일본어 주석 코퍼스(tagged corpus)의 구축 방법에 대하여

真島知秀* · 金曙泳**
majima@youngdong.ac.kr · yuiyu@korea.ac.kr

<要 旨>

本稿は日本語タグ付きコーパス(tagged corpus)の構築方法について、高麗大学李漢燮教授研究室で作業が進められている、日韓並列コーパス構築の際の問題点を交えながら述べたものである。本稿ではまずテキストコーパスと音声コーパスの違い、また生コーパスとタグ付きコーパスとの違いについて説明し、各研究分野におけるさまざまな形態のコーパスについても見てきた。そしてタグ付きコーパスがなぜ必要なのか、その必要性について述べた。タグ付きコーパスの実際の構築方法についても詳しく触れ、各段階における注意点やどうすれば有用なコーパスを構築することができるのかを考えてみた。

そして、タグ付け作業時における手作業と自動化の各方法を比較し、その長所と短所について正確で効率的という観点から検証してみた。その結果、最も現実的な方法としては、まず各種のタグ付け支援プログラムを利用して、全体的な加工を行なった後、手作業でプログラムの誤認識を修正し、編集していくという方法であった。しかしプログラムを使うこの方法では、プログラム側の仕様によって、求めている作業結果が得られないものも多く、常に自分の研究目的に合ったプログラムを探さなければならないという難点がある。しかし全ての作業を一つのプログラムに依存するのではなく、各々のプログラムの持つ特徴や長所などを把握し、様々なプログラムを組み合わせて自動化作業をすることができれば、より早く、より正確に、より大量のタグ付きコーパスを構築することができるということを確認した。

コーパスの価値を決める均衡性、妥当性、一貫性を維持したタグ付きコーパスを構築するためには、様々なプログラムに関する知識と、それをうまく組み合わせて活用することのできる能力が、これからは必要不可欠になっていくであろう。

Keyword : 원시 코퍼스, 주석 코퍼스, 텍스트 코퍼스, 음성 코퍼스

1. 들어가며

본 발표는 일본어 주석 코퍼스(tagged corpus)의 구축 방법과 그 문제점에 대하여 언급한 것이다.

코퍼스(corpus) 연구는 1960년대에 미국에서 시작된 컴퓨터를 이용한 언어연구의 한 분야이다. 1964년에 미국 Brown대학에서 최초의 영어 코퍼스[1]가 등장된 이후, 현재는 컴퓨터의 보급에 의해 영어뿐만 아니라 한국어[2], 일본어[3], 중국어[4] 등 각국의 언어를 대상으

* 永同大学校 日語学
** 高麗大学校大学院 日語学

1) Brown Corpus(The Brown University Standard Corpus of Present-Day American English). 미국 영어의 언어자료로서 1964년 Brown 대학에서 완성되었다. 1961년도에 미국에서 발간된 출판물 중에서 15의 장르에 걸쳐 무작위로 추출된 텍스트 500편(1편 약 2,000語)으로 된 총 100만語의 기계 가독형의 언어자료이다.

2) 한국과학기술원(KAIST) 국어정보베이스 용례 검색 시스템 등이 있음.

3) 京都大学 텍스트 코퍼스. 일본 毎日新聞의 1995년 1월 1일부터 17일까지의 신 기사(약 20,000文), 1월부터 12월까지의 사설 기사(약 20,000文).

4) 대만 중앙연구원 언어학연구소의 500만어 漢語 구어 코퍼스 등이 있음.

로 한 코퍼스 구축이 활발하게 이루어지고 있다. 이와 같이 코퍼스를 이용한 연구가 전 세계적으로 널리 알려지면서 각종 코퍼스의 구축이 진행되었는데, 이러한 배경으로는 코퍼스가 언어학의 연구 방법론 면에서 획기적인 변화를 가져왔기 때문이라고 할 수 있다.

코퍼스의 대표적인 특징으로는 단시간에 많은 연구 자료의 분석이 가능하도록 하는 "시간의 절약성"이라는 점과, 컴퓨터 가독형으로 만들어진 코퍼스 자료의 특성상, 연구자간 상호적인 자료교환을 보다 용이하게 하는 "자료의 공유성"이라는 점을 들 수 있다. 그리고 코퍼스는 앞으로의 연구 성과에 따라서 더 다양한 분야에서 이용할 수 있다는 "무한의 가능성"을 가진 자료이기도 하다. 이러한 이유로 현재는 코퍼스를 새로운 연구 분야로 삼아 그 활용 방법에 관심을 갖고, 실제로 언어학 연구에 이용하는 연구자가 점점 증가하고 있는 추세이다.

현재 한국에서는 21세기 세종계획의 일환으로 국어 기초자료 구축 사업이 진행되고 있으며, 고려대학교 이한섭 교수 연구실에서는 한일 병렬 코퍼스의 구축을 담당하고 있다[5]. 현재까지 많은 텍스트 자료를 수집하였으며, 세종계획의 지침에 따라 기초 편집 작업이 이루어지고 있다. 앞으로는 이 자료를 특수 목적으로 가공하는 작업에 들어갈 예정이나 작업을 하는데 있어서 시간적, 기술적 등 여러 가지 문제가 있는 것도 사실이다.

이에, 본고에서는 이한섭 교수 연구실에서 진행 중인 일본어 주석 코퍼스의 구축 방법과 지금까지 드러난 문제점, 그리고 그 해결 방안에 대하여 고찰하고자 한다.

2. 코퍼스의 종류

코퍼스란 인간의 언어행동에서 실제로 발화되거나 표기된 말을 컴퓨터 가독형 파일로 변화시킨 데이터의 집합을 말한다. 코퍼스는 그 사용 목적에 따라 다양한 형태가 있으며, 컴퓨터에 입력된 문서 파일은 물론, 음성 파일도 코퍼스라 할 수 있다.

우선, 코퍼스는 그 형태에 따라서 크게 텍스트 코퍼스와 음성 코퍼스로 나눌 수 있는데, 전자인 텍스트 코퍼스란 "어떤 문서를 컴퓨터에 전자화된 텍스트로 저장하고 컴퓨터에서 처리할 수 있는 형태로 구성한 것의 모음" 을 말한다. 그리고 후자인 음성 코퍼스는 "음성 언어를 전자화된 음성 파일로 저장하고 컴퓨터에서 처리 할 수 있는 형태로 만든 음성 자료의 모음" 을 말한다.

그리고, 앞서 언급한 텍스트 코퍼스와 음성 코퍼스를 주석(tag) 부착 여부에 따라서 "원시 코퍼스(raw corpus)"와 "주석 코퍼스(tagged corpus)"로 나눌 수 있다. 원시 코퍼스는 단순히 텍스트를 컴퓨터에 입력한 자료를 말하며, 주석 코퍼스는 원시 코퍼스에

---

5) 이한섭 교수 연구실에서는 한일 병렬 코퍼스 팀을 구성하여 2001년도부터 한일 병렬 코퍼스 구축을 진행중이다. 2001년에 13만 어절, 2002년에 18만 어절의 병렬 코퍼스가 구축되었으며, 2003년에는 29만 어절이 구축될 예정이다. 아울러 2003년에는 5만어절의 형태소 정보가 부착된 주석 코퍼스도 구축될 예정이다.

이용 목적에 따라 각종 부가 정보를 부여한 것이다. 이 두 가지 코퍼스의 형태적 차이를 비교해보면 다음 <표1>, <표2>와 같다.

<표1> 원시 코퍼스(raw corpus)

たばこ増税に伴って１日、ほとんどの銘柄が１箱２０円値上げされた。厚生労働省内には、全面禁煙の喫茶店がオープン、東京都豊島区では公共施設が禁煙となった。今年５月の健康増進法施行以来、愛煙家は肩身が狭くなる一方で、たばこ離れは一層進みそうだ。

<표2> 주석 코퍼스(tagged corpus)

たばこ[たばこ/nou] 増税[増税/nou] に[に/par] 伴っ[伴う/ver] て[て/par] １[１/nou] 日[日/nou] 、[、/sym] ほとんど[ほとんど/adv] の[の/par] 銘柄[銘柄/nou] が[が/par] １[１/nou] 箱[箱/nou] ２０[２０/nou] 円[円/nou] 値上げ[値上げ/nou] さ[する/ver] れ[れる/ver] た[た/par_ver] 。[。/sym] 厚生労働省[厚生労働省/nou] 内[内/nou] に[に/par] は[は/par] 、[、/sym] 全面[全面/nou] 禁煙[禁煙/nou] の[の/par] 喫茶店[喫茶店/nou] が[が/par] オープン[オープン/nou] 、[、/sym] 東京都[東京都/nou] 豊島区[豊島区/nou] で[で/par] は[は/par] 公共施設[公共施設/nou] が[が/par] 禁煙[禁煙/nou] と[と/par] なっ[なる/ver] た[た/par_ver] 。[。/sym] 今年[今年/nou] ５月[５月/nou] の[の/par] 健康増進法[健康増進法/nou] 施行[施行/nou] 以来[以来/nou] 、[、/sym] 愛煙家[愛煙家/nou] は[は/par] 肩身[肩身/nou] が[が/par] 狭く[狭い/adj] なる[なる/ver] 一方[一方/nou] で[で/par] 、[、/sym] たばこ[たばこ/nou] 離れ[離れ/nou] は[は/par] 一層[一層/adv] 進み[進む/ver] そう[そう/nou] だ[だ/par_ver] 。[。/sym]

동일한 신문 기사 내용에서 만든 <표1>와 <표2>의 코퍼스를 비교해보면, 단순히 문서 내용만 입력된 <표1>의 원시 코퍼스와 달리 <표2>의 주석 코퍼스의 경우, 문서 내용을 형태소 단위로 띄어쓰기 하여 각각의 형태소에 품사정보를 부여한 것임을 알 수 있다. <표2>와 같은 주석 코퍼스는 어떤 문서에 나타나는 단어를 품사 단위로 추출하고자 할 때 유용하게 사용할 수 있을 것이다. 예를 들어 명사만 추출할 경우 컴퓨터의 찾기 기능을 이용하여 "nou"를 검색하면, 그 파일 안에 있는 모든 명사를 추출할 수 있으며, 같은 방법으로 동사는 "ver", 조사는 "per", 부사는 "adv", 등으로 검색하면 원하는 정보를 손쉽게 얻을 수 있는 것이다[6]. 이러한 이용 방법은 <표1>의 원시 코퍼스에서는 불가능하다.

3. 코퍼스의 활용분야

주석 코퍼스는 원시 코퍼스와 달리 어떤 특정한 분야에서 사용하는 것을 목적으로 가

6) 이 코퍼스는 주석 코퍼스의 유용성을 설명하기 위해 만들어진 간이 코퍼스이며, 실제 언어 연구에 이용하기 위해서는 연구 목적에 맞게 더 세밀한 주석 기술과 형식의 검토가 필요하다. 이 코퍼스에서는 일본 학교문법에 의거하며 명사를 "nou", 동사를 "ver", 형용사를 "adj", 조사를 "per", 부사를 "adv", 조동사를 "per_ver"로 약식 표기하여, 구두점은 "sym"로 표기하고 있다. 그리고 동사나 형용사 등 활용하는 것은 괄호 안에 그 기본형을 기술하여, 단어의 빈도 조사에도 곧 바로 이용할 수 있는 형태로 가공되어 있다.

공된 코퍼스로, 그 특성상 연구 분야에 따라 여러 형태의 코퍼스가 있다. 여기에서는 주석 코퍼스가 어떤 연구 분야에서 유용하며 각 연구 분야에 따라 어떠한 종류의 주석 코퍼스가 있는지 소개하고자 한다.

3.1 문법 연구 분야

문법 연구 분야에서는 단어의 형태 변화와 그 의미·기능을 다루는 형태론, 문을 구성하는 단어의 배열 규칙과 그 기능을 연구하는 통사론, 더 나아가서 방대한 문맥에서 문과 문 사이의 언어적 구성 관계를 연구하는 텍스트론까지 그 연구 범위를 넓힐 수가 있다. 문법 분야에서 필요한 정보를 부여한 주석 코퍼스를 예시하면 다음 <표3>과 같다.

<표3> 격 관계 및 구문 구조 등, 문법 정보에 대해 기술한 코퍼스

思う[思う/動詞-自立-;N1{が/は}N2{を}-;]
なる[なる/動詞-自立-;N1{が/は}N2{に}-;]
いい[いう/動詞-自立-;N1{が/は}N2{に}N3{と}-;]
あげる[あげる/動詞-自立-;N1{が/は}N2{を}N3{に}-;]
もらう[もらう/動詞-自立-;N1{が/は}N2{に/から}N3{を}-;]

<표3>은 문법 분야에서 활용이 가능한 격 관계를 부여한 코퍼스의 한 예로, 격 관계나 구문 구조 등을 연구 대상으로 할 경우, 유용하게 사용할 수 있을 것으로 생각된다. 그러나 문법 정보를 기술한 코퍼스라 하더라도 그 연구 목적에 따라 여러 형태가 있다. 예를 들어 態(Voice) 연구, 相(aspect) 연구, 叙法(modality) 연구 등 각각의 연구 목적에 따라 주석 코퍼스의 정보 기술 형태가 달라질 수 있을 것이다.

3.2 어휘 연구 분야

다음으로 어휘 연구 분야의 코퍼스에 대하여 알아보기로 한다. 어휘 연구 분야는 코퍼스를 사용하는데 있어서 가장 기본적이고 활용도가 높은 것으로 생각된다. 다만, 일본어와 같은 띄어쓰기를 하지 않는 언어 체계에서는 문을 형태소 단위로 구분하여 단어를 추출하는 과정이 반드시 필요하므로, 텍스트를 단순히 전산화한 원시 코퍼스가 아닌 형태소 분석 코퍼스로 분석하게 된다. 이 형태소 분석 코퍼스에는 활용이 있는 각 단어를 기본형으로 고치고, 각 형태소마다 품사정보·활용정보·문 구조 정보 등, 각종 정보를 부여하는 것이 유용한 코퍼스 데이터가 된다. 어휘 연구 분야의 코퍼스를 예시하면 다음 <표4>와 같다.

<표4> 和語, 漢語, 混種語, 외래어 등, 일본어 어종 정보를 기술한 코퍼스

新しい[新しい/形容詞-自立-**和語**]
パソコン[名詞-固有名詞-一般-**外来語**]
を[を/助詞-格助詞-一般]
電気店[名詞-一般-**漢語**]
で[で/助詞-格助詞-一般]
買っ[買う/動詞-自立-**和語**]
て[て/助詞-接続助詞]
電子メール[名詞-固有名詞-一般-**混種語**]
を[を/助詞-格助詞-一般]
指導教官[名詞-一般-**漢語**]
に[に/助詞-格助詞-一般]
送り始め[送り始める/動詞-自立-**和語**]
た[た/助動詞]
。[。/記号-句点]

현재 여러 형태의 코퍼스가 각 연구 기관 혹은 개인에 의해 개발되고 있는 가운데 가장 기본적인 코퍼스는 형태소 정보를 기술한 것으로, 일본어 문장을 형태소 단위로 분할하여 각각에 해당되는 품사 정보를 부여한 것을 말한다. <표4>는 이 품사 정보가 부여된 코퍼스에 일본어 어종 정보까지 표시한 것이다. 이와 같은 코퍼스를 사용하면 단순 단어 빈도 조사에서는 알 수 없는 각 단어의 어종을 조사할 수 있으며, 어떤 표현에서 어떠한 단어를 사용했는지에 대해 어종 면으로부터의 사용 실태를 연구할 수 있을 것이다.

3.3 음성 연구 분야

다음은 음성 연구 분야의 코퍼스에 대하여 살펴보기로 한다. 지금까지 알아본 코퍼스 연구 분야는 책이나 소설 등 문자화된 자료를 입수하여 이것을 컴퓨터에 입력해 구축하는 텍스트 코퍼스의 사용이 일반적이었으나, 문자 대신 소리로 정보를 전달하는 음성 연구 분야에서도 코퍼스의 이용이 가능하다.

우선 녹음된 음성 데이터를 모은 것을 음성 코퍼스라고 하는데, 이 음성 코퍼스를 문자화하여 텍스트 코퍼스로 컴퓨터에 입력하는 작업을 거치게 되면 음성을 체계적으로 연구하는데 큰 도움이 될 것이다. 음성 연구를 하는데 있어서는 단순히 음성 코퍼스만 사용하는 것보다 텍스트 코퍼스를 병용하는 것이 각 코퍼스의 활용도가 훨씬 높아지게 된다. 그 이유에 대하여 간략하게 설명하면 다음 <표5>의 음성 연구용 코퍼스의 구축 단계를 들 수 있다.

<표5> 음성 연구용 코퍼스의 구축 단계 및 실제 구축 예

【1단계】 음성 코퍼스를 텍스트 코퍼스로 문자화
え一、私が生まれた場所はですね、あの一、神奈川県の川崎市でして、あの一、父が、えーと、

【2단계】 음성 코퍼스의 실제 발음 방법을 표시
エー | ワタクシ | ガ | ウマレ | タ | バショ | ワ | デス | ネ | アノー | カナガワ | ケン | ノ | カワザキ | シ | デシ | テ | アノー | チチ | ガ | エート

【3단계】 음성 코퍼스에 표출된 발화시의 각종 정보의 부여
(Fエー) | ワタクシ<H> | ガ | ウマレ | タ | バショ | ワ | デス | ネ | (Fアノー) | カナガワ | ケン | ノ | (W カワザキ ; カワサキ) | シ<H> | デシ | テ | (F アノー) | チチ | ガ | (F エート)

우선 1단계에서 음성 코퍼스를 텍스트 코퍼스로 문자화하는데 이 작업을 함으로 음성 코퍼스의 전체 내용 및 문맥의 파악과 원하는 표현이 대략 어디에 있는지 알 수 있게 된다. 이 텍스트 코퍼스와 음성 코퍼스를 컴퓨터의 각종 프로그램[7]을 이용하여 연결시키면, 더욱 효율적으로 음성 연구를 할 수 있을 것이다.

다음으로 2단계에서는 1단계에서 문자화한 텍스트 코퍼스의 실제 발음을 표기한 것이다. 문법 연구나 어휘 연구 분야와 달리 음성 연구 분야에서는 어떤 발음으로 발화했는가에 중점이 두어짐으로, 텍스트 코퍼스에 실제 발음 방법을 표기하는 것이 바람직하다. 그렇게 함으로 음성 연구에 필요한 정보를 단시간에 정확하게 얻을 수 있을 것이다.

마지막으로 3단계에서는 음성 코퍼스에 표출된 발화시의 각종 정보를 부여한 것이다[8]. 이와 같은 정보까지 문자화되어 있으면, 어떤 사람의 발화시에 있어서 특징이나 문제점 등을 개관할 수 있다. 이러한 발화시의 정보를 부여하는 것도 연구 분야에 따라 달라질 수 있으며, 코퍼스 구축의 설계시의 면밀한 검토를 요한다.

3.4 담화 분석 연구 분야

다음으로 담화 분석 분야의 코퍼스에 대하여 알아보기로 한다. 담화 분석이란 의사소통의 실제적인 단위인 담화의 구조를 분석하여, 인간 의사소통의 모든 면을 포함한 회화 상의 수행(competence)을 지배하는 원리를 연구하는 것인데, 코퍼스에 의해 구축된 방대한 텍스트 자료와 음성자료의 분석을 통해 객관적 담화처리 모형(모델)개발 등을 할 수 있을 것이다. 담화 분석 분야의 코퍼스를 예시하면 다음 <표6>과 같다.

---

7) 이 분야 연구에 대해서는 민광준(2002)에서 그 활용 방법에 대하여 상세히 언급되어 있다.

8) 이 코퍼스의 출처는 "開放的融合研究「話し言葉工学」による「日本語話し言葉コーパス (モニター版2002) 」"의 용례 샘플이며, 3단계에서 부여한 각종 표기의 설명은 다음과 같다.
【F】 발화와 발화 사이에 보이는 의미 없는 말
【H】 비정상적인 모음의 장음화
【W】 비표준적인 발음으로 발화된 것으로 (;) 표시는 그 표준적인 발음을 나타냄.

<표6> 담화 분석 연구용 코퍼스의 예

[<未知情報要求><関係有>]
01 : 01 : 970-01 : 04 : 6300015 L : {Fと}部屋はどうしましょうか
[<未知情報要求><関係有>]
01 : 05 : 650-01 : 07 : 1700016 R : 部屋の空き具合はどうでしょうか
[<未知情報応答><0016>]
01 : 07 : 840-01 : 19 : 6200017 L : {Fえと}月曜日は{Fえ(と)}朝の9時から12時まで会議室おなじく月曜日の14時から16時まで小会議(し)つが使用可能です
[<未知情報応答／依頼><0015>]
01 : 20 : 290-01 : 22 : 0200018 R : {Dでは}小会議室でお願いし(ます)
[<肯定・受諾><0018>]
01 : 22 : 620-01 : 23 : 3000019 L : 分かりました

<표6>[9]은 사회 언어학적 측면에서 담화를 분석하기 위해 만들어진 코퍼스이며, 어떤 관계의 사람이 어떤 장면에서 어떤 발화를 했는지를 고찰하기 위하여 설계되어 있음을 알 수 있을 것이다. 앞서 소개한 문법 연구나 어휘 연구, 그리고 음성 연구 분야의 코퍼스와는 상당한 차이를 보인다. 이 코퍼스의 경우도 원 데이터는 음성으로 된 것이므로, 음성 연구 분야의 코퍼스와 같은 방식으로 대화 내용이 녹음된 음성 코퍼스를 문자화하여 텍스트 코퍼스로 작성하는 과정이 필요하다.

3.5 교육 연구 분야

다음으로 교육 연구 분야의 코퍼스에 대하여 살펴보기로 한다. 이 코퍼스는 일본어 학습자의 작문이나 회화에 나타난 오용례를 수집하고 올바른 표현으로 수정하여, 오용의 유형 및 원인을 분석하기 위해 구축된 것이다. 코퍼스의 데이터에는 오용례 분석과 함께 학습자의 모국어, 연령, 성별, 일본어 학습력 등을 게재하여, 조사 대상자의 학습 환경을 고려한 오용례 분석을 가능하도록 하는 것이 바람직하다. 교육 연구 분야의 코퍼스를 예시하면 다음 <표7>과 같다.

9) 荒木雅弘(2000)의 코퍼스 샘플.

<표7> 교육 연구 분야용 코퍼스의 예

@Begin
@Participants : GAK EPD34003
@Language of GAK : English
@Study History of GAK : 日本語２０１
@Sex of GAK : male
@Age of GAK : ?
@Date: 1996
@Location : D University, USA
@Writion : H, keyboard
@Type : class assignment
@Coder : CHEN, WEN-MIIN, Ohso, Mieko
*GAK : ＫＳさんへ、
*GAK : はじめまして、Kさん。
*GAK : 私の名前はＲＣです。
*GAK : このメッセージははじめての電子メールですから、<ちっと>[*] <簡単言葉>[*] です。
%err : ちっと = ちょっと ;
簡単言葉 = 簡単な言葉 ;
簡単な言葉 = 簡単なメッセージ ;
*GAK : すみませんね。
*GAK : 今、私は シカゴのＤ大学で勉強しています。
*GAK : でも、<先年>[*] の８月から１２月まで大阪のＫ大学で勉強しました。
%err : 先年 = 去年 ;

<표7>은 일본어 학습자가 쓴 전자 메일의 내용을 오용 분석 연구용으로 편집하여 구축한 코퍼스의 예이다[10]. 코퍼스 파일의 상단 부분에는 학습자의 인적사항을 표기하여 어느 수준의 학습자인지를 구분할 수 있도록 되어 있다. 문장 속에 나타나는 괄호"<>"로 된 표현이 오용으로 판단된 부분이며, "%err" 부분에 올바른 표현이 제시되어 있다.

이러한 코퍼스가 인터넷을 통해서 공개되면 각 지역에 있는 일본어 교사들에게 유익한 연구 자료가 될 뿐만이 아니라 일본어 학습자에게도 독학으로 일본어를 배울 수 있는 기회를 마련해줄 것으로 예상되어, 그 이용 가치가 높다.

4. 일본어 주석 코퍼스(tagged corpus)의 구축

일본어 주석 코퍼스의 구축 방법은 우선 자료를 선정한 다음, 자료의 입력, 그리고 입력 자료의 확인 및 수정 과정을 거친 후, 마지막으로 주석 부여 작업에 들어가게 된다. 그럼 자료 선정 과정부터 순서대로 살펴보도록 한다.

10) 이 코퍼스는 일본 文部省 과학 연구비에 의한 특별 추진 연구『日本語の普遍性と個別性に関する理論的および実証的研究(1985-1989年度)』의 일환인「外国人学習者の日本語誤用例の収集・整理と分析」의 연구 결과 샘플이다.

4.1 자료선정

코퍼스를 구축하기 위한 가장 첫 단계는, 어떠한 목적으로 어떠한 분야를 대상으로 할 것인가 하는 코퍼스의 영역을 설정 한 후 이에 적합한 텍스트의 카테고리를 설정하여 그에 맞는 텍스트를 선정하는 것이다. 앞서 언급한대로 어떤 목적과 어떠한 분야를 대상으로 하는가에 따라 그 자료선정의 범위와 대상은 확연히 달라지나, 일반적으로는 다음과 같은 원칙으로 자료선정을 하는 것이 좋다.

그것은 코퍼스 영역에 맞는 범위 안에서 치우침 없는 자료를 선정하여 "균형성"을 가질 수 있도록 하고, 그 자료가 텍스트 카테고리에 타당한가 여부와 질과 양을 고려하여 "타당성"과 "신뢰성"을 가질 수 있도록 자료를 선정하는 것이다. 이 "균형성", "타당성", "신뢰성"을 충족시키는 자료를 선정하는 것이 유용하고 활용도가 높은 코퍼스를 구축하는데 있어서 중요한 첫 단계가 된다.

4.2 자료의 입력

자료의 입력은 텍스트 코퍼스인가 음성 코퍼스인가에 따라 크게 둘로 나누어 볼 수 있는데, 전자인 텍스트 코퍼스는 선정된 자료를 키보드를 이용해 직접 입력하거나, OCR(광학식 문자 판독기) 소프트웨어와 스캐너를 이용하여 입력한다. 그리고 후자인 음성 코퍼스는 직접 음성을 녹음하거나, 각종 음향 매체 등으로 녹음되어 있는 음성을 수집하여 음성 파일로 입력한다. 그리고 이렇게 수집한 음성자료는 향후 원활한 검색과 사용을 위해, 문자화하여 텍스트 코퍼스와 같이 컴퓨터에 입력하는 방법도 있다.

4.3 입력 자료 확인 및 수정

구축된 코퍼스의 신뢰도 및 타당성은 정확한 입력 작업과 교정 작업에 달려 있다고 해도 과언이 아니다. 코퍼스 구축시 다양한 원인에 의해서 오류가 발생하기 마련이므로 이 오류를 정확하게 찾아내어 수정하여야 한다.

4.4 주석(tag)의 부여

주석(tag)이란 것은 완성된 코퍼스에서 원하는 정보를 추출하기 위하여 텍스트의 임의의 단위에 언어정보를 부여하는 코드로서, 구축한 원시 코퍼스를 어떠한 목적으로 사용하는가에 따라 적절한 태그셋[11]에 맞추어 주석 작업을 해 주어야 한다. 이때에 부여

11) 태그셋(tag set)이란 주석을 부여하기 위해 고안된 작업 규칙과 주석 표기 방식의 일람이다. 일반적으로 주석 코퍼스의 구축 시에는 부여하고자 하는 주석에 모든 설명을 기술할 수 없으므로, 약식 표기로 표현하게 된다. 예를 들어 품사 정보를 부여할 때 동사를 "ver"로 표기하거나, 일본어 표현으로서 적절치 않은 부분을 "*"로 표기하여, 혹은 코퍼스 속에 나타나는 일련번호는 무엇을 의미하는가 등을 약식 표기로 기술한다. 주석 코퍼스의 구축은 모두 이 태그셋의 규칙에 의거하여 진행된다. 이와 같이 규칙을

되는 언어 정보로는 문법정보, 형태소 정보, 구문정보, 활용정보 등이 있다. 또한, 연구 목적에 따라 임의로 설정한 태그셋으로 부가정보를 부여 할 수도 있다.

4.4.1 왜 주석(tag)이 필요한 것인가

그런데 코퍼스를 구축할 때 복잡하고 반복적인 주석 부여 작업까지 왜 해야 할 것인가에 대한 의문도 생길 것이다. 이 부분에 대해서는 앞서 언급한 바와 같이 원시 코퍼스에서는 불가능한 것을 주석 코퍼스로 가능하게 해주기 때문이다.

인간은 어떤 텍스트를 자기 머리로 품사 단위로 인식하고 문의 구조를 분석하여, 어떤 상황에서 표출된 표현인지 생각할 수 있으나 컴퓨터에서는 그것이 불가능하다. 원시 코퍼스는 컴퓨터에 있어서는 단순히 문자의 나열에 불과하기 때문이다. 그러나 원시 코퍼스에 각종 주석을 부여하여, 컴퓨터가 인식할 수 있도록 명시적으로 표기되어 있는 주석 코퍼스가 있다면, 컴퓨터는 인간의 힘으로 분석하기에는 도저히 불가능한 대량의 데이터를 정확하게 그리고 순식간에 분석할 수 있는 능력을 가지게 된다.

컴퓨터를 언어 연구 목적으로 유용하게 사용하기 위해서는 원시 코퍼스를 컴퓨터가 인식할 수 있는 형태로 가공하여, 텍스트 속에 어떤 정보가 들어있는지를 명시하는 표지가 필요하다. 이와 같이 주석의 부여는 컴퓨터에 입력된 문서 데이터의 이점을 최대한 높이기 위해서 필요한 작업이라 할 수 있다.

4.4.2 주석의 항목

4.4.2.1 범용 주석

주석의 항목이란 주석으로 부여하는 기술 내용을 나타내며, 주석 부여 작업에 들어가기 전에는 어떤 언어 정보를 부여하는가에 대해서 고려할 필요가 있다. 주석은 코퍼스의 사용 목적에 따라 "범용 주석"과 "특수 주석"으로 나눌 수 있는데, "범용 주석"은 많은 사용자가 사용할 수 있도록 기본적인 언어 정보만 부여한 것을 말한다. 주석으로 부여하는 기본적인 언어 정보로서는 품사 정보를 들 수 있으며 이 종류의 주석을 예시하면 다음 <표8>과 같다.

<표8>범용 주석의 예

私[私/名詞]
は[は/助詞]
昨日[昨日/名詞]
図書館[図書館/名詞]
で[で/助詞]
勉強[勉強/名詞]
を[を/助詞]
し[する/動詞]
まし[ます/助動詞]
た[た/助動詞]
。[。/記号]

<표8>의 범용 주석을 살펴보면 언어 분석을 하는데 기본적인 품사 정보만 부여되어 있음을 알 수 있다. 범용성이 있다는 의미에서 어

세워 일관성 있게 작업하는 것이 코퍼스 구축의 대원칙이므로, 태그셋은 코퍼스의 설계서와 같은 중요한 것으로 볼 수 있다.

떤 연구를 하기 위해서도 비교적 활용이 쉽게 설계된 주석을 말한다. 그러므로 이 코퍼스는 그대로 자기 연구에 이용하거나 혹은 자기 연구 목적에 맞게 주석을 약간 추가·수정한 후 사용이 가능할 것이다.

4.4.2.1 특수 주석

한편, 특수 주석이란 어떤 특수 분야의 연구자가 사용하기 위해 특수 언어 정보를 부여한 것을 말한다. 범용 주석과 달리 범용성이 적으며, 특수 목적에만 사용이 한정된 코퍼스가 이에 해당된다. 예를 들어 앞서 예시한 <표6>의 담화 분석용 코퍼스도 이에 해당되는데, 이 종류의 코퍼스는 단어의 종류나 문의 구조를 나타내는 표지가 없으며, 화자간의 간계나 장면 설명을 나타내는 주석만 부여되어 있음이 알 수 있다. 이 형태의 주석 부여 방식은 담화 분석용으로 한정된 특수 주석이라 할 수 있다. 그밖에도 자기의 연구 목적에 맞게 임의로 주석을 설정하여 구축된 코퍼스에 특수 주석이 많은데, 이는 얼마든지 부가 정보를 부여할 수 있다는 주석 코퍼스 특징의 일면을 보여주고 있다.

4.4.3 주석 부여 방법

다음으로 주석 부여 방법에 대하여 알아보기도 한다. 주석을 부여하는 방법에 대해서는 수작업 방법과 자동 방법의 두 가지가 있다. 어느 방법이든 장단점이 있으므로, 각각의 방법에 대해 숙지하여 어느 과정에서 어떤 방법을 택하는 것이 가장 효율적인가를 고려할 필요가 있다.

4.4.3.1 수작업 방법

수작업 방법은 일련의 작업을 키보드로 직접 입력하는 작업 방법이다. 이 방법은 주석 부여 작업에 바로 들어갈 수 있으며, 주석 내용의 중간 변경도 비교적 쉽게 할 수 있다는 장점이 있다. 수작업으로 원시 코퍼스를 가공하여 품사 정보가 부여된 주석 코퍼스를 구축하고자 할 경우의 작업 과정을 설명하면 다음 <표9>와 같다.

<표9> 수작업 주석 부여의 과정

(1) 원시 코퍼스 (원문)	私は昨日図書館で勉強をしました。
(2) 형태소 단위로 분할 (띄어쓰기 하기)	私 は 昨日 図書館 で 勉強 を し ま し た 。
(3) 품사 정보의 부여	私[名詞] は[助詞] 昨日[名詞] 図書館[名詞] で[助詞] 勉強[名詞] を[助詞] し[動詞] ま[助動詞] た[助動詞] 。[記号]
(4) 활용이 있는 動詞, 助動詞 등은 기본형으로 수정	私[名詞] は[助詞] 昨日[名詞] 図書館[名詞] で[助詞] 勉強[名詞] を[助詞] **し[する/動詞]** **ま[ます/助動詞]** た[助動詞] 。[記号]
(5) 기술 방법에 일관성을 유지하기 위해 괄호 안의 기술 형식을 통일화함 (예 : 표출형[기본형/품사정보])	私[私/名詞] は[は/助詞] 昨日[昨日/名詞] 図書館[図書館/名詞] で[で/助詞] 勉強[勉強/名詞] を[を/助詞] し[する/動詞] ま[ます/助動詞] た[た/助動詞] 。[。/記号]

우선 (1)의 원시 코퍼스를 준비한다. 다음으로 (2)와 같이 형태소 단위로 띄어쓰기를 한다. 여기서 어느 부분에 분할하는가에 따라 주석 기술 정보도 달라진다. (2)의 띄어쓰기 작업이 끝나면 (3)과 같이 품사 정보를 부여하게 된다. 여기에서는 품사 용어를 일본 "学校文法"의 용어를 그대로 따르기로 한다. 그런데 <표9>에서는 일본 "学校文法"에 의거해 띄어쓰기 단위를 설정해 품사 정보를 부여하고 있으나, 이 띄어쓰기 단위는 연구자 자신이 임의로 설정해도 무방하며, 품사의 용어 사용에 대해서도 임의로 정할 수 있다. 다음으로 (4)와 같이 활용이 있는 품사는 기본형으로 수정한 것을 병기한다. 왜냐하면, 품사의 기본형은 어휘 연구를 하든 문법 연구를 하든 언어 현상을 분석할 때 반드시 필요한 정보이기 때문이다. 이 (2)~(4)까지의 작업이 가장 중요하고 기본적인 작업이 된다. 그리고 마지막 단계 (5)에서 주석 코퍼스의 표기 형식을 일관성 있게 모든 품사에 (4)와 같이 형태소와 품사 정보를 병기하는 것이 좋다. 이때 활용하지 않은 형태소는 괄호 안에 중복 표기가 되나, 컴퓨터로 처리하기 위해서는 편집 방식도 일관성 있게 가공하는 것이 앞으로의 활용 시에도 이점이 많다.

이상으로 수작업 방법시에 있어서 형태소 정보의 부여 과정에 대하여 순서대로 설명하였다. 그러나 이러한 작업을 통하여 대규모 코퍼스를 구축하려면, 우선 시간이 많이 소요되며, 작업 시에도 처음에 설정한 규칙과 다른 띄어쓰기 단위를 분할해버리거나 주석 설계 시에 없었던 품사 정보 용어를 사용하거나 하여 일관성을 유지하기 어렵다는 단점이 있다.

4.4.3.2 자동 작업 방법

다음으로 자동 작업 방법에 대하여 언급하면, 일련의 작업을 기계가 대신 해주는 방

법으로, 수작업 방법 시에 발생하는 오입력이나 일관성 유지의 문제를 상당 부분 해결할 수 있다는 장점이 있다. 요즘은 여러 연구 기관이나 개인이 개발한 주석 부여 지원 프로그램들이 인터넷을 통해 유상·무상으로 공개되어 있는데, 이런 프로그램들을 이용하면 수작업으로 몇 시간, 며칠 소요되었던 작업을 순식간에 그것도 일관성과 정확성을 유지하면서 대신 수행해줄 수 있다.

일본어 형태소 분석 프로그램은 "JUMAN[12]"이나 "茶筌(chasen)[13]" 등이 있으며, 일본어의 문 구조를 해석하는 프로그램은 "日本語構文解析システムKNP[14]"가 유용하다. 또한 주석 부여 지원 프로그램으로서는 "VisualMorphs[15]"가 공개되어 있으며 복잡한 작업 과정을 상당 부분 기계가 처리함으로, 정확도 높은 주석 부여를 가능하게 해준다.

그러나 수작업 방법과 달리 자동 작업 방법의 경우 주석 내용의 중간 변경이 어려우며 사전에 면밀한 주석 방식의 설계가 필요하게 된다. 왜냐하면 자동 작업의 경우 모든 텍스트에 대해 일괄적으로 처리가 이루어짐으로, 주석 설계상에 오류가 있을 경우 텍스트 전체에 문제가 생기기 때문이다. 자동 작업을 택할 경우는 프로그램의 동작 환경이나 분석 결과 등 프로그램의 특정에 대해 미리 숙지하여 주석 설계를 하는 과정이 중요하며, 작업에 착수하는데 까지 상당한 연구 기간이 필요하다는 것도 감안해야할 것이다.

5. 문제점 및 해결 방안

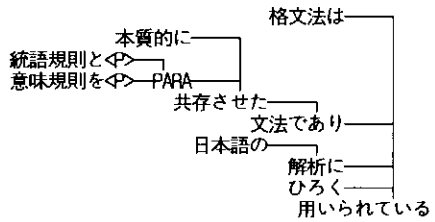
주석 코퍼스를 구축하는데 있어서는 여러 가지 문제점이 있으나 여기에서는 수작업 방법과 자동 작업 방법으로 나눠 그 문제점을 고찰하여 해결 방안에 대해 언급하도록 한다.

---

12) 東京大学大学院 情報理工学系研究科 電子情報学専攻 西田・黒橋研究室에서 공개되어 있다.
http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

13) 奈良先端科学技術大学院大学 情報科学研究究科「自然言語処理講座」松本研究室에서 개발된 일본어 형태소 분석기이며, 종래의 일본어 형태소 분석기 "JUMAN"을 개량하여 분석 속도를 크게 향상시키고 있다. 프로그램은 http://chasen.aist-nara.ac.jp/에서 무상으로 입수할 수 있다.

14) 京都大学情報学研究究科 知能情報学専攻 知能メディア講座言語メディア研究에서 개발된 일본어 구문을 분석하는 시스템이며, 예를 들어 다음과 같은 분석 결과를 얻을 수 있다. 프로그램은 http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html에서 무상으로 입수 가능하다.
【일본어 원문】 格文法は本質的に統語規則と意味規則を共存させた文法であり日本語の解析にひろく用いられている



15) 奈良先端科学技術大学院大学 情報科学研究究科「自然言語処理講座」松本研究室
http://chasen.aist-nara.ac.jp/vm/index.html.ja

우선 수작업 방법의 경우, 작업에 많은 시간이 소요된다는 문제점이 있다. 작업량이 A4용지 몇 페이지 정도면 수작업 방법으로 가능한 분량이지만, 몇 백 페이지, 몇 천 페이지가 되면, 인력으로 작업하기에는 도저히 불가능한 정도의 분량이 된다. 만약 이 방대한 작업을 수작업 방법으로 하게 되면, 많은 시간이 소요된다는 것을 짐작할 수 있을 것이다. 또한, 수작업 방법으로는 앞서 언급한 바와 같이, 오입력이나 일관성 있는 주석 방식의 유지가 어렵다는 것도 큰 단점이 된다.

다음으로 자동작업 방법의 경우, 오인식을 하거나 인식이 못하는 단어가 나타나는 등 기계적 처리에 있어서 완벽한 자동화가 어렵다는 문제점이 있다. 컴퓨터가 발달한 현재에도 인간의 언어를 기계적으로 완벽하게 처리하는 것은 불가능하다. 인간의 언어는 국제사회 등에서 시사용어가 만들어지거나, 어떤 사회집단에서 새로운 은어가 생기거나 하여 계속 변화가 이루어지게 된다. 이러한 매일 같이 새로 만들어지는 표현까지 완벽하게 인식하는 프로그램은 없다. 따라서 기계가 오인식하거나 인식을 못 하는 부분에 대해서는 결국 수작업으로 해야 한다는 것이다. 또한, 부여하고자 하는 주석 내용에 따라 자동작업 프로그램이 달라져야 하는 문제가 있다. 일반적으로 자동작업 프로그램은 어떤 정해진 작업을 위해 설계되어 있으므로, 그 이외의 사용자가 원하는 작업까지는 대신 해줄 수 없다는 것이다. 예를 들어 형태소 단위로 주석을 부여하고자 할 경우, 각 프로그램마다 형태소 분류 기준이 다르므로, 자기가 원하는 작업을 하기 위해서는 그에 적합한 프로그램을 찾을 필요가 있다는 것이다.

이와 같이 여러 문제점을 해결하는 방안으로써, 자동 작업을 바탕으로 하되 최종적으로 수작업으로 확인하는 방법이 있다. 수작업 시의 문제점에서도 알아본 것과 같이 수작업으로 방대한 양의 작업을 정확하게 그리고 일관성을 유지하면서 진행하는 것은 사실상 불가능하다. 따라서 1차적으로 자동작업 프로그램이 작업을 담당하여 그 작업 결과를 수작업으로 확인 및 수정해나가는 것이 제일 적합하고 현실적인 방법이라 생각된다.

또한, 각종 프로그램을 병용하면서 사용 목적에 맞는 주석 코퍼스를 구축하는 것도 중요하다고 생각된다. 자동작업 프로그램은 어떤 정해진 작업을 하기 위해서 설계되어 있으나, 각각의 프로그램이 가지는 작업 특징이나 장점을 파악하여 여러 프로그램을 복합적으로 사용하는 연구가 진행된다면, 단시간에 정확한 대량의 주석 코퍼스가 완성될 것이다. 그렇게 함으로써 균형성과 타당성 그리고 신뢰성을 갖춘 유용한 주석 코퍼스를 구축할 수 있을 것으로 생각된다.

6. 마치면서

이상으로 일본어 주석 코퍼스의 구축 방법과 각 방법에서 발생하는 문제점, 그리고

그 문제점의 해결 방안에 대하여 살펴보았다.

코퍼스는 대량의 데이터의 처리 및 공유를 가능하게 하며, 그 활용 방법과 분야에 있어서 많은 가능성을 가지고 있다. 따라서, 타당성과 신뢰성 그리고 균형성을 갖춘 텍스트 자료를 선정하여 코퍼스를 구축한다면 어휘, 음성, 담화 분석, 교육 등 많은 분야에서 유용하게 사용될 수 있을 것이다.

그러나 앞으로 코퍼스의 구축에 있어서 정확성과 효율성을 위해 기계적 처리에 관한 연구가 좀 더 진행되어야 할 것으로 생각된다. 앞으로 여러 연구기관에서 유용한 코퍼스가 공개되어 일반인에게도 이를 간편하게 이용할 수 있는 날이 올 때까지 코퍼스 구축에 관한 유익한 정보도 계속적으로 공개되어야 하며, 질과 양을 갖춘 코퍼스 구축을 위한 꾸준한 노력이 필요할 것이다.

◀ 参考文献 ▶

김홍규 · 강범모(1996), 「고려대학교 한국어 말모둠 1(Korea-1 Corpus): 설계 및 구성」, 『한국어학 3』, 한국어학회, pp.233-258

문화관광부(1999), 『21세기 세종계획 연구보고서 -전자사전 개발- 』, 문화관광부

문화관광부(2002), 『21세기 세종계획 국어 특수자료 구축』, 문화관광부

민광준(2002), 「제7차 교육과정 중 · 고등학교 일본어 교과서 텍스트 데이터베이스의 작성과 활용 방안」, 『외국어교육(Foreign Languages Education) 9(2)』, 한국외국어교육학회, 327-341

이한섭(1997), 「어휘조사 단위에 대한 연구 -일본 국립국어연구소의 각종 어휘조사를 중심으로-」, 국립국어연구원 보고서

이한섭(2000), 「일본어 코퍼스 구축에 관한 기본 구상」, 『언어정보 3』, 고려대 언어정보 연구소, pp.33-51

이한섭(2003), 「일본어 연구에 있어서 컴퓨터의 활용에 대하여」 『일본어학연구 7』, 한국일본어학회, pp.1-10

유민아(2003), 「한일 병렬 코퍼스 구축의 실제와 문제점」, 『일본어학연구 7』, 한국일본어학회, pp.109-124

荒木雅弘(2000)「音声対話コーパスへの自動タグ付け技術の開発対話システム自動構築への応用」産業技術研究助成事業平成１２年度採択成果報告予稿集, 産業開発業務部 研究助成課

国立国語研究所(2002), 「開放的融合研究『話し言葉工学』による『日本語話し言葉コーパス（モニター版2002）』」

寺村秀夫(1990), 「日本語学習者の日本語誤用例集」（科学研究費 特別推進研究「日本語の普遍性と個別性に関する理論的及び実証的研究」代表者井上和子、分担研究「外国人学習者の日本語誤用例集、整理及び分析」資料）

■ 투　　고 : 2003. 9. 30

■ 심　　사 : 2003. 10. 25

■ 심사완료 : 2003. 11. 20