이베로아메리카 제12권 2호 [2010. 12] : 155~188

온라인 스페인어 코퍼스 분석 및 언어학적 활용 방안 연구

신 자 영*
(Shin, Ja-Young)

〈Abstract〉

El análisis del corpus del español y su aplicación en el estudio lingüístico

El presente estudio consiste en analizar los varios tipos del corpus del español y sus aplicaciones en el estudio lingüísitco y en la educación del español como segunda lengua. El corpus es definido como el conjunto de ejemplos reales de uso de una lengua. La lingüística del corpus es la subdisciplina de la lingüística que estudia la lengua a través del corpus. Este tipo de aproximación contrasta con el método de investigación de la gramática generativa que tiende a estudiar la lengua a través de la intuición lingüística del hablante.

En este estudio estudiamos los varios tipos del corpus del español en línea tales como CREA, *el corpus del español diseñado por Mark Davies y el corpus de estructura sintáctica de la Universidad de Santiago de Compostela*. Analizamos la estructura, el sistema de anotaciones y el método del uso de cada corpus. También analizamos los corpus fuera de línea tales como *CORLEC (Corpus*

* 연세대학교 언어정보연구원
E-mail: jayoungshin@yonsei.ac.kr

Oral de Referencia de la Lengua Española Contemporánea), Corpus de Referencia de la Lengua Española en la Argentina, Corpus de Referencia de la Lengua Española en Chile y CHIEDE, entre otros. Presentamos cómo descargar y archivar el corpus en la computadora personal y cómo acceder al dato a través del programa de concordancia. Por último, estudiamos el método de aplicación del corpus en varios campos de la lingüística. Especialmente examinaremos cómo los datos de frecuencia y de colocación léxica pueden ser utilizados en la lexicología y en la compilación del diccionario. También examinamos otras posibles utilizaciones del corpus en la sintaxis, así como en la educación del español como segunda lengua.

[Key Words: Corpus/ Corpus Monolingüe/ Programa de Concordancia]
[주제어: 코퍼스/ 단일어 코퍼스/ 용례색인 프로그램]

Ⅰ. 서론

최근 언어학의 동향을 살펴보면 객관적인 언어 수행 자료를 근거로 철저하게 귀납적인 방법에 의하여 다양한 언어 현상을 분석하고 이를 외국어 교육, 언어 연구, 자연언어 처리 등에 응용하는 연구들이 상당한 진보를 거듭하고 있음을 볼 수 있다. 특히 언어의 구체적 수행을 연구하기 위해서는 정확한 언어자료의 확보가 필수적인데, 이를 위해서는 질적, 양적으로 객관적 언어자료를 모으는 것이 중요하다. 기존 언어학에서 원어민의 직관에 의존한 언어 자료의 판별에 많은 부분 의존하였으

나, 언어 수행의 연구에서는 코퍼스라는 거대한 언어 자료의 모음을 통하여 양적, 질적 객관성을 확보하는데 중점을 두고 있다.

특히 한국에서 스페인어를 연구하는 입장에서 다양한 스페인어 자료의 확보는 모든 연구자들의 고민거리일 수밖에 없다. 이러한 시점에서 코퍼스는 그 동안 객관적 자료의 확보를 위하여 고민하였던 많은 연구자들의 갈증을 해소할 수 있는 매우 중요한 자료의 원천이라고 할 수 있다. 또한 외국어 학습의 측면에서 볼 때, 코퍼스는 실제 언어자료의 충분한 입력(input)을 제공하여 스페인어 교육의 한계를 많은 부분에서 극복할 수 있도록 한다.

본 연구에서는 코퍼스 언어학의 개념 및 연구 범위를 살펴보고, 일반에게 자료가 공개되어 있는 온라인 혹은 오프라인 스페인어 코퍼스의 특징 및 장단점을 분석하고 언어학 연구 및 외국어 교육, 사전편찬 및 번역학 등에 이를 활용할 수 있는 방안을 모색하는 것을 목적으로 한다. 특히 인터넷상 검색 프로그램과 함께 공개되어 일반인이 쉽게 접근이 가능한 대표적 스페인어 코퍼스인 Real Academia의 CREA와 미국 Brigham Young 대학의 Mark Davies 교수에 의하여 개발된 corpus del español의 구성 및 활용법을 제안할 것이며, 그 밖에 공개된 스페인어 코퍼스 자료 목록 및 그 내용 분석을 하고 이를 콩코던스 프로그램(용례색인 프로그램)을 통하여 사용하는 방법을 제시하게 될 것이다.

Ⅱ. 코퍼스 언어학의 목적 및 방법론

1. 코퍼스 언어학의 연구 목적

코퍼스 언어학(lingüística del corpus)은 대량의 '실제언어' 혹은 실제 언

어의 샘플을 이용하여 귀납적으로 언어를 연구하는 응용언어학의 한 분야이다. 생성문법의 경우 보편문법을 강조하며, 인간들이 경험 이전에 언어에 대하여 알고 있는 초기 상태 이론을 중점적으로 다루게 되고, 모국어 화자가 무의식적으로 가지고 있는 내적 문법(gramática innata)을 연역적인 방법에 기초하여 설명적 타당성을 추구한다. 따라서 생성언어학에서는 모국어 화자의 직관에 많은 부분 의존하여 이론적 가설을 검증하는 방법론을 취하게 된다. 생성문법은 보편 문법 및 설명적 적절성을 추구하므로, 이상적 화자의 직관에 의존하므로, 때로는 기술적 적절성이 문제가 되는 경우가 있는데, 이는 동일한 언어 속에 존재하는 변이 현상은 생성 문법 이론 정립과는 거리가 있기 때문이다.

반면에, 코퍼스 언어학의 관심은 언어 구조의 복잡성이며, 철저하게 귀납적 방법에 의존하여 언어의 기술적 타당성에 중점을 두고 이를 통하여 설명적 타당성을 달성하는데 그 목표를 둔다. 코퍼스 연구는 언어의 수행, 즉, 언어에 대한 실제 사용에 기반을 두며, 언어의 기능 및 사용 빈도를 중요시 한다. 이처럼 생성 문법과 코퍼스 언어학은 언어기술을 하는 관점 및 추구하는 목표에 있어서 근본적인 차이를 보인다. 그러나, 상호 보완적인 면이 없는 것이 아니다. 우선 원어민의 직관에 의존한 생성 언어학의 연구에 다양한 원어민 직관의 집합체라고 할 수 있는 코퍼스의 자료를 활용한다면, 기술적 타당성을 더할 수 있다. 코퍼스 언어학의 목적 또한 단순히 기술적 타당성에 그치는 것이 아니라, 다양한 통계 기술 및 이론 언어학적 기초를 통하여 언어 현상에 대한 설명적 타당성을 더하는 방향으로 이루어지고 있는 것이 실정이다.

언어 연구에 있어서 어떤 방법론을 취하는가의 여부는 연구의 목적과 대상이 무엇인가에 달려 있다. 특히, 자연언어처리, 대조언어학, 언어변이 연구, 번역학, 사전편찬, 외국어 교육과 같이 언어 수행과 밀접하게

관련된 연구 분야에서는, 실제 언어 수행에 나타나는 구체적 자료에 기초하므로 코퍼스를 통하여 중요한 자료적 기반을 제공받을 수 있다. 외국어 교육의 측면에서 문법서나 제한된 교실 수업에서 다 파악하기 어려운 언어 수행과 관련된 구체적인 자료들을 코퍼스를 통하여 확보할 수 있으며, 소수의 모국어 화자의 직관에 의존한 연구의 한계를 코퍼스를 활용하여 극복할 수 있다. 특히, 한국에서 스페인어 교육을 하는 입장에서 학습자에게 충분한 입력(imput)의 부족 문제로 야기되는 외국어 학습의 한계를 코퍼스를 통하여 극복할 수 있으며, 연구자에게는 객관적이고 균형적 언어자료의 확보를 가능하게 하여 언어학 연구의 실증적 기반을 제공한다.

2. 코퍼스의 종류

코퍼스는 구축대상이 단일 언어를 대상으로 하는가 혹은 두개 이상의 언어를 대상으로 하는가에 따라서 단일 언어 코퍼스(Corpus monolingüe)와 다중언어 코퍼스(Corpus multilingüe)로 구분된다.

단일 언어 코퍼스(Corpus monoligüe)는 한 개의 언어로 이루어진 코퍼스를 뜻한다. 본 논문에서 다룰 코퍼스도 스페인어 단일 언어 코퍼스에 해당한다. 다중언어 코퍼스의 가장 대표적 유형으로 병렬 코퍼스(Corpus paralelo)를 들 수 있는데, 병렬 코퍼스란 한 언어의 텍스트와 그것을 다른 언어로 번역한 텍스트가 정렬된 형태로 이루어지는 다중언어 코퍼스이다. 병렬 코퍼스는 동일 내용의 복수 언어 코퍼스로서, 예를 들면 한국어 원문과 스페인어 번역문(혹은 스페인어 원문의 한국어 번역문)을 문장 단위로 대응 정렬시켜 데이터베이스를 만든 것을 말한다. 병렬 코퍼스는 데이터 자체가 복수 언어를 대상으로 만들기 때문에 언어의 대조 연구나 기계번역, 이중언어 사전 구축, 언어 교육(작문 교육, 회화 교

육) 등에 유용하게 사용될 수 있다.

다중언어 코퍼스는 번역의 방향, 텍스트의 성격, 언어의 수 등에 따라 다양한 유형이 있을 수 있다. 〈표 1〉은 다중언어 코퍼스의 유형을 나타낸다.

〈표 1〉 다중언어 코퍼스의 종류

코퍼스 종류	예시
병렬/번역 코퍼스 (corpus paralelo/ corpus de traducción)	한국어 원문/스페인어 번역문
	스페인어 원문/한국어 번역문
	한국어 원문/다국어 번역문
비교 코퍼스 (corpus comparativo)	한국어 원문/한국어 번역문
	같은 주제/장르에 대한 한국어 원문 vs. 스페인어 원문

미주 및 유럽에서는 서구어 사이의 대응을 기반으로 한 병렬 코퍼스 구축되어 왔다. 주로 유럽을 중심으로 영어-노르웨이어, 영어-프랑스어, 영어-독일어 병렬 코퍼스 구축이 주도되었다. 최근 유럽연합의 참여 국가 언어를 중심으로 한 다국어 병렬 코퍼스가 여러 방면에서 구축되고 있다. 최근에는 구축된 병렬 코퍼스를 활용하기 위한 사용자 인터페이스 도구, 텍스트 정렬 방법 등이 다양하게 연구되고 있다. 병렬 코퍼스를 이용한 응용분야가 갈수록 늘어나는 만큼, 구축 및 응용 도구 개발도 더욱 가속화 될 전망이다.

또한 학습자의 중간언어를 테이터베이스화 한 것을 학습자 중간언어 코퍼스(corpus de interlengua)라고 하며, 스페인어를 모국어로 하는 한국어 학습자의 중간언어 코퍼스나 한국어를 모국어로 하는 스페인어 학습

자 코퍼스를 그 예로 들 수 있다.

3. 코퍼스 사용의 도구

코퍼스가 사용자들에게 유용한 것이 되기 위해서는 코퍼스에 주석(anotaciones)이 포함되어 있어야 한다. 이러한 주석의 종류로는 텍스트에 대한 기술적 정보, 즉, 텍스트의 특징, 장르, 참고 문헌적 정보, 유형, 단어수, 기술자 정보 등을 포함하며, 주로 파일 헤더 정보에 포함된다.[1] 코퍼스에 있는 모든 단어에 품사 및 형태소 정보 등를 포함하는 품사표기 및 단어 수준을 넘어서는 주어, 목적어, 절 등을 나타내는 문법 구조 문법표기 등도 코퍼스에 유용한 주석 정보가 된다. 3장에서 살펴볼 스페인어 코퍼스에서도 이러한 위에서 언급한 코퍼스 주석이 어느 정도까지 포함되어 있는 가에 따라 그 코퍼스의 유용성이 결정된다.

코퍼스를 사용하여 언어를 분석하기 위해서는 구체적인 용례를 추출할 수 있는 프로그램인 콩코던스 프로그램 혹은 용례색인 프로그램이 필요하다. Word Smith, Monoconc 등이 대표적인 프로그램으로 특정한 조건과 함께 용례를 추출하거나, 일정 텍스트 내의 단어 빈도, 혹은 특정 단어 단어를 중심으로 다른 단어와의 문장 내에서 공기 빈도를 보여주는 연어(colocación) 검색, 키워드 검색 등을 가능하게 해 준다. 또한 간단한 통계 프로그램이 결합되어 있어, 도출된 결과가 유의미한 결과인지 통계적 검증이 가능하도록 해준다.

코퍼스는 전산화된 자료이므로, 대개는 아스키ASCII 형식으로 표준화된 저장 체계를 가진다. 그러나 아스키 형식의 단점은 제한된 문자 집합을 가지고 있어 스페인어와 같은 특수 기호를 모두 표시하기 어렵다는

1) 주석을 표기하는 방식으로는 TEI(Text Encoding Initiative), XML(Extensible Markup Language), SGML(Starndard Generalized Language) 등이 있다.

점이다. 따라서 파일 형태로 저장된 스페인어 코퍼스를 Monoconc과 같은 용례색인 프로그램으로 읽기 위해서는 텍스트 파일로 변환하여, 인코딩을 유니코드 특히 UTF-8 형식으로 변환하여야 특수문자 등이 손상되지 않은 상태로 읽어 올 수 있다.[2] 또한, 코퍼스에서 추출한 연구 결과가 유의미한 결과인지 단순히 우연의 결과인지를 결정하기 위해서는 특정한 통계학적 검증이 필요한데, SAS, SPSS, SYSTAT, G와 같은 통계 프로그램을 사용할 수 있다.

Ⅲ. 스페인어 코퍼스

본 장에서는 인터넷 상으로 접근이 가능한 스페인어 단일 언어 코퍼스에 대한 분석을 하고자 한다. 코퍼스에 기반한 분석의 결과가 연구 대상인 언어에 대한 객관적 자료가 될 수 있기 위해서는 코퍼스는 대표성(representativeness) 및 균형성(balancedness)의 요건을 갖추어야 한다. 즉 코퍼스가 언어의 특성을 객관적으로 잘 반영할 수 있기 위해서는 다양한 장르의 표본들을 일정한 조합으로 포함하여야 한다. 또한 코퍼스의 분석 결과가 타당성을 가질 수 있도록 일정 규모 이상의 크기가 확보되어야 한다. 코퍼스 전체의 규모도 중요하나, 표본이 되는 각각의 텍스트의 길이와 성별, 나이를 포함한 사회언어학적 변인을 어떻게 반영할 것인가 하는 점도 중요하다. 이러한 대표성과 균형성을 갖추기 위하여 적절한 장르의 비율 및 규모 대하여는 코퍼스 구축시 다양한 샘플링을 통하여 적절한 구조를 결정하게 된다. 본 장에서는 각각의 스페인어 코퍼

2) 용례색인 프로그램에서 아스키파일을 바로 열기 위해서는 파일열기 창에서 모든 파일 열기를 선택하면, 읽어 올 수 있다. 그러나 스페인어의 특수기호들을 읽어오지 못하는 경우가 생길 수 있다.

스의 규모 및 구성 장르의 분포 및 구조, 품사, 구문 주석 여부 및 활용도의 측면에서 분석하여 보기로 하겠다.

1. 스페인어 온라인 코퍼스

본 장에서는 스페인어의 대표적 온라인 코퍼스를 소개하고 그 장단점을 비교 분석하도록 한다. 온라인 코퍼스란 구축된 코퍼스를 인터넷상 용례 검색 프로그램인 콩코던스 프로그램과 함께 제공되어 사용자들이 웹상에서 원하는 자료를 바로 찾을 수 있도록 제공되는 코퍼스이다. 이러한 온라인 코퍼스는 자료를 따로 다운받아서 각자의 컴퓨터에 저장하고 콩코던스 프로그램을 적용하여 용례를 추출하는 과정이 따로 필요없다는 점에서 편리하다. 그러나 사용자 자신이 자료를 가공하기에는 어렵고, 인터넷상 제공되는 기능 안에서만 코퍼스를 사용해야 하는 단점이 있다.

가장 대표적인 코퍼스로는 스페인 한림원(Real Academia)의 현대 스페인어 데이터베이스인 CREA이다.

1) CREA : Corpus de referencia del español actual.
〈http://www.rae.es〉

Real Academia에서 구축한 대표적 스페인어 코퍼스로 1억2천500만 어절 규모의 균형 코퍼스이다. 스페인의 텍스트 50%와 중남미 텍스트 50%의 비율로 되어 있으며, 90%문어 10%의 구어 코퍼스로 구성되어 있다. 시대적으로는 1975년부터 2000년대까지의 자료들이 주종을 이룬다. 다음과 같은 장르가 반영된 균형 코퍼스라고 할 수 있다.

〈표 2〉 CREA의 구성 장르 및 분포

Distribución de los textos del *CREA* por grandes áreas temáticas	
1. Ciencia y Tecnología	10,125%
2. Ciencias sociales, creencias, pensamiento	13,5%
3. Política y Economía	13,5%
4. Artes	10,125%
5. Ocio y vida cotidiana	10,125%
6. Salud	10,125%
7. Ficción	22,5%

CREA는 온라인상 콩코던스 프로그램이 함께 장착되어 사용할 수 있는 코퍼스로 이를 통하여 원하는 단어 혹은 구의 용례추출이 가능하다. 이 밖에도 특정 어휘의 지역적 사용 분포, 매체(책, 신문, 잡지, 구어 등)에 따른 검색 및 작가별 문헌별 검색이 가능하며, 장르별 검색도 가능하다. 다음은 CREA의 검색창을 보인 것으로, 위에서 언급한 여러 가지 검색 조건으로 용례 및 빈도 검색이 가능하다.

〈그림 1〉 CREA의 검색창

또한 검색창 아래쪽으로 빈도 목록(Lista de frecuencia)가 제공되어 CREA 전체에 나타난 어휘의 절대 빈도(frecuencia absoluta)및 상대 빈도(frecuencia normalizada)[3]를 text파일 형태로 내려받을 수 있다. 이러한 어휘 빈도 목록은 한 어휘의 모든 변이형을 하나의 단위로 하여 산출된 것이므로 형태소 분석 및 품사 분석을 거친 후에만 스페인어 기초 어휘 목록 및 등급별 어휘 목록 구축에 활용할 수 있다.

언어 자료를 그대로 모아 놓기만 하고 형태소 및 품사 주석 등을 달지 않은 코퍼스를 원시 코퍼스(Corpus primas)라고 한다. CREA는 가장 기초적 헤더정보(지역, 텍스트 장르, 출전) 정도만 기술된 가장 원시 코퍼스라고 할 수 있으며, 실제 언어 연구에 사용되기 위해 필요한 품사 및 형태소 분석, 구문 정보에 대한 주석이 전혀 설계상 고려되지 않은 미가공 상태의 코퍼스라는 단점이 있다.

형태 변화가 없는 부사, 접속사, 전치사 등과 같은 범주나 명사나 형용사와 같이 매우 규칙적인 범주의 검색에는 어려움이 없으나, 동사와 같이 다양한 굴곡형을 가진 어휘의 검색은 매우 어렵다. 실제로 한 동사의 용례를 찾기 위해서는 그 형태적 변화형을 일일이 찾아야 하는 단점이 있다. 형태소 분석이 되어 있다면 레마(lema)형인 [ir]형으로 모든 용법을 볼 수 있는데, CREA에서는 이러한 검색이 불가능하다. 따라서 실제 연구를 위해서는 찾은 모든 예문들을 텍스트 파일로 따로 저장하여 형태소분석, 품사 및 구문분석 표지 등을 일일이 붙여서, 사용해야하는 단점이 있다. 또한 특정 어휘의 지역별 사용 빈도에 있어서도 코퍼스의 지역적 자료의 균형이 다르고 절대 출현 빈도만이 제공되므로, 사용 빈도의 지역적 분포를 파악하기에 쉽지 않은 단점이 있다.

---

3) 절대 빈도는 전체 코퍼스에서 특정 어휘의 출현 횟수를 말하며, 상대빈도는 100만 어절당 출현하는 어휘의 횟수를 나타낸다.

2) Corpus del español : http://www.corpusdelespanol.org

미국의 Brigham Young 대학의 Mark Davies 교수가 개발한 스페인어 코퍼스로 13세기부터 20세기에 걸친 1억 어절 규모의 스페인어 코퍼스이다. 크게 학술(académico), 뉴스(noticia), 소설(ficción), 구어(oral)의 네 장르를 포함하며, 검색에 편리한 tool이 장착되어 있다. 다음 〈그림 2〉는 corpus del español의 검색창을 확대한 것이다.

〈그림 2〉 corpus del español의 검색창

특정 단어, 구, 접사의 등의 용례 검색이 가능하며(lista 보이기 기능), 검색 결과를 그래프(gráfico) 옵션을 통하여 년도, 장르별 분포 차트로도 볼 수 있다. 또한 연도별, 장르별 검색 조건을 제한할 수 있으며, 두 개의 단어를 다른 단어들과의 연어 관계 분석을 하여 유사한 두 단어가 사용 분포에 있어서 어떻게 다른지 구분할 수 있다. 그 밖에도 N-gram, Mutual Information과 같은 간단한 통계 프로그램이 함께 있어, 검색 결과 나타난 수치를 해석하는데 도움을 줄 수 있다.

이 코퍼스는 형태소 분석이 되어 있고, 품사 태깅이 되어 있어서, 검

색이 용이하며, 품사별 검색 및 연어, 빈도수 검색이 가능하다. 또한 단어를 형태소로 분리하여 검사가 가능하므로, 파생 형태론의 연구에 있어서도 유용하게 활용될 수 있다. 그 밖에도 두개의 어휘에 대한 연어 검색을 동시에 실행하여, 각각의 어휘가 자주 결합하는 어휘를 비교하여 그 의미 및 용법을 대조하는 기능도 있으며, 시소러스를 바탕으로 한 유사어 검색이 기능도 갖추고 있다. 코퍼스를 사용한 언어 연구 방법의 구체적인 사례에 대하여는 3.3.장에서 언급하도록 하겠다.

corpus del español은 코퍼스 언어학에서는 가장 모범이 되는 온라인 코퍼스로 인식될 정도로 아주 잘 설계된 코퍼스이다. Mark Davies 교수는 스페인어 이외에도 영어의 다양한 코퍼스를 온라인상 편리하게 자료검색을 할 수 있도록 설계하였고, 포루투갈어 코퍼스도 구축하였다.[4].

3) Base de datos sintácticos del español actual : http://www.bds.usc.es/

산티아고 콤포스텔라 대학에서 구축한 145만 어절 규모의 코퍼스인 ARTHUS (El Archivo de textos hispánicos de la Universidad de Santiago de Compostela)를 바탕으로 160,000 문장에 대한 구문분석 코퍼스이다. 구문분석 코퍼스의 특징은 스페인어의 동사를 중심으로 가능한 모든 문형을 품사별, 문장성분 및 기능별로 분석하였다는 점이다. 가령 'ayudar' 동사로 검색을 할 경우 ayudar 동사의 코퍼스 상 나타나는 모든 용례의 문장

4) Corpus of Contemporary American English (COCA ; 4억천만 어절규모의 1990- 2010 기간의 미국 영어 코퍼스), Corpus of Historical American English (COHA ; 4억 어절 규모의 1810-2000 기간동안의 영어자료), British National Corpus (1억어절 규모의 1980-1993 기간의 영국영어 코퍼스), TIME Corpus (1억어절 규모의 1923년 부터 현재까지의 타임지 기사 코퍼스), Corpus do Português (4천5백만 어절의 1300-1900년대까지의 포루투갈어 코퍼스) 등이며, 간단한 로그인 절차를 걸친 모든 연구자들에게 공개된다.

성분 분석과 그 분포를 보여준다.

다음은 검색 결과이다. 이 결과로 보면 ayudar 동사의 총 247용례에서 나타나는 문형의 비율을 각각의 예문과 함께 볼 수 있다.

〈표 3〉 ayudar 동사 구문분석 검색 결과

Activa	S	14	5.67%
Activa	S SP	21	8.50%
Activa	SD	128	51.82%
Activa	SD AD	1	0.40%
Activa	SD SP	75	30.36%
Pasiva	S A	8	3.24%
(S:주어, D: 직접목적어, AD: 부사구, SP:보어, A: 행위자 보어)			

이 코퍼스의 특징은 기존의 코퍼스와는 달리 통사적 정보에 대한 기술이 되어 있어, 특정한 동사를 중심으로 이 동사가 구성할 수 있는 가능한 모든 문형에 대한 정보를 그 분포와 함께 검색할 수 있다. 특히, 기본 문형 정보에서 상세 문형 정보[5]를 문장 의 구성 요소들의 통사, 형태적 기능 뿐 아니라, 문장 구성 요소들에 대한 좀 더 세분된 의미정보(명사의 경우 유정, 무정, 부사구의 경우 위치, 원인 등)까지 분석되어 있으

5) 코퍼스의 검색창에는 subesquema (하위 구조 혹은 세부 구조)로 명시되어 있다. 위의 표3에서 ayudar가 직접목적어를 취하는 능동 구문으로 쓰이는 구조의 세부 구조를 검색하면 주어와 목적어의 유정/무정 자질에 따라 다음과 같은 네 가지 세부 유형의 결과를 얻을 수 있다.

San Dan (F=110; 85.94% s/esq.)
San Dnan (F=3; 2.34% s/esq.)
SnanDan (F=11; 8.59% s/esq.)
SnanDnan (F=4; 3.13% s/esq.)

며, 각각 그 분포와 예문을 볼 수 있다. 또한, 특정 문형을 선택하여 그 문형으로 실현되는 동사와 그 용례를 검색할 수 있으며, 전치사를 중심으로 결합 가능한 동사와 실현 가능한 문형을 용례 및 빈도와 함께 검색이 가능하다. 본 코퍼스는 통사 구조 연구에 매우 유용한 자료를 제공하여 준다.

2. 스페인어 오프라인 코퍼스

3.1.에서 살펴본 온라인 코퍼스의 경우 용례색인 프로그램이 함께 장착되어 사용자가 인터넷에 접속 하여 원하는 자료를 검색하는 것을 가능하게 한다. 오프라인 코퍼스의 경우 구축된 코퍼스를 압축된 아스키 파일 형태로 연구자들에게 제공되어, 연구자가 개인 컴퓨터에 저장한 후 Word Smith나 Monoconc과 같은 콩코던스 프로그램을 사용하여 필요한 자료를 추출할 수 있도록 한 것이다. 파일 형태로 제공되는 스페인어 코퍼스의 대표적 예로는 스페인 마드리드 아우또노마 대학의 Marcos Marín 교수를 중심으로 한 언어정보 연구소(laboratorio lingüísitco de informática)에서 구축한 코퍼스로 이 연구소의 웹사이트의 자료실에서 무료 혹은 제한적으로 제공받을 수 있다.[6].

대부분의 자료는 아스키파일 형태로 장르별로 폴더로 저장되어 있으며, 이를 텍스트 파일로 변환하고 인코딩을 유니코드나 UTF-8로 변환하여 ANTCONC이나 Word Smith와 같은 콩코던드 프로그램을 이용하면 용례추출 및 빈도, 연어 검색 등이 가능하다. 아스키파일을 직접 콩코던스 프로그램으로 열 수도 있으나, 스페인어 특수 문자 등이 인식되지 않는 경우가 있다.

---

6) 다음 인터넷 주소를 참고할 것. http://www.lllf.uam.es/ESP/Recursos.html

오프라인 코퍼스중 웹상에서 무료로 내려받기가 가능한 코퍼스들을 분석하면 다음과 같다.

〈표 4〉 스페인어 오프라인 코퍼스 목록

코퍼스명	특 징
CORLEC Corpus Oral de Referencia de la Lengua Española Contemporánea	1,100,000 어절의 구어 전사 코퍼스
Corpus de Referencia de la Lengua Española en la Argentina	2,000,000 어절의 아르헨티나 스페인어 문어 코퍼스
Corpus de Referencia de la Lengua Española en Chile	2,000,000 어절의 칠레 스페인어 문어 코퍼스
Spanish Treebank Corpus	스페인 신문에서 추출한 1,500 문장에 대한 통사 주석이 첨가된 코퍼스
CHIEDE Corpus de Habla Infantil Espontánea del Español	60,000 어절규모의 유아 언어 구어 코퍼스

코퍼스의 요건 중 가장 중요한 것은 자료의 균형성 및 대표성으로 다양한 장르의 언어가 균형적으로 반영되어야 한다. CORLEC의 경우 다음과 같은 장르가 포함되었다.

〈표 5〉 CORLEC의 장르 구성

장르		구성비율
Administrativos y políticos		5.6%
Científicos		3.3%
Conversacionales o familiares		24.5%
Educativos		5.3%
Humanísticos		5.6%
Instrucciones (megafonía)		0.6%
Jurídicos		3.2%
Lúdicos (concursos, etc.)		5.6%
Periodísticos:	Debates	8.5%
	Deportes	5.3%
	Documentales	2.6%
	Entrevistas	15.6%
	Noticiario	6.6%
	Publicitarios	2.8%
	Religiosos	1.1%
	Técnicos	3.9%

구어 코퍼스를 중심으로 다양한 장르에서 대화문을 코퍼스로 구축하였다는 점에서 의미가 있다. 특히 구어 코퍼스는 화용론 및 담화 분석 연구의 자료로 매우 요긴하게 쓰일 수 있다.

CORLEC는 TEI 표준에 따라 코퍼스의 기술적 정보를 다음과 같이 표기하였으며, 음성 전사의 원칙에 의거하여 다양한 표기 기호를 사용하였다. 다음 〈표 6〉은 CORLEC의 태그 정보 및 텍스트에 대한 기술적 정보를 주석으로 나타내고 있다. 대화의 출처 및 장소, 키워드, 화자의 직업, 나이 등의 기본 정보가 표시되어 있으며, 텍스트는 품사나 구문 분석이

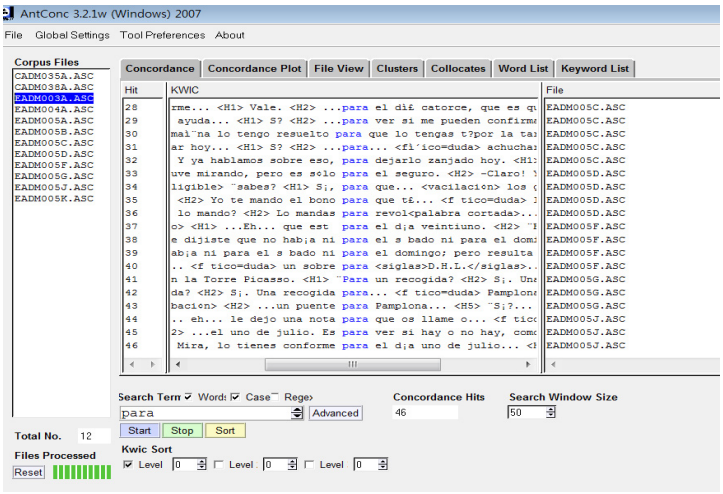
되지 않은 미가공 상태로 제공된다.

〈표 6〉 CORLEC의 헤더 정보 및 텍스트 내용 예

〈ACON021A.WPT〉 〈También en cinta n§8〉
〈22-5-91〉
〈fuente=conversación telefónica〉
〈localización=Madrid〉
〈términos=plan, tenis, estudiantes extranjeros, trabajo, exámenes, agobio, fiesta, canciones, idioma alemán, lengua, libros, morfología〉
〈H1=Mujer, profesora de español para extranjeros en la Universidad, (filóloga), 23 años〉
〈H2=Mujer, filóloga, estudiante, 23 años〉

〈texto〉
〈H1〉 Digo: "Bueno". Ya se me ha estropea〈(d)〉o el plan.
〈H2〉 Jo.
〈H1〉 &ieclm;Me he podido quedar! Si me pod〈palabra cortada〉... Me hubiese podido quedar, pero tendría que haber vuelto, ¿no? y ya me ha parecido demasiado. Digo: "Bah, ya... ya jugaré otro día con él". Porque me... al irme, me dice Alberto, dice: "¿no te quedas a jugar un poco más?"
〈H2〉 Pero, ¿ése Alberto es... es un alumno tuyo... extranjero?
〈H1〉 Sí.
〈H2〉 ¿Y de dónde es, llamándose Alberto?
〈H1〉 Italiano.
〈H3〉 Ah, italiano. Claro, claro.
〈H1〉 Di〈palabra cortada〉... tiene, fíjate, se llama Alberto y tiene un apellido catalán.
〈H2〉 〈risas〉 ¿Y eso?
〈H1〉 ¿Eh?
.....(중략)
〈/texto〉

CORLEC의 경우 제공되는 압축파일 내에는 각각의 장르를 이름으로 한 폴더 내에 적게는 수십 개 많게는 수백 개의 코퍼스가 아스키 파일 형태로 저장되어 있다. 이를 용례색인 프로그램인 Monoconc을 이용하여 담화표지사 'pues'의 용례를 추출한 화면이 〈그림 3〉이다. 추출한 용례의 결과는 따로 text파일로 저장하여 사용할 수 있다.



〈그림 3〉 Monoconc을 이용한 CORLEC의 용례 추출 예.

이처럼, 오프라인에서 아스키파일 형태로 제공되는 코퍼스는 개인 연구자로서는 혼자서 구축하기 힘든 방대한 자료를 손쉽게 접근 가능하게 해준다는 점에서 시간과 노력을 최대한 절약할 수 있으며, 이미 균형성 및 대표성이 검증된 스페인어 자료를 바탕으로 객관적인 연구를 가능하게 한다.

〈표 4〉에서 제시된 다른 아르헨티나 스페인어 참조 코퍼스(Corpus de Referencia de la Lengua Española en la Argentina) 및 칠레 스페인어 참조 코퍼스(Corpus de Referencia de la Lengua Española en Chile)도 같은 방법으로 다운하여 저장한 후 ANTCONC이나 Word Smith와 같은 공개된 용례 검색 소프트웨어를 실행하면 사용이 가능하다. 〈표 7〉과 〈표 8〉에서는 각 코퍼스의 구성 장르의 비율을 보여준다.

〈표 7〉 Corpus de Referencia de la Lengua Española en la Argentina의 구성 비율

장르	어절수	비율
Humanísticos	433,295	21.66 %
Escolares	184,019	9.20 %
Literarios	181,822	9.09 %
Científicos	322,317	16.11 %
Jurídicos	126,561	6.30 %
Técnicos	135,831	6.79 %
Periodísticos	560,000	28.00 %
Comerciales	65,124	3.25 %

〈표 8〉 Corpus de Referencia de la Lengua Española en Chile의 구성비율

de textos científicos.	15 %
de textos económicos comerciales.	10 %
de textos escolares.	15 %
de textos humanísticos	5 %
de textos histórico-jurídicos.	10 %
de textos literarios.	10 %
de textos periodísticos.	15 %
de textos publicitarios	5 %
de textos técnicos.	15 %

이상에서 살펴본 코퍼스들은 품사나 구문분석의 주석이 없으므로, 전장의 'corpus del español'과 같은 품사나 구문분석이 정보를 이용한 검색은 불가능하다. 그러나 파일 형태로 저장하여 스페인어의 품사나 구문분

석 주석을 위한 태그셋을 정하고 주석을 달 수 있는 방법을 개발하는 것이 가능하다. 자동으로 주석을 달 수 있는 프로그램 개발하거나 텍스트의 규모가 크기 않을 경우에는 찾아 바꾸기 기능을 이용하여 수동으로 주석을 달 수도 있다. 품사 태깅, 구문 분석 등이 이루어져 있으면, 훨씬 더 유용한 코퍼스로서 기능을 하게 되므로, 다양한 연구에서 코퍼스를 재사용할 수 있다.

스페인어의 품사 및 구문분석 태그셋의 예를 볼 수 있는 것이 스페인어 Tree Bank 프로젝트이다. 스페인어 Tree Bank 프로젝트는 신문 코퍼스에서 추출한 1600개 문장을 대상으로, 품사태깅 세트와 자동 구문분석 도구를 개발하여 스페인어 문장에 대한 품사, 통사, 의미 정보 및 문장 성분 구조를 분석한 주석을 부착한 것으로 연구 목적의 비영리적 사용의 경우 일정한 절차를 걸쳐 자료 공개가 가능하다. 트리뱅크의 구조는 표면 통사구조를 밀접하게 반영하여 스페인어에 적절한 태그셋을 개발한 것이다. 다음 〈그림 4〉는 "Eligieron a Tomás presidente."라는 문장에 대한 Tree Bank 태그 구조를 보여준다.

"Eligieron a Tomás presidente"

```
(S
  (NP * SUBJ PL P3)
  (VP TENSED PAST IND PL P3
    (V "<Eligieron>" "elegir" TENSED PAST IND PL P3)
    (PP A OBJ1
      (PREP "<A>" "a")
      (NP ID-1
        (N "<Tomás>" "Tomás" PROPER)))
    (PRED-COMPL REF-1 MASC SG
      (N "<presidente>" "presidente" MASC SG))))
```

〈그림 4〉 Spanish Tree Bank의 태그 예시

CHIEDE는 스페인어 유아 언어 구어 전사 코퍼스로 스페인어의 모국어 습득 단계 연구에 유용하게 사용될 수 있다. 제1언어 습득 과정을 구축한 코퍼스와 더불어 추후 한국인 학습자를 대상으로 한 스페인어 학습자 중간언어 코퍼스 구축에 기초적 자료를 제공해 줄 수 있을 것이다. 특히, 다양한 모국어 배경을 가진 개인들이 발전시키는 중간언어의 연구를 위하여, 다양한 모국어를 배경으로 하는 학습자 스페인어 중간언어 코퍼스의 구축이 필수적이다.

Ⅳ. 스페인어 코퍼스의 활용 방안

본 장에서는 2장에서 언급한 코퍼스를 언어학의 다양한 영역에서 활용하는 방법에 대하여 살펴보기로 한다. 코퍼스 자료의 활용 범위는 너무나 방대하여, 이를 일일이 언급하기는 어렵다. 본 장에서는 분야별로 코퍼스를 활용할 수 있는 범위를 언급하며, 간략한 예시를 보이기로 한다. 코퍼스의 사용의 가장 큰 장점은 특정한 언어 표현에 대한 객관적 용례를 대량으로 추출할 수 있다는 점과, 장르별 사용 빈도를 추출할 수 있다는 점이다.

용례 추출의 경우 그 단위가 접두사, 접미사와 같이 형태론적 연구의 대상 뿐 아니라, 어휘, 구문, 문형까지 다양한 단위의 예문을 원어민의 직관을 빌리지 않고 대량으로 추출할 수 있으므로, 어휘론, 형태론, 통사론, 화용론 등 다양한 영역의 언어 연구에서 활용이 가능하다. 추출된 예문들은 그 사용역에 따라 분류를 할 수 있는데, 크게는 구어와 문어에 나타나는 특징을 구분하거나, 지역적 변이, 그 밖의 사회적 변이에 따른 언어 사용의 분포를 볼 수 있게 한다.

1. 어휘 연구와 코퍼스

어휘론은 전통적으로 경험적, 실증적 연구에 기초한 연구 영역이므로, 코퍼스 언어학과 매우 밀접한 관계가 있다. 특히, 사전 편찬 분야에서 코퍼스의 어휘 빈도는 표제어 선정의 기준이 될 수 있으며, 사전에서 기술되는 용례는 대부분 코퍼스에서 추출한 전형적 예문을 직접 혹은 가공하여 사용한다. 일반적으로 사전 편찬에서 어휘의 선정은 신뢰할만한 대용량 코퍼스에서의 빈도에 기초하는데, 이러한 어휘 빈도 추출을 위해서는 코퍼스에 형태소 분석 주석이 첨가되어 있어야만 그 원형별 빈도의 추출이 가능하다. 3장 1절에서 언급하였던 CREA의 빈도 목록은 형태소 분석이 이루어지지 않은 채로 주어지므로, 연구자가 이를 다시 분석하여 원형(lema)별 빈도를 구해야 한다. 반면에 corpus del español과 같이 형태소 분석이 이루어진 코퍼스에서는 어휘의 원형별 빈도 추출이 가능하다.[7)]

사전 편찬과 관련하여 코퍼스의 중요한 역할 중에 하나는 연어(colocación) 정보의 추출이다. '연어'란 특정 어휘를 중심으로 관습적으로 높은 결합관계를 보이는 어휘를 뜻한다. 'corpus del español'의 연어 검색 기능을 이용하여 명사 'análisis'와 높은 빈도로 결합하는 동사를 검색하면 다음 (1)과 같은 결과를 얻을 수 있다. 이와 함께 'análisis'와 자주 결합하는 형용사를 검색한 결과는 다음 (2)와 같다.

(1) análisis와 고빈도 결합을 보이는 동사
: hacer, realizar, haber, requerir, desarrollar, aplicar, dar, completar, detallar

7) corpus del español의 빈도 목록은 유료로 구매 가능하다.

(2) análisis와 고빈도 결합을 보이는 형용사
: químico, espacial, político, cuantitativo, crítico, estadístico, cualitativo, profundo, espectral, objetivo, técnico, exhaustivo, completo, general, teórico, sistemático

이러한, 고빈도 결합은 사전에서 구단위 예문 선정에 매우 유용하게 쓰일 수 있다. 특히, 이해 사전의 기능에서 더 나아가 학습자들이 능동적으로 어휘를 쓰도록 도와주는 표현 사전의 기능을 가진 경우에는 이러한 구단위 예문이 풍부하게 제공되는 것이 매우 중요하다.
이러한 연어 결합은 뜻풀이로는 구분할 수 없는 유의어간 어휘의 사용을 이해하는데 중요한 단서를 제공하여 준다. 즉, 뜻풀이가 유사한 두개 이상의 어휘의 유의어 구분을 위하여 공기되는 어휘의 결합관계를 살펴보면 두 어휘가 의미적으로는 유사하나 실제 사용에서 어떻게 구분되는지를 알 수 있다. 예를 들어, 유사어 검색을 통하여 정도를 나타내는 부사 'completamente'에 대한 다음과 같은 유사어 들을 추출하면 다음 (3)과 같다.

(3) completamente의 유사어: totalmente, absolutamente, enteramente, plenamente, cumplidamente, íntegramente

이중 completamente와 totalmente의 사전적 정의를 Real Academía 사전을 기초로 기술하면 다음과 같다.

(4)
completamente
1. adv. m. Cumplidamente, sin que nada falte.
totalmente
1. adv. m. Enteramente, del todo.

(4)에서 예시한 두 어휘의 사전적 정의는 매우 유사하여, 사전 정의로 두 어휘의 구체적 쓰임을 구분하기는 매우 어렵다. 그렇다면, 두 어휘를 대상으로 각각 자주 결합하여 연어를 이루는 형용사를 코퍼스에 근거하여 선정하여 비교할 수 있다. 이 검색 결과를 정리하면 (5)와 같이 completamente와 상대적으로 높은 결합관계를 보이는 형용사, 두 부사와 유사한 연어 관계를 보이는 형용사, totalmente와 현저하게 높은 결합을 보이는 형용사로 분류하여 볼 수 있으며, 두 단어의 차이를 사용 분포와 관련하여 정의할 수 있게 된다.

(5) a. completamente와 상대적으로 높은 빈도를 보이는 형용사 : armado, tranquilo, solo, feliz, borracho, olvidado, inútil, abandonado, perdido, igual, oscuro, desnudo, libre, aislado.
b. completamente/totalmente와 유사한 결합관계를 보이는 형용사: desconocido, tranquilo, independiente, distinto, contrario, nuevo, extraño, absurdo, etc.
c. totalmente와 상대적으로 높은 빈도를 보이는 형용사 : inaceptable, imparcial, inesperado, claro, independiente, etc.

이처럼 코퍼스를 통하여 얻어지는 언어의 실제 수행과 관련된 구체적 자료는 사전 예문 기술 뿐만 아니라, 외국어 교육, 특히 표현 교육을 위한 중요한 자료들을 기술할 수 있도록 한다.

2. 코퍼스와 통사론

통사론에서 코퍼스의 활용은 크게 두 가지 방향으로 볼 수 있다. 우선 코퍼스 데이터를 기반으로 순순하게 귀납적이고 계량적인 통사 연구에

활용될 수 있으며, 기존의 통사 이론에서 제시된 가설에 대한 검증의 도구로서도 활용될 수 있다.

코퍼스의 통사적 활용에 대한 예시를 위하여 스페인어의 수동구문과 관련한 코퍼스 추출 가능한 정보를 corpus del español 및 ARTHUS 코퍼스를 기준으로 정리하면 다음과 같다.

1) 수동구문으로 쓰일 수 있는 동사의 빈도별 목록 및 그 용례
2) por 수동구문과 por 생략 수동구문의 동사 분포 및 빈도, 용례
3) ser 수동구문과 se 수동 구문 사용 동사 분포 및 용례
4) 수동구문 사용역 비교
5) 수동 구문의 역사적 사용 분포 비교

사용역 분포는 코퍼스 구축시 어떤 사용역을 구분하여 자료를 정리하였는가에 따라 복구할 수 있는 정보의 차이가 있다. CREA의 경우 지역적 변이가 코드화 되어 있으나, 형태소 분석 및 구문 분석 태깅이 되어 있지 않은 관계로, 수동문과 같은 특정 구문에 대한 지역적 변이 검사는 어렵고, 특정 어휘 사용에 대한 지역적 분포만이 비교 가능하다.

사용역에 대한 검색의 예로 스페인어의 '[ser] [과거분사] por'로 구성되는 수동구문과 '[estar] [현재분사]'의 진행형 구문의 코퍼스상 분포를 알아보았다. 〈표 9〉에서 보는 바와 같이, ser 수동구문은 문체적으로 학술적인 글 및 뉴스에서 상대적으로 높은 빈도로 사용됨을 알 수 있으며, '[estar] [현재분사]'의 진행형 구문은 구어에서의 사용이 다른 장르와 확연하게 빈도 차이를 보임을 알 수 있다.

〈표 9〉 수동구문과 진행형의 사용역 비교

	[ser] [VPS*] por 수동태				[estar] [VPP*] 진행형			
장르	학술	뉴스	소설	구어	학술	뉴스	소설	구어
절대빈도	3349	2603	774	415	888	3984	4755	10175
상대빈도	669.81	524.40	162.27	98.04	177.6	802.62	996.88	2403.7
순위	1	2	3	4	4	3	2	1

통사론 및 문법 연구에서 코퍼스의 사용으로 인하여 전통적인 문어 중심의 문법체계에서 벗어나 구어 문법에 대하여 새로운 시각을 가지는 계기가 되었다. 전통적으로 문법은 대부분 문어에 기초하여 기술되어 왔다. 그러나 코퍼스의 구축과 함께 전통적 문법 체계가 실제 언어의 사용과 일치하지 않는 부분들이 있다는 것이 지적되었다. 문어 및 구어에 나타나는 상이한 문법 체계를 기술하고, 이러한 구어 및 문어 문법을 통합하여 특정 언어의 참조 문법 (Gramática de Referencia)을 기술하는 노력이 활발하다. 참조 문법은 기술에 있어서 풍부한 용법에 기초하여 특정 구문의 분포 및 빈도 정보를 포함하여, 스페인어의 다양한 구조들의 형식뿐 아니라 구어, 문어 등과 같은 다양한 장르에서 그 빈도 및 활용에 대한 폭넓은 정보를 제공할 수 있다.

3. 외국어 교육과 코퍼스

외국어 교육과 관련하여 코퍼스에서 추출한 어휘 빈도는 스페인어의 교재 편찬에서 등급별 기본 어휘 선정에 기초 자료로 쓰일 수 있다. 다시 말하면 외국어 교육시 교육 대상의 항목의 우선순위를 정해줄 수 있다는 것이다. 사전 표제어 선정이나 등급별 어휘 목록 선정 작업은 코퍼

스 빈도 및 전문가의 판정과 같은 여러 단계의 판정 절차를 걸치는 것이 일반적이다. 이러한 정제된 전체 목록이 확보되지 않은 단계에서도 코퍼스의 빈도는 교재 집필시 어휘 선정의 우선순위를 정하는 데 참고가 된다.

가령 초급 학생을 대상으로 교재 편찬에서 '-mente'로 끝나는 스페인어 부사를 기술하기 위해서는 가장 흔히 쓰이는, 즉 고빈도 부사를 선정하여 우선적으로 교육하는 것이 더 효율적일 것이다. 이를 위하여 corpus del español에서 wild card 기능을 이용하여 '-mente'로 끝나는 부사의 용례 및 빈도를 찾은 것이 〈표 10〉이다.[8)]

〈표 10〉 *corpus del español*에 나타난 스페인어 부사의 사용빈도

빈도순	부사	빈도
1	solamente	4755
2	realmente	4091
3	especialmente	3559
4	precisamente	3439
5	completamente	2667
6	inmediatamente	2460
7	finalmente	2357
8	verdaderamente	2150
9	principalmente	2128
10	perfectamente	2016

고빈도 부사 중에서 몇 개 항목을 선정할 것인가는 전체 초급 어휘 빈

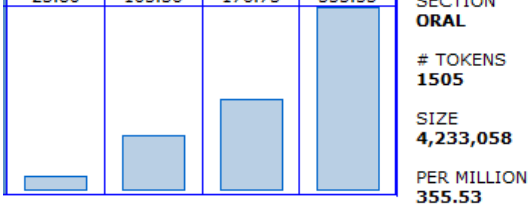
8) corpus del español의 경우 고빈도 100개까지 검색이 가능하며, 각각의 용례를 추출할 수 있고, 사용역(학술, 소설, 언론, 구어)에 따른 빈도 산출도 가능하다. 본 논문에서는 편의상 빈도 10까지만 보인 것이다.

도에서 부사가 차지하는 상대적 빈도를 고려하면, 초급 교재에서 선정하여야 할 최종 목록이 확정될 것이다.

선정된 어휘의 사용과 관련하여, 코퍼스를 이용하여 사용역 정보를 함께 교육할 수도 있을 것이다. 〈표 10〉에서 가장 높은 빈도를 보이는 부사 'solamente'에 대한 사용역 정보를 차트로 보이면 다음과 같다.

〈표 11〉 solamente의 사용역 비교

ACAD	NEWS	FICT	ORAL
119	523	843	1505
23.80	105.36	176.73	355.53



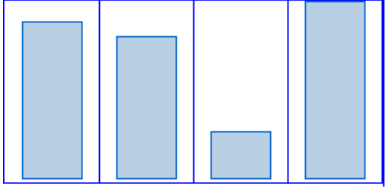
〈표 11〉에 의하면 부사 'solamente'는 전체 코퍼스 장르 중 구어 및 문학작품에서 상대적으로 많이 쓰이며, 학술 및 뉴스 분야에서는 매우 낮은 사용 빈도를 보임을 알 수 있다. 총 4,233,058어절 구어 코퍼스에서 100만 어절 기준으로 355.53의 상대빈도를 나타내며, 학술 분야에서는 23.80의 상대빈도를 보임을 알 수 있다.[9)]

이와는 대조적으로 solamente와 의미적으로 유사한 exclusivamente의 분포를 보면 〈표 12〉와 같다.

9) 각 장르별 코퍼스의 규모가 다르므로 장르별 절대 출현 횟수보다는 100만어절 기준으로한 출현횟수인 상대빈도가 더 의미있는 수치이다.

〈표 12〉 exclusivamente의 사용역 비교

ACAD	NEWS	FICT	ORAL
187	168	52	178
37.40	33.85	10.90	42.05



SECTION
1700s

TOKENS
56

SIZE
9,816,663

PER MILLION
5.70

exclusivamente는 학술, 뉴스, 구어에서 사용 분포의 차이가 상대적으로 적으며, 문학 작품과 같은 허구적 글에서는 매우 낮은 사용분포를 보임을 알 수 있다. 외국어 교육에서 이러한 사용역에 대한 정보는 실제 스페인어의 수행과 관련된 정보 중의 하나로, 적절한 맥락에서 적절한 어휘를 사용하도록 교육하는 지침이 될 수 있다.

코퍼스는 한국에서 스페인어를 배우는 학습자들, 특히 살아있는 스페인어에 접촉할 기회가 별로 없는 학습자들에게 실제로 쓰이는 살아있는 예문을 제공하는 중요한 역할을 할 수 있다. 이렇듯 코퍼스는 학습 목표어에 대한 충분한 입력을 제공하는 기능을 한다. 특히, 코퍼스를 활용한 외국어 교수법에 대한 많은 연구 가운데 학습자 스스로가 코퍼스를 검색해서 어휘, 의미, 문법을 분석할 수 있는 가능성도 모색되고 있다. 이 과정에서 학생들은 자신들이 가지고 있었던 예상이나 직관에 문제가 있음을 알게 되며, 스스로가 가설을 세워 검증하는 연구자적 학습 방식을 터득하게 된다. 이것이 이른바 '데이터 중심 학습법 (DDL: data-driven learning)(Johns: 1991, 1994)의 요체인데, 학습 과정에서 학습자 스스로 의식고양 (consciousness raising) 연습과정을 통하여 문제와 해결책을 발견할 수 있도록 유도하는 것이다. 학생들은 이렇게 코퍼스를 통해서 실제

로 사용되는 외국어 표현에 손쉽게 접할 수 있고 또 이러한 방법을 통해서 알고자 하는 어휘 항목이나 문법 구조를 찾아내게 된다.

Ⅴ. 결론

코퍼스 사용의 기술적인 문제에도 불구하고 언어 연구에 있어서 객관적인 연구를 위하여 양적, 질적으로 균형 잡힌 코퍼스의 필요성은 점점 더 강조되고 있다. 이미 구축된 대용량의 코퍼스를 이용하지 않더라도, 연구의 객관성을 확보할 만큼 균형 잡힌 코퍼스를 스스로 구축하고 이를 바탕으로 연구를 실행하는 것도 가능한 일이다. 특히 스페인어 연구에 있어서 언어 수행의 구체적 자료에 근거하여 실증적 연구의 기초가 되는 것이 코퍼스이며, 소수의 원어민의 직관만으로는 전부 파악하기 어려운 언어 정보에 대한 거시적 연구를 가능하게 한다. 실제로 인간의 직관이나, 사전적 정보로 정확하게 기술되었다고 하는 자료도 코퍼스의 자료와 비교하였을 때에 상이한 패턴이 나오기도 한다.

대부분의 대규모의 코퍼스는 사전 편찬과 함께 발달되었다. 사전 편찬 과정은 점차 실제 쓰이는 언어를 반영하는 기술적 측면이 강조되고 있으므로 코퍼스의 활용도는 절대적이다. 특히 추후 한국어-스페인어, 스페인어-한국어 병렬코퍼스가 구축된다면, 이는 한-서, 서-한 이중언어 사전의 기초적 자료가 될 수 있다.

본 논문에서는 코퍼스 언어학의 목적 및 연구 방법론을 살펴보았으며, 스페인어 연구에 활용할 수 있는 온라인 코퍼스 및 자료가 무료로 배포되는 기구축 대용량 코퍼스를 소개하고 이용하는 방법에 대하여 살펴보았다. 본 논문은 국내의 스페인어 연구자들에게는 아직 생소한 스페인어 코퍼스의 소개 및 사용 방법을 제시하여, 추후 코퍼스를 활용 한 스페인

어 교육 및 연구에 활용할 수 있는 출발점을 제시한다는 점에서 그 의의를 찾을 수 있을 것이다.

참고문헌

강범모(2004),『코퍼스와 어휘 데이터베이스』, 서울; 월인.

강범모(2003),『언어 · 컴퓨터 · 코퍼스 언어학 : 컴퓨터를 이용한 국어 분석의 기초와 이론』, 서울; 고려대학교 출판부.

박기성 역(2008),『대조언어학과 번역학의 코퍼스 기반 방법론 연구』, S. Granger 외 (eds.) (2003) *Corpus based approaces to contrastive linguistics and translation studies,* 서울; 동인.

신자영(2010), "병렬코퍼스 및 학습자 코퍼스를 이용한 중간언어 연구 방법론",『언어 사실과 관점』 제 25집, 연세대학교 언어정보연구원: 71-88.

안동환 역 2008),『코퍼스영어학』, Charles F. Meyer, *English corpus linguistics : an introduction.* 서울 : 한국문화사.

Altengerg, B. & Granger, S. (eds.) (2002), *Lexis in Contrast : corpus based approach,* Amsterdam ; Philadelphia : J. Benjamins.

Davies, Mark. (2002-) Corpus del Español (100 million words, 1200s-1900s). http://www.corpusdelespanol.org.

G. Guilquin, S. (2008), "Combining contrastive and interlanguage analysis to apprehend trasfer: detection, explanation, evaluation", G. Guilquin, S. Papp and M.B. Diez-Bedmar (eds.) *Linking up contrastive and learner corpus reasearch.*

Gilquin, G., Papp, S., Díez-Bedmar, M. B. (eds.) (2008), *Linking up contrastive and learner corpus research,* Amsterdam ; New York, NY : Rodopi.

Granger, J. Lerot, S. Petch-Tysonx(eds.) (2003), *Corpus-based approaches to contrastive linguistics and translation studies,* Amsterdam ; New York : Rodopi.

Johns, Tim (1991), "Should You be Persuaded - Two Samples of Data-Driven Learning Materials", *English Language Journal* 4. Birmingham: Birmingham University.

Johns, Tim, (1994), "From handout to handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning", in Odlin, T. (ed.) *Perspectives on Pedagogical Grammar,* Cambridge: CUP.

Ludeling, A. & Kyto, M. (eds.) (2008), *Corpus Linguistics : An Introductional Handbook,* Berlin ; New York : Walter de Gruyter.

McEnery, T., Xiao, R. & Tono, Y.(2006), *Corpus-Based Language Studies, an advanced resource book,* London and New York: Routeledge.

Sara Laviosa (2002), *Corpus-based translation studies,* Amsterdam ; New York : Rodopi.

Sinclair, John McHardy (eds.) (2004), *How to use corpora in language teaching,* Amsterdam ; Philadelphia : J. Benjamins.

Stig Johansson and Signe Oksefjell Rodopi (eds) (1998), *Corpora and cross-linguistic research: theory, method and case studies,* Amsterdam ; Atlanta, GA: Rodopi.

St. John, E. (2001), "A case for using a parallel corpus and concordancer for beginners of a foreign language", *Language learning and*

Technology 5(3): 185-203.

Base de datos sintácticos del español actual http://www.bds.usc.es/

REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. Corpus diacrónico del español. 〈http://www.rae.es〉

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. Corpus de referencia del español actual. 〈http://www.rae.es〉

http://web.bham.ac.uk/johnstf/paracon.htm

http://www.ruf.rice.edu/~barlow/pc.htm

Corpus Oral de Referencia del Español Contemporáneo

Corpus Lingüístico de Referencia de la Lengua Española en Argentina

Corpus lingüístico de referencia de la lengua española en Chile

CHIEDE El Corpus de Habla Infantil Espontánea del Español

▌논문투고일자: 2010년 10월 31일
▌심사완료일자: 2010년 11월 29일
▌게재확정일자: 2010년 12월 10일