치 개발 말뭉치 관련 지침』

Abeillé, A. (2003) Treebanks: Building and Using Parsed Corpora, Kluwer Academic Publishers.

Bond, F., E. Nichols, D. S. Appling, and M. Paul. (2008) "Improving Statistical Machine Translation by Paraphrasing the Training Data," Proceedings of the 5th International Workshop on Spoken Language Translation, 150-157.

Koehn, P. (2005) "Europarl: A Parallel Corpus for Statistical Machine Translation." MT Submit.

Koehn, P., M. Federico, W. Shen, N. Bertoldi, O. Bojar, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, R. Zens, A. Constantin, C. C. Moran, and E. Herbst (2007) "Moses: Open Source Toolkit for Statistical Machine Translation." Proceedings of the ACL 2007 Demo and Poster Sessions, 177-180.

Och, F. J. and H. Ney (2003) "A Systematic Comparison of Various Statistical Alignment Models." Computational Linguistics 29(1): 19-51.

Papineni, K., S. Roukos, T. Ward, and W-J Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation." Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 311-318.

Song, S. and F. Bond (2009) "Online Search Interface for the Sejong Korean-Japanese Bilingual Corpus and Auto-interpolation of Phrase Alignment." Proceedings of the 3rd Linguistic Annotation Workshop, 146-149.

〈 要 旨 〉

NARA ver. 2.0: Sejong 韓日並列コーパス検索環境

並列コーパスは比較言語学、言語教授法、自然言語処理などの分野において、有効なコーパスとして、その重要性が日に日に増している。しかし、ほとんどの研究者が並列コーパスを自身の研究に利用しにくいという側面があるのも事実である。これは平行コーパスを手軽に使うためのツールの無さゆえである。本稿の目的は、2007年までに構築されたSejong韓日並列コーパス（形態素注釈バージョン）を言語研究者、教育者、そして学習者が手軽に検索できるオンラインシステムとして紹介し、その開発過程において考慮・工夫した点を再確認することである。NARAと名付けられた本検索環境はデータベースをもとに、一般的なインターネット検索ページに似たインターフェイスであり、オンライン上で「いつでも」「どこでも」「だれでも」が利用できるという特徴がある。さらに、研究者の必要と便宜に応じるために、検索方法の多様化、検索結果の別途保存も可能にした。

▣ 송상헌(**宋相憲**)

Dept. of Linguistics, Univ. of Washington
sanghounsong@gmail.com

■ 投 稿 日：2009年　9月 30日
■ 審査開始：2009年　11月 18日
■ 審査完了：2009年　12月　4日
■ 掲載確定：2009年　12月　7日