

국립국어원 2019-01-04

발 간 등 록 번 호
11-1371028-000761-01

일상 대화 말뭉치 구축

사업 책임자
나 윤 정

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘일상 대화 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 5월 9일 ~ 2019년 11월 8일

2019년 11월 8일

사업 책임자: 나운정((주)메트릭스코퍼레이션)

사업 수행기관

(주)메트릭스코퍼레이션

사업 책임자

나운정

사업 참여자

조일상, 이영미, 박래희, 안재준, 안수정,
황민수, 이유림, 한금만, 박두진, 이혜진,
하지영, 김주연, 윤지은, 이은미, 채운철,
이형주, 서라벌, 이경화, 박혜수, 이영경,
신현주, 김도현, 김태경, 임유종, 백경미

〈사업 수행자〉 (주)메트릭스코퍼레이션

사업 책임자	나윤정((주)메트릭스코퍼레이션)
	조일상((주)메트릭스코퍼레이션)
	이영미((주)메트릭스코퍼레이션)
	박래희((주)메트릭스코퍼레이션)
	안재준((주)메트릭스코퍼레이션)
	안수정((주)메트릭스코퍼레이션)
	황민수((주)메트릭스코퍼레이션)
	이유림((주)메트릭스코퍼레이션)
	한금만((주)메트릭스코퍼레이션)
	박두진((주)메트릭스코퍼레이션)
	이혜진((주)메트릭스코퍼레이션)
	하지영((주)메트릭스코퍼레이션)
	김주연((주)메트릭스코퍼레이션)
사업 참여자	윤지은((주)메트릭스코퍼레이션)
	이은미((주)메트릭스코퍼레이션)
	채윤철((주)메트릭스코퍼레이션)
	이형주((주)메트릭스코퍼레이션)
	서라벌((주)메트릭스코퍼레이션)
	이경화((주)메트릭스코퍼레이션)
	박혜수((주)메트릭스코퍼레이션)
	이영경((주)메트릭스코퍼레이션)
	신현주((주)메트릭스코퍼레이션)
	김도현((주)메트릭스코퍼레이션)
	김태경(한양대학교)
	임유중(한양대학교)
	백경미(한양대학교)

요 약 문

4차 산업혁명 대비 기반 기술의 개발 및 인공지능 기술의 산업의 발로로 대규모 고품질 우리말 자원에 대한 수요가 증대되고 있다. 본 사업은 국어 자원의 활용도와 가치를 높일 수 있도록 민간에서 활용 가능한 공공재로서의 국어 말뭉치 확대를 위한 기초 자료를 마련하고자 하는 데 그 목적이 있다. 이를 위한 본 사업의 주요 목표는 일상 대화 말뭉치를 구축하여 대화 시스템 개발 등 민간에서의 자유로운 활용을 지원하고자 하는 것이다.

사업 내용: 일상 대화 말뭉치 구축 사업은 특정 주제에 대해 두 사람이 자유롭게 나누는 일상 대화를 녹음하고 정제하며, 그 음성 자료를 전사하여 원시 말뭉치를 구축하고, 대상 자료에 대한 메타 정보를 구축하는 것을 주요 내용으로 한다. 구축된 음성 자료 분량은 정제 후 1,000시간이다.

음성 녹음: 2,000쌍(총 4,000명)이 30분간 특정 주제에 대해 나누는 자연스러운 일상 대화를 녹음했다. 일상 대화의 주제는 군대, 게임, 휴일, 자동차, 영화, 건강/다이어트, 방송/연예, 스포츠/레저, 먹거리, 자연/휴양림, 국가/지역, 연애/결혼, 경제/재테크, 만화, 정치, 문화 등 16개를 최종적으로 선정하였다. 화자는 성별, 연령, 지역별 인구 특성을 고려하여 선정하였다. 화자에게는 본 사업의 목적을 다시 한 번 설명하고 이용 허락 계약을 체결하고 녹음 시 유의 사항을 고지한 후 녹음 장비를 착용하고 녹음을 진행하였다. 녹음은 주로 좌담회실과 회의실 등 주변 소음이 최소화된 별도의 공간에서 진행되었다. 두 명의 화자는 각자 헤드셋 마이크를 착용하고 발화하였으며, 화자별로 샘플 녹음 후 본 녹음을 실시하는 방식으로 진행하였다. 음성 파일은 16KHz 표본화, 16bit 양자화 선형 PCM으로 저장하였다.

음성 자료 전사: 1차적으로는 음성 파일을 자동 전사 시스템을 활용해 전사하고 이를 전문 속기사가 2차적으로 보완하는 방식으로 진행하였다. 2차 전문 속기사는 전사 규칙에 따라 발화자 표시 변경, 전사 단위와 전사 기호 반영, 문장 기호 반영, 들리는 대로 전사, 끊어진 단어 교정, 숫자 및 로마자 한글화, 띄어쓰기, 한글 맞춤법, 겹침 발화, 발화자 중복 표시 등을 수정하였다.

원시 말뭉치 구축 및 메타 정보 구축: 전사 파일을 대상으로 녹음 날짜, 대화 주제, 화자 간 관계, 화자 정보(성, 연령대, 직업, 출생지, 주 성장지 등) 등의 메타 정보가 기록된 헤더 정보를 생성 및 부착하고, 발화 단위로 마크업 하는 등 XML 형식으로 변환하는 작업을 수행함으로써 자료 활용도를 높이고자 하였다.

음성 자료 정제: 프라트(Praat)의 스크립트 기능을 활용해 1차적으로 음성을 자동 분할하고, 2차적으로 음성 정제자가 음성을 직접 들으면서 전사 단위에 따라 분할 경계를 수정하고 개인 정보를 비식별 처리하였다.

데이터베이스 구축 및 관리 시스템 개발 운영: 음성 파일, 전사 파일, 이용 허락 계약서 파일, 정제 파일, 메타 정보 등의 관리를 위한 데이터베이스를 구축하고 지침에 따라 파일명이 자동으로 생성되는 웹 관리 시스템을 개발 운영하였다. 데이터베이스와 관리 시스템은 정보망 보안 관리 규정에 따라 관리하고 장애 대책 및 주기적 백업 등을 통해 자료를 안정적으로 관리하였다.

사업의 기대 효과: 일상 대화 말뭉치 구축 사업을 통해 일상 대화 말뭉치가 음성 인식 등 언어 인공 지능 기술 개발 기반 자료로 활용될 수 있는 발판을 마련할 수 있을 것으로 기대된다. 성별, 연령, 지역이 다양한 화자가 특정한 주제에 대해 나눈 대화의 수집을 통해 일상 대화에 나타난 다양한 언어 실현 양상을 이해하고, 나아가 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용하고자 한다.

주요어: 일상 대화, 원시 말뭉치, 주제별 대화, 자연스러운 대화, 음성 정제

차 례

제1장 사업 개요

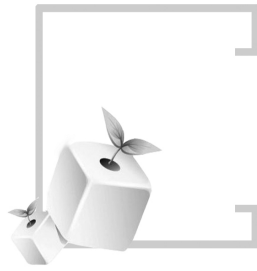
1. 사업 목적	3
2. 사업 수행 범위	4
3. 사업 수행 절차	5

제2장 사업 수행

1. 대화 주제 선정	9
2. 교육	10
3. 일상 대화 녹음	23
4. 음성 자료 전사	35
5. 원시 말뭉치 구축 및 메타 정보 구축	42
6. 음성 정제	47
7. 데이터베이스 구축 및 운영	51

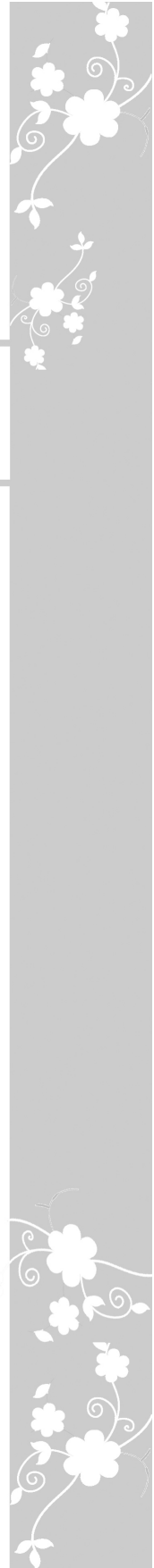
제3장 사업 수행 결과

1. 주제별 수집 결과	61
2. 참가자 특성별 수집 결과	62
3. 정책 제언	71



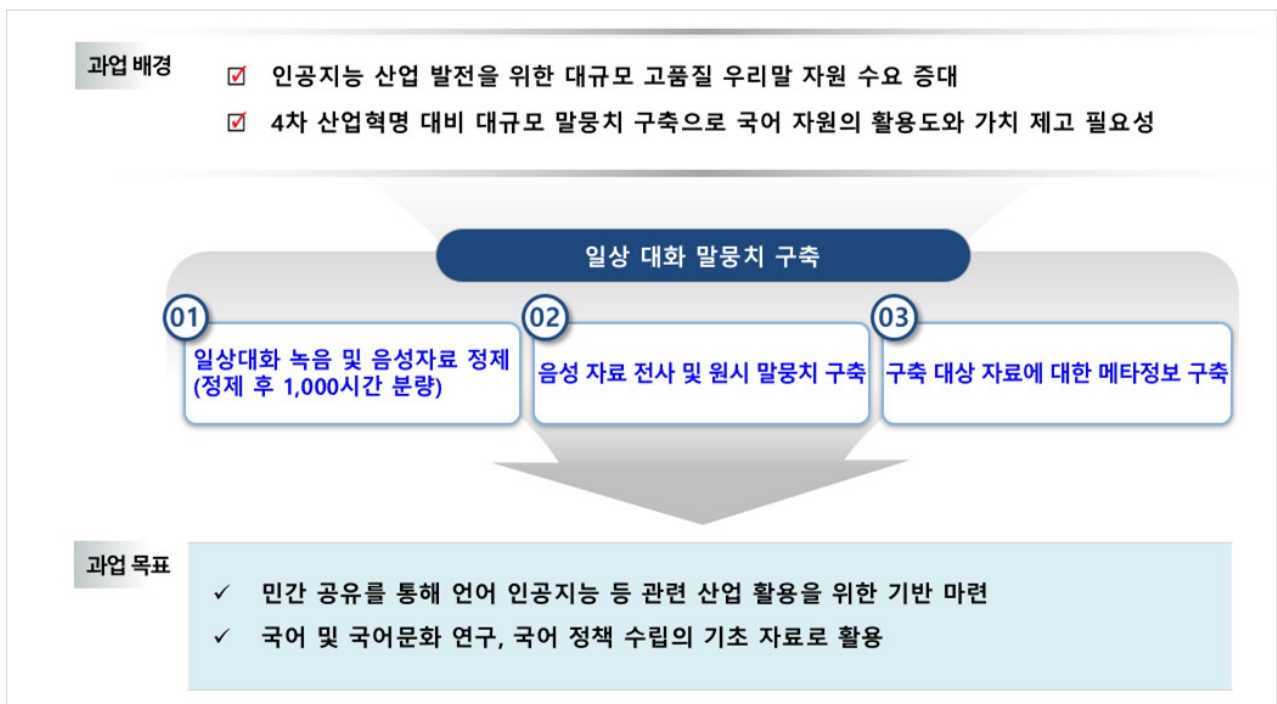
제 1 장

사업 개요



1. 사업 목적

4차 산업혁명 대비 기반 기술의 개발 및 인공지능 기술의 산업의 발로로 대규모 고품질 우리말 자원에 대한 수요가 증대되고 있다. 본 사업은 국어 자원의 활용도와 가치를 높일 수 있도록 민간에서 활용 가능한 공공재로서의 국어 말뭉치 확대를 위한 기초 자료를 마련하고자 하는 데 그 목적이 있다. 이를 위한 본 사업의 주요 목표는 일상 대화 말뭉치를 구축하여 대화 시스템 개발 등 민간에서의 자유로운 활용을 지원하고자 하는 것이다.



<그림 1> 사업 추진 배경 및 목표

2. 사업 수행 범위

본 사업의 수행 범위는 크게 네 부분으로 나눌 수 있다. 첫째는 특정 주제에 대해 두 명의 화자가 자유롭게 30분간 대화하는 상황을 전체 1,000시간 분량 녹음하는 것이다. 둘째는 음성 자료를 전사 지침에 따라 전사하는 것이다. 셋째는 전사 결과물에 헤더 정보 등을 부착하여 원시 말뭉치 형태로 가공하고, 구축된 자료에 대하여 녹음 날짜, 대화 주제, 화자 정보, 화자 간 관계 등 메타 정보를 구축하는 것이다. 넷째는 음성 자료를 전사 단위에 따라 분할하여 정제하는 것으로, 이때 대화 주제와 무관한 대화는 제외하고 개인 정보는 비식별 처리한다.

사업 수행을 위해 데이터베이스를 구축하고 관리 시스템을 개발해 산출물을 관리하고 효율적으로 운영하였다. 데이터베이스와 관리 시스템 외 각종 산출물은 관리자 및 접근 권한이 있는 자 외에는 접근이 불가능하도록 했으며, 정기적 백업을 진행하였다.

<표 1> 사업의 범위와 과업 내용

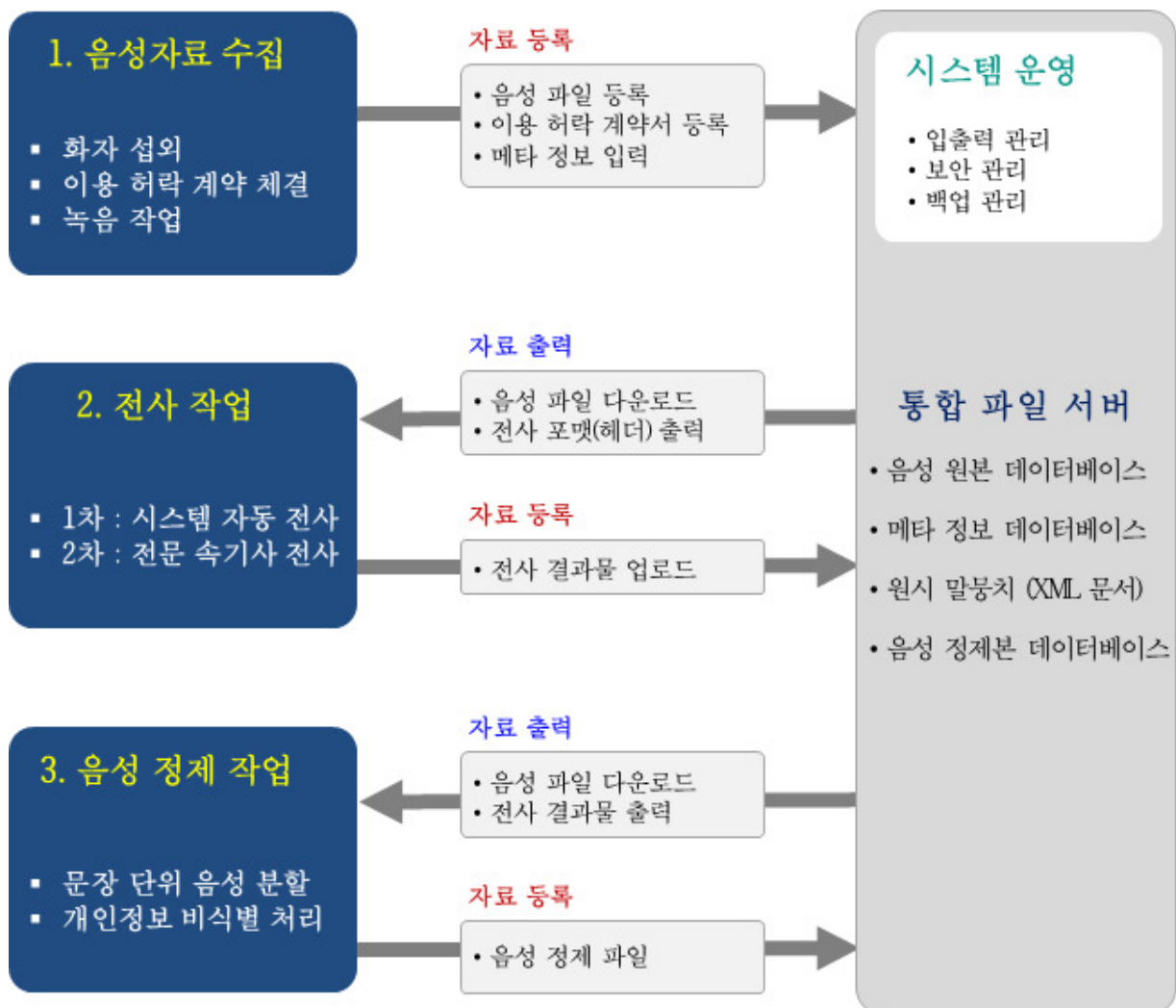
사업의 범위	과업 내용
음성 녹음	<ul style="list-style-type: none"> • 두 명의 화자가 특정 주제에 대해 자유롭게 대화하는 음성 녹음 • 성별×연령×지역별 인구 비율을 고려하여 화자를 모집 • 한 화자당 최대 녹음 시간을 30분으로 제한하며, 동일한 화자가 중복되지 않도록 하기 위해 2,000쌍(4,000명)을 모집 • 두 명의 화자가 자유롭게 대화할 수 있는 장소에서 진행하며, 발화가 겹치지 않도록 화자 간 거리가 적정 수준이 되도록 함 • 각자 헤드셋 마이크를 착용하고 발화하고 음성의 최대 샘플값이 10,000~20,000이 되도록 음량을 조절함. • 16KHz 표본화, 16bit 양자화 선형 PCM으로 저장 • 녹음된 음성 자료 및 전사 결과물에 대해 화자와 이용 허락 계약을 체결
음성 자료 전사	<ul style="list-style-type: none"> • 음성 자료의 전사는 발화된 그대로 전사하는 것을 원칙으로 함. • 세부적인 전사 규칙은 전사 지침을 따름.
원시 말뭉치 및 메타 정보 구축	<ul style="list-style-type: none"> • 전사 결과물에 대해 헤더 정보 부착 등의 표지 부착 작업을 수행하여 원시 말뭉치 형태로 가공 • 메타 정보에는 녹음 날짜, 대화 주제, 화자 정보, 화자 간 관계가 포함됨.
음성 정제	<ul style="list-style-type: none"> • 전사 단위로 음성 분할 • 대화 주제와 무관한 대화는 제외하여 정제 • 음성 구간이 잘리지 않도록 정제하며 음성 구간 앞, 뒤에 200msec 이상의 휴지가 포함되도록 정제

3. 사업 수행 절차

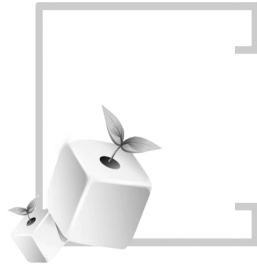
먼저, 모집된 화자를 대상으로 이용 허락 계약 체결 후 음성 자료를 수집하고 메타 정보를 수집한다. 수집된 음성 자료를 통합 파일 서버에 메타 정보, 이용 허락 계약서와 함께 등록한다.

다음으로 1차적으로 서버에 등록된 음성 자료를 시스템에서 자동으로 전사 후 자동 전사된 결과를 전문 속기사가 전사 규칙에 맞게 전사되었는지 확인하고 수정 후 최종 전사 파일을 통합 파일 서버에 등록한다. 그리고 통합 파일 서버에서 전사 파일과 메타 정보를 합하고 표지를 부착해 XML 형태로 변환한다.

마지막으로 음성 자료와 XML 형태로 변환된 전사 파일을 통합 파일 서버에서 내려 받아 전사 단위와 일치하고 정제 규칙에 맞도록 음성을 정제한다.

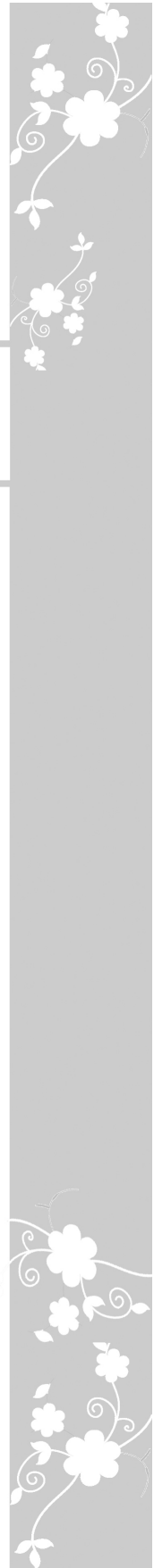


<그림 2> 사업 수행 절차



제 2 장

사업 수행



1. 대화 주제 선정

대화의 주제는 쉽고 편한 것으로 화자가 흥미를 가지고 대화할 수 있는 16개를 최종 선정하였다. 특정 주제에 편중되지 않도록 주제별로 균일하게 125쌍으로 진행하였으나, 모집 과정에서 화자들의 선호도가 낮은 만화, 정치, 문학 등의 주제는 최소 70쌍 이상이 되도록 조정하여 진행하였다.

대화 주제는 화자가 직접 선택하도록 했으며, 주제에 대한 이해를 돕고자 세부 주제를 예시로 들어 주제 선택이 용이하도록 하였다.

<표 2> 일상 대화 주제 및 최소 목표 모집 수

주제	세부 예시 주제	최소 목표 (쌍)
군대	군대 경험 등	125
게임	게임 종류, 게임 방법, 게임 경험 등	125
휴일	휴일, 휴일 여행 등	125
자동차	자동차, 자동차 관리, 자동차 종류, 위반, 주차 등	125
만화	만화, 만화영화, 만화 작가, 웹툰 등	70
영화	영화, 영화인, 영화관, 영화제 등	125
정치	정치인, 정치적 이슈, 선거 등	70
건강/다이어트	질병, 다이어트, 건강관리, 건강 검진 등	125
방송/연예	연예인, 드라마, 방송 프로그램, 연예계 이슈 등	125
스포츠/레저	스포츠, 운동선수, 올림픽, 운동 경기 관람, 운동 등	125
먹거리	음식, 음료, 요리법, 요리사, 맛집 등	125
자연/휴양지	산, 공원, 바다, 국내외 휴양지 등	125
국가/지역	나라, 세계 도시, 국내 도시	125
문학	작가, 책, 오디오북 등	70
연애/결혼	연애, 결혼, 배우자, 애인, 결혼 생활, 데이트 등	125
경제/재테크	경제, 재테크, 자산 관리, 부동산, 금융/증권, 보험, 대출 등	125

2. 교육

2.1. 녹음 진행 요원 선발 및 교육

본 사업은 총 7개 지역(서울, 대전, 대구, 부산, 광주, 강원, 제주)에서 전국 16개 지역 거주자를 대상으로 2,000쌍의 녹음을 진행하는 대규모 사업인 만큼 녹음을 위해 전국적으로 총 25명이라는 많은 인원이 투입되었다. 지역별로 투입되는 관리자 및 진행 요원은 적정 자격 조건에 의해 우선 선발된 자 중 집체 교육을 이수하고 평가를 통과한 자를 최종적으로 선발하여 녹음 결과물의 품질이 균일하게 유지될 수 있도록 하였다.

우선, 각 지역의 진행 요원은 유사 사업 수행 경험자 및 좌담회 수행 경험자 중 전문 녹음 장비 작동 경험이 있는 자를 우선 선발하였다. 지역별 비표본 오차 방지를 위해 선발된 녹음 진행 요원을 대상으로 집체 교육을 진행하였다. 교육은 1차 교육 후 진행 요원의 업무가 예상보다 많다고 판단한 후 인력을 충원해 2차 교육을 실시하였다. 1차 집체 교육 시에는 11명이, 2차 집체 교육 시에는 14명이 최종 선발되어 전체 25명의 관리자 및 진행 요원이 선발되었다.

<표 3> 진행 요원 선발

구분	선발 기준 및 운영 내용	
선발 기준	<ul style="list-style-type: none"> • 좌담회 진행 경험이 3년 이상인 자 • 전문 녹음 장비 작동 경험이 있는 자 • 좌담회 진행 수행 능력 평가 등급 A~A+ • 최종 교육 이수 및 평가 통과자 	
투입 인원	<ul style="list-style-type: none"> • 초기 11명 교육 후 사업 초기 추가 인력 투입 • 전체: 관리자 7명, 진행 요원 18명 	
진행 요원 역할	<ul style="list-style-type: none"> • 진행 요원 1 <ul style="list-style-type: none"> - 화자 안내 - 화자 인적사항 확인 - 저작권 이용 허락 계약 체결 - 녹음 종료 후 사례비 지급 	<ul style="list-style-type: none"> • 진행 요원 2 <ul style="list-style-type: none"> - 녹음 진행 개요 설명 - 녹음 장비 작동 - 주제 이탈 및 녹음 관리 - 녹음 장비 관리

교육 내용은 크게 4단계로, 사업 배경 및 목적, 진행 시 유의 사항 등의 이론 교육과 녹음기 작동, 헤드셋 마이크 착용, 녹음 방법 등의 실사 교육, 화자 응대 태도, 불만 사항 발생 시 대처 방법 등의 CS 교육, 화자 정보 관리 등의 보안 교육으로 나누어 진행하였다.

교육 후 실제 녹음 장소와 동일한 환경에서 지역별로 2인 1조가 녹음 준비부터 진행, 완료까지 직접 실습하는 롤-플레이를 실시하였다. 롤-플레이 평가 결과가 적정 수준 이상인 자를 최종 녹음 진행 요원으로 선발하였다. 적정 수준에 미달인 자는 재교육을 실시해 다시 평가를 거쳐 선발하였다. 2차 평가에서 미달인 자는 최종 선발에서 제외되었다.

이러한 과정을 통해 최종 선발된 녹음 진행 요원들은 보안 서약서 작성 후 본 녹음 진행에 참여하였다. 지역별로는 전체 녹음 분량의 50% 이상을 녹음하는 서울은 6명(진행 요원 5명, 관리자 1명), 다음으로 분량이 많은 부산은 4명(진행 요원 3명, 관리자 1명), 기타 지역은 3명(진행 요원 2명, 관리자 1명)이 투입되었다.

<표 4> 진행 요원 교육

구분	내용
교육 일시 및 장소	<ul style="list-style-type: none"> 2019년 6월 28일 (주)메트릭스코퍼레이션 1층 대회의실
교육자	<ul style="list-style-type: none"> 박래희, 윤지은(주)메트릭스코퍼레이션
참석인원	<ul style="list-style-type: none"> 1차 집체 교육: 11명 2차 집체 교육: 7명
교육 내용	<ul style="list-style-type: none"> 사업의 목적 및 결과 활용 방안 절차 대화 주제 화자 대상자 선정 가이드 녹음 환경 및 녹음기 사용법 녹음 방법 대화 시 주의 사항 롤-플레이 평가 보안 교육 질의응답

4 단계 교육 시스템

1단계 이론 교육

- 사업의 배경 및 목적, 세부 진행 내용
- 진행 시 유의 사항 등

2단계 실사 교육

- 녹음기 작동법, 헤드셋 마이크 착용법
- 녹음 방법, 녹음 시 유의 사항

3단계 CS 교육

- 화자(참석자) 응대 태도 및 기본 자세
- 불만 사항 대처 방법 사전 숙지

4단계 보안 교육


- 개인 정보 노출에 대한 중요성
- 화자 정보 관리 방안

실제 녹음 상황과 동일한 환경에서 롤-플레이

<그림 3> 4단계 교육 시스템

8. 녹음기 볼륨

- 1) 테스트 녹음 후 어도비 오디오미터로 아래 그림과 같이 음성 샘플 최대값이 1만~2만이 되는지 확인
- 2) 테스트 음성이 평균 1만이 넘지 않는 경우 녹음기 볼륨을 높여 다시 테스트
- 3) 녹음기 오른쪽 고 녹음
- 4) 평균 음성 샘플 생합 (PEAK)



9. 대화 시 유의사항

- 1) 목소리 크기는
- 2) 마이크는 만지
- 3) 상대방의 말이
- 4) 웃음, 맞장구 등
- 5) '말 끝임', '말
- 6) 얼버무리지 않
- 7) 대화 중 5초 0

10. 자료 저장 및

- 1) 컴퓨터에 파일
- 3) 관리자 사이트
- 4) 별도의 사용설

말뭉치 구축사업 실사 교육 자료

1. 사업 목적

◎ 4차 산업혁명 대비 기반 기술 개발 및 인공지능 기술 개발, 활용을 위한 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치를 제고하고 민간 공유를 통해 언어 인공지능 등 관련 산업 활용을 위한 기반을 마련하고 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용하고자 함

2. 대화 주제

◎ 대화 주제는 다음의 16가지 중 한가지로 진행

◎ 대화의 주제는 기본적으로는 1개의 메인 주제로 진행하되, 대화 중 다른 주제가 같이 나오는 것은 인정한다.

(ex. 휴가 대화 중 음식에 대한 이야기가 나올 수 있음)

군대	만화	방송/연예	국가/지역
게임	영화	스포츠/레저	문학
투표	정치	먹거리	연애/결혼
자동차	건강/다이어트	자연/휴양지	경제/재테크

3. 실사 프로세스

설외 -> 동의서 -> 유의 사항 전달 -> 녹취 테스트 -> 녹취 -> 녹취 확인

4. 실사 준비물

- 1) 녹음기
- 2) 헤드셋 마이크 2대
- 3) 이어폰 2개 (연장선 포함)

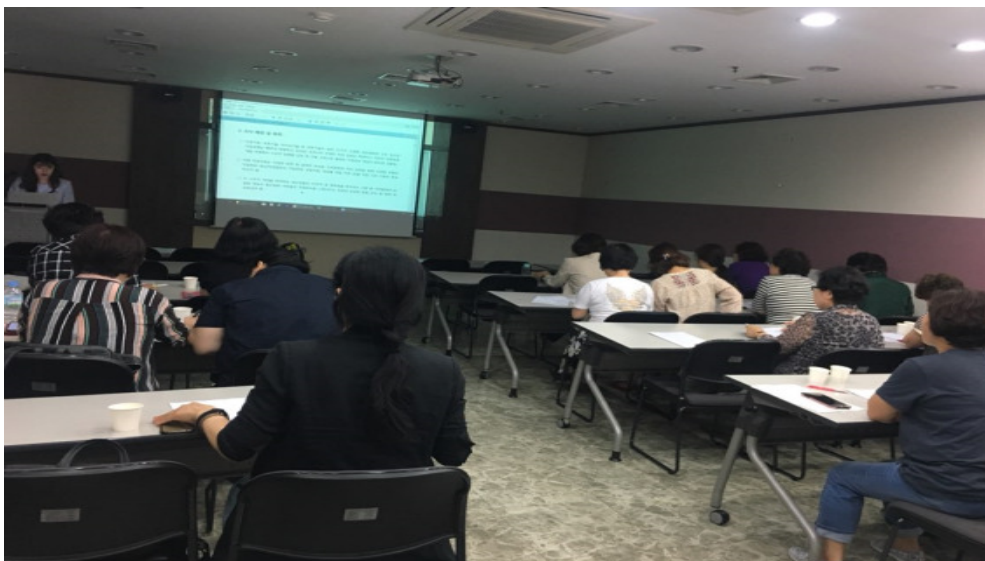
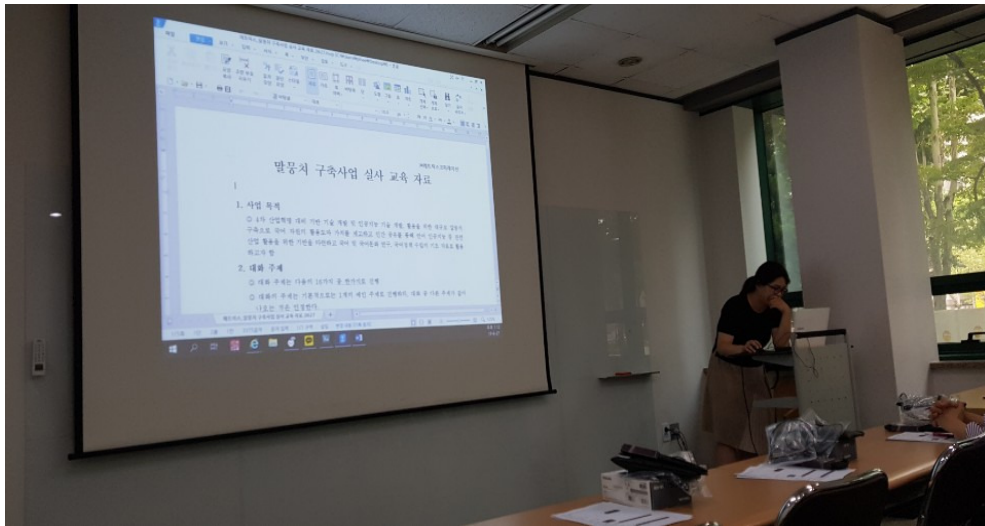
5. 녹음 환경

- 1) 가토*세토 크기가 3m*3m의 조용한 회의실
- 2) 독립된 공간으로 녹음소리 외의 잡음이 들어가지 않는 공간
- 3) 화자 간 파티션 등으로 마이크 개입을 최소화 함
- 4) 녹음 장소에는 화자들이 볼 수 있게 '대화 시 유의사항'을 붙여 놓는다.

6. 녹음 방법

- 1) 녹음기와 마이크 연결 후 녹음기와 마이크 모두의 전원을 켜다.
- 2) 마이크는 입에서 손가락 2~3개가 들어갈 정도의 거리로 하고 입과 같은 높이로 한다.
- 3) 'RECODE' 버튼을 한 번 누르면 버튼이 깜빡이고 볼륨을 테스트 할 수 있다.
- 4) 'RECODE'를 한 번 더 누르면 화면에 'REC'가 나타나고 녹음이 시작됨
- 5) 장소/시간/사람마다 소리의 크기가 다르게 녹음되므로 반드시 참여자별로 테스트 녹음을 진행해야 함
- 7) 테스트 녹음은 반드시 음성 샘플 폭을 확인해야 함
(음성 샘플 폭이 1만~2만이 되도록 마이크 볼륨 조절)

<그림 4> 녹음 진행 요원 교육 자료 일부



<그림 5> 진행 요원 교육 사진

2.2. 전사자 선발 및 교육

본 사업은 정제 기준 1,000시간, 한 쌍의 대화당 30분, 전체 2,000쌍이라는 많은 양의 녹음을 짧은 기간 동안 전사해야 하므로 많은 전사자를 투입하였다. 2차 수동 전사자로는 유사 사업 경험이 있는 자, 좌담회 및 인터뷰 속기 경험이 풍부한 전문 속기사를 선발하였다.

전사를 위해 사업 기간 동안 약 80명을 교육하였으며, 하루에 1인당 약 2개의 음성 파일을 전사하는 것을 기준으로 일평균 30명 정도가 투입되었다. 모든 전사자는 사업 수행 기관이 직접 관리 교육하였다.

전사자 간의 전사 단위, 전사 품질 수준의 차이를 최소화하기 위해 전사자 대상 집체 교육을 실시하였다. 사전 평가 시 1차 시스템에 의한 자동 전사의 정확도가 70% 정도였던 것을 감안해 1차 집체 교육 후 14명, 2차 집체 교육 후 26명을 전사에 투입하였다. 그러나 실제 작업 시 1차 시스템에 의한 자동 전사의 정확도가 평균 30% 이하 수준으로 낮고 30분 단위 음성 파일 한 개 전사 시 4시간 이상이 소요되어 중간에 탈락자가 많이 발생하게 됨에 따라 수시 교육을 통해 40명 정도를 더 투입하였다.

전사자 교육은 전사 지침과 유의 사항, 한글 맞춤법 위주로 진행되었으며, 그 외 사업 배경 및 목적, 사업의 절차와 방법, 전사 절차 등을 교육하였다. 마지막으로 문화체육부 고시 한글 맞춤법 해설을 전달하는 방식으로 진행되었다.

이러한 과정을 통해 최종 선발된 전사자들은 보안 서약서 작성 후 전사에 참여하였다. 교육 받은 전사자는 샘플 전사 후 전사를 교정 받아 두어 차례 수정 전사 후 본 전사에 투입되었다. 이 과정에서 전사 교육 이수자의 10% 이상이 샘플 전사 후 전사를 포기하거나 탈락했다.

<표 5> 전사자 선발

구분	선발 기준 및 운영 내용
선발 기준	<ul style="list-style-type: none"> 좌담회 및 인터뷰 속기 경험 3년 이상인 자 유사 사업 수행 경험자 전사 교육 이수 및 샘플 평가 통과한 자
투입 인원	<ul style="list-style-type: none"> 초기 14명 투입했으나 진행 중 중도 탈락으로 추가 인력 투입 전체 80명 투입, 일평균 30명 정도 진행
운영	<ul style="list-style-type: none"> 수도권 외 지역은 해당 지역 출신 전사자에게 해당 지역 음성 자료를 우선 배정함.

<표 6> 전사자 교육

구분	내용
교육 일시 및 장소	<ul style="list-style-type: none"> • 2019년 6월 28일 매트릭스 1층 대회의실
교육자	<ul style="list-style-type: none"> • 박래희(주)메트릭스코퍼레이션), 임유종(한양대학교)
참석인원	<ul style="list-style-type: none"> • 1차 집체 교육: 14명 • 2차 집체 교육: 26명, 수시 교육: 40명
교육 내용	<ul style="list-style-type: none"> • 사업의 배경 및 목적, 사업의 절차와 방법 • 전사 절차 • 전사 지침 및 유의 사항 • 한글 맞춤법 주요 내용 • 문화체육부 고시 한글 맞춤법 해설 전달 • 보안 교육 • 질의응답

말뭉치 구축사업 전사 교육 자료

1. 사업 목적

4차 산업혁명 대비 기반 기술 개발 및 인공지능 기술 개발, 활용을 위한 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치를 제고하고 민간 공유를 통해 언어 인공지능 등 관련 산업 활용을 위한 기반을 마련하고 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용하고자 함

2. 대화 주제

◎ 대화 주제는 다음의 16가지 중 한가지로 진행

문대	만화	방송/연예	악기/지역
게임	영화	스포츠/레저	문학
취업	정치	먹거리	언어/일본
자동차	건강/다이어트	자연/휴양지	경제/해태크

3. 사업 절차와 방법

(1) 대화 녹음 - 5인 대화로 녹음함

(2) 자동(1차) 전사 - 대화 녹음 자료를 음성 인식 프로그램으로 처리하는 1차 전사 작업

(R) 일상생활 이벤트에 대해서 유가 계획 있으세요.

(L) 우리는 때마다 여름 때 이제 방학 때 맞춰서 해외로 나갔었는데 올해는 아직 특별한 계획이 없네. 작년 같은 경우는 우리 코타키나탈로 갔다 왔을 거다 칠수 씨는 뭐 어디가 주를 위하여 여행 이라든지 한 되게 좋은데 새로운 것을

(R) 아 저는 아버지께서 지리교사 하시거든요. 그래서 전국 지도를 그냥 안 보고 다니세요. 그래서 되게 좋은데 볼 국내에서 많이 데려가 주셨는데 작년에는 거제도 갔고요 올해 초에 겨울에도 통영 쪽에 갔었는데 너무 풍경이 좋았어요. 근데 올해는 어디 갈지 아직 안 정했는데 얼마도 또 안 가면 새로운 곳을 데리고 가실 거 같아요.

(L) 아 그러면 여름 때마다 여름에는 이제 그 지리 속이 안 머릿속의

(R) 내가 없습니다.

그때 인천공항 말하는 거잖아. 근데 우리 싱가포르 도착했을 때 완전 다 예드* 뭐야 다 녹초가 되가지고 막 핸드폰만 보다가 시간 언제 가지 이라고 기다렸잖아 너도 막 피곤해 가지고 막 가질래 있고

- 짧은 대답 이후에 곧바로 발화가 이어진 경우에는, 대담에 마침표가 있어도, 대담과 후행 발화 사이에 엔터 표시를 하지 않아도 됨.

참 발화: 응. 약간 그래 가지구 놀았어.

전사: 응. 약간 그래 가지구 놀았어.

- 느낌표나 쉼표는 사용하지 않는다. 문장이 완전히 종결이 되었을 때는 마침표를 사용한다.

뉴스통의 영커보리링을 시작하겠습니다.

최악의 패시브 플레이가 진행한 라디오 시사 토크였습니다.

5.4. 발화 검침

- 검침 발화는 표시하지 않고 시간 순서에 따라 적는다. 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다. (‘응, 어.’ 등 짧은 맞장구 발화의 경우도 마찬가지로 줄 바꿈).

1:그래서 그랬는데 이번에 여의도에 갔었는데

여의도 거기 붐뚱 했잖아요.

2:응중로

1:예.

5.5. 단어(내용 전사)

- 발화 내용은 기본적으로 철자법 수준의 전사를 한다. 다만, 구어의 발음 특성, 개인의 발음 특성, 지역적인 특성 등에 의해 철자법대로 소리나지 않는 발음(표준 발음이 아닌 경우)에 대해서는 발음나는 대로 적는다.

[자 상담소에는 어떤 걸 기대하고 왔어요?]

- 모음의 변화, 수의적 경음화 등을 반영하여 적는다.

어쨌든, 늦두록, 외우, 꼬집어래도.....

- 모음 약화 현상에 의한 이형체는 반영하지 않는다. 예를 들어 의문사 ‘뭐?’가 ‘마’로 모음 이 약화되어 둘레도 ‘뭘?’로 적는다. (‘이것도’>‘이거’와 같은 자음 약화 현상은 예외)

- 숫자나 기호, 영문 등도 발음에 따라 한글로 적는다.

[오늘 제 동생이 이제 하나 오백 원이라고 사 가지고 왔더라구]

5. 전사 지침

5.1. 기본 원칙

- 발화 내용은 기본적으로 한글 맞춤법에 따라 전사하는 것을 기본 원칙으로 하며 띄어쓰기나 한글 맞춤법에 따른다.

- 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우(표준 발음이 아닌 경우) 발음나는 대로 적는다.

5.2. 화자 표시

1) - CH1은 화자 1로 표시하고, CH2는 화자 2로 표시한다. (원록 화자(L)는 CH1, 오펜록 화자(R)는 CH2임. 화자 번호와 발화 사이는 띄어 쓰지 않음.)

<자음(1차) 전사 형식>

(L) 이제 시작할까요?

(R) 네.

<수동(2차) 전사 형식>

1:아예 시작할까요? (O)

1: 아예 시작할까요? (X)

2:네 (O)

2: 네 (X)

5.3. 전사 단위

- 기본 전사 단위는 큰 절 단위(t-unit)나 문장 단위가 되도록 하며, 문장이 끝나지 않더라도, 쉼(pause)이 1초 이상 들어가면 줄을 바꿈. (단, 짧은 대답, 말 끊임(단어), 군말 뒤에 서는 이어서 음)

참 발화: 내가 어제 집에 가서 칠수를 만났는데 그런 얘기 못 들었어.

전사: 내가 어제 집에 가서 [칠] (줄 바꿈)

칠수를 만났는데 [칠] (줄 바꿈)

그런 얘기 못 들었어. [칠] (줄 바꿈)

1:자 호환한데 과학실 수업입니다.

자 오늘 우리가 공부할 문제

지난 시간에 얘기했던 걸 바탕으로 떠올려 봅시다.

어떤 걸 공부할 거 같아요?

1:나는 이제 드*

2:비행기 안에서보다 오히려 비행기 밖에 공항 안에서 기다릴 때가 더 재밌었던 거 같애. 막 음식 먹고 막 b1할 얘기하고.

그리고 그 맨세점에서 막 산 거 들고 이러는 거

더 재밌었던 거 같아.

1:그때는 근데, 우리가 여행가기 전이어서 막

기본에 신 나가지고 돌피가지고 그런 거 같은데

어떻게 이제 크림 같나 아니야.

이거도 오리저날 제주도 갈일이 아니야.

5.6. 끊어진 단어(단어가 불완전하게 발화된 경우)

- 발화된 대로 그대로 전사하고, ‘*’를 붙여 정상적인 단어와 구별할 수 있게 한다. 불완전하게 발화된 단어(어절)가 둘 이상인 경우 어절마다 ‘*’를 붙인다(수정 발화, 반복 발화에 표시하는 것은 아님).

아~ 생크림 크림은 베* 크림은 베이커리 생크림이 좀 맛있고.

전* 전* 전통이라고 우리가 흔히 얘기할 때

(비교) 그 이후에는 식민지 당국과의 그* 다이알로그가 있을 있었을 것이구요

5.7. 띄어쓰기 (> 띄어 맞춤법 참조)

- 합성명사의 경우 대부분 붙여 쓴다. (예) 우리나라(O), 우리 나라(X)

- 외래명사는 띄어 쓴다. 수를 적을 때는 만 단위로 띄어 쓴다(예: 십이억 삼천백만 팔백구 불 등).

- 본용언과 보조용언도 띄어 쓴다.

5.8. 축약형의 표기

- 구어에서는 발음의 축약 현상이 많이 나타나는데, 두 음절이 한 음절 사잇소리가 된다가 나, 두 음절이 한 음절 겹소리기가 되는 것 등이다. 구어 발음에서는 발음되는 음절수와 표기상의 음절수를 맞추는 것이 원칙이므로 축약형의 경우 모두 표기에 반영한다.

[일부, 그나간.....]

- 모음의 축약형의 경우 대부분 현재 국어의 모음 체계상 표기할 글자가 존재하지만, 반줄 소리된 /i/, /u/의 표기는 문제가 된다. /i/ /u/가 반줄소리가 되어 /i/ /u/와 축약되는 현상은 구어에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 구어 전사에서는 ‘*’를 사용해서 연결해 준다.

[자극*어, 바뻐*어, 바뻐*어*어, ...]

5.9. 군말

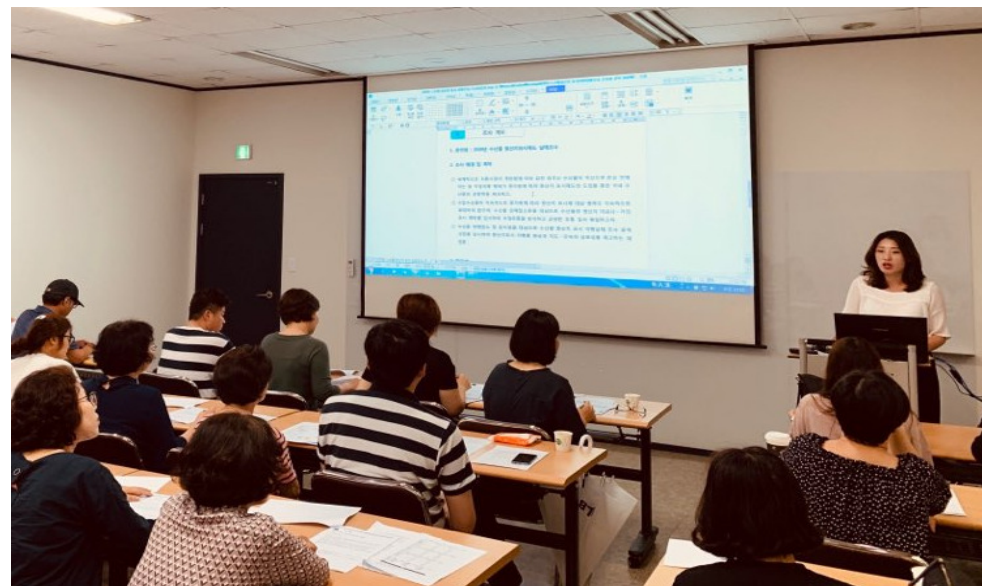
- ‘아, 그, 저, 아, 어, 오, 유’ 등 군말은 불결표(-)를 이용하여 표시한다(주로 머뭇거림의 표기로 사용되는 아~, 그~, 저~, 어~, 아~, 예~, 오~, 음~ 등이 해당됨. 인제, 이제, 그 날, 무슨, 어떤 등은 붙이지 않음).

- 억양과 운율에 의해서만 구분이 가능할 경우는 반드시 전사 단계에서 표시해 준다.

[같은 경우에 눈썹 그~ 어~ 연구는 데이션 국가라는 거하구 직결되는 과정이요.]

5.10. 잘 들리지 않는 부분

- 잘 들리지 않는 부분은 () 안에 전사한다. 잘 들리지 않는 부분에 문장부호가 있는 경우는 부호 다음에 괄호를 달는다.



<그림 7> 전사자 교육 사진

2.3. 음성 정제 교육

음성 정제는 녹음된 음성 파일을 전사 단위대로 자르고 개인 정보 및 불필요한 내용을 정제하는 과정으로 전사와 마찬가지로 많은 시간이 요구되는 작업 단계이다. 음성 정제는 과거 음성 정제 경험이 있는 자, 프로그램에 익숙한 자를 우선 대상으로 선발하였다.

음성 정제를 위해 사업 기간 동안 전체 약 60명을 교육하였으며, 하루에 1인당 약 3개의 음성 파일을 정제하는 것을 기준으로 일평균 20~25명 정도가 투입되었다. 모든 음성 정제자는 사업 수행 기관이 직접 관리 교육하였다.

음성 정제 품질의 균질화를 위해 음성 정제자를 대상으로 전체 집체 교육을 실시하였다. 1차 교육 시 30명 정도를 교육하였다. 이탈자가 많이 발생해 사업 중간에 수시 교육을 통해 30명 정도를 더 투입하였다.

음성 정제자 교육은 실기 중심의 교육으로 진행되었으며, 사업 배경 및 목적, 프라트(Praat) 프로그램 이용 방법, 파일 분할 방법, 개인 정보 비식별 처리 방법, 유의 사항 등을 교육하였다. 이러한 과정을 통해 최종 선발된 음성 정제자들은 보안 서약서 작성 후 음성 정제에 참여하였다. 교육 받은 정제자는 샘플 정제 후 두어 차례 교정 후 본 정제에 정식으로 투입되었다.

<표 7> 음성 정제 내용

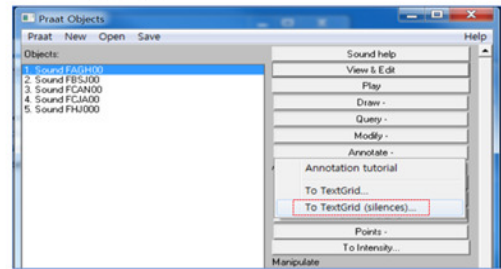
구분	음성 정제 교육 내용
사전 준비	<ul style="list-style-type: none"> • 프라트(Praat) 프로그램 설치 환경 및 설치 방법
파일 분할	<ul style="list-style-type: none"> • 프라트(Praat) 프로그램에서 분할 값 설정 방법 • 텍스트그리드(TextGrid)와 원본 음성 연결 방법 • 파일 분할 저장 방법 • 파일 분할 저장을 위한 스크립트 실행 방법
개인 정보 비식별 처리	<ul style="list-style-type: none"> • 개인 정보 비식별 처리 대상 정의 및 비식별 방법 • 묵음 처리 방법
유의 사항	<ul style="list-style-type: none"> • 인사말, 대화와 무관한 대화 분할에서 제외 • 발화가 겹치는 경우 처리 방법

음성 정제 교육자료

2019. 07. 09.(화) 14:00-16:00
한양대학교 에리카캠퍼스 국제문화관 321호

WAV 파일 분할 저장

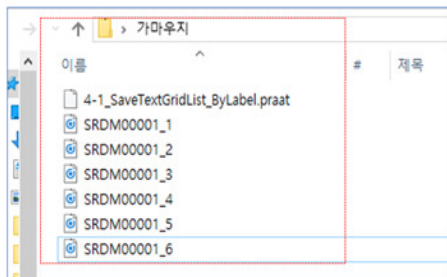
1. Praat 실행
2. Segmentation을 수행할 사운드 파일 불러오기
3. Segmentation을 수행할 사운드 파일에 대해 To TextGrid (silences)... 선택



WAV 파일 분할 저장

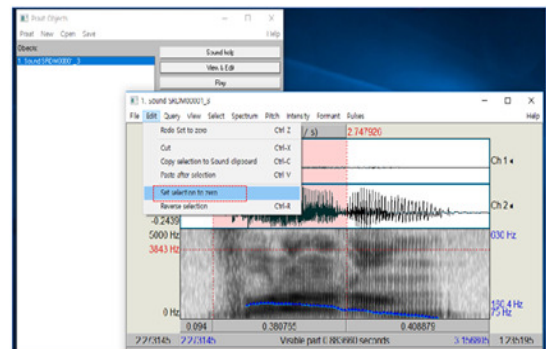
11. 분할 결과 확인

- 스크립트(4-1) 파일과 같은 경로에 저장됨.
- '원본 파일명_문장번호.wav'로 자동 저장됨.



개인정보 비식별 처리

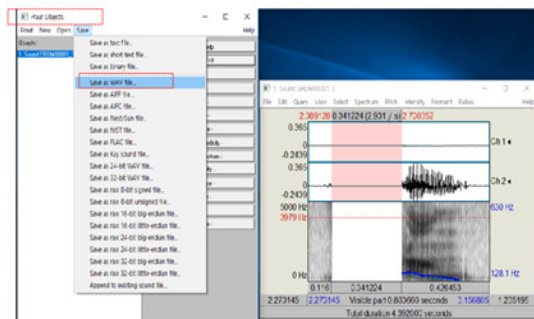
4. 목음 처리 edit – set selection to zero



개인정보 비식별 처리

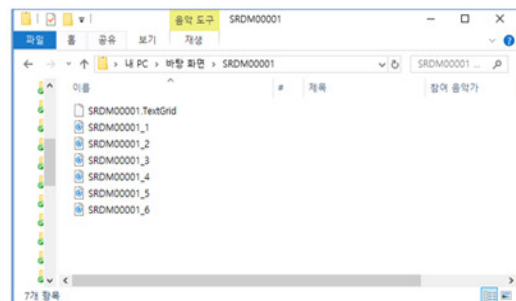
5. 목음 처리한 wav 파일 저장(파일 덮어쓰기)

Praat object – Save – Save WAV file



음성 분할 결과를 제출

1. 음성 파일명과 동일한 폴더명으로 폴더 생성
2. TextGrid 파일, 음성 분할 파일



<그림 8> 음성 정제자 교육 자료 일부



<그림 9> 음성 정제 교육 사진

2.4. 보안 교육

본 사업에 참여하는 관리자, 진행 요원, 전사자, 음성 분석자 등 모든 사업 참여자 대상으로 보안 교육을 실시하였다. 진행 요원, 전사자, 음성 분석자는 해당 집체 교육 시행하였으며, 그 외 참여 인력은 별도의 시간에 교육을 진행하였다. 교육 내용은 개인 정보 보호 교육, 산출물·자료 관리 교육, 녹음 장비 및 PC 등 장비 보안 교육, 웹 서버 및 데이터베이스 관리 교육 등이다.

<표 8> 보안 교육 내용

구분	보안 교육 내용
개인 정보 보호	<ul style="list-style-type: none"> • 사업 진행 중 알게 되는 화자의 인적 사항에 대한 비밀 보장 • 화자의 대화 청취 중 알게 되는 개인 사생활 및 사적 의견에 대한 비밀 보장
자료에 대한 보안 관리	<ul style="list-style-type: none"> • 사업 수행 과정에서 생산되는 모든 산출물 관리 방안 • 인터넷 자료 공유 사이트 및 상용 메신저 사용으로 인한 해킹 위험 방지 • 전자 우편 이용 시 파일 암호화 방안 • 선정 질문지, 이용 허락 계약서 등의 출력물 및 산출물은 사무실 내 시건 장치가 된 보관함에 보관
사무실·장비에 대한 보안 관리	<ul style="list-style-type: none"> • 모든 참여 인력의 PC는 비밀번호를 설정하고, 백신 프로그램을 설치 • 퇴근 시 녹음 장비, 외장 하드는 사무실 내 시건 장치가 된 보관함에 보관 • 인가된 USB 및 휴대용 저장 매체만 사용 가능함.
웹 서버 및 데이터베이스 관리 교육	<ul style="list-style-type: none"> • 구축된 웹 시스템과 데이터베이스는 주기적으로 백업 • 웹 시스템은 관리자에 의해 인증된 사람만 접근 가능하며 로그인 후 일정시간 사용하지 않을 시 보안을 위해 자동 로그아웃 • 데이터베이스는 개인 키와 공공 키를 동시에 사용해 보안 강화



<그림 10> 보안 교육 사진

3. 일상 대화 녹음

3.1. 화자 구성 및 모집

모집 대상은 전국 16개 시·도에 거주하는 만 19세 이상 성인 남녀로, 성별, 연령, 지역 등의 비율이 편중되지 않도록 사전에 참가자 할당표를 설계해 모집하였다. 참가자 할당표의 모집단으로는 2017년 통계청 인구 총조사를 기준으로 활용하였다. 지역은 현 거주지가 아닌 주 성장지 기준으로 세종시를 제외한 16개 지역(서울, 인천, 대전, 대구, 부산, 울산, 광주, 경기, 충북, 충남, 경북, 경남, 전북, 전남, 강원, 제주)으로 할당하였으며, 연령은 10세 단위로 분할하여 할당하였다. 2012년에 출범한 세종시는 주 성장지가 그에 해당하는 대상자를 찾기 어려워 해당 할당을 충남과 충북으로 분할하였다. 16개 지역별로 비례 할당 후 각 지역별로 성별×연령별로 균등 할당을 하는 방법을 이용하였다.

화자 모집에 있어 최대한 인구 비율에 맞도록 진행하고자 했으나, 일부 할당은 그 대상을 찾기가 어려워 지역은 권역 전체로 할당을 맞추고 각 지역별로는 기본 할당량의 최소 50%를 맞추어 진행하였다. 성별×연령별 할당 역시 남자 40~60대 이상, 여자 50~60대 이상은 기본 할당량의 최소 50%를 맞추어 진행하였다.

화자 모집은 섭외된 대상자(화자1)가 대화할 상대 화자(화자2, 지인)와 동반하는 방식으로 진행했기 때문에 할당표는 섭외된 대상자 기준으로 설계되었다.

<표 9> 화자 할당표 설계 기준

구분	화자 할당표 설계 기준
모집단	<ul style="list-style-type: none"> 통계청, 2017 인구 총조사 기준 활용
고려 변수	<ul style="list-style-type: none"> 성별: 남자/여자 연령대: 20대/30대/40대/50대/60대 이상 지역(주 성장지 기준): 서울/인천/대전/대구/부산/울산/광주/경기/충북/충남/경북/경남/전북/전남/강원/제주 * 세종시의 경우, 2012년 출범한 세종시를 주 성장지로 하는 대상을 찾기 어려워 제외함.
배분 방법	<ul style="list-style-type: none"> 제공된 비례 배분
표본 할당	<ul style="list-style-type: none"> 지역별: 비례 할당 성별×연령별: 균등 할당

<표 10> 섭외자 기준 성×연령×지역별 할당표

단위: 명		남자					여자					합계		
		20대	30대	40대	50대	60대 이상	20대	30대	40대	50대	60대 이상	권역별	기본 할당	최소 할당
수도권	서울	24	24	23(12)	23(11)	23(12)	24	24	24	23(12)	23(11)	630	235	(118)
	경기	27	27	26(13)	26(13)	26(13)	27	27	27	26(13)	26(13)		265	(133)
	인천	13	13	13(6)	13(7)	13(6)	13	13	13	13(7)	13(8)		130	(65)
충청권	대전	11	11	11(6)	11(5)	11(6)	11	11	11	11(5)	11(6)	335	110	(55)
	충북	11	10	11(5)	10(5)	10(5)	10	11	10	10(5)	10(5)		103	(52)
	충남	12	13	12(6)	12(6)	12(6)	13	12	12	12(6)	12(6)		122	(61)
경북권	대구	12	12	12(6)	12(6)	12(6)	12	12	12	12(6)	12(6)	245	120	(60)
	경북	13	13	12(6)	12(6)	12(6)	13	13	13	12(6)	12(6)		125	(63)
경남권	부산	14	14	14(7)	14(7)	14(7)	14	14	14	14(7)	14(7)	355	140	(70)
	경남	14	14	14(7)	13(6)	13(7)	14	14	13	13(6)	13(7)		135	(68)
	울산	8	8	8(4)	8(4)	8(4)	8	8	8	8(4)	8(4)		80	(40)
호남권	광주	9	9	9(4)	9(5)	9(4)	9	9	9	9(5)	9(4)	290	90	(45)
	전북	10	10	10(5)	10(5)	10(5)	10	10	10	10(5)	10(5)		100	(50)
	전남	10	10	10(5)	10(5)	10(5)	10	10	10	10(5)	10(5)		100	(50)
강원	강원	9	9	9(4)	9(5)	9(4)	9	9	9	9(5)	9(4)	90	90	(45)
제주	제주	6	6	5(3)	5(2)	5(3)	6	6	6	5(3)	5(2)	55	55	(28)
합계		203	203	199	197	197	203	203	201	197	197	2,000	2,000	(1,003)

대규모 화자 모집을 위해 주로 사업 수행 기관에서 자체 보유하고 있는 온라인 패널을 활용하였다. 사업 수행 기관의 온라인 패널을 대상으로 이메일, SMS, 홈페이지를 통해 일상 대화 말뭉치 구축 참여 모집 공고를 게시하고 온라인 선정 질문지를 통해 적합한 대상자를 선정하였다. 1차 모집된 대상자(화자1)는 자신과 함께 대화를 나눌 지인(화자2)을 함께 동반하는 것을 원칙으로 하여 모집하였다.

이렇게 1차 모집된 화자를 전문 연락원이 전화 통화를 통해 적합한 대상자가 맞는지 다시 한 번 확인하고, 화자가 원하는 첫 번째 희망 대화 주제에 대해 그 모집 현황을 확인한 다음 녹음 가능한 날짜를 협의해 최종 대상자로 모집하였다. 만약 첫 번째 희망 주제의 모집이 완료된 경우에는 두 번째 혹은 세 번째 희망 주제로 참여 가능한지 확인한 다음 녹음 날짜 협의를 진행하였다.

기타 방법으로는 CLT(Central Location Test) 기법을 활용해 거리에서 직접 모집하거나 추천, 지역 커뮤니티 공지 등의 다양한 방법으로 화자를 모집하였다.

녹음이 진행될수록 성별×연령별×지역별 할당 외에 주제 할당까지 맞는 화자를 찾는 것이 쉽지 않았다. 특히, 50~60대 이상, 도 지역 거주자, 일부 주제는 모집이 어려워 해당 지역 거리에서 직접 대상자를 모집하거나 추천 받는 방식으로 모집하였다. 또한 일부 화자는 녹음 전 확인 전화 시, 녹음을 거절하거나 통화가 계속적으로 이루어지지 않기도 했으며, 녹음 당일 녹음 장소에 오지 않기도 하는 등의 이유로 화자 모집은 녹음 한 달 전부터 시작해 녹음 종료 3일 전까지 계속 진행되었다.



<그림 11> 온라인 패널 대상 모집 공고



www.metrix.co.kr



[범국민 일상 대화 말뭉치 구축 프로젝트를 위한 참여자 모집 안내문]

안녕하십니까?
저희는 국립국어원의 의뢰를 받아 '일상대화 말뭉치 구축' 프로젝트를 수행하고 있는 리서치 전문 업체 (주)메트릭스입니다.

국민들의 일상 대화를 수집하고 분석하는 '일상대화 말뭉치 구축' 프로젝트를 참여자를 모집하고자 합니다.
본 조사의 목적은 국민들의 일상대화 수집을 통해 대규모 언어 자원을 구축하여, 이를 국어 연구 및 자연언어처리 기술개발 등을 위해 사용하고자 하는 데 있습니다.

조사 참여자들은 저희가 마련한 장소에 지인과 함께 오셔서 관심주제에 대해 30분 정도의 일상적이고 자연스러운 대화를 나누시면 됩니다. 나누시는 대화내용은 음성자료 수집 및 분석을 위해 녹취됩니다.

녹취된 자료는 통계법 33조에 따라 비밀이 보장되며, 귀하가 만드신 저작물은 개별 단어 및 문장, 텍스트에 대한 정보 추출과 분석에만 사용됩니다.



www.metrix.co.kr

문 1. 귀하께서는 평소 지인들과 대화를 많이 나누는 편입니까?

☐ 예

☐ 아니오



www.metrix.co.kr

문 5. 그렇다면 귀하께서 만16세까지 주로 성장하신 지역은 어디입니까?

☐ 서울 ☐ 인천 ☐ 대전
☐ 대구 ☐ 부산 ☐ 울산
☐ 광주 ☐ 경기 ☐ 세종
☐ 충북 ☐ 충남 ☐ 경북
☐ 경남 ☐ 전북 ☐ 전남
☐ 강원 ☐ 제주



www.metrix.co.kr

문 9. 다음 중 귀하께서 관심 있는 주제는 무엇일까요?
순서대로 3가지만 선택해 주십시오.

☐ 군대
☐ 휴일
☐ 만화
☐ 정치
☐ 방송/연예
☐ 먹거리
☐ 국가/지역
☐ 연애/결혼

☐ 게임
☐ 자동차
☐ 영화
☐ 건강/다이어트
☐ 스포츠/레저
☐ 자연/휴양지
☐ 문학
☐ 경제/재테크

[다음문항으로](#)

<그림 12> 온라인 대상자 선정 질문지

<표 11> 선정 질문 내용

구 분	상세 선정 질문
인구 특성	<ul style="list-style-type: none"> • 성별 • 연령 • 직업 • 학력
지역	<ul style="list-style-type: none"> • 출생지 • 주 성장지 • 현 거주지
기타 정보	<ul style="list-style-type: none"> • 관심 주제 • 화자간의 관계

3.2. 녹음 환경

녹음은 전국 7개 지역(서울, 대전, 대구, 부산, 광주, 강원, 제주)에서 동시에 진행되었으며, 두 명의 화자가 자유롭게 대화할 수 있도록 주변 소음이 적은 좌담회실, 회의실 등 별도의 장소를 마련하여 화자 간 거리가 1m 이상 떨어진 공간에서 녹음을 진행하였다. 장소가 여의치 않은 경우 녹음 기간 동안 유료 회의실이나 별도 공간을 대여해서 진행하였다.



▲ 서울 녹음실 1



▲ 서울 녹음실 2



▲ 서울 녹음실 3



▲ 광주 녹음실



▲ 대구 녹음실



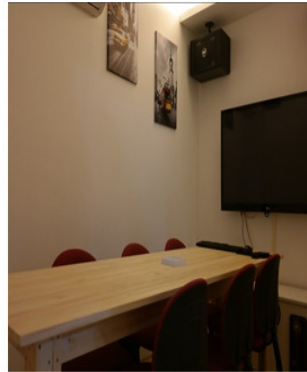
▲ 부산 녹음실



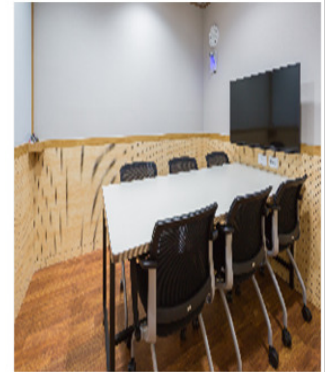
▲ 대전 녹음실 1



▲ 대전 녹음실 2



▲ 제주 녹음실 (대여)



▲ 강원 녹음실 (대여)

<그림 13> 지역별 녹음실



<그림 14> 녹음 지역

녹음은 2채널로 진행되었으며, 두 명의 화자는 각자 헤드셋 마이크를 착용하고 음성의 최대 샘플값이 10,000~20,000 사이가 되도록 음량을 조절하였다.

각 지역에 마련된 녹음 장소와 녹음 장비가 세팅된 환경을 직접 확인하고 수정 보완하였으나 진행 중 발생하는 외부 자동차 소리 및 에어컨 소리, 돌발음 등의 소음은 통제하기 힘들었다.

이렇게 구축된 녹음 환경에서 발화한 샘플을 국립국어원에 3차례 검증 받아 최종 적합한 환경을 확인 후 본 녹음을 진행하였다.



<그림 15> 녹음 장비

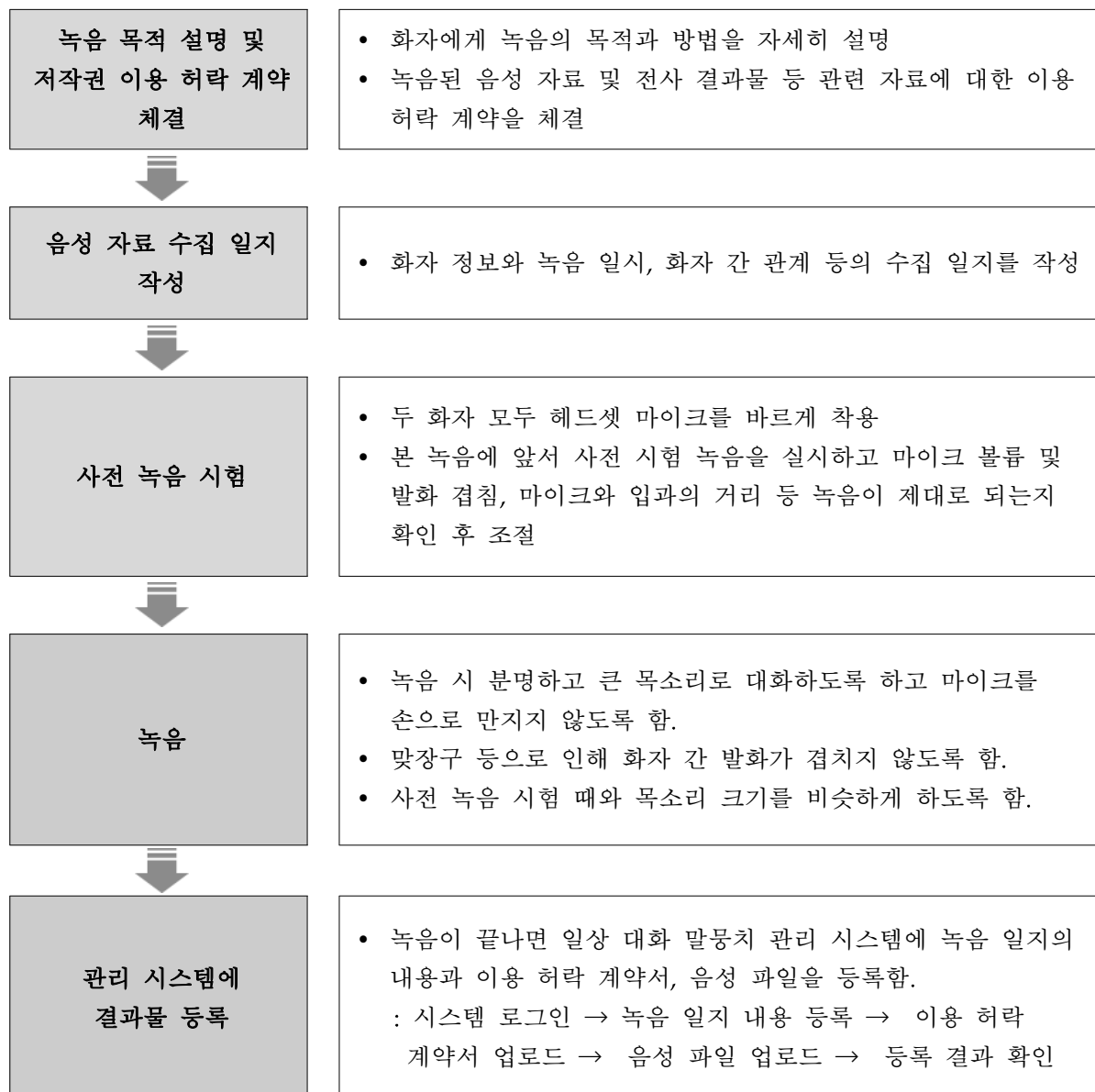
3.3. 녹음 진행

3.3.1. 녹음 절차

약속된 시간에 화자가 녹음 장소에 도착하면 진행 요원은 화자 두 명 모두의 개인 정보를 확인하고 녹음된 음성 자료 및 전사 결과물 등 관련 자료에 대한 이용 허락 계약을 체결한다. 그리고 녹음 시 유의 사항에 대해 충분히 전달한 다음 녹음을 진행하였다. 대화는 최대 30~35분으로 제한했으며, 동일한 화자가 두 개 이상의 녹음에 중복해서 참여하지 않는 것을 원칙으로 하여 진행하였다.

녹음은 다음과 같은 순으로 진행하였다.

<표 12> 녹음 절차



3.3.2. 이용 허락 계약 체결 및 수집 일지 작성

녹음 전 참가자에게 본 사업의 목적과 개인 정보 보호 준수에 대해 충분히 설명하고 이용 허락 계약을 체결하였다. 이용 허락 계약서는 결과물인 음성 파일, 전사 파일, 그 변형물에 대한 복제권, 전송권, 배포권, 2차적 저작물 작성권에 대해 국립국어원에서 활용하는 것을 허락한다는 내용이다.

사전에 녹음 날짜를 확인하는 통화 시, 이용 허락 계약서에 대해 설명하기 때문에 참석한 대부분의 화자들이 이용 허락 계약서에 동의는 했으나, 현장에서 다소 많은 분량과 계약서 내용 중 ‘계약’, ‘권리자의 의무’, ‘손해배상권’ 등의 용어들로 인해 부담스러워하는 화자도 있었다. 간혹 일부 화자는 계약서 내용을 확인한 후 녹음을 거절하고 돌아간 경우도 있으며, 어떤 화자는 녹음 후 다시 거절 의사를 표명해 녹음을 비롯한 관련 파일을 삭제하는 경우도 있었다.

<p>국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서</p> <p>저작자 및 저작권 이용허락자 _____ (이하 "권리자"이라 함)와 저작권 이용자 국립국어원(이하 "이용자"이라 함)은 아래 저작물에 관한 저작권산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.</p> <p style="text-align: center;">다 음</p> <p>제1조 (계약의 목적) 본 계약은 저작권산권 이용허락과 관련하여 권리와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.</p> <p>제2조 (계약의 대상) 본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상저작물")에 대한 저작권산권 중 당사자가 합의한 권리로 한다.</p> <p>저작물: 저작자: 종별: <input checked="" type="checkbox"/> 어문저작물 권리: <input checked="" type="checkbox"/> 복제권, <input checked="" type="checkbox"/> 전송권, <input checked="" type="checkbox"/> 배포권, <input checked="" type="checkbox"/> 2차적저작물작성권</p> <p>※ 저작권 이용허락 대상 권리의 내용</p> <ol style="list-style-type: none"> 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 대상저작물의 음성을 청취·전사한 텍스트를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물의 음성을 청취·전사하여 텍스트로 변형하고 그 텍스트를 복제·변형(목차·마리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적 비언어적 정보 부착 등)하는 일 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물을 제공·배포받은 학 <p>본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.</p> <p>제8조 (계약의 해지)</p> <ol style="list-style-type: none"> 당사자는 현재지점 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다. 당사자는 상대방이 정당한 이유 없이 본 계약을 위반하는 경우에 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부 의사를 표시하였거나 위반 사항의 성격상 시정이 불가능하다는 것이 명백히 인정되는 경우에는 위와 같은 최고 없이 계약을 해지할 수 있다. 본 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다. <p>제9조 (손해배상)</p>	<p>계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물의 음성을 청취·전사한 텍스트 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일</p> <p>제3조 (이용허락 기간) 대상저작물의 이용허락 기간은 계약 체결일로부터 2040년 12월 31일까지로 하며, 권리자가 이용허락을 중지하고자 하는 의사를 밝히지 아니하면 이용허락이 5년 단위로 자동 갱신된다. 권리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락을 중지하여야 하며, 그렇지 아니하면 이용허락 내용이 유지된다.</p> <p>제4조 (권리자의 의무)</p> <ol style="list-style-type: none"> 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작권산권을 이용할 권리를 제3조의 기간 동안 비독점적으로 허락한다. 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있지 않은 경우 이용자가 요청하면 이용허락자는 대상저작물의 저작권산권을 등록한 후 위 의무를 이행한다. 권리자는 대상저작물에 제3자의 이용허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다. 권리자는 대상저작물의 저작권산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 한다. <p>제5조 (이용자의 권리 및 의무)</p> <ol style="list-style-type: none"> 이용자는 대상저작물을 제3조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다. 이용료는 설정하지 아니한다. 이용자는 관례적으로 저작자 및 저작권산권자의 성명 등 표시를 허용하는 대상저작물을 이용하는 경우, 그 저작자 및 저작권산권자의 성명 등을 표시하여야 한다. 이용자는 대상저작물의 이용함에 있어서 저작권침해를 침해하지 아니한다. 다만, 제2조에 따른 목적에 한하여 제2조에 따른 변형을 할 수 있으며, 대상저작물의 본질적인 내용이 아니다. <p>제13조 (기타부속합의)</p> <ol style="list-style-type: none"> 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의를 작성할 수 있다. 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다. <p>제14조 (계약의 해석 및 보완) 본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.</p> <p>제15조 (계약 효력 발생일) 본 계약의 효력은 계약 체결일로부터 발생한다.</p> <p style="text-align: right;">년 월 일</p>
---	--

<그림 16> 이용 허락 계약서

이용 허락 계약 체결 후 진행 요원은 두 명의 화자 모두의 화자 정보(이름, 성별, 생년월일, 연령대, 직업, 출생지, 거주지, 주 성장지)와 대화 주제, 화자 간 관계 등을 수집 일지에 작성하였다.

녹음 날짜		년 월 일					
주제 1				주제 2			
화자 간 관계				기타			
대상자	화자1 (CH 1)	이름	성별	연령대	생년월일	직업	학력
		출생지	주 성장지	거주지	비고		
	화자2 (CH 2)	이름	성별	연령대	생년월일	직업	학력
		출생지	주 성장지	거주지	비고		

<그림 17> 음성 자료 수집 일지

3.3.3. 녹음 진행

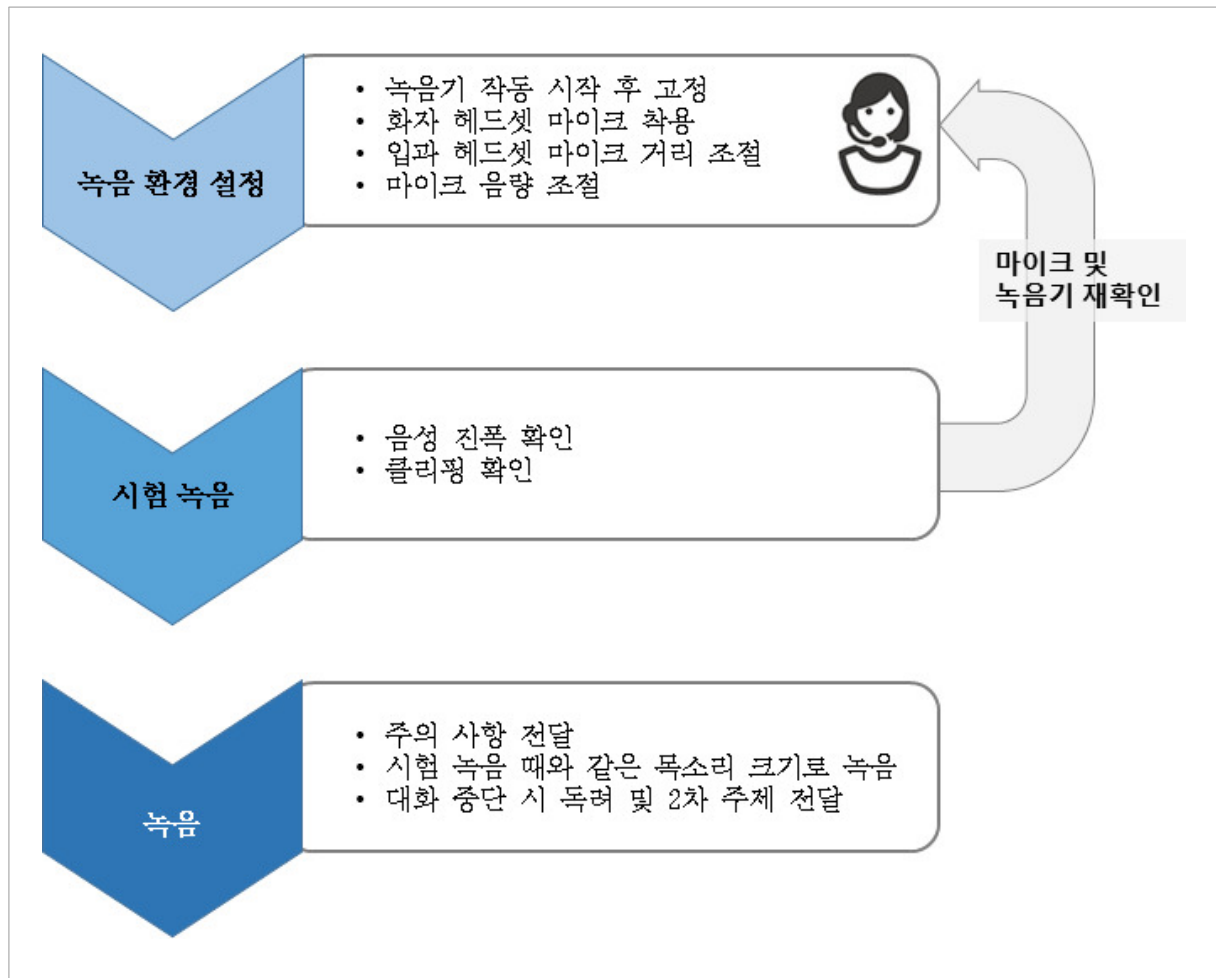
진행 요원은 녹음기를 지정된 장소에 고정하고, 화자별로 헤드셋 마이크를 바르게 착용했는지 확인 후 사전 시험 녹음을 3~5분 정도 진행하였다. 반드시 실제와 같은 목소리 크기로 시험 녹음을 한 다음 녹음 상태를 확인했다. 녹음이 제대로 되는지, 음량은 적당한지, 클리핑이 발생하지 않는지 등을 확인한 다음 본 녹음을 시작했다.

만약 녹음 상태가 올바르지 못하다면 헤드셋 마이크와 입과의 거리, 마이크 음량 등을 조절한 후 다시 시험 녹음을 진행하였다. 클리핑이 발생한 경우에는 전체적으로 마이크 음량을 낮추고 시험 녹음을 다시 진행했다.

본 녹음 시 분명하고 큰 목소리로 대화하도록 하고 사전 시험 녹음 때와 비슷한 목소리 크기로 대화하도록 했다. 또한 마이크를 손으로 만지지 않도록 했으며, 맞장구 등으로 인해 화자 간 발화가 최대한 겹치지 않도록 하는 등 녹음 시 주의 사항을 다시 한번 전달하고 본 녹음을 시작했다.

대부분의 화자는 녹음 시작 후 처음 몇 분 동안은 대화가 자연스럽게 못하고 녹음에 대해 많이 의식했으나 5분 정도 지나 녹음이 진행될수록 본연의 일상 대화가 자연스럽게 이루어졌다.

녹음 중 주제에서 벗어난 대화가 지속될 경우 또는 대화가 끊어질 경우 진행 요원은 녹음을 잠시 중단하고 주제와 관련된 대화를 하도록 요청했다. 만약 하나의 주제에 대해 대화가 더 이상의 대화가 힘들다고 판단되는 경우 진행 요원은 화자가 선택한 두 번째 관심 주제를 제시하고 대화를 지속하도록 했다.



<그림 18> 녹음 절차

녹음을 완료한 후 진행 요원은 녹음 원본을 16bit 양자화 선형 PCM 파일로 PC에 저장한 다음 수집 일지에 작성한 화자 관련 정보, PDF로 변환된 이용 허락 계약서, 음성 녹음 파일을 일상 대화 말뭉치 관리 시스템에 등록하였다.

일상 대화 말뭉치 구축
관리 사이트

로그인

MetriX
www.metrix.co.kr

LOGIN

ID

PW

로그인

조사 지원 사이트입니다.

- 본 서비스를 이용하시려면 로그인 하셔야 합니다.
- 아이디와 패스워드를 입력하세요
- 본시스템은 인가된 분만 사용할수 있습니다. 불법으로 사용시에는 법적 제재를 받을 수 있습니다.

▲ 관리 시스템 로그인

34.85.46.170

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

관리자용 **등록 및 수정**

화자정보및 음성파일등록 전사자료등록

“클릭”

▲ ‘화자 정보 및 음성 파일 등록’ 메뉴 선택

음성 파일 선택

찾아보기...

대화 정보

녹음 날짜	주제1	주제2	주제3	화자간 관계	기타
yyyyymmdd (예: 2019060)	선택				대화 상황을 이해하는데 필요한 사항을 적어주세요. (장소, 주변소음, 분위기 등)

화자 정보

	이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	동의서 업로드 (pdf 파일만 등록 가능)	비고/특이사항
화자1 (왼쪽:메인)		남	yyyyymmdd (예: 2019060)	선택	선택	선택	선택	선택	선택	찾아보기...	
		여									
화자2 (오른쪽)		남	yyyyymmdd (예: 2019060)	선택	선택	선택	선택	선택	선택	찾아보기...	
		여									

관리 전송

▲ 화자 정보, 음성 파일, 이용 허락 계약서 등록

(계속)



Home

파일이름	SDRW00000001.wav		다운로드	찾아보기...	음성 파일 업로드	담당자	sp_metrix8	담당자 선택	sp_metrix8 ▼	선택	녹음확정	● 확정	확정	녹음파일 등록자	metrix1
	이름	성별	생년월일	직업		학력	연령대	출생지	주 성장지	거주지	비고/특이사항				
화자1 (CH 1)	박	남	1980-02-19	사무 종사자		대졸	30~39세	대구	대구	경북	직업 역무원				
화자2 (CH 2)	기	남	1983-11-08	사무 종사자		대졸	30~39세	대구	대구	대구	회사원				

파일이름	SDRW00000002.wav		다운로드	찾아보기...	음성 파일 업로드	담당자	sp_metrix8	담당자 선택	sp_metrix8 ▼	선택	녹음확정	● 확정	확정	녹음파일 등록자	metrix1
	이름	성별	생년월일	직업		학력	연령대	출생지	주 성장지	거주지	비고/특이사항				
화자1 (CH 1)	임	남	1995-05-07	학생		대학교 재학	20~29세	경북	경북	대구					
화자2 (CH 2)	임	남	1995-12-09	학생		대학교 재학	20~29세	경기	대구	대구					

파일이름	SDRW00000003.wav		다운로드	찾아보기...	음성 파일 업로드	담당자		담당자 선택	sp_metrix8 ▼	선택	녹음확정	● 확정	확정	녹음파일 등록자	metrix1
	이름	성별	생년월일	직업		학력	연령대	출생지	주 성장지	거주지	비고/특이사항				
화자1 (CH 1)	영	남	1994-08-06	기술자 종사자(장치/기계 조작 및 조립 종사자)		고졸	20~29세	경북	경북	대구	기관사				
화자2 (CH 2)	권	남	1992-03-15	기술자 종사자(장치/기계 조작 및 조립 종사자)		고졸	20~29세	경북	경북	대구	기관사				

파일이름	SDRW00000004.wav		다운로드	찾아보기...	음성 파일 업로드	담당자		담당자 선택	sp_metrix8 ▼	선택	녹음확정	● 확정	확정	녹음파일 등록자	metrix1
	이름	성별	생년월일	직업		학력	연령대	출생지	주 성장지	거주지	비고/특이사항				
화자1 (CH 1)	이	여	1995-07-18	무직/취업준비생		대졸	20~29세	대구	대구	대구					
화자2 (CH 2)	정	여	1996-01-02	무직/취업준비생		대졸	20~29세	경북	경북	대구					

▲ 등록 결과 확인

<그림 19> 관리 시스템에 화자 정보, 음성 파일, 이용 허락 계약서 등록 과정 예시

4. 음성 자료 전사

4.1. 전사 규칙

발화 내용은 기본적으로 한글 맞춤법에 따라 전사하는 것을 원칙으로 하며, 띄어쓰기도 한글 맞춤법을 따라 전사하였다. 단, 구어의 발음 특성이나 개인적인 발음 특성, 방언 등으로 인해 표준 발음과 차이가 있는 경우에는 들리는 대로 전사하였다.

전사 시 적용한 규칙은 크게 다음과 같다.

- 화자 표시 방법
- 전사 단위
- 발화 겹침
- 내용 전사 (모음 약화, 발음 처리, 숫자 · 영문 전사)
- 끊어진 단어(단어가 불완전하게 발화된 경우)
- 띄어쓰기
- 축약형의 표기
- 군말 (이, 그, 저, 아, 어, 오, 음)
- 잘 들리지 않는 부분
- 전사자의 설명
- 비언어적 의사 표현과 기타 소리들(@웃음, @목청, @박수, @노래)
- 익명성 보장을 위한 전사 (사람 이름, 부정적 내용 시 상호명)

4.2. 전사 절차

일상 대화 말뭉치 관리 시스템에서 음성 파일을 내려받아 음성 인식 텍스트 변환 프로그램을 이용해 1차 자동 전사 후 2차로 전문 속기사가 전사 규칙에 따라 발화자 표시 변경, 전사 단위, 맞춤법, 띄어쓰기 등을 수정 보완하는 방식으로 진행하였다.

1차 자동 전사

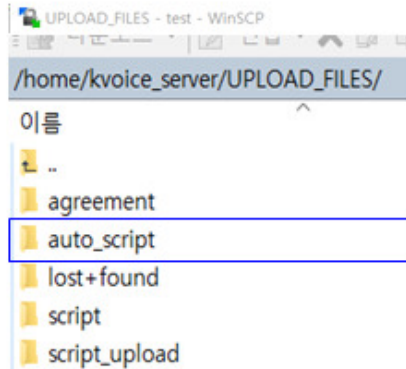
Step 1

말뭉치 관리 사이트에서
음성 파일 다운로드

파일이름	SDRW00000001.wav			다운로드	
	이름	성별	생년월일		
화자1 (CH 1)	박	남	1980-02-19	사무 종사자	
화자2 (CH 2)	기	남	1983-11-08	사무 종사자	

Step 2

시스템을 통한 자동 전사
(음성 -> TEXT)



Step 3

말뭉치 관리 사이트에
자동 전사 파일 업로드

2차 수동 전사

Step 4

말뭉치 관리 사이트에서
자동 전사 파일 다운로드

Step 5

전문 속기사를 통한
수동 교정 전사(UTF-8 Type TXT)

- 맞춤법 띄어쓰기
- 전사 단위 전사 지침 등 확인

Step 6

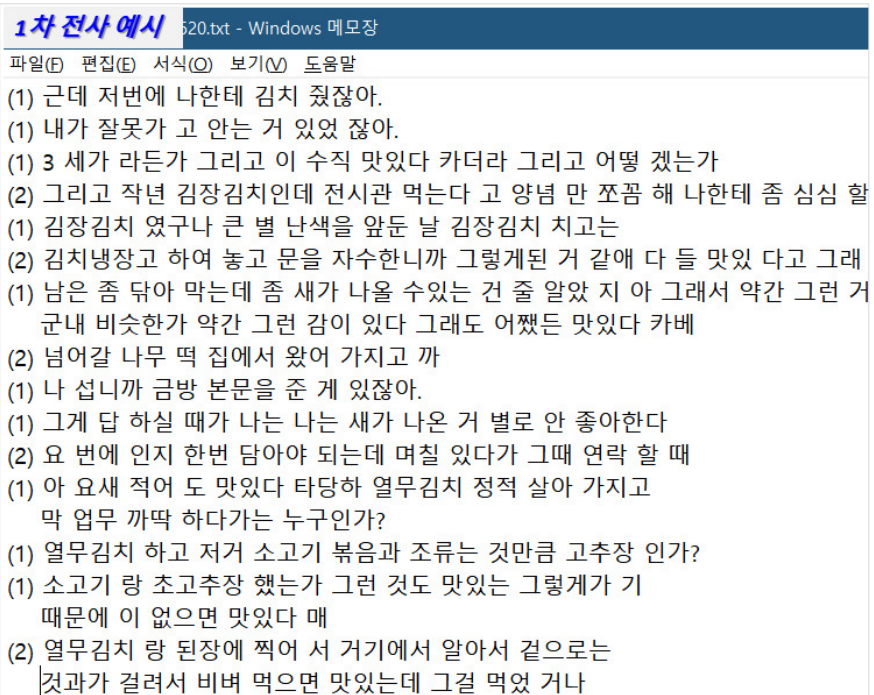
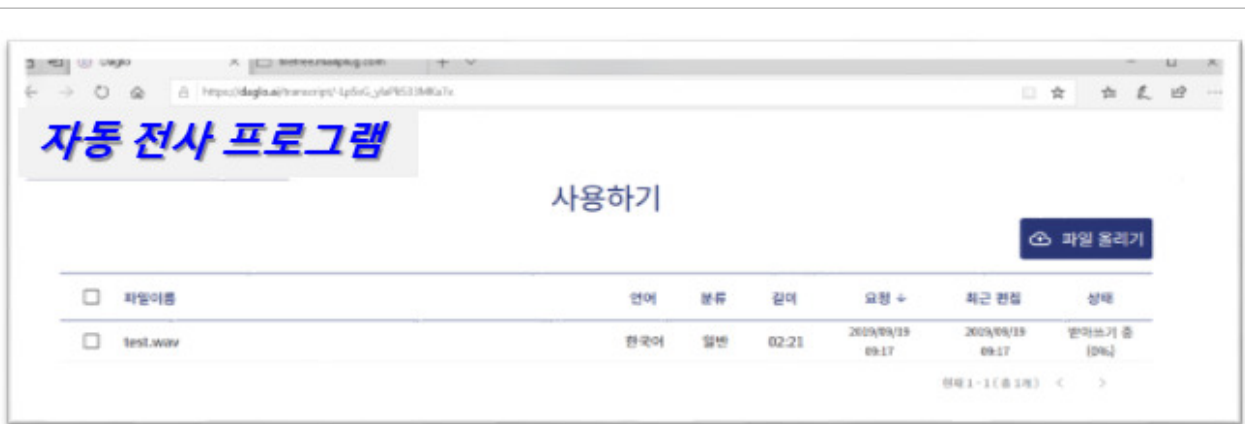
말뭉치 관리 사이트에
수동 전사 파일 업로드

파일이름	주제1	주제2	주제3	자동전사파일	수동전사파일		
SDRW00000001.wav	자동자			전사 파일 없음		찾아보기...	업로드
다운로드	이름	성별	생년월일	직업	학력	연령대	출생지
화자1 (CH 1)	남		1980-02-19	사무 종사자	대졸 30~39세		대구

<그림 20> 전사 절차

4.3. 1차 자동 전사

음성 인식 시스템을 통한 1차 자동 전사는 30분 녹음 기준으로 30분의 전사 시간이 소요되고 정확도가 사전 테스트 파일 기준 80% 이상으로 나타나 전체적인 전사 시간이 단축될 것으로 기대했다. 그러나 본 사업의 음성 파일 전사 결과, 화자가 표준어를 구사하며 정확한 발음을 하는 경우에는 전사의 정확도가 높지만, 일반적으로는 화자의 발화 특성(발음, 숨쉬기)에 따라 맞춤법, 띄어쓰기 등 전사의 정확도가 20~50% 정도 수준으로 낮았다. 특히, 들리는 대로 전사, 끊어진 단어 전사 등은 이루어지지 않았으며, 방언이나 축약형 표기 같은 경우 모두 표준어로 전사하거나 전혀 다른 단어로 전사되었다.



<그림 21> 1차 자동 전사 결과 예시

4.4. 2차 수동 전사

2차 전문 속기사에 의한 수동 전사는 1차 자동 전사에서 전사하지 못하는 경우 (발화자 표시 변경, 전사 단위와 전사 기호 반영, 문장 기호 반영, 들리는 대로 전사, 끊어진 단어 교정, 숫자 및 로마자 한글화, 띄어쓰기, 한글 맞춤법, 겹침 발화, 들리지 않는 부분 전사, 군말 처리, 발화자 중복 표시 등)에 대해 추가·수정 전사하는 단계이다.

2차 수동 전사는 30분 음성 파일 기준으로 4시간 이상이 소요되어, 1인당 일평균 2개 정도의 음성 파일을 전사하였다. 일평균 약 30명 이상의 전사 인력이 투입되어 진행되었다. 표준어보다는 방언이, 연령이 높을수록, 남자보다는 여자의 음성 파일을 전사하는데 상대적으로 시간이 더 많이 소요되었다. 방언이 심한 화자의 음성 파일의 경우는 전사하는데 2~3일 정도가 소요되기도 했으며, 특히, 제주도 60대 이상 여자 화자의 음성 파일은 그 정도가 너무 심해 제주도 방언 전문가에게 전사를 직접 요청하기도 했다.

2차 수동 전사는 기본적으로 한글 맞춤법에 따라 발화 내용을 전사하는 것을 기본 원칙으로 했으며, 띄어쓰기도 한글 맞춤법에 따랐다. 전사 단위 구분이 가장 어려웠으며, 그 다음으로 들리는 대로 전사와 띄어쓰기가 어려웠다.

전사 단위는 발화 중 쉬는 구간이 있는 경우 구분하였으나, 특히 발화 속도가 매우 빠르고 쉬는 구간이 없이 말하는 화자의 경우에는 전사 단위의 구분이 쉽지 않았다. 이런 경우 중복 발화나 머뭇거림 부분에서 전사 단위를 구분하여 처리하였다.

들리는 대로 전사는 화자가 발화 시 습관적으로 여러 차례 빠르게 중복 발화를 하는 경우 중복 발화 횟수 파악이 어려워 여러 차례 다시 듣기를 해야 했으며, 발화 시작 시 단어를 작은 목소리로 발화하는 경우가 많아 전사가 어려웠다.

2차 수동 전사 시 맞춤법 및 띄어쓰기는 국립국어원의 『표준국어대사전』을 활용하였다. 관리자는 전사자의 전사 결과물에 틀린 부분이 없는지, 전사 규칙은 제대로 준수했는지 등을 확인한 다음 틀린 부분이 있다면 수정 요청하여 수정한 후, 완료된 전사 파일은 관리 시스템에 등록하였다.

2차 수동 전사에서 주로 작업한 내용은 다음과 같다.

<표 13> 전사 지침

화자 표시	<ul style="list-style-type: none"> • 제외된 사람을 1로 표시했으며, 그들이 대부분 주 발화자 • 전사 파일에는 화자와 대화만 표시(화자 정보는 최종 원시 말뭉치에만 표시함.)
전사 단위	<ul style="list-style-type: none"> • 기본적으로는 절 단위(t-unit)이나 문장 단위가 되도록 함. • 가급적 쉽이 있는 구간에서 구분함. • 전사 단위는 3초 이상을 넘지 않도록 함.
문장 기호	<ul style="list-style-type: none"> • 억양에 따라 의미가 달라지는 경우는 마침표와 물음표를 사용해 구분 • 느낌표나 쉼표 등은 사용하지 않음.
발화 겹침	<ul style="list-style-type: none"> • 겹침 발화는 표시하지 않고 시간 순서대로 적음. • 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화 구분
단어	<ul style="list-style-type: none"> • 기본적으로는 한글 맞춤법을 따르지만 구어의 특성, 개인의 발음 특성, 방언 등 표준 발음이 아닌 경우는 소리 나는 대로 적음. • 모음의 변화, 수의적 경음화 등은 소리 나는 대로 적음(예: 쉼주, 쪼금). • 약화 현상에 의한 경우는 철자법에 따라 적음(예: 머 → 뮌). • 숫자나 기호, 영문자 등을 사용하지 않고 발음에 따라 한글로 적음.
끊어진 단어	<ul style="list-style-type: none"> • 단어가 불완전하게 발화된 경우, 발화된 그대로 전사하고 ‘=’를 붙여 정상적인 단어와 구분 * 단, 수정 발화, 반복 발화는 ‘=’를 표시 하지 않음.
띄어쓰기	<ul style="list-style-type: none"> • 합성 명사의 경우 대부분 붙여 쓰고, 의존명사는 띄어 씀. • 수를 적을 때는 만 단위로 띄어쓰기 • 본 용언과 보조용언 띄어쓰기 • 띄어쓰기 및 맞춤법은 국립국어원의 『표준국어대사전』을 참고함.
축약형의 표기	<ul style="list-style-type: none"> • 축약형의 경우 소리 나는 그대로 표기 • 모음 축약형은 ‘를’ 표시해 두 음소를 연결함(예: 사귀어, 바귀어).
담화 표기	<ul style="list-style-type: none"> • 아, 그, 저 등 별다른 의미가 없고 주로 머뭇거림이나 발화 습성으로 나타나는 단어는 단어 뒤에 ~를 표시함.
잘 들리지 않는 부분	<ul style="list-style-type: none"> • 잘 들리는 않는 부분 중 음절 수가 구분이 되는 경우 음절 수 만큼 xx를 표시 • 잘 들리지 않는 부분은 ()안에 전사 • 전혀 들리지 않으면 ... (마침표3개)로 전사
준음성의 표기	<ul style="list-style-type: none"> • 준음성은 @웃음, @목청, @박수, @노래 4가지로 국한 • 노래는 가사 없이 @노래로만 표기
의명성 보장	<ul style="list-style-type: none"> • 대화 내용 중 사람 이름이나 부정적 의미의 상호는 모두 n1, n2, ...로 전사하며 지칭하는 대상이 다를 경우 n1, n2로 구분하여 표기
일반 대화 관련 지침	<ul style="list-style-type: none"> • 대화 중 제 3자의 목소리, 전사자의 설명 등은 { }로 처리해 적고 향후 XML에서 <note>~</note>로 일괄 변환함.

- 1차 전사 예시**
- (1) 근데 저번에 나한테 김치 줬잖아.
 (1) 내가 **화자 표시** 안는 거 있었잖아.
 (1) 3 세가 라든가 그리고 재수없는 이 수직 맛있다 카더라
 그리고 어떨 겠는가
 (2) 그리고 작년 김장김치인데 전시관 먹는다 고 양념 만
 쪼끔 해 나한테 좀 심심 할 거 같은가 봐
 (1) 김장김치 였구나 큰 별 난색을 앞둔 날 김장김치 치고는
 (2) 김치냉장고 하여 놓고 문을 자주한니까
 그렇게된 거 같애 다 들 맛있 다고 그래. 2년 연속으로
 (1) 남은 좀 닦아 막는데 좀 새가 나올 수있는 건 줄 알았 지
 아 그래서 약간 그런 거는 있더라 군내 비슷한가
 약간 그런 감이 있다 그래도 어쨌든 맛있다 카베
 (2) 넘어갈 나무 떡 집에서 왔어 가지고 까
 (1) 나 섭니까 금방 본문을 준 게 있잖아.
 (1) 그게 답 하실 때가 나는 나는 새가 나온 거 별로 안 좋아한다
 (2) 그러면 요 번에 인지 한번 담아야 되는데 며칠 있다가 그때 연락
 할 때
 (1) 아 요새 적어 도 맛있다 타당하 열무김치 정적 살아 가지고 막
 업무 까딱 하다가는 누구인가?
 (1) 열무김치 하고 저거 소고기 볶음과 조류는 것만큼 고추장 인가?
 (1) 소고기 랑 초고추장 했는가 그런 것도 맛있는 그렇게가 기
 때문에 이 없으면 맛있다 매
- 전사 단위**

- 2차 전사 예시**
- 1:언니 저 **의명정보장** 김치 줬잖아. **픽어쓰기**
 내가 잘못 갖고 왔는 거 있었잖아 새로 나왔던 거
 그거 재수없는 맛있다 카더라 그게 어떻게 한 거야?
 2:그거 작년 김장김치인데 설 쉬고 먹는다고 양념을
 조금 해 냈더니 좀 심심하니 그렇던가 봐. **잘 들리지 않는 부분**
 1:아 김장김치였구나 그런데 별로 안 ... 던데 김장김치치고는
 2:김치냉장고 넣어 놓고 문을 자주 안 여니까 **숫자 표시**
 그렇게 된 것 같아 다들 맛있다고 그래. **이 년** 연속으로
 1:응 나는 좀 담가 냈는데 좀 새가로워 진 건 줄 알았지
 아 그래서 약간 저런 건 있더라 군내 비슷한 거
 약간 그런 건 있더라 그래도 애들은 맛있다 카대
 2:또 먹으려면 또 집에 있어 와서 가지고 가
 1:응 그러고 나서 그~ 언니가 금방 버무려 준 거 있잖아
 그게 더 맛있더라 나는 나는 새로운 거 별로 안 좋아한다
 2:그면 요번에 김치 한 번 담아야 하는데 며칠 있다가 그때 연락할게
 1:어 요새 **축약형의 말** 카더라 열무김치
 쪼쪼 썰어 가지고 뭐하고 먹었다 카더라 누군가 **군말**
 열무김치하고 저거 아~ 소고기 이렇게 볶은 거
 고추장인가 소고기랑 고추장 있는 거 그런 것도 **맛있=** 맛있다카대
 그렇게 해가 비벼 먹었더니 역수로 맛있다카대 **불완전 발화**

<그림 22> 2차 수동 전사 결과 예시

Home

음성파일번호 이름

검색

(1) 녹음 파일 다운로드

(2) 자동(1차)전사파일 다운로드

파일이름	주제1	주제2	주제3	자동전사파일		수동전사파일		자동전사여부	수동전사여부
SDRW00000000.wav	만화			파일 선택 선택된 파일 없음 다운로드		업로드		Yes	Yes
<div style="position: absolute; top: 225px; left: 135px; background-color: #0056b3; color: white; padding: 2px 5px;">다운로드</div>									
이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	비고/특이사항
화자1 (CH 1)	김철수	남	판매/영업 종사자	대학원재학/졸업이상	40~49세	서울	서울	부산	
화자2 (CH 2)	김미미	여	학생	대학원재학/졸업이상	20~29세	부산	서울	서울	

파일이름	주제1	주제2	주제3	자동전사파일		수동전사파일		자동전사여부	수동전사여부
SDRW00000002.wav	게임	군대	군대	파일 선택 선택된 파일 없음 업로드		업로드		No	No
<div style="position: absolute; top: 375px; left: 715px; background-color: #0056b3; color: white; padding: 5px; border: 1px solid #0056b3;">(3) 전사 파일 업로드</div>									
이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	비고/특이사항
화자1 (CH 1)	이빈	남	경영/관리직	초등학교 졸업 이하	20~29세	서울	서울	서울	
화자2 (CH 2)	이빈친구	여	경영/관리직	초등학교 졸업 이하	20~29세	서울	서울		

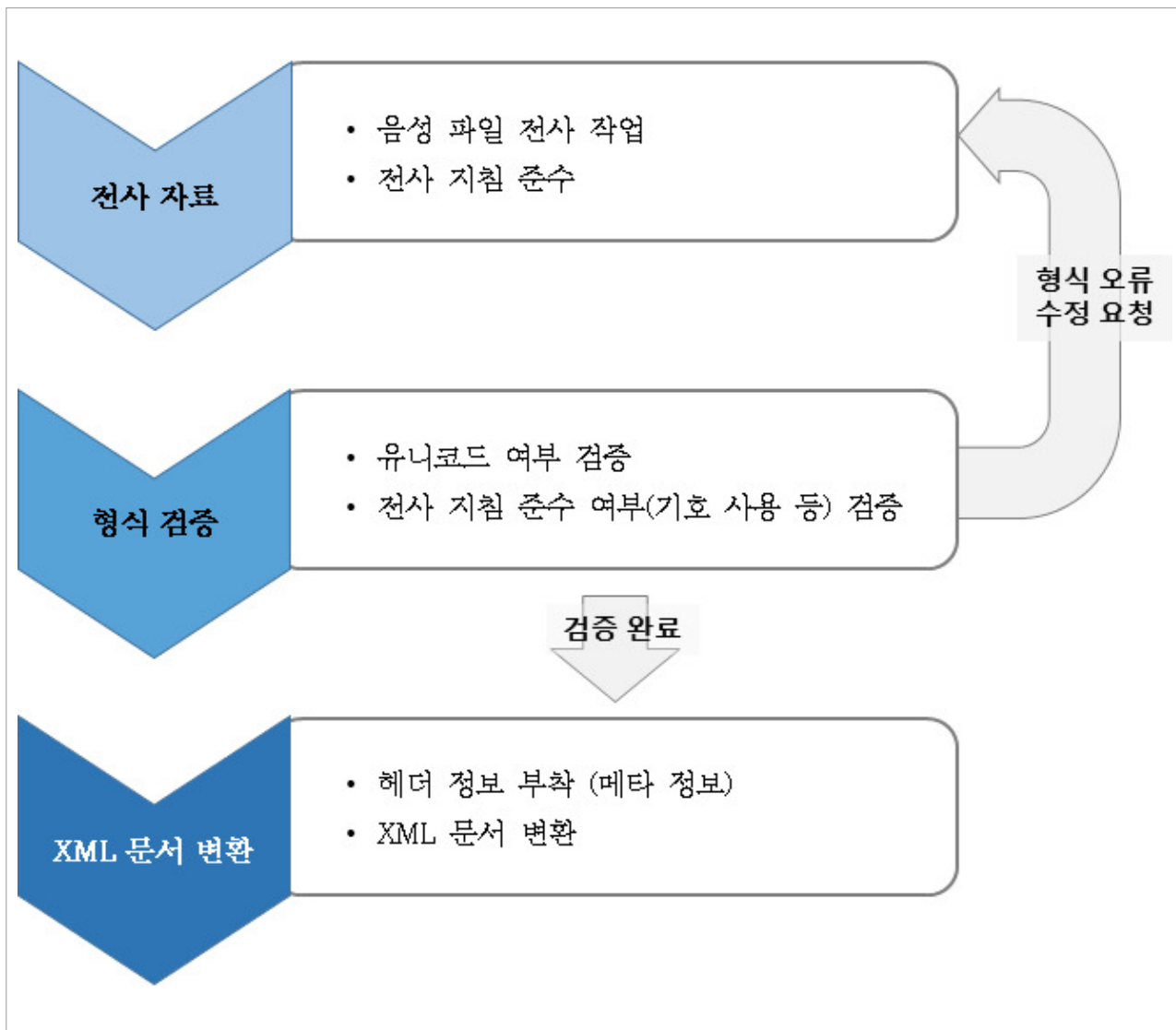
<그림 23> 녹음/전사 자료 다운로드 및 2차 수동 전사 자료 업로드 화면

5. 원시 말뭉치 구축 및 메타 정보 구축

5.1. XML 문서 변화 및 검증 절차

대화 녹음 전사 파일을 대상으로 메타 정보가 기록된 헤더 정보를 생성 및 부착하고, 발화 단위 마크업 등 XML 문서 형식으로 변환하는 작업을 수행함으로써 자료의 활용도를 높이하고자 하였다.

전사 파일을 XML 문서로 변환하기 전, 전사 파일이 전사 지침을 준수하였는지, 문자 인코딩은 유니코드로 되어 있는지 등을 검증 후 잘못된 부분은 전사자에게 수정 요청하고 형식이 제대로 갖춰진 전사 파일에 대해서만 헤더 정보를 부착하여 XML 문서로 변환하였다.



<그림 24> XML 문서 변환 및 검증 절차

5.2. 형식 검증 및 XML 문서 변환

XML 문서 변환은 한양대에서 자체 개발한 웹 서버용 XML 자동 변환 프로그램을 이용하여 웹 서버에 업로드 된 전사 파일과 메타 정보를 자동으로 XML 문서로 변환하고 헤더 정보를 생성·부착하였다.

XML 문서로 자동 변환하기 전, 자체 개발한 형식 검증 프로그램(KoreanXml.jar)을 이용함으로써 작업의 오류를 낮추고 효율성을 높이하고자 하였다. 형식 검증 프로그램을 이용해 단순한 형식 오류는 자동으로 수정했으며, 파일 형식(유니코드 형식) 및 전사 형식을 검증을 통해 오류가 없으면 XML 자동 변환 프로그램을 실행하고 오류가 나타나면 전사 요원에게 수정 요청하는 방식으로 진행했다.

SDRW00000326.txt -> 인코딩 형식이 UTF-8이 아닙니다.
 SDRW00000522.txt -> 화자 표시 오류
 SDRW00000525.txt -> 허용되지 않은 준음성 표기(@)가 있습니다.
 ==> 어~ 비도 오고 @웃
 SDRW00000598.txt -> 허용되지 않은 준음성 표기(@)가 있습니다.
 ==> 그니까는 이제 앞으로는 자율주행차도 많이 나올 거고 어 그리고@기침
 SDRW00000949.txt-> 포함되면 안 되는 기호(<, >, &)가 포함되어 있습니다.
 ==> 1:그러니까 군대 가서 나도 우리 아들 군대 저기 의정부>

<그림 25> 수정 요청 예시

XML 파일명은 아래와 같은 방식으로 부여하였다.

<표 14> 파일명 부여 방식

첫째 자리 :문어와 구어의 구분	둘째 자리 :매체 및 장르 대분류	셋째 자리 :말뭉치 유형 구분	넷째 자리 :구축년도	8자리 일련번호
S: 구어 말뭉치	D: 사적 대화	RW: 원시 말뭉치	19	#####

- SDRW1900000001.sjml 원시 말뭉치 첫 번째 파일
- SDRW1900000001.pcm 음성 원본 첫 번째 파일
- SDRW1900000001-00001.pcm 음성 원본 첫 번째 파일의 정제본 첫 번째 파일

<그림 26> 파일명 예시

5.3. XML 문서 구조와 형식

원시 말뭉치 파일의 확장자는 SJML이며, 문자 인코딩은 유니코드(UTF-8)이다. 기본 구조와 형식은 다음과 같다.

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
  <header>
    <fileInfo>
      <fileId>SDRW1900000001</fileId>
    </fileInfo>
  </header>
  <text>
    ....
  </text>
</SJML>
```

<그림 27> SJML 기본 구조

<표 15> 헤더(메타 정보를 담는 요소) 구조

<fileInfo>	<fileId>	파일의 고유 식별자(말뭉치 파일명)
	<annoLevel>	주석 수준: 원시
	<sourceDesc>	원자료에 대한 약술: 녹음하여 전사, 녹화하여 전사
	<class>	구축 계획에 따른 장르 분류: 사적 대화
	<subclass>	구축 계획에 따른 하위 장르 분류: 일상대화
	<tokenSize>	어절 수
<sourceInfo>	<title>	제목: 2인 일상대화 #####
	<author>	개인발화자
	<publisher>	개인대화녹음
	<date>	녹음 날짜
	<topic>	대화 주제
<profileInfo>	<personId> 외	화자 아이디(id) 화자 성별(sex) 화자 연령(age) 화자 직업(occupation) 출생지(bplace) 주 성장지(gplace) 현 거주지(city) 학력(education) 비고/특이 사항(note)
	<setting>	화자 간 관계(relation) 대화 장소, 대화 상황 정보 등 구어 사용 맥락에 대한 약술(situation)

<표 16> 본문(음성 전사 정보 포함) 마크업

XML 기호	내용	수동 전사 기호	수동 전사 예시	XML 문서 변환 예시
<u>	화자 표시 및 발화 번호 표시 · who(화자 정보) · n(발화 번호)	엔터	2: 근데 근데 얘기 엄청 귀여워.	<u who="P2" n="14">근데 근데 얘기 엄청 귀여워.</u>
<trunc>	끊어진 단어	=	전=	<trunc>전</trunc>
<unclear>	잘 들리지 않아 추정하여 전사한 경우	()	하나라구(더 힘들어)	하나라구 <unclear>더 힘들어</unclear>
	잘 들리지 않아 전사를 못한 경우 너무한 거 같더라.	<unclear/> 너무한 거 같더라.
	잘 들리지 않는 음절	x	진짜 xx해야 되겠더라.	진짜 <unclear>xx해야</unclear> 되겠더라.</u>
<vocal desc>	준음성과 기타 소리	@웃음	<ul style="list-style-type: none"> • @웃음 • @목청 • @박수 • @노래 	<vocal desc="laughing"/> <vocal desc="목청가다듣는소리"/> <vocal desc= "applauding"/> <vocal desc= "singing"/>
<anon type>	익명성 보장을 위한 마크업	n: 사람이름 s: 주민번호 c: 카드번호 a: 주소 t: 전화번호	신촌에 n는 진짜 맛없어. 그때 n1랑 n2랑 나랑 갔잖아	신촌에 <anon type= "name"/>는 진짜 맛없어. 그때 <anon type= "name" n="1"/>랑 <anon type="name" n="2"/>랑 너랑 나랑 갔잖아.

```

<?xml version="1.0" encoding="UTF-8"?>
<SJML>
<header>
  <fileInfo>
    <fileId>SDRW1900000086</fileId>
    <annoLevel>원시</annoLevel>
    <sourceDesc>녹음하여 전사</sourceDesc>
    <class>사적대화</class>
    <subclass>일상대화</subclass>
    <tokenSize>2577</tokenSize>
  </fileInfo>
  <sourceInfo>
    <title>2인 일상대화 00000086</title>
    <author>개인발화자</author>
    <publisher>개인대화녹음</publisher>
    <date>2019.07.10</date>
    <topic>휴일</topic>
  </sourceInfo>
  <profileInfo>
    <personId id="171" sex="여" ... 개인 정보 생략">P1</personId>
    <personId id="172" sex="여" ... 개인 정보 생략">P2</personId>
    <setting>
      <relation>친구</relation>
      <situation>(14~30분)중간중간 층간 소음(발소리),</situation>
    </setting>
  </profileInfo>
</header>
<text>
  <u who="P2" n="1"><anon type="name" n="1"/> 너 이제 방학했네.</u>
  <u who="P2" n="2">뭐 하고 지내꺼?</u>
  <u who="P1" n="3">이제 방학이니까 아무래도 휴일이니까</u>
  <u who="P1" n="4">놀러가야겠지</u>
  <u who="P1" n="5">아마 여행 같은 걸로.</u>
  <u who="P1" n="6">너는?</u>
  <u who="P2" n="7">나두 여행 가고 싶은데.</u>
  <u who="P2" n="8">너 이번에 어디 여행 가꺼?</u>
  <u who="P1" n="9">아 나는 이번에 휴일이 아무래도 길다 보니까</u>
  <u who="P1" n="10">유럽 여행 갈려고 하고 있어.</u>
  <u who="P2" n="13">그냥 밖에 돌아다니고 싶어.</u>
  ... 중략 ...
</text>
</SJML>

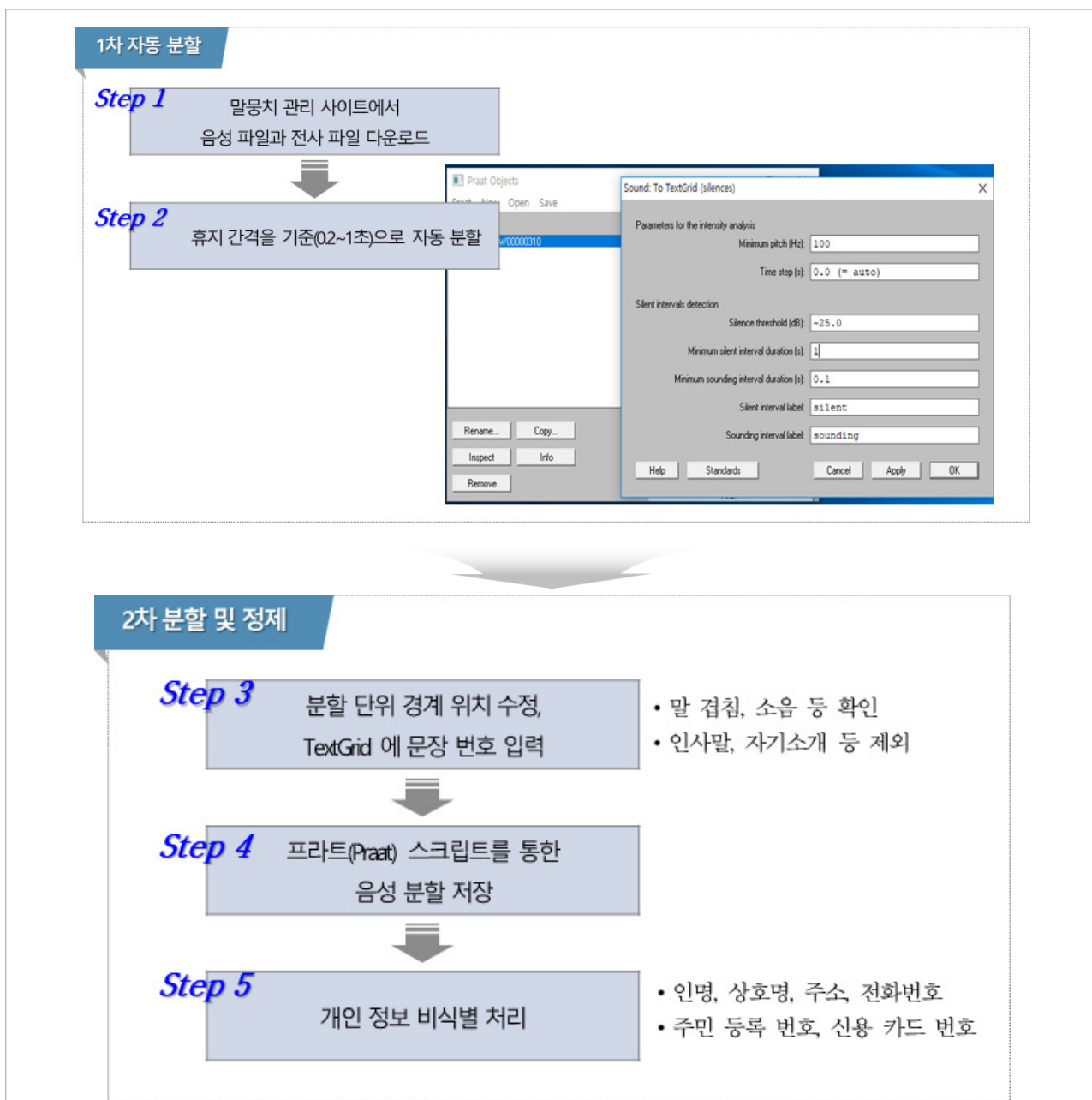
```

<그림 28> XML 문서 변환 예시

6. 음성 정제

6.1. 음성 정제 절차

음성 정제는 음성을 전사 단위에 따라 분할하여 저장하는 작업이다. 일상 대화 말뭉치 관리 시스템에서 음성 파일과 전사 파일을 내려받아 1차 자동 분할 작업 후, 음성 정제 작업자가 음성 파일을 직접 청취하면서 전사 파일(XML 파일)의 전사 단위에 따라 분할 경계를 수정 후 최종 정제본을 관리 사이트에 등록하는 순으로 진행되었다.



<그림 29> 음성 정제 절차

6.2. 음성 분할

음성 정제 작업자가 음성을 직접 듣고, 전사 파일(XML 파일)의 전사 단위와 프라트의 텍스트그리드의 경계를 일치시킨 후 텍스트그리드에 문장 번호를 입력했다. 파일 분할 저장을 위해 프라트 스크립트를 실행하고 음성 분할 결과를 저장했다. 이때 음성 구간 앞·뒤에 200msec의 휴지가 포함되도록 저장했다. 또한 음성 구간 앞·뒤에 잡음이 포함된 경우에는 잡음 외에 200msec 이상의 휴지가 포함되도록 했다.

음성 정제 시 대화 주제와 무관한 대화(예: 인사말, 자기소개 등)는 제외하고 정제했으며, 인명, 부정적 맥락에서 사용된 상호명, 주소, 주민 등록 번호, 신용 카드 번호, 주소, 전화번호 등의 개인 정보는 익명성 보장을 위해 묵음으로 비식별 처리하였다.

전사 파일 (XML 문서)

<u who="P2" n="1">우리 녹음 시작하까?</u>

<u who="P2" n="2">니가 먼저 해.</u>

<u who="P1" n="3">되게 <voice desc="laughing"/> 어색하다.</u>

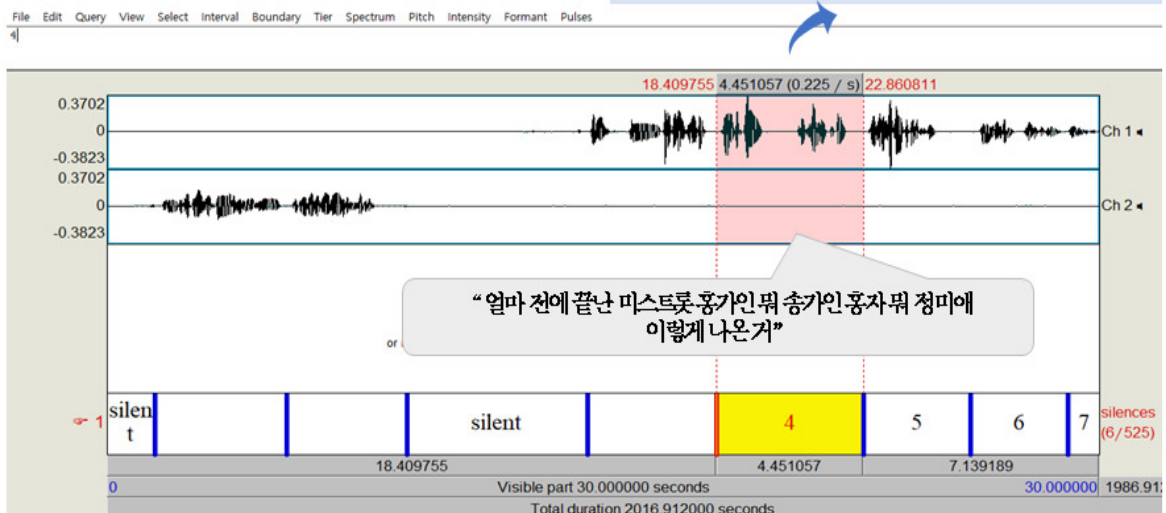
<u who="P1" n="4">얼마 전에 끝난 미스트롯 홍가인 뭐 송가인 홍자 뭐 정미에 이렇게 나온 거</u>

<u who="P1" n="5">재미있게 봤지.</u>

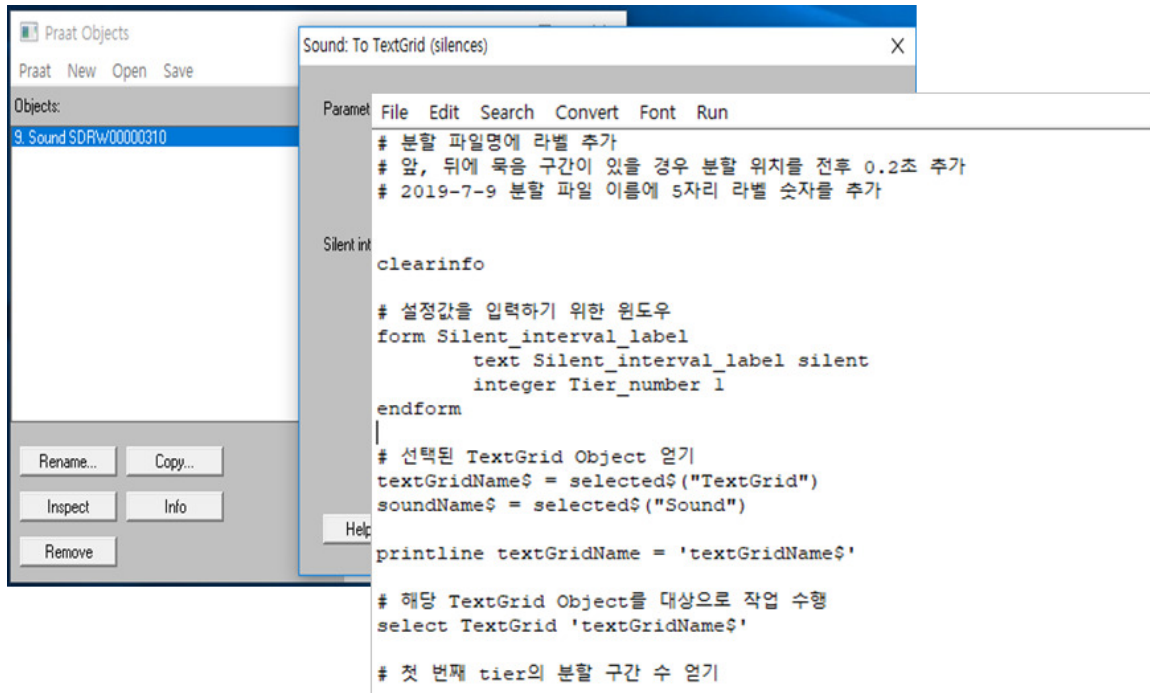
<u who="P1" n="6">그런데 거기서 송가인이 요새 엄청 여기저기 많이 프로그램에 엄청 많이 나오더라고</u>

프라트 스크립트를 이용한 음성분할 화면

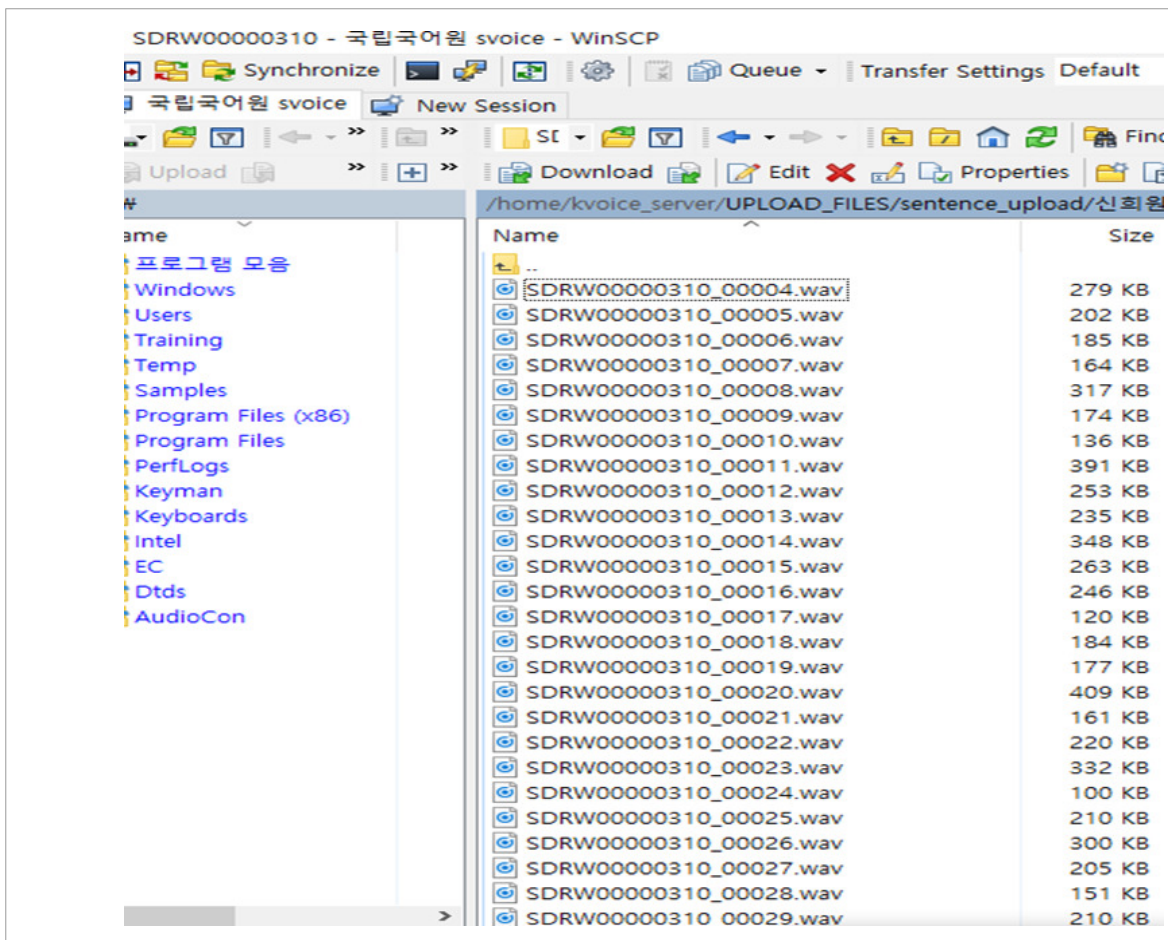
음성 구간 앞·뒤에 200msec 이상의 휴지 포함



<그림 30> 전사 단위와 텍스트그리드 문장 번호 일치

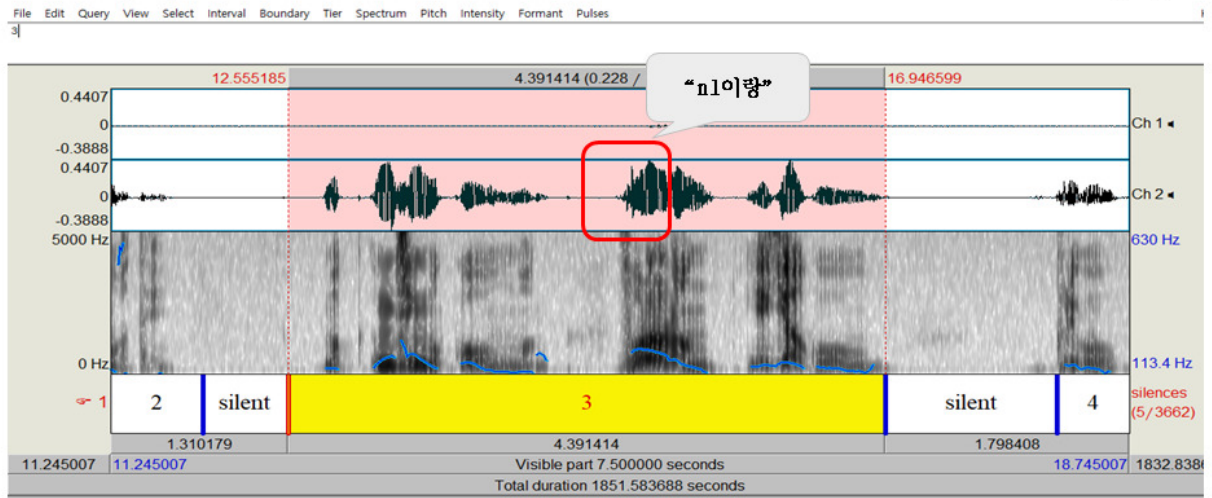


<그림 31> 프라트 스크립트 예시

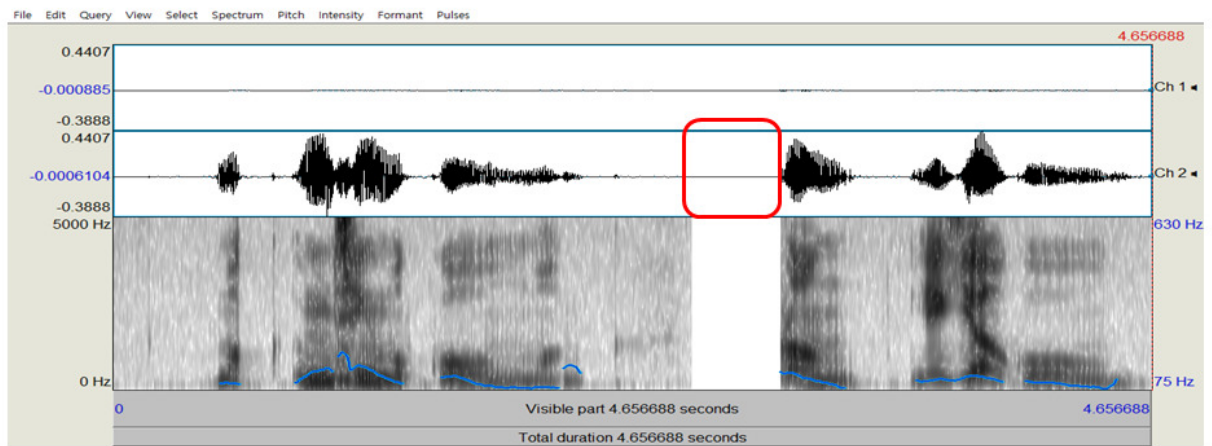


<그림 32> 문장 단위 음성 파일 저장 폴더 화면 예시

목음 처리 전



목음 처리 후



<그림 33> 개인 정보 비식별 처리 예시

7. 데이터베이스 구축 및 운영

7.1. 일상 대화 말뭉치 관리 시스템 구축

본 사업은 서울 및 6개 지역에서 총 2,000쌍의 대화 녹음이 진행되며, 산출물(음성 원본, 정제본, 메타 정보 파일, 1차·2차 전사 파일, XML 말뭉치 파일, 이용 허락 계약서)의 종류도 많아 지역별, 과정별로 자료를 관리하는 것이 무엇보다도 중요하다. 이러한 자료들을 효율적으로 관리하기 위해 일상 대화 말뭉치 관리 시스템(입출력 시스템)을 구축하여 운영하였다.

일상 대화 말뭉치 관리 시스템에 녹음 자료와 메타 정보, 이용 허락 계약서를 동시에 등록하도록 함으로써 자료의 누락 및 오류를 최소화하고자 하였으며, 자료의 유출을 방지할 수 있도록 하였다. 또한 각 지역별, 과정별 담당자들이 동시에 작업 수행이 가능하고, 과정 간 자료의 이동 시간을 줄일 수 있어 자료 수집 기간을 단축할 수 있었다.

입출력 시스템에 한양대에서 자체 개발한 XML 변환 프로그램을 연동하여 메타 정보가 포함된 원시 말뭉치 파일(XML 파일)을 자동으로 생성하도록 함으로써 사업 진행의 효율성을 높였다. 데이터베이스 구축은 데이터베이스 구축 관련 지침을 준수하여 설계하였다.



<그림 34> 일상 대화 말뭉치 구축 관리 시스템 로그인 화면 및 메뉴

자료의 안정적 관리와 보안을 위해 관리자와 작업자의 이용 권한을 다르게 하였다. 또한 음성 등록자, 전사자, 음성 정제자는 개인별로 각각의 계정을 부여하여 자신의 역할에 해당하는 메뉴만 보이도록 했으며, 자신에게 할당된 자료만 보이게 하여 자료의 보안을 유지하고 효율적으로 작업할 수 있도록 하였다.

파일 수정은 관리자와 해당 파일 등록자만이 수정할 수 있도록 하였다.

전체 진행사항											
전체 군대 게임 휴일 자동차 만화 영화 정치 건강/다이어트 방송/연예 스포츠/레저 먹거리 자연/휴양지 국가/지역 문학 연애/결혼 경제/재테크	확인										
	수집 상황										
	남성					여성					총합
	29세	30~39세	40~49세	50~59세	60세 이상	20~29세	30~39세	40~49세	50~59세	60세 이상	
	99	56	40	36	121	138	124	101	72	880	
	4	3	1	1	26	10	3	2	0	57	
	32	24	45	27	71	57	77	82	47	520	
	38	27	24	17	59	58	75	82	49	489	
	38	40	27	22	165	57	105	79	30	686	
	0	1	1	1	2	0	3	0	1	14	
합계	83	27	26	34	26	80	75	91	92	46	580
경기	59	31	15	11	9	75	37	31	45	29	342
세종	0	2	2	0	1	0	0	6	2	1	14
충북	8	7	0	3	2	18	9	8	15	11	81
충남	12	8	7	8	4	15	8	9	11	10	92
경북	4	4	0	2	2	9	5	1	6	3	36
경남	6	5	4	1	0	5	7	7	4	3	42
전북	1	2	0	1	0	5	2	7	1	0	19
전남	1	4	0	2	1	0	4	2	1	0	15
강원	41	10	14	8	20	31	20	24	15	18	201
제주	16	6	4	7	5	15	11	16	17	11	108
총합	577	317	223	215	174	697	498	589	555	331	4176

<그림 35> 일상 대화 말뭉치 구축 관리 시스템 진행 상황 확인 화면(관리자용)

수집 자료 검색

Home

전체 엑셀파일 다운로드

음성 파일 정보로 검색

음성 파일 아이디

숫자만 입력. (예: 파일명이 SDRW00000023 이면 23 입력)

검색

녹음 날짜

yyyyymmdd (예: 2019060)

검색

주제1

군대

검색

참여자 정보로 검색

이름

검색

성별

☐ 남
☐ 여

검색

직업

경영/관리직

검색

학력

초등학교 졸업 이하

검색

연령대

20~29세

검색

출생지

서울

검색

주 성장지

서울

검색

거주지

서울

검색

녹음파일 확정

검색

전사자명

검색

녹음등록자

검색

검색결과 엑셀파일 다운로드

<그림 36> 일상 대화 말뭉치 구축 관리 시스템 자료 검색 및 참가자 정보 다운로드 화면(관리자용)

파일이름	SDRW00000001.wav	다운로드	찾아보기...	음성 파일 업로드	담당자	sp_metrix8	담당자 선택	sp_metrix8	선택
	이름	성별	생년월일	직업	학력	연령대	출생지	sp_metrix9	지 비고/특이사항
화자1 (CH 1)	박	남	1980-02-19	사무 종사자	대졸	30~39세	대구	sp_metrix10	직업 역무원
화자2 (CH 2)	기	남	1983-11-08	사무 종사자	대졸	30~39세	대구	sp_metrix11	회사원
								sp_metrix12	
								sp_metrix13	
								sp_metrix14	
								sp_metrix15	
								sp_metrix16	
								sp_metrix17	
								sp_metrix18	
파일이름	SDRW00000002.wav	다운로드	찾아보기...	음성 파일 업로드	담당자	sp_metrix8	담당자	sp_metrix19	거주지 비고/특이.
	이름	성별	생년월일	직업	학력	연령대	출생지	sp_metrix20	대구
화자1 (CH 1)	임	남	1995-05-07	학생	대학교 재학	20~29세	경북	sp_metrix21	
화자2 (CH 2)	임	남	1995-12-09	학생	대학교 재학	20~29세	경기	sp_metrix22	대구
								sp_metrix23	
								sp_metrix24	
								sp_metrix25	
								sp_metrix26	
								sp_metrix27	
파일이름	SDRW00000003.wav	다운로드	찾아보기...	음성 파일 업로드	담당자		담당자 선택	sp_metrix28	비고/특이사항
	이름	성별	생년월일	직업	학력	연령대	출생지	sp_metrix29	기관사
화자1 (CH 1)	양	남	1994-08-06	기술자 종사자(장치/기계 조작 및 조립 종사자)	고졸	20~29세	경북	sp_metrix30	기관사
화자2 (CH 2)	권	남	1992-03-15	기술자 종사자(장치/기계 조작 및 조립 종사자)	고졸	20~29세	경북	sp_metrix31	기관사
								sp_metrix32	
								sp_metrix33	
								sp_metrix34	
								sp_metrix35	
								sp_metrix36	
								sp_metrix37	
파일이름	SDRW00000004.wav	다운로드	찾아보기...	음성 파일 업로드	담당자		담당자 선택	sp_metrix38	비고/특이사항
	이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지 비고/특이사항
화자1 (CH 1)	이	여	1995-07-18	무직/취업준비생	대졸	20~29세	대구	대구	대구
화자2 (CH 2)	정	여	1996-01-02	무직/취업준비생	대졸	20~29세	경북	경북	대구

<그림 37> 일상 대화 말뭉치 구축 관리 시스템 전사자 지정 화면(관리자용)

음성 파일 선택

대화 정보

녹음 날짜	주제1	주제2	주제3	화자간 관계	기타
yyyyymmdd (예: 2019060)	선택				<div> 대화 상황을 이해하는데 필요한 사항을 적어주세요. (참소, 주변소음, 분위기 등) </div>

화자 정보

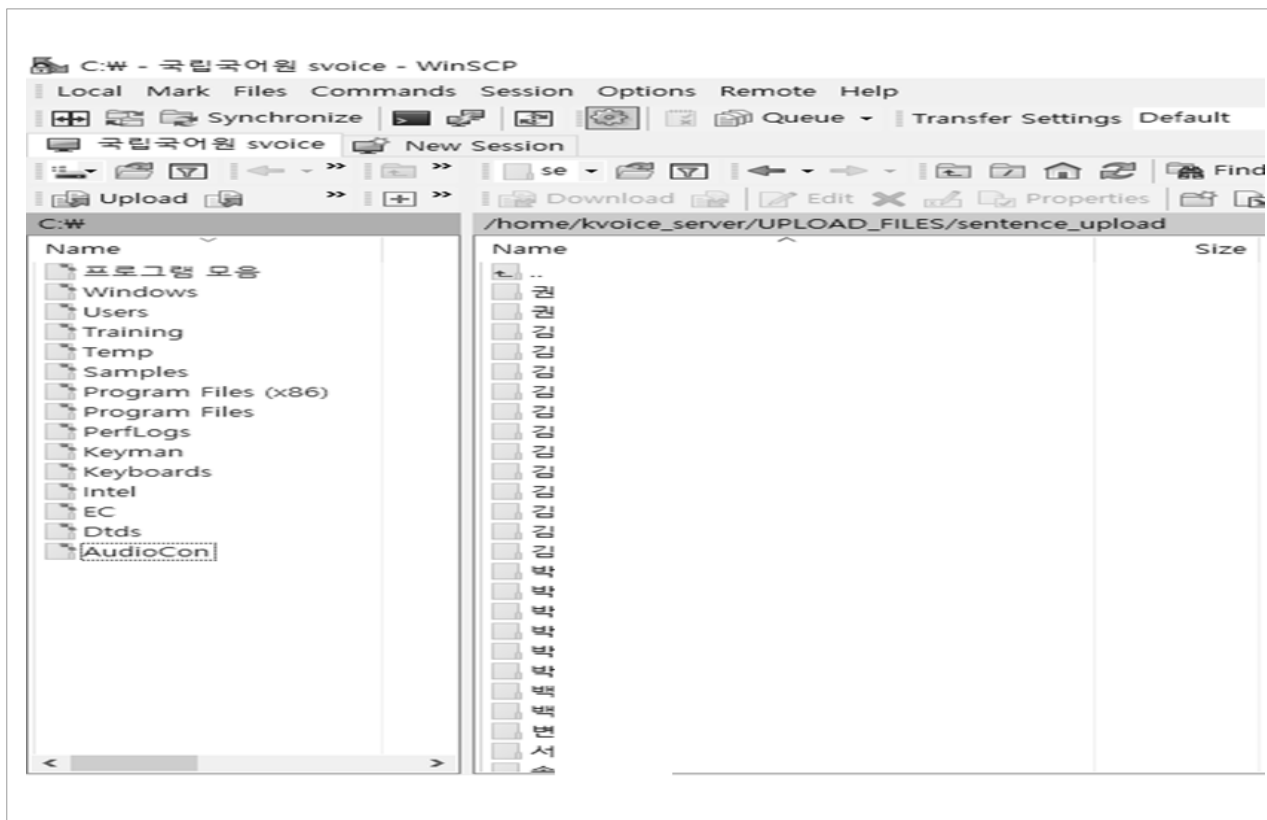
	이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	동의서 업로드 (pdf 파일만 등록 가능)	비고/특이사항
화자1 (왼쪽:메인)		○ 남 ○ 여	yyyyymmdd (예: 2019060)	선택	선택	선택	선택	선택	선택	찾아보기...	구체적 직업, 발화 속도, 서투리 등
화자2 (오른쪽)		○ 남 ○ 여	yyyyymmdd (예: 2019060)	선택	선택	선택	선택	선택	선택	찾아보기...	구체적 직업, 발화 속도, 서투리 등

<그림 38> 일상 대화 말뭉치 구축 관리 시스템 화자 정보 및 음성 파일 등록 화면(녹음 진행 요원용)

파일이름	주제1	주제2	주제3	자동전사파일		수동전사파일		자동전사여부	수동전사여부	녹음파일 등록자
SDRW00000001.wav 다운로드	자동차			전사 파일 없음		찾아보기...	업로드	No	Yes	metrix1
	이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	비고/특이사항
화자1 (CH 1)	박	남	1980-02-19	사무 종사자	대졸	30~39세	대구	대구	경북	직업 역무원
화자2 (CH 2)	기	남	1983-11-08	사무 종사자	대졸	30~39세	대구	대구	대구	회사원

파일이름	주제1	주제2	주제3	자동전사파일		수동전사파일		자동전사여부	수동전사여부	녹음파일 등록자
SDRW00000002.wav 다운로드	게임			전사 파일 없음		찾아보기...	업로드	No	Yes	metrix1
	이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	비고/특이사항
화자1 (CH 1)	임	남	1995-05-07	학생	대학교 재학	20~29세	경북	경북	대구	
화자2 (CH 2)	임	남	1995-12-09	학생	대학교 재학	20~29세	경기	대구	대구	

<그림 39> 일상 대화 말뭉치 구축 관리 시스템 전사 자료 등록 화면(작업자용)



<그림 40> 일상 대화 말뭉치 구축 관리 시스템 음성 정제 등록 화면(음성 정제자용)

음성 파일 정제 시 파일의 양이 많고 파일의 용량이 큰 관계로 개인 키와 공개 키를 이용한 시스템 로그인 방법을 통해 자료를 등록하였다.

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

Home

같은 참여자 정보로 기록된 파일 SDRW00000001.wav, 이 이미 존재합니다.

음성 파일 선택 (wav 파일만 등록 가능)

찾아보기...

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

Home

같은 참여자 정보로 기록된 파일 SDRW00000001.txt, 이 이미 존재합니다.

파일이름	주제1	주제2	주제3	자동전사파일	수동전사파일
SDRW00000001.wav	자동차			전사 파일 없음	
다운로드					찾아보기... 업로드

<그림 41> 음성 파일 및 전사 파일 중복 등록 시 오류

국립국어원 - 검색 x 국립국어원 표준국어대사전

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

수집 자료 검색

Home

참여자 정보로 검색

이름 임 검색

파일이름	주제1	주제2	주제3	자동전사파일	수동전사파일	자동전사여부	수동전사여부	녹음파일 등록자	담당자	참가자 정보 수정
SDRW00000002.wav	게임			전사 파일 없음	다운로드	No	Yes	metrix1	sp_metrix8	
	이름	성별	생년월일	직업	학력	연령대	출생지	주 성장지	거주지	비고/특이사항
화자1 (CH 1)	임	남	1995-05-07	학생	대학교 재학	20~29세	경북	경북	대구	
화자2 (CH 2)	임	남	1995-12-09	학생	대학교 재학	20~29세	경기	대구	대구	

<그림 42> 음성 파일 및 화자 정보 수정 화면

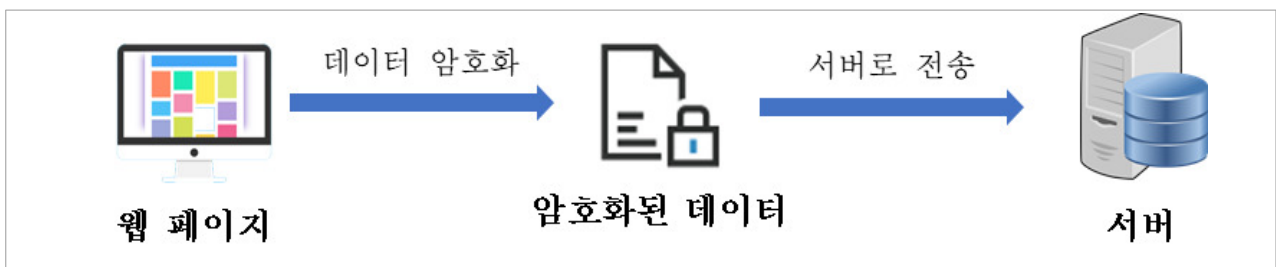
7.2. 관리 시스템 보안

계정 허락 절차를 통해 승인된 사용자만 웹 페이지 이용이 가능하며, 스프링 시큐리티(Spring security)를 이용하여 로그인한 사용자만 이용이 가능하도록 하였다. 계정별 권한을 다르게 부여하여 권한별로 접근할 수 있는 기능에 제한을 두었으며, 각각의 사용자는 자신이 열람 가능한 정보만 확인할 수 있도록 하여 불필요한 데이터 접근을 최소화하였다.



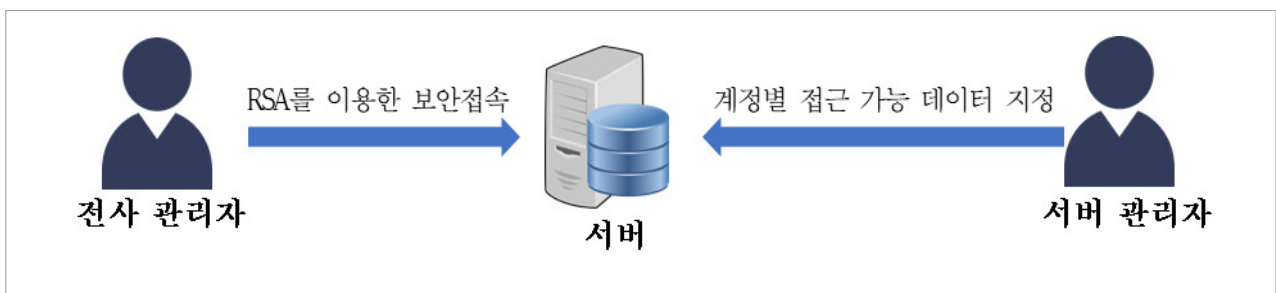
<그림 43> 웹 페이지 보안

HTTPS 프로토콜을 이용하여 암호화된 데이터 통신을 하였다.

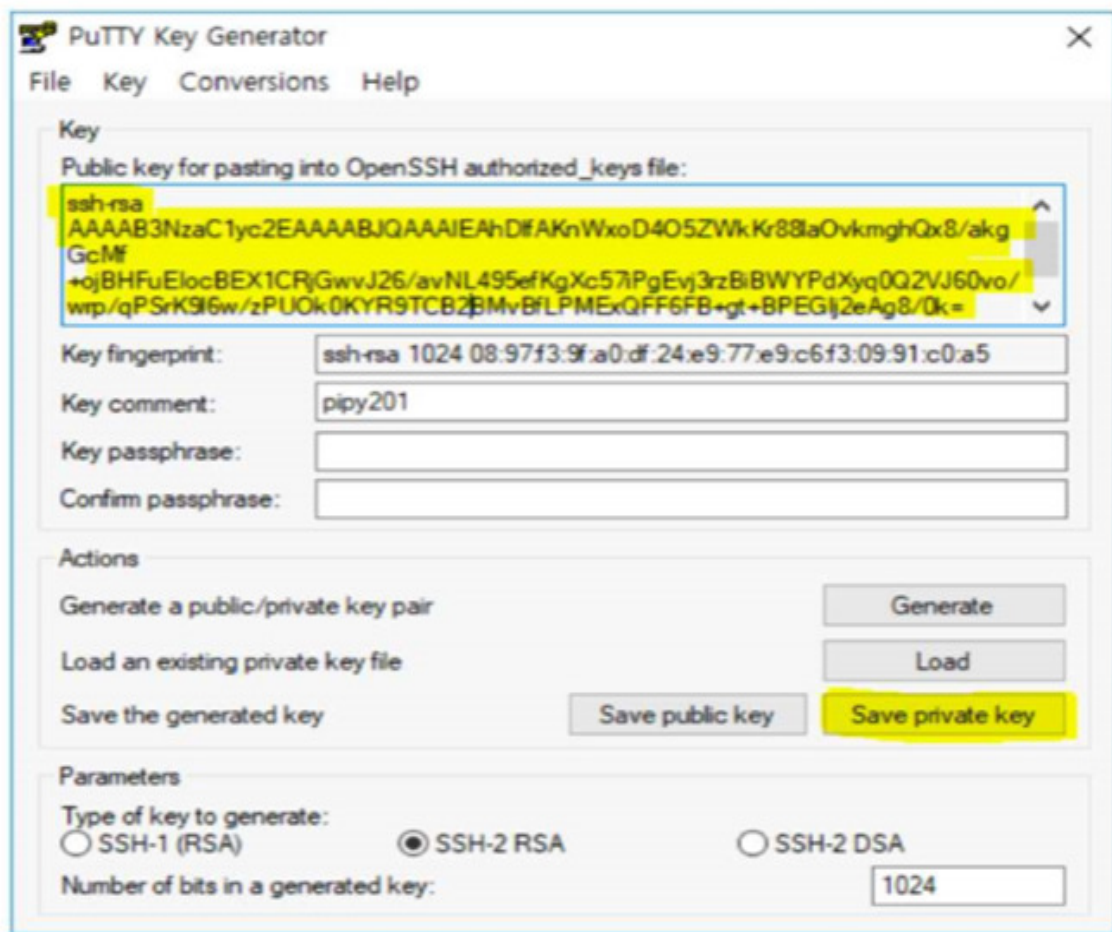


<그림 44> 인터넷 통신 보안

서버 접근 시 계정의 비밀번호 유출 방지를 위해 RSA 기반의 공개 키-개인 키를 이용한 로그인 방식만 허용하였다. 계정별 데이터 접근 권한을 부여하여, 불필요한 접근을 최소화하였다.

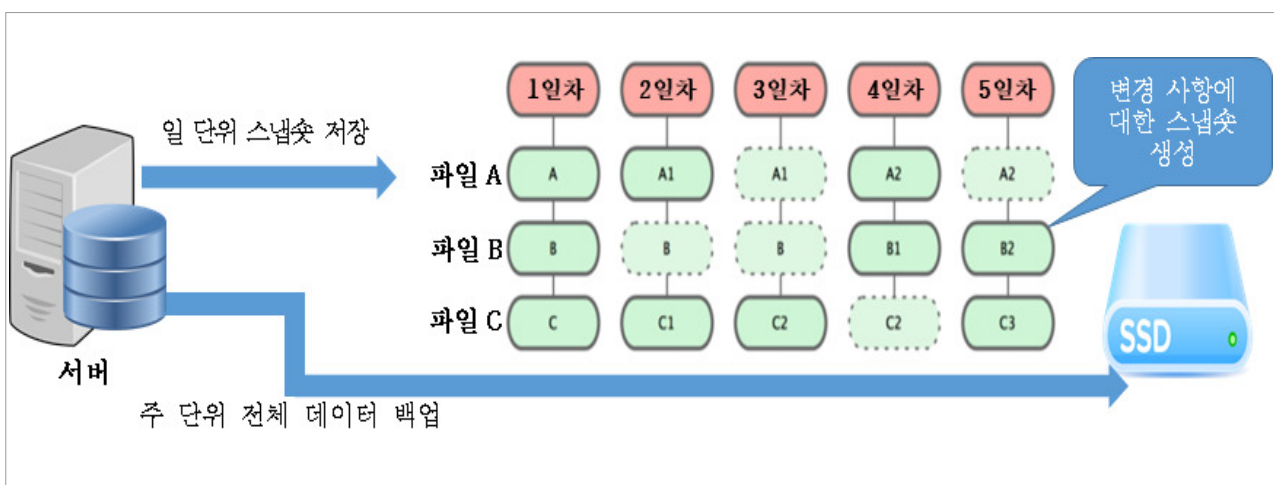


<그림 45> 서버 보안



<그림 46> 개인 키와 공개 키 생성 화면

서버의 데이터가 손실되는 위험을 방지하기 위해 매주 토요일 새벽 SSD에 수집한 데이터를 백업하여 관리하였으며, 일 단위로 서버 이미지의 스냅샷을 저장하였다.



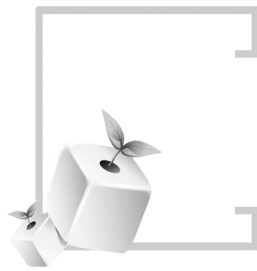
<그림 47> 서버 보안

이러한 시스템은 국가 정보보안 기본지침(국가정보원, 2019. 3.), 국가·공공기관 용역업체 보안관리 가이드라인(국가정보원, 2014. 3.), 문화체육부 개인정보보호지침(훈령 제342호, 2018. 6. 11.)의 보안정책 및 지침을 준수하여 개발하였다.

기타 본 사업수행 과정에서 생산되는 모든 산출물은 지정된 PC에 저장·관리하고 본 사업을 수행하지 않는 자에게는 제공·대여·열람하지 못하도록 하였다. 작업자의 PC에는 비밀번호를 설정하였으며 인가받지 못한 USB나 휴대용 저장 매체는 사용하지 못하도록 했다. 또한 P2P, 웹 하드, 상용 메일이나 메신저를 통해 자료 공유를 하지 않도록 했다. 그 외에도 정부가 제정·공포한 관계 제 법규(지침)를 준수하였다.

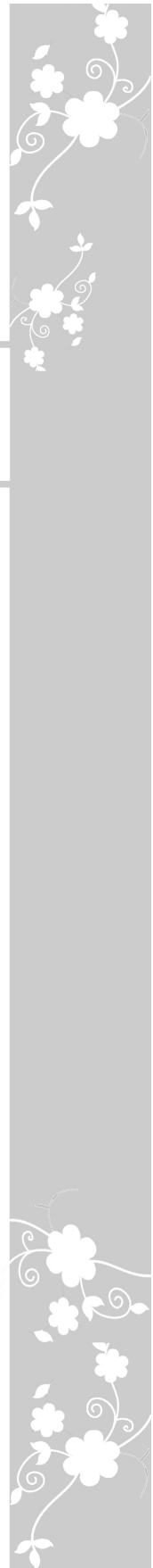
<표 17> 작업장 및 장비 등에 대한 보안 관리 방안

구분	작업장 및 장비 등에 대한 보안 관리
단말기 보안	<ul style="list-style-type: none"> • 외부 파일 전송 차단(웹하드, 메신저, 메일 등) • IP주소 형식의 웹사이트 및 보안 우려 사이트 차단 • PC 로그인, 화면 보호기 비밀번호 설정 • USB, 외장 하드 등 저장 매체 관리(매체 인증, 차단, 로그 관리) • 공유 폴더 차단 • CD/DVD, 무선 랜, 블루투스 차단 • 화면 캡처 차단
문서/파일 보안	<ul style="list-style-type: none"> • 생성 또는 입수 시 관리 번호 생성 • 관리 대상 문서의 생성, 이동/이관, 폐기 시 자료 관리 대장 기재 • 모든 관리 대상 문서에 비밀번호 생성 • 단계별 작업 완료 시 작업자 보관 문서/파일 폐기 <ul style="list-style-type: none"> - 디지털 파일은 책임자 입회하에 폐기 후 보안 소프트웨어로 확인 - 종이 문서는 문서 폐기 기관 위탁. 폐기 과정 폐쇄 회로 TV로 촬영 - 소량 종이 문서는 세절기로 책임자 입회하에 직접 폐기
출입 보안	<ul style="list-style-type: none"> • 외부로 연결 가능한 모든 출입구에 폐쇄 회로 TV 설치 • 인가된 직원만 출입 가능한 사원증 타각 출입 통제 시스템



제 3 장

사업 수행 결과



1. 주제별 수집 결과

주제별 수집 결과, 모든 주제가 최소 목표치를 초과하여 수집되었다. 선호가 낮은 문학은 95쌍으로 가장 적게 수집되었으며, 군대가 156쌍으로 가장 많이 수집되었다.

<표 18> 주제별 수집 결과

주제	최소 목표 (쌍)	수집 결과 (쌍)	목표 대비 수집률 (%)
군대	125	156	124.8
게임	125	125	100.0
휴일	125	129	103.2
자동차	125	129	103.2
만화	70	108	154.3
영화	125	127	101.6
정치	70	110	157.1
건강/다이어트	125	127	101.6
방송/연예	125	128	102.4
스포츠/레저	125	127	101.6
먹거리	125	127	101.6
자연/휴양지	125	128	102.4
국가/지역	125	131	104.8
문학	70	95	135.7
연애/결혼	125	127	101.6
경제/재테크	125	126	100.8
합계		2,000쌍	100.0%

2. 참가자 특성별 수집 결과

2.1. 인구 특성별 수집 결과

인구 특성별 수집 결과, 섭외자 기준으로 목표치에 맞게 모두 수집하였다. 목표 대비 수집 비율을 살펴보면, 남·여 20~30대와 여자 40~50대는 초과 수집하였다. 반면, 60대 이상은 성별에 관계없이 최소 목표치의 50%는 초과했으나 100%에는 미치지 못하였다. 지역별로는 녹음 장소가 있는 지역은 목표치를 초과 수집했으며, 그 외 지역은 최소 목표치 50%를 초과해 85% 이상을 수집하였다.

<표 19> 성×연령×지역별 목표 대비 수집률(기준: 섭외자, 단위 : %)

구분		남자					여자					합계	
		20대	30대	40대	50대	60대 이상	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	104.2	133.3	121.7	78.3	65.2	141.7	104.2	108.3	134.8	82.6	107.7	100.0
	경기	174.1	133.3	53.8	50.0	50.0	125.9	107.4	100.0	103.8	61.5	96.6	
	인천	130.8	107.7	69.2	53.8	53.8	130.8	107.7	100.0	115.4	61.5	93.1	
충청권	대전	118.2	100.0	63.6	109.1	81.8	127.3	100.0	145.5	109.1	81.8	103.6	100.0
	충북	109.1	110.0	54.5	90.0	60.0	110.0	100.0	110.0	100.0	100.0	94.2	
	충남	91.7	100.0	58.3	125.0	91.7	100.0	100.0	100.0	141.7	108.3	101.6	
경북권	대구	191.7	116.7	91.7	58.3	66.7	141.7	108.3	108.3	116.7	91.7	109.2	100.0
	경북	100.0	92.3	66.7	66.7	50.0	100.0	100.0	107.7	125.0	100.0	91.2	
경남권	부산	235.7	100.0	100.0	57.1	50.0	285.7	100.0	100.0	85.7	42.9	115.7	100.0
	경남	128.6	78.6	78.6	53.8	61.5	157.1	100.0	100.0	107.7	46.2	91.9	
	울산	125.0	100.0	75.0	62.5	50.0	112.5	100.0	100.0	87.5	50.0	86.3	
전라권	광주	177.8	111.1	66.7	77.8	55.6	144.4	122.2	122.2	88.9	122.2	108.9	100.0
	전북	140.0	90.0	80.0	110.0	70.0	100.0	100.0	120.0	140.0	80.0	103.0	
	전남	110.0	100.0	70.0	70.0	80.0	100.0	100.0	100.0	100.0	60.0	89.0	
강원	강원	133.3	100.0	100.0	66.7	100.0	122.2	100.0	100.0	100.0	77.8	100.0	100.0
제주	제주	100.0	100.0	80.0	100.0	80.0	100.0	100.0	100.0	120.0	120.0	100.0	100.0
합계		138.4	108.4	77.9	73.6	64.5	135.0	103.4	107.0	112.2	77.2	100.0	100.0

<표 20> 성×연령×지역별 수집 결과(기준: 섭외자, 단위: 명)

구분		남자					여자					합계	
		20대	30대	40대	50대	60대 이상	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	25	32	28	18	15	34	25	26	31	19	253	630
	경기	47	36	14	13	13	34	29	27	27	16	256	
	인천	17	14	9	7	7	17	14	13	15	8	121	
충청권	대전	13	11	7	12	9	14	11	16	12	9	114	335
	충북	12	11	6	9	6	11	11	11	10	10	97	
	충남	11	13	7	15	11	13	12	12	17	13	124	
경북권	대구	23	14	11	7	8	17	13	13	14	11	131	245
	경북	13	12	8	8	6	13	13	14	15	12	114	
경남권	부산	33	14	14	8	7	40	14	14	12	6	162	355
	경남	18	11	11	7	8	22	14	13	14	6	124	
	울산	10	8	6	5	4	9	8	8	7	4	69	
전라권	광주	16	10	6	7	5	13	11	11	8	11	98	290
	전북	14	9	8	11	7	10	10	12	14	8	103	
	전남	11	10	7	7	8	10	10	10	10	6	89	
강원	강원	12	9	9	6	9	11	9	9	9	7	90	90
제주	제주	6	6	4	5	4	6	6	6	6	6	55	55
합계		281	220	155	145	127	274	210	215	221	152	2,000	2,000

참가자 전체의 인구 특성별 현황을 살펴보면, 성별×연령별로는 여자 20대, 여자 40대가 가장 많이 수집되었으며, 남자 60대 이상이 가장 적게 수집되었다. 지역별로는 서울이 가장 많이 수집되었으며, 울산이 가장 적게 수집되었다.

<표 21> 성×연령×지역별 수집 결과(기준: 화자 전체, 단위: 명)

구분		남자					여자					합계	
		20대	30대	40대	50대	60대 이상	20대	30대	40대	50대	60대 이상	지역별	권역별
수도권	서울	67	67	47	31	22	97	97	96	73	45	642	1,193
	경기	76	44	16	14	15	84	51	34	37	23	394	
	인천	20	18	9	7	7	30	20	18	15	13	157	
충청권	대전	36	16	13	20	12	45	31	42	39	14	268	623
	충북	15	11	6	11	8	20	19	23	22	17	152	
	충남	19	14	10	23	14	26	22	20	30	25	203	
경북권	대구	51	28	14	9	10	47	32	40	45	27	303	516
	경북	18	14	11	14	7	23	20	32	41	33	213	
경남권	부산	85	22	23	11	11	124	36	63	46	18	439	745
	경남	30	14	15	10	12	32	26	31	30	10	210	
	울산	15	10	7	6	4	12	13	15	10	4	96	
전라권	광주	31	11	8	11	7	33	25	31	24	20	201	640
	전북	25	10	11	16	12	24	28	41	45	28	240	
	전남	25	11	11	11	10	27	30	36	31	7	199	
강원	강원	27	11	14	7	13	25	19	22	18	19	175	175
제주	제주	13	6	5	7	5	18	12	14	16	12	108	108
합계		553	307	220	208	169	667	481	558	522	315	4,000	4,000

2.2. 주제별×연령별/성별 수집 결과

화자 전체의 연령별 주요 대화 주제를 살펴보면, 20대는 게임과 군대, 30대는 자동차와 휴일, 40대는 건강/다이어트와 경제/재테크, 50대는 자연/휴양지와 먹거리, 60대 이상은 정치, 먹거리에 대한 주제를 많이 선택하여 대화하였다.

<표 22> 주제별×연령별 수집 결과(기준: 화자 전체, 단위: 명)

주제	20대	30대	40대	50대	60대 이상	총합계
건강/다이어트	38	32	82	56	46	254
게임	176	46	17	8	3	250
경제/재테크	13	53	79	65	42	252
국가/지역	56	49	64	52	41	262
군대	155	59	36	39	23	312
만화	112	58	31	13	2	216
먹거리	48	34	51	70	51	254
문학	61	32	35	44	18	190
방송/연예	76	59	59	39	23	256
스포츠/레저	80	43	42	47	42	254
연애/결혼	106	56	47	19	26	254
영화	129	46	37	35	7	254
자동차	43	72	57	59	27	258
자연/휴양지	37	54	41	79	45	256
정치	18	29	54	62	57	220
휴일	72	66	46	43	31	258
합계	1,220	788	778	730	484	4,000

화자 전체의 성별로 주요 대화 주제를 살펴보면, 남자는 군대와 게임, 스포츠/레저, 여자는 방송/연예와 건강/다이어트, 휴일, 먹거리 등에 대한 주제를 많이 선택하여 대화하였다.

<표 23> 주제별×성별 수집 결과(기준: 화자 전체, 단위: 명)

주제	남자	여자	총합계
건강/다이어트	45	209	254
게임	173	77	250
경제/재테크	84	168	252
국가/지역	78	184	262
군대	213	99	312
만화	69	147	216
먹거리	48	206	254
문학	45	145	190
방송/연예	39	217	256
스포츠/레저	147	107	254
연애/결혼	55	199	254
영화	84	170	254
자동차	131	127	258
자연/휴양지	68	188	256
정치	127	93	220
휴일	51	207	258
합계	1,457	2,543	4,000

2.3. 화자 관계별 수집 결과

화자 간 관계별 현황을 살펴보면, 친구 관계가 전체의 55.3%로 가장 많았으며, 다음으로 부부(14.1%), 부모/자녀 관계(7.6%) 등의 순으로 많이 수집되었다. 부부, 부모/자녀, 형제/자매, 기타 가족을 가족 관계라고 하면 가족 관계로 참석한 쌍은 28.7% 이다.

일부 참석자는 이웃사촌이나 모임/동아리 지인, 교회 지인, 직장 동료 등 친한 사이를 친구 관계로 밝히기도 해 친구 관계의 화자 쌍이 다른 관계보다 많이 나타났다.

다른 화자 간 관계와 달리 기타 관계는 구체적 관계를 밝히지 않았다.

<표 24> 화자 관계별 수집 결과 (기준 : 쌍, 단위 : 쌍)

관계	그룹 수	비율
친구	1,105	55.3
부부	282	14.1
부모/자녀	151	7.6
형제/자매	115	5.8
연인	96	4.8
직장 동료	35	1.8
이웃사촌	30	1.5
모임/동아리 지인	18	0.9
대학 선후배	15	0.8
교회 지인	7	0.4
고향 선후배	3	0.2
사제 관계	1	0.1
기타 가족	24	1.2
기타	118	5.9
합계	2,000쌍	100.0%

2.4. 직업별 수집 결과

참가자 전체의 직업별 현황을 살펴보면, 전업주부가 가장 많이 수집되었으며, 다음으로 학생, 사무 종사자 순으로 많이 수집되었다.

<표 25> 직업별 수집 결과(기준: 화자 전체, 단위: 명)

직업	화자 수	비율
전업주부	1,048	26.2
학생	750	18.8
사무 종사자	626	15.7
판매/영업 종사자	260	6.5
전문가 및 관련 종사자	240	6.0
서비스 종사자	227	5.7
경영/관리직	99	2.5
기술자 종사자	71	1.8
기능원 및 관련 기능 종사자	47	1.2
단순 노무 종사자	36	0.9
농업/임업/어업 종사자	11	0.3
군인	3	0.1
무직/취업 준비생	340	8.5
기타	242	6.1
합계	4,000명	100.0%

2.5. 학력별 수집 결과

참가자 전체의 학력별 현황을 살펴보면, 대학교 졸업이 가장 많이 수집되었으며, 다음으로 고등학교 졸업, 대학교 재학 순으로 많이 수집되었다.

<표 26> 학력별 수집 결과(기준: 화자 전체, 단위: 명)

학력	화자 수	비율
초등학교 졸업 이하	32	0.8
중학교 졸업	78	2.0
고등학교 졸업	837	20.9
대학교 재학	718	18.0
대학교 졸업	2,108	52.7
대학원 재학/졸업 이상	227	5.7
합계	4,000명	100.0%

2.6. 출생지별 수집 결과

참가자 전체의 출생지별 현황을 살펴보면, 서울이 가장 많으며, 다음으로 부산, 대구, 전남 등의 순으로 많이 수집되었다.

<표 27> 출생지별 수집 결과(기준: 화자 전체, 단위: 명)

출생지	화자 수	비율
강원	206	5.2
경기	247	6.2
경남	223	5.6
경북	221	5.5
광주	204	5.1
대구	289	7.2
대전	241	6.0
부산	472	11.8
서울	743	18.6
울산	88	2.2
인천	82	2.1
전남	281	7.0
전북	217	5.4
제주	102	2.6
충남	212	5.3
충북	172	4.3
합계	4,000명	100.0%

2.7. 주 성장지별 수집 결과

참가자 전체의 주 성장지별 현황을 살펴보면, 서울이 가장 많으며, 다음으로 부산, 경기의 순으로 많이 수집되었다.

<표 28> 주 성장지별 수집 결과(기준: 화자 전체, 단위: 명)

주 성장지	화자 수	비율
강원	175	4.4
경기	394	9.9
경남	210	5.3
경북	213	5.3
광주	201	5.0
대구	303	7.6
대전	268	6.7
부산	439	11.0
서울	642	16.1
울산	96	2.4
인천	157	3.9
전남	240	6.0
전북	199	5.0
제주	108	2.7
충남	203	5.1
충북	152	3.8
합계	4,000명	100.0%

2.8. 현 거주지별 수집 결과

참가자 전체의 현 거주지별 현황을 살펴보면, 서울이 가장 많으며, 다음으로 부산, 광주, 대전의 순으로 많이 수집되었다.

<표 29> 현 거주지별 수집 결과(기준: 화자 전체, 단위: 명)

현 거주지	화자 수	비율
강원	195	4.9
경기	342	8.6
경남	42	1.1
경북	34	0.9
광주	533	13.3
대구	437	10.9
대전	485	12.1
부산	664	16.6
서울	876	21.9
세종	14	0.4
울산	14	0.4
인천	57	1.4
전남	14	0.4
전북	19	0.5
제주	102	2.6
충남	91	2.3
충북	81	2.0
합계	4,000명	100.0%

3. 정책 제언

본 사업은 정제 기준 1,000시간의 일상 대화를 녹음하고 전사하고 정제하는 대규모 사업이다. 사업 진행 중 발생한 주요 문제점 및 개선 사항을 살펴보면 다음과 같다.

첫 번째로 각 주제별로 대주제와 세부 주제를 제시하고 하나의 세부 주제에 대해 이야기를 마친 다음 두 번째 세부 주제로 넘어가도록 할 필요가 있다. 지금의 주제는 주제별로 너무 광범위한 대화를 나눌 수 있다. 예를 들어 ‘정치’라는 주제는 정치인, 국내 정치, 세계 정치, 현 정부, 과거 정부, 현재 정치적 이슈 등 정치와 관련된 많은 대화를 순서 없이 이야기할 수 있다. 너무 광범위한 주제로 대화를 나누는 자료보다는 세부적인 주제에 대하여 대화를 나누는 자료를 수집하는 것이 인공 지능 등에서의 활용도가 높을 것이다.

두 번째로 본 사업에서는 화자의 연령층을 20대~60대 이상으로 구성하여 수집하였으나, 사실상 60대 이상은 대부분 60대로 구성되었다. 70대 이상의 경우 약속한 날짜에 지정된 녹음 장소에 방문하여 녹음하는 것이 힘들고(특히, 도 지역 거주자), 방문을 하더라도 자유로운 일상 대화를 나누기가 사실상 힘들기 때문에 수집이 잘 이루어지지 않았다. 그러나 노령 인구가 증가하는 만큼 70대 이상 화자의 대화를 수집하고 그 특성을 파악하는 것은 매우 중요한 일이다. 따라서 향후 진행 시 그들이 쉽게 참여하고 자연스럽게 대화할 수 있는 환경을 조성해 노령 참가자의 모집을 늘릴 필요가 있다.

향후 진행 시 이러한 문제점을 보완한다면 일상 대화 말뭉치 구축 사업이 국어 및 국어 문화 연구, 4차 산업 대비 기반 기술 발전에 보다 실효성 있는 사업이 될 것이라고 사료된다.

<Abstract>

Construction of the Korean daily conversation corpus

With the development of foundation technology in anticipation of the fourth industrial revolution and the emergence of the artificial intelligence technology industry, the demand for large-scale, high-quality Korean language resources continues to rise further by the day. Thus, the purpose of this project is to prepare basic data for the expansion of the corpus of the Korean language available for public use to enhance the utilization and value of Korean language resources. To this end, the main objective of this project is to establish a corpus of daily conversation to support its free use by the private sector, such as the development of a dialogue system.

Project details: The project to build a corpus of daily conversations is primarily comprised of recording and refining daily conversations freely shared by two people on a particular topic, building a raw corpus by transcribing such speech data, and establishing meta information regarding the target materials. The established speech data amount to 1,000 hours of work.

Speech recording: The project recorded casual daily conversations shared by 2,000 pairs of participants (a total of 4,000 participants) on certain topics for 30 minutes. With regard to the topics for the daily conversations, the following 16 different subjects were selected: military service, games, holidays, automobiles, movies, health/diet, broadcasting/entertainment, sports/leisure, food, nature/recreation forest, country/region, dating/marriage, economy/managing finances, comics, politics, and culture. The speakers were selected in light of their gender, age and the regional characteristics of the population. The speakers were reminded of the objective of this project once again, entered into a contract to allow the use of their conversations, were notified of the necessary precautions to take during the recording sessions, and then wore the recording equipment to carry out the recording. The recording was conducted mainly in a separate space with minimal background noise, such as a discussion room and a conference room. Two speakers each wore a headset microphone and each session was conducted by recording a sample first before moving on to recording the main conversation. Speech files were saved as the

samples at a rate of 16kHz and 16bit PCM with linear quantization.

Transcription of speech data: At first, the speech files were transcribed using an automatic transcription system, which were then supplemented by a professional stenographer. The professional stenographer modified the displayed speakers, applied transcription units and symbols, reflected the symbols in sentences, transcribed what is heard, corrected split sentences, localized the numbers, and Roman alphabets, and corrected word-spacing, spelling, overlapping speeches, and the overlapping display of the speakers.

Establishment of a raw corpus and meta information: The headers with meta information such as the recording date, conversation topic, relationship between the speakers, speaker's information (gender, age group, occupation, place of birth, place where he/she mainly grew up, etc.) were created and attached to the transcribed files, and such information was converted into the XLM format by marking it up for each speaker to enhance the utility of information.

Refinement of speech data: First, the Praat script was run to automatically segment each speech file, and subsequently, the speech refiner listened to the speech to modify the segmentation boundaries according to the transcription unit and processed the speech to no longer be personally identifiable.

Construction of a database and the development and operation of a management system: A database to manage speech files, transcribed files, files containing contracts to allow the speech use, preprocessed files, meta information, etc. was built, and a web management system that automatically generates the file names under the guideline was then developed and operated. The database and management system were managed in accordance with the regulations on the information network security management, and the data was reliably managed through countermeasures for failures and periodic backup.

Expected project outcome: The project to build a corpus of daily conversations is expected to lay the foundation for a corpus of daily conversations to be used for the base materials for the development of artificial technologies such as speech recognition. Through the collection of conversations made between the speakers with a varying gender, age and region on particular topics, we seek to enhance the

understanding of various aspects of the languages realized in daily conversations and, by extension, use them as the basic data for the study of the Korean language and culture, and the establishment of the Korean language policy.

Keywords: Daily conversation, conversation by subject, raw corpus, casual conversation, speech refinement

Project Director: Na Yun-jung(Metrix Corporation)

사업 책임자	나운정((주)메트릭스코퍼레이션 리서치부문 부사장)
사업 참여자	조일상((주)메트릭스코퍼레이션 리서치부문 대표)
	이영미((주)메트릭스코퍼레이션 리서치부문 부장)
	박래희((주)메트릭스코퍼레이션 리서치부문 팀장)
	안재준((주)메트릭스코퍼레이션 리서치부문 차장)
	안수정((주)메트릭스코퍼레이션 리서치부문 과장)
	황민수((주)메트릭스코퍼레이션 리서치부문 대리)
	이유림((주)메트릭스코퍼레이션 리서치부문 대리)
	한금만((주)메트릭스코퍼레이션 리서치부문 주임)
	박두진((주)메트릭스코퍼레이션 리서치부문 이사)
	이혜진((주)메트릭스코퍼레이션 리서치부문 차장)
	하지영((주)메트릭스코퍼레이션 리서치부문 차장)
	김주연((주)메트릭스코퍼레이션 리서치부문 차장)
	윤지은((주)메트릭스코퍼레이션 리서치부문 과장)
	이은미((주)메트릭스코퍼레이션 리서치부문 주임)
	채윤철((주)메트릭스코퍼레이션 리서치부문 팀장)
	이형주((주)메트릭스코퍼레이션 리서치부문 과장)
	서라별((주)메트릭스코퍼레이션 리서치부문 부장)
	이경화((주)메트릭스코퍼레이션 리서치부문 대리)
	박혜수((주)메트릭스코퍼레이션 리서치부문 사원)
	이영경((주)메트릭스코퍼레이션 리서치부문 주임)
	신현주((주)메트릭스코퍼레이션 리서치부문 팀장)
	김도현((주)메트릭스코퍼레이션 리서치부문 대리)
	김태경(한양대학교 창의융합교육원 교수)
	임유종(한양대학교 미래문화연구소 연구부교수)
	백경미(한양대학교 국제문화대학 한국어언어문학과 강사)
담당 연구원	이승재(국립국어원 언어정보과장)
	홍혜진(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9757

인쇄일: 2019년 11월 8일

발행일: 2019년 11월 8일

인 쇄: (주)타라그래픽스

※ 이 책은 국립국어원의 용역비로 수행한 ‘일상 대화 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.

