# 여름코퍼스아카데미 일정 및 강의개요

(1) 일자: 2014년 7월 30일(수) - 31일(목)
(2) 시간: 오전 10:00 - 오후 18:00
(3) 장소: 한국해양대학교 국제대 컴퓨터실 102호, 230호
(4) 주최: 한국외국어교육학회, 한국코퍼스언어학회
(5) 주관: 해양영어교육연구센터, 코퍼스언어학연구회

| 날짜 | 시간 | 세션 장소 | 내용 | 강연자 |
|---|---|---|---|---|
| 30일 | 10:00-14:300<br>12:00-13:00<br>(점심) | A-1:<br>102호 | 코퍼스입문과 WordSmith 실습 | 장세은교수(한국해양대) |
| | | B-1:<br>230호 | 온라인 코퍼스를 활용한 강의 및 연구 | 권혁승교수(서울대) |
| | 14:30-18:00 | A-2:<br>102호 | Antconc를 사용한 소규모<br>코퍼스의 연구와 활용 | 남대현교수(울산과학기술대) |
| | | B-2:<br>230호 | 코퍼스자료 중심 교육수업 실습 | 홍신철교수(부산외대) |
| 31일 | 10:00-14:300<br>12:00-13:00<br>(점심) | A-3:<br>102호 | 병렬코퍼스 기초 | 김정렬교수(한국교원대) |
| | | B-3:<br>230호 | 무료 소프트웨어만<br>(e.g. pdftotext, Notepad++,<br>Senna, Antconc)을 이용한<br>코퍼스구축과<br>코퍼스데이터 분석 | 류미림교수(한국해양대) |
| | 14:30-18:00 | A-4:<br>102호 | 코퍼스와 어휘교육 | 신동광박사<br>(한국교육과정평가원) |
| | | B-4:<br>230호 | NLP 태깅 및 통계 연어 측정방법 | 김재훈교수(한국해양대) |

**과목: A-1: 102호**
**제목: 코퍼스입문과 WordSmith 실습 (기초)**
**강사: 장세은 (한국해양대)**

누구나 코퍼스언어학에 대한 이해를 할 수 있도록 코퍼스정의와 역사에서 자신이 직접 소규모 코퍼스를 구축하여 키워드분석, 연어분석, 코퍼스시각화까지 두루 살펴보는 이론과 실습을 병행하는 코퍼스입문과정이다.
시간별 강의와 실습 요약은 아래와 같다.

10:00~11:00 제1강, 제2강, 제3강 (이론)
11:00~12:00 제4강, 제5강 (실습포함)
13:00~14:30 제6강, 제7강 (실습포함)

**제1강 코퍼스기본:**
코퍼스의 정의와 역사, 코퍼스의 유형, 코퍼스 소프트웨어와 웹활용 프로그램 소개, 코퍼스 디자인 및 코퍼스구축과정
**제2강 코퍼스접근방법:**
코퍼스언어학은 방법론인가 아니면 이론인가? 코퍼스언어학이 답해야 하는 두 가지 기본 연구질문은? 코퍼스언어학이 할 수 있는 것과 없는 것은?
**제3강 코퍼스분석과 언어이론 강독** (Meyer, 2004. English Corpus Linguistics. Cambridge: Cambridge University Press. pp.1-29)
**제4강 코퍼스기반 사례연구:**
영문학작품(Jane Austen의 Pride and Prejudice, Charles Dickens의 전체작품)
**제5강 코퍼스구축방법:**
Gutenberg 웹 출처, 텍스트 크리닝(Notepad++, Regular Expressions), 소규모코퍼스구축실습(Melville의 10개 작품과 Moby Dick, 19세기 당대의 문학작품)
**제6강 WordSmith Tools 사용방법:**
(설치, Wordlist, Keywords, Concord, Index, Cluster, Keyclusters): 기본통계읽는 방법(TTR, STTR), Python을 활용한 TTR 분석, Keyness정의와 Loglikelihood에 의한 Keyness 산출방법, keyword list 추출방법, Concordance와 collocation 추출방법, Index 생성과 clusters와 keyclusters 추출
**제7강 코퍼스의 시각화:** Word Cloud, Wordle, and Collocational Networks

**과목: B-1: 230호**
**제목: 온라인 코퍼스를 활용한 강의 및 연구 (기초/중급)**
**강사: 권혁승교수 (서울대)**

# 온라인 코퍼스를 활용한 강의 및 연구

2000년대에 접어들면서 코퍼스언어학은 코퍼스 종류의 다양화, 코퍼스 규모의 대형화, 코퍼스 사용의 용이성에 따라서 언어학 관련 연구에서 양적 성장과 더불어 괄목할만한 질적 성장이 이루어졌다. 코퍼스 연구 초기 단계인 1960년대와 1970년대의 Brown Corpus와 LOB Corpus와 함께 1990년대 최대 코퍼스의 하나인 British National Corpus의 구축에 이어 2008년부터는 Corpus of Contemporary American English를 비롯한 다양한 대현 코퍼스가 인터넷에서 무료로 제공되기 시작하였다. 이를 기반으로 예전에는 불충분한 자료로 접근했던 각종 영어 연구 주제에 대해 대규모 코퍼스를 기반으로 다각적인 측면에서 영어 자료를 분석할 수 있게 되었다. 본 워크샵의 전반부에서는 코퍼스언어학의 기반을 이루는 기본적인 개념과 철학을 설명하고 후반부에서는 강의 및 연구를 위하여 온라인 코퍼스를 어떻게 활용할 수 있는지를 다양한 영어 분석의 예를 들면서 소개하고자 한다. 강의에서 사용할 온라인 코퍼스는 http://corpus.byu.edu에서 제공하는 Corpus of Contemporary American English (COCA), Corpus of Historical American English (COHA), British National Corpus (BNC) 등이며 누구나 무료로 접속하여 사용할 수 있다.

과목: A-2: 102호
제목: Antconc를 사용한 소규모 코퍼스의 연구와 활용 (기초)
강사: 남대현교수 (울산과학기술대)

# Research and practice with small corpora using AntConc

While large corpora such as the British National Corpus or the Corpus of Contemporary American English have been invaluable resources for language teaching and research, they also have been found problematic as they often provide either too much and too complex linguistic details, or they offer too little that is relevant to the needs of specific groups of learners. A response to this concern can be found in the development of small or specialist corpora, and their exploitation for pedagogic purposes. Through the analysis of such small corpora, it is possible for teachers to begin to develop curriculum specifications for ESP/EAP courses, and for researchers to investigate linguistic features in certain disciplines. In this workshop, you will have the opportunity to develop your own pedagogic corpus for classroom purposes and to practice analyzing corpus data for research purposes. Although no previous experience of classroom applications of corpora is required, but it will be useful to bring with you an idea of the kinds of students and research you are interested in.

## 1. Requirements

Corpora: Although sample corpora will be provided for the practice session, participants are encouraged to bring their own text materials, such as student writing texts, collection of specialist texts (research articles, administrative documents etc) or fiction texts.

**2. Concordancer program**: Bring your own laptop with AntConc installed. The concordance is freely available at
-> http://www.antlab.sci.waseda.ac.jp/software.html)

## 3. Outcomes

By the end of the 3 hour workshop, participants will be able to generate wordlists, keyword list, and edited concordances which can be used as the basis for classroom materials and research practice.

과목: B-2: 230호
제목: 코퍼스자료 중심 교육수업 실습 (중급)
강사: 홍신철교수 (부산외대)

# How to apply corpus data for language pedagogy: Focusing on Date-Driven Learning (DDL)

Corpora have been considered a source of native speakers' language use. For this reason, they are often applied to language pedagogy under the assumption that learners can benefit from experiencing authentic language use. However, it is also true that EFL learners and teachers hesitate to use complicated and massive amounts of corpus data. Besides, the complex nature of such corpus data, another crucial reason for wavering is a lack of systematic methodology to guide learners and teachers in applying corpus data to language learning and teaching. In fact, many studies have focused on the positive potential and possibilities of corpus-based approaches in terms of how they contribute to developing learners' awareness of how language is actually used. However, few studies have presented systematic guidelines for how corpus data can be used or analysed in EFL contexts. For this reason, the purpose of the presentation is to present guidelines which EFL learners and teachers adopt for their language learning and teaching from a theoretical perspective. To this end, the presentation first discusses pedagogically useful aspects of corpus data which represent authentic language use (language normality) in terms of 'what to learn'. To do this, three aspects (collocation, colligation, and semantic prosody) are argued. Second the presentation demonstrates specific methodologies for applying corpus data in terms of 'how to learn'. For this, it presents two specific examples of corpus-based methodologies (Data-Driven Learning) which are based on guidelines for EFL contexts.

**과목: A-3: 102호**
**제목: 병렬코퍼스의 기초 (기초)**
**강사: 김정렬 (한국교원대학교)**

## 1. 병렬코퍼스

병렬코퍼스는 코퍼스의 종류를 나누는 기준 중의 하나인 단일어 코퍼스와 다국어 코퍼스의 분류에서 나오는 코퍼스의 한 종류이다. 코퍼스는 수집의 대상이 음성언어인가 문자언어인가에 따라서 음성코퍼스와 문어코퍼스로 나뉘고 코퍼스 구축의 목적이 일반적인 언어현상의 연구에 있는지 아니면 특수한 목적을 가진 (예를 들면 학습자 코퍼스나 경제나 법률코퍼스) 코퍼스에 따라서 일반코퍼스와 특수코퍼스, 코퍼스에 주석이 붙었느냐 아니면 주석이 붙지 않은 생코퍼스냐에 따라서 태그코퍼스와 태그전코퍼스로 나뉘고, 또한 시간적으로 동시대 언어를 중심으로 코퍼스를 구축했느냐 아니면 시간적으로 일정한 기간내의 시대적 코퍼스를 통시적으로 구성했느냐에 따라서 공시적 코퍼스와 콩시적 코퍼스로 구분된다. 이에 더불어 코퍼스가 하나의 언어로 구성되어 있느냐 두 개 이상의 언어로 구성되어 있느냐에 따라서 단일어 코퍼스와 다국어 코퍼스로 나뉘어지고 후자를 병렬코퍼스라고 부른다. 이유는 일반적으로 다국어 코퍼스가 유용하게 이용되려면 언어요소 단위로 어휘, 문장, 단락 단위로 병렬성이 있어야 하기 때문에 이를 병렬코퍼스라고 부른다.
병렬코퍼스는 텍스트의 동일한 선정준거를 통해서 구축된 병렬 비교코퍼스, 단어와 단어의 짝, 문장 대 문장 혹은 문단 대 문단으로 언어의 요소별 단위에 따라서 정밀도를 달리하면서 구축되어 있다. 이에 더불어서 번역본 병렬코퍼스의 경우는 방향이 단방향인데 반하여 양방향 병렬코퍼스가 있는데 이는 번역된 문서를 다소 원어로 하여 재번역한 형태와 동일한 준거를 통해서 선정된 병렬 유사코퍼스를 양방향으로 번역하여 구성할 수 있다.

## 2. 강의내용

1) 병렬코퍼스의 종류, 2) 병렬코퍼스의 활용,
3) 코퍼스의 구축, 4) TravitaAligner

## 3. 병렬코퍼스를 활용한 연구 사례

1) 어휘 대조분석: 신어대조분석, 의미장 대조분석, 전문어휘구조 등
2) 메타포 연구: 병렬코퍼스에 나타난 메타포의 구조와 사용에 대한 대조분석
3) 대조 담화분석

## 4. ParaConc 실습

1) ParaConc 사용법
2) 세종코퍼스 활용 실습

**과목: B-3: 230호**

**제목: 무료 소프트웨어만(e.g.pdftotext, Notepad++, Senna, Antconc)을 이용한 코퍼스구축과 코퍼스데이터 분석 (중급)**

**강사: 류미림교수 (한국해양대)**

본 강의는 인터넷상에서 무료로 다운받아 사용할 수 있는 소프트웨어들만을 사용해서, corpus를 구축하고 그기에 tagging하고 untagging하고, Antconc, WordSmith, Notepad++ 등에서 Regular expressions(Regexs)를 사용하여 데이터에서 언어사용의 일정한 패턴을 찾는 방법을 보여 주고자 한다.

**Step1:** 먼저, 데이터 수집의 첫 단계로 pdftotext 프로그램과 htmlastext 프로그램을 이용하여 pdf 파일로 된 자료와 웹사이트와 같이 html 형식으로 된 자료를 일반 텍스트로 전환하는 방법을 보여준다.

**Step2:** 이렇게 텍스트 파일로 전환된 자료를 텍스트 에디팅 소프트웨어인 Notepad++를 이용하여 cleansing 하고 줄 나누기를 한다.

**Step3:** Gotagger 프로그램이나 Senna 프로그램 등을 이용하여 corpus에 tagging 한다. 다음, Notepad++에서 tagging된 데이터 자료에서 taggers를 제거한다(BNC Sampler를 이용).

**Step4:** 다음, 데이터에서 언어의 패턴을 찾는데 이용할 수 있는 Regular expressions을 익힌다. 마지막으로, Regular expressions를 이용하여 데이터에서 특정 패턴을 찾는다.

**과목: A-4: 102호**
**제목: 코퍼스와 어휘교육 (기초)**
**강사: 신동광 (한국교육과정평가원)**

본 강좌는 코퍼스 입문반으로서 코퍼스에 대한 전반적인 개념 및 역사와 함께 교실에서의 다양한 활용 방식을 어휘 분석에 초점을 두고 진행됩니다. 또한 쉬운 설명과 함께 다양한 코퍼스 분석 프로그램의 소개 및 실습도 포함되어 있습니다. 본 강좌에서 소개되는 자료 및 프로그램은 수강자의 활용도를 제고하기 위해 모두 무료 프로그램으로 선정되었습니다. 대표적으로 어휘분석의 고전이라고 할 수 있는 Paul Nation의 Range Program은 어휘수준 분석 및 어휘목록 개발에 활용되며 교재선택이나 학생들의 어휘수준 측정의 자료로도 활용가능합니다. Laurence Anthony의 AntWordProfiler는 교재분석, 수준별 교재개발 및 영어 읽기 문항 출제 시 지문의 어휘 난이도 관리에 매우 유용한 프로그램입니다. Tom Cobb의 코퍼스 자료 웹사이트인 Lextutor에는 다양한 프로그램이 탑재되어 있지만 본 강좌에서는 온라인 오류 피드백을 위한 concordance lines의 하이퍼링크 방법만을 소개할 예정입니다. 끝으로 국내에서는 잘 알려지지는 않았지만 베를린의 Freie University에서 개발한 TextSTAT는 웹사이트 주소를 입력하거나 텍스트 파일 형태의 코퍼스를 업로드하는 것으로 하나의 맞춤형 코퍼스의 제작을 도와주는 프로그램입니다. 어휘목록 개발과 같은 고난위도의 실습은 강좌의 성격상 제외되었지만 본 강좌는 실제 활용을 위한 실습 위주의 실용적인 교육 내용으로 구성되어 있습니다. 다음의 <표>는 본 강좌의 세부 교육 내용과 일정을 정리한 것입니다.

< 세부 교육 내용 및 일정 >

| 교육 내용 | 교육 방식 | 사용 프로그램 |
|---|---|---|
| 코퍼스의 개념 및 역사 | 강의 | - |
| 어휘분석을 위한 어휘 개념 | 강의 및 실습 | Range program |
| 교재 어휘 수준 분석 및 수준별 교재 선택 방법 | 강의 및 실습 | Range program, AntWordProfiler |
| 어휘평가 검사지개발 | 강의 및 실습 | Range program |
| 어휘교육에 대한 오해 | 강의 | - |
| 점심 | | |
| 코퍼스 분석을 통한 학습자 오류 분석 | 강의 및 실습 | AntConc |
| 코퍼스 분석 프로그램을 이용한 학습자 오류 관리 및 피드백 | 강의 및 실습 | Lextutor concordancer, AntConc, Range program |
| 학습자 맞춤형 온라인-오프라인 혼합형 코퍼스 구축 및 활용 | 강의 및 실습 | TextSTAT |
| 질의응답 및 정리 | | |

**과목: B-4: 230호**
**제목: NLP 태깅 및 통계 연어 측정방법 (중급)**
**강사: 김재훈교수 (한국해양대)**

본 강의는 컴퓨터프로그램에 익숙하지 않은 학습자도 누구나 쉽게 프로그램 실습을 따라할 수 있도록 준비하였다. 주요 강의 내용은 연어를 통계적인 관점에서 살펴보고 연어추출 방법론을 학습하고, Python을 활용한 자연언어처리를 통해서 코퍼스를 스스로 태깅해 본다.

**1강:** 통계적인 연어 추출방법을 (1)빈도 (2) 평균 및 변수, (3) 가설검증방법의 (t-test,카이스퀘어테스트), (4) 정보이론(Pointwise Mutual Information).

**2강:** 태깅을 하기 위하여 Python기초를 학습한다. 자연언어처리(Natural Language Processing, NLP)를 통해서 코퍼스를 다루는 방법을 실습을 통해서 학습한다. 먼저, Python를 다루기 위해 필요한 기초 개념을  PPT를 통하여 학습하고 실습해 본다. 다음으로 Python 프로그램의 자연언어처리 모듈인 NLTK를 활용하여 코퍼스를 직접 태깅 한다.