인공지능 데이터 구축·활용 가이드라인
- 한국어 방언 AI 데이터 -

인공지능 데이터 구축	사업 총괄	Saltlux
	데이터 설계	경북대학교 산학협력단
	원천데이터 수집 및 정제	Saltlux, SORIZAVA, ESTsoft, DQ, EXOTECH, INFINIQ, 도담에스엘, COREINFORMATION, DGRAIB, TimeSoft, 나라지식정보, 사람과 숲, RD
	데이터 가공	
	데이터 검수	
	크라우드 소싱	
	저작도구 개발	Saltlux
	AI모델 개발	Saltlux
	응용 서비스 개발	Saltlux
	품질 관리	b2en
가이드라인 작성	솔트룩스	박재원 이사
가이드라인 버전	버전 1.1.5 작성일 2021. 01. 18	

목 차

1. 데이터 명세 정보

1.1 데이터 정보 요약

AI 학습용 데이터 구축량 : 한국어 방언(강원도, 경상도, 전라도, 제주도, 충청도)의 총 250만 문장의 학습데이터 구축

데이터 이름	한국어 방언 AI 데이터	
활용 분야	- 연구분야: 음성 발화, 음성 인식, NLU, NLG를 포함한 NLP 전분야 - 산업분야: 온라인 심리상담, 고객상담 챗봇, 스마트 스피커 등	
데이터 요약	- 한국어 방언 데이터 구성(안) : 한국어 방언(강원도, 경상도, 전라도, 제주도, 충청도)의 총 250만 문장의 학습데이터	
데이터 출처	- 강원도, 경상도, 전라도, 제주도, 충청도 연령별 화자	
데이터 이력	배포버전	버전 1.5
	개정이력	신규
	작성자/ 배포자	박재원 / ㈜솔트룩스

1.2 데이터 포맷

수집된 원천 데이터를 체계적으로 정제하여, 불필요한 데이터를 제거하고 데이터 셋을 구축하는데 필요한 형태로 저장하며 JSON 형식으로 구성

- 방언 전사포맷 '필수값 여부' 항목이 'Y'인 경우 학습데이터(JSON)에 필수 값(Value)이 반드시 존재해야 함
- 학습데이터 생성 (json 파일)시 방언 전사포맷과 동일한 '수준1,2,3', 동일한 항목 '순서'를 고려하여 생성이 필요함

단계	수준 1	수준 2	수준 3	타입	필수값 여부	설명
획득, 정제	id			string	Y	AI 학습데이터 파일 아이디 (수동부여)
	metadata			object		AI 학습데이터 파일 메타 정보
		title		string	Y	AI 학습데이터 파일 제목
		creator		string	Y	구축자: 솔트룩스
		distributor		string		배포자: 솔트룩스
		year		string	Y	구축년도: 2020
		category		string		분류: 구어 > 사적 대화 > 일상 대화
		annotation_level		array(string)		분석 층위: 원시
		sampling		string		샘플링 방식: 본문 전체
		author		string		저작권자: 개인 발화자
		publisher		string		발행자: 개인 발화 녹음
		date		string	Y	녹음일자: YYYYMMDD
		topic		string	Y	대화 주제

	speaker			array(object)		화자 정보
		id		string	Y	화자 아이디(회사ID 0001)
		name		string		이름
		age		string	Y	연령
		occupation		string	Y	직업
		sex		string	Y	성별
		birthplace		string	Y	출생지
		principal_residence		string	Y	주 성장지
		current_residence		string	Y	현 거주지
		education		string	Y	학력
	setting			object		환경 정보
		relation		string	Y	화자 간 관계
가공	utterance			array(object)		발화 정보
		id		string	Y	발화 아이디
		form		string		방언 전사
		standard_form		string		표준어 대응쌍 부착
		dialect_form		string		방언 문장
		speaker_id		string		화자 아이디
		start		num	Y	발화 시작 시간(소수점 2자리까지)
		end		num	Y	발화 종료 시간(소수점 2자리까지)
		note		string		전사자 기타 메모
		eojeolList		array(object)		방언 어절 단위 정보
			id	num	Y	
			eojeol	string		방언 어절
			standard	string		표준어 어절
			isDialect	boolean		방언 어절 여부

1.3 어노테이션 포맷

```
{
    "id":    "SDRW2000000001",
    "metadata": {
        "title": " 경상도 방언 AI 학습데이터 SDRW2000000001",
        "creator": "솔트룩스",
```

```
        "distributor": "솔트룩스",
        "year": "2020",
        "category": "경상도 방언 > 사적 대화 > 일상 대화",
        "annotation_level": [
        "원시"
        ],
        "sampling": "본문 전체"
    },
    "document": [
        {
            "id":    "SDRW2000000001.1",
            "metadata": {
                "title":    "2인 일상 대화",
                "author":    "개인 발화자",
                "publisher":    "개인 발화 녹음",
                "date":    "20190711",
                "topic":    "자동차",
                "speaker":    [
                    {
                        "id":    "SD1900011",
                        "age":    "30대",
                        "occupation":    "사무 종사자",
                        "sex":    "남성",
                        "birthplace":    "대구",
                        "pricipal_residence":    "대구",
                        "current_residence":    "경북",
                        "education":    "대졸"
                    },
                    {
                        "id":    "SD1900012",
                        "age":    "30대",
                        "occupation":    "사무 종사자",
                        "sex":    "남성",
                        "birthplace":    "대구",
                        "pricipal_residence":    "대구",
                        "current_residence":    "대구",
                        "education":    "대졸"
                    }
                ],
                "setting":    {
                "relation":    "동료"
                }
            },
            "utterance": [
                {
                    "id": "SDRW2000000001.1.1.1",
                    "form":    "안녕하세요.",
                    "original_form":    "안녕하세요.",
                    "speaker_id":    "SD1900011",
                    "start":    30.56600,
                    "end":    32.48262,
                    "note":    ""
                },
                {
                    "id":    "SDRW2000000001.1.1.2",
                    "form":    "아~ xx님 오랜만입니다.",
                    "original_form":    "아~ ((xx님)) 오랜만입니다.",
                    "speaker_id":    "SD1900012",
```

```
                "start":    33.12500,
                "end":    34.1543323,
                "note":    ""
            },
```

1.4 데이터 구성

산출물 업로드시 폴더명 구성을 설명과 예시를 통해 참여기관과 공유를 하였고, 동일한 폴더구조로 산출물 및 파일을 관리

말뭉치 유형 구분	방언권(지역) 분류	구축 회사명 (2자리)	구축년도	일련번호 (6자리)
D: 방언 말뭉치	G: 강원도 K: 경상도 J: 전라도 Z: 제주도 C: 충청도	예) DQ: ㈜디큐 …	20	000001

1.5 데이터 통계

1.5.1 데이터 구축 규모

과제명	주요 내용	데이터 구축량	데이터 형식
한국어 방언 발화 데이터 (강원도)	방언(강원도)을 사용하는 일상 대화를 인식하여 음성을 문자로 실시간으로 변환하고, 텍스트를 방언 음성으로 합성할 수 있는 기술 개발을 위한 방언 발화 데이터셋 구축	• 조용한 환경에서 ***2,000명*** 이상의 화자가 발화한 성별, 연령별 적정 길이의 ***3,000시간*** 이상의 음성 데이터셋 • 원본 표준어 텍스트 및 방언 특성을 고려하여 그대로 전사한 텍스트 ***50만건***	• 원본형태 : • 화자가 구분된 담화 텍스트 말뭉치 • 학습용 데이터 형태 : • 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋
한국어 방언 발화 데이터 (경상도)	방언(경상도)을 사용하는 일상 대화를 인식하여 음성을 문자로 실시간으로 변환하고, 텍스트를 방언 음성으로 합성할 수 있는 기술 개발을 위한 방언 발화 데이터셋 구축	• 조용한 환경에서 ***2,000명*** 이상의 화자가 발화한 성별, 연령별 적정 길이의 ***3,000시간*** 이상의 음성 데이터셋 • 원본 표준어 텍스트 및 방언 특성을 고려하여 그대로 전사한 텍스트 ***50만건***	• 원본형태 : • 화자가 구분된 담화 텍스트 말뭉치 • 학습용 데이터 형태 : • 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋
한국어 방언 발화 데이터 (전라도)	방언(전라도)을 사용하는 일상 대화를 인식하여 음성을 문자로 실시간으로 변환하고, 텍스트를 방언 음성으로 합성할 수 있는 기술 개발을 위한 방언 발화 데이터셋 구축	• 조용한 환경에서 ***2,000명*** 이상의 화자가 발화한 성별, 연령별 적정 길이의 ***3,000시간*** 이상의 음성 데이터셋 • 원본 표준어 텍스트 및 방언 특성을 고려하여 그대로 전사한 텍스트 ***50만건***	• 원본형태 : • 화자가 구분된 담화 텍스트 말뭉치 • 학습용 데이터 형태 : • 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋

한국어 방언 발화 데이터 (제주도)	방언(제주도)을 사용하는 일상 대화를 인식하여 음성을 문자로 실시간으로 변환하고, 텍스트를 방언 음성으로 합성할 수 있는 기술 개발을 위한 방언 발화 데이터셋 구축	• 조용한 환경에서 ***2,000명*** 이상의 화자가 발화한 성별, 연령별 적정 길이의 ***3,000시간*** 이상의 음성 데이터셋 • 원본 표준어 텍스트 및 방언 특성을 고려하여 그대로 전사한 텍스트 ***50만건***	• 원본형태 : • 화자가 구분된 담화 텍스트 말뭉치 • 학습용 데이터 형태 : • 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋
한국어 방언 발화 데이터 (충청도)	방언(충청도)을 사용하는 일상 대화를 인식하여 음성을 문자로 실시간으로 변환하고, 텍스트를 방언 음성으로 합성할 수 있는 기술 개발을 위한 방언 발화 데이터셋 구축	• 조용한 환경에서 ***2,000명*** 이상의 화자가 발화한 성별, 연령별 적정 길이의 ***3,000시간*** 이상의 음성 데이터셋 • 원본 표준어 텍스트 및 방언 특성을 고려하여 그대로 전사한 텍스트 ***50만건***	• 원본형태 : • 화자가 구분된 담화 텍스트 말뭉치 • 학습용 데이터 형태 : • 방언 발화된 음성 데이터가 맵핑된 텍스트와 음성 데이터셋

데이터 종류	포함 내용	제공 방식
음성 데이터셋	총 15,000 시간 정제된 음성데이터 - 도별 3,000 시간	wav 포맷 파일
전사한 텍스트 데이터셋	총 250건의 원본 표준어 텍스트 및 방언 특성을 고려한 이중전사 텍스트 - 도별 50만건	JSON 포맷 파일

2. 데이터 구축 가이드

2.1 데이터 구축 개요

-AI 학습용 데이터 구축량 : 한국어 방언(강원도, 경상도, 전라도, 제주도, 충청도)의 총 250만 문장의 99.9% 고품질 학습데이터 구축 및 AI 응용서비스 개발

-데이터 구축 프로세스는 한국어 방언 인공지능 학습용 데이터 구축 → 데이터 품질 관리 및 검증방안 → AI 데이터 활용 응용서비스 개발로 진행

(고품질 학습데이터) 한국어 방언 (5개도) 구축 총 3,000시간의 음성 데이터 수집 및 50만 문장전사	• 원천 데이터 수집 시 정제 (사전녹음, 녹음수행, 완료파일청취 등) • 학습데이터 설계 및 대화 주제 선정 (표준어와 방언 매핑) • 음성 녹음 화자 구성 및 절차 수립 (각 세부과제 별 인구통계기준) • 음성 수집 비율과 수집 도구 (대면/비대면, 녹음도구/화상 녹취)
(학습데이터 품질관리) 4단계 품질 공정 및 도구	• 고품질 학습용 데이터 확보를 위한 검수 방안 제시 • 참여기업 및 세부책임, 수행기관, 품질기관 검수 → 99.9 %품질 • 음성전사 저작도구 활용한 투입작업 인력 품질관리 • 음성 녹음 지원자 모집 및 수집 평가 (크라우드소싱, 데이터품질평가)
(AI 응용서비스) 서비스 적용사례 4가지 개발	• 각 도별 5개+ 통합 1개 (음성인식,합성,기계번역,일상대화모델) • 4가지 서비스 적용사례 제시 • 솔트룩스 AI클라우드 3년간 무상제공 → 학습된 AI모델 활용

2.2 문제정의

2.2.1 임무 정의

○ 데이터 경제로의 패러다임 변화

- 4차 산업혁명 시대로 급속 진입하면서 제조업, 서비스업 중심의 한국경제는 도태의 위기에 직면하게 됨. 특히 코로나19로 인한 극심한 경기 침체와 함께 데이터 경제로의 패러다임 전환이라는 이중 과제를 해결해야 하는 시점
- 데이터를 기반으로 한 인공지능의 시대가 도래함 에 따라 인공지능 시대의 석유라고 일컫는 기초 데이터의 국적 차원의 확보 및 제공이 글로벌적인 경쟁력 확보의 필수 요소이며, 데이터 확보가 이루어져야 비로소 디지털 시대로의 전환기를 맡은 수많은 기업과 스타트업 그리고 국가 공공 행정 서비스의 미래 선도형 경제 실현이 가능한 시점에 도달함
- 이미 다른 선진국에서는 미래 경쟁력을 좌우하는 데이터의 중요성을 인식, 데이터 산업 활성화를 위해 국가 차원의 선제적인 전략 수립과 정책 투자 확대 등 데이터 경쟁에 돌입함
- 구글, 아마존 등 글로벌 IT 대기업은 빅데이터의 축적과 함께 다양한 AI 혁신기술을 공개하며 수많은 형태의 새로운 산업과 서비스 영역을 개척하며 선보이고 있어 벌써부터"디지털 독과점"이란 비판을 받고 있는 수준으로 앞서 나가고 있음

○ 한국어 방언 데이터가 필요한 이유

- 모든 디지털 산업의 기초가 될 데이터는 80% 이상이 텍스트, 음성, 영상 등으로 되어 있음. 이중 음성, 텍스트 데이터는 인공지능 A.I를 학습시키기 위한 기술인 NLP(Natural Language Processing)의 핵심적인 부분을 차지함
- 그러나 한국어 말뭉치를 비롯한 원천 데이터 구축은 선진국과 글로벌기업 대비 걸음마 수준이며, 관련 업체 수요에도 부응하지 못하는 열악한 수준임. 따라서 디지털 뉴딜에서의 한국어와 한국어 방언 데이터의 수집 및 구축 사업은 인공지능 학습용 데이터 구축 사업의

가장 근간이 되는 중요한 부분이라 할 수 있음

2.2.2 데이터 구축 유의사항

○ 저작권 이용 허락 계약

- 관리자는 화자에게 사업 설명 및 녹음의 목적을 설명하고 저작권 이용 허락 동의서를 체결

○ 동의서 체결 프로세스화

- 동의서 체결이 되지 않을 경우는 화자에 참여 시키지 않음

한국어 방언 AI 데이터 구축 및 활용 저작권 이용허락 동의서

한국정보화진흥원은 대규모 언어 자원을 구축하여, 이를 국어 연구 및 자연언어처리 기술 개발 등을 위해 사용하고자 합니다. 이에 귀하의 저작물의 활용에 대한 승인을 구하고자 합니다. 귀하가 만드신 저작물은 개별 단어 및 문장, 텍스트에 대한 정보 추출과 분석에 쓰이며, 귀하의 저작물을 이용하여 구축한 방언 데이터가 귀하의 저작·출판에 관한 어떠한 권리에도 손상을 입히지 않을 것임을 약속드립니다. 모쪼록 21세기 4차 산업 혁명 시대에 귀하의 저작물을 우리말 정보 처리 발전의 기초가 되는 국가적 언어 자원의 구축에 유용하게 활용할 수 있도록 아래와 같이 협조하여 주시면 감사하겠습니다.

-

2.3 획득·정제

2.3.1 원시데이터 선정

○ 환경 구성 및 모집

- 녹음 환경 및 화자 모집

녹음 품질 및 데이터 표본의 다양화를 위해 녹음지역을 전국에 있는 지사 사무실을 활용하여 진행하며, 두 명의 화자가 서로 자유롭고 편안한 상황에서 대화할 수 있도록 1 조 2 인 단위로 화자를 모집하여 자료 수집을 진행

• 녹음 환경 구성

- 두 명의 화자가 편안하게 이야기 할 수 있는 사무실 환경 마련
- 녹음실은 외부와 차단된 상태로 대화에 참여한 두 명만이 대화할 수 있도록 구성
- 화자는 각각 헤드셋 마이크를 착용하고 발화
- 상대방의 목소리가 들어가지 않도록 적정거리 유지

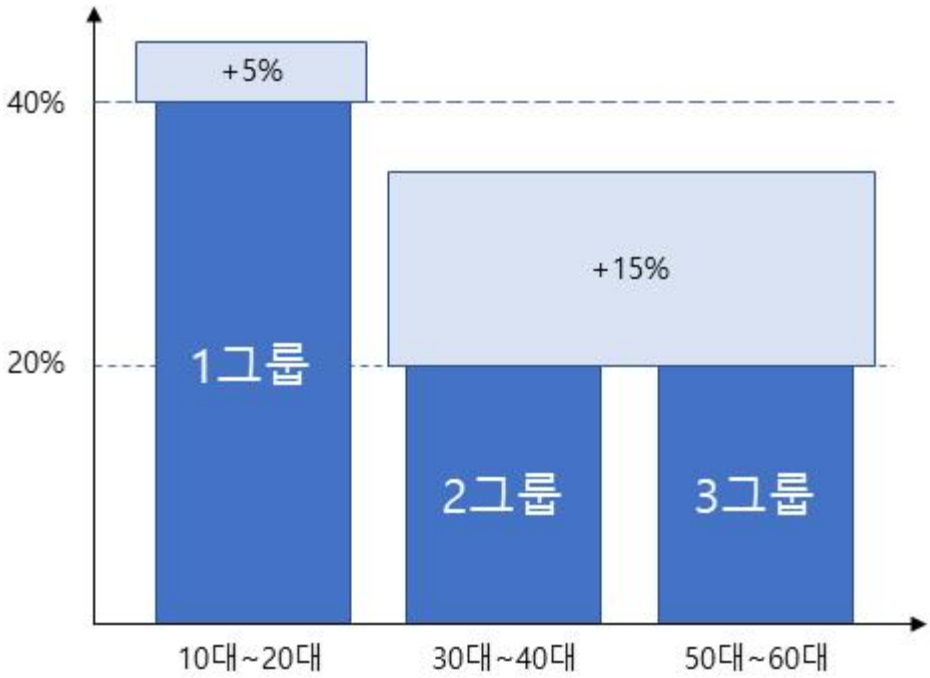
• 녹음 화자 모집

- 특정 성별, 연령, 지역 등이 편중되지 않도록 사전 협의하여 진행
- 한 화자당 최대 녹음시간은 가능한 약 30분으로 하고 동일 화자가 중복 참여하지 않도록 제한하나, 동일 주제가 아닐 경우에는 허용
- 녹음 화자 모집 시 최초 2인 1조로 신청자를 최우선으로 하며, 1인이 개별 신청했을 경우 비슷한 연령대 및 관심사를 구분하여 조 편성
- 주제에 따라 1인 녹음, 3인 이상 녹음을 허용

화자별 분류방법	세부 내용
연령별	1그룹(10대~20대), 2그룹(30~40대), 3그룹(50대 이상)
지역별	강원도, 경상도, 전라도, 제주도, 충청도

• 녹음 화자 구성

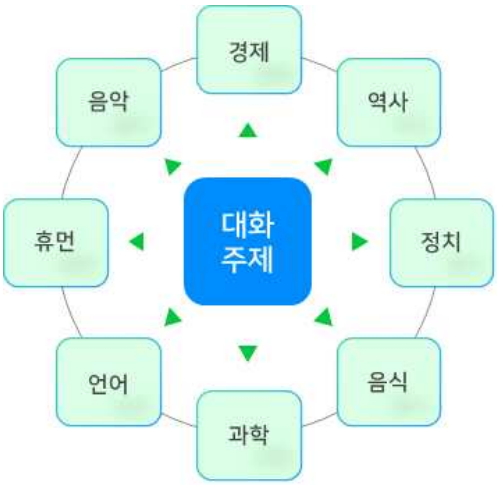
- 연령별 그룹 기준으로 분류
- 총 3개의 그룹으로 구성되며 1그룹은 10대~20대, 2그룹은 30~40대, 3그룹은 50대 이상으로 정하였으며 실제 녹음이 불가능하다고 판단되는 0~9세 / 70세 이상의 대상자는 제외이나 녹음이 가능할 경우 3그룹에 포함하여 진행
- 최종 인력 배분은 NIA(한국정보화진흥원) 협의 후 진행
- 화자 비율은 1그룹 40%, 2그룹 20%, 3그룹을 20%를 최소 비율로 하고 최대 비율은 1그룹은 45%, 2그룹과 3그룹은 합하여 15%를 구축함
- 단, 2그룹과 3그룹의 편차는 5% 이상을 넘지 않게 함



- 대화 주제 분류

  녹음 화자가 편중된 내용을 발언하지 않도록 체계적인 주제와 자료를 제시하여 화자가 원활하고 편안한 환경에서 대화를 나눌 수 있도록 하며, 녹음 상황임을 인지하지 않고 자연스럽게 참여할 수 있도록 환경 조성

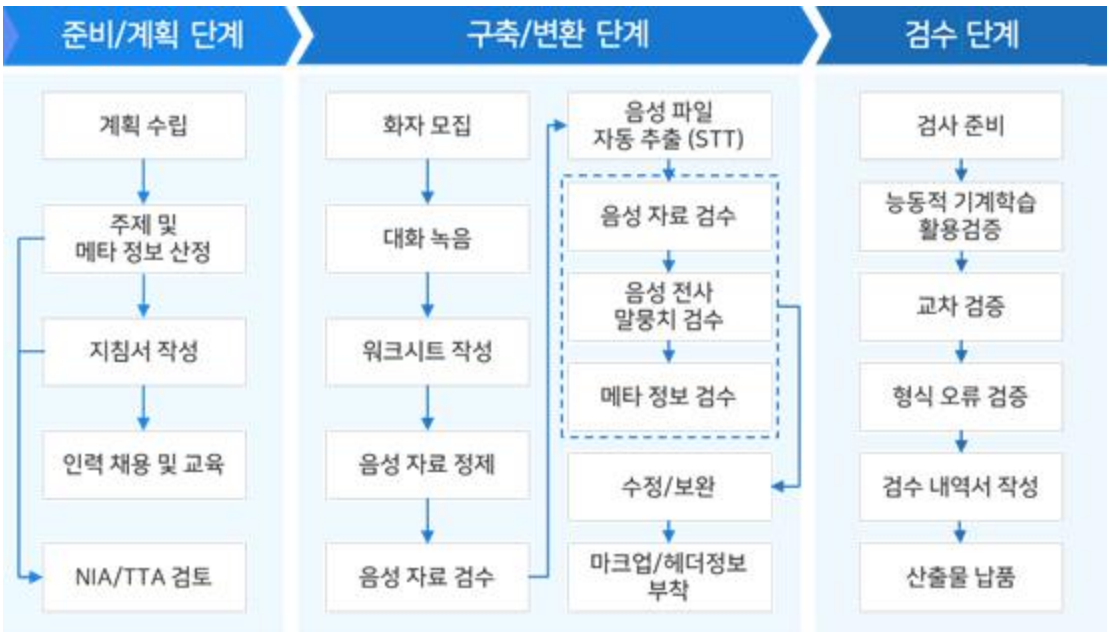
  • 대화 주제 목록 (예시)



  • 대화 주제 제시 자료 - 사진, 그림, 멀티미디어 등 (예시)

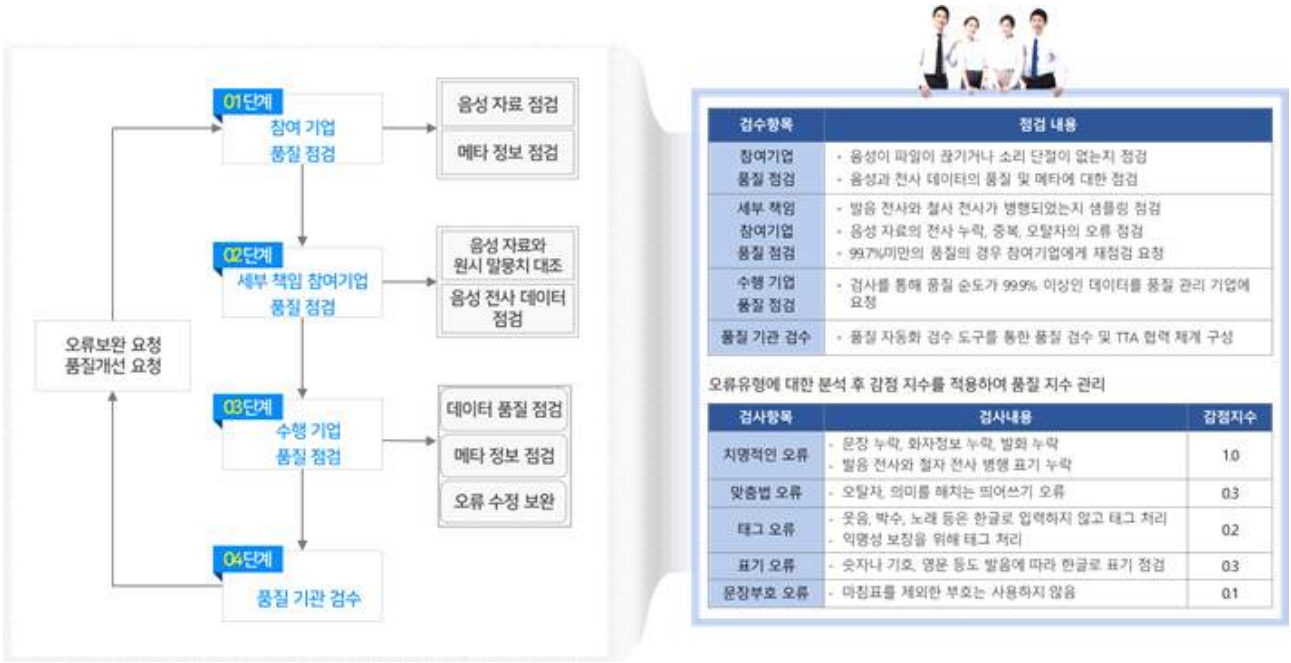
2.3.2 획득·정제 절차

한국정보화진흥원의 데이터베이스 구축방법론(Ver.4)을 적용하여 음성 녹음, 이중 전사, 원시 말뭉치 구축에 대한 대상 자료별 공정 태스크와 주요 활동 절차를 표준화하여 효율적인 학습용 데이터셋 구축 체계를 확보하고 한국어 방언 AI데이터 구축에 적합하도록 자료의 특성을 고려하여 준비/계획 단계, 구축/변환단계, 검수단계의 3단계 공정을 설계한다.



2.3.5 획득·정제 도구

음성 및 이중전사 데이터의 고품질로 생성하기 위하여, 단가는 낮지만 품질 관리가 어려운 크라우드소싱 방식의 데이터생성 작업을 보완하기 위한 검수 프로세스의 정립은 데이터셋 구축에 매우 중요한 의미를 갖는다. 이 데이터셋에서는 4단계 검수 체계를 구축하였으며, 가장 하위 레벨에는 크라우드 워커들이 작업한 결과물을 전사규칙 가이드라인에서 제시한 형식에 맞는지 체크하는 참여기업에 1차 검수자가 있으며, 이들이 검수한 결과물에 대해서 재검수하는 도별 책임기관 2차 검수자가 있습니다. 이렇게 만들어진 데이터셋을 전체적으로 들여다보며 데이터셋의 밸런스나 가이드라인의 적절성 등 품질 확보를 위한 총괄기업인 ㈜솔트룩스에서 5년 이상 학습데이터 구축 및 품질검수를 수행한 지식큐레이션 팀이 3차 검수를 수행하여 음성 및 전사 텍스트에 대한 품질을 확보하며, 최종 4차 검수는 참여기관인 ㈜비투앤에서 정확도 및 유효성 품질 검증을 통하여 고품질 학습 데이터셋의 품질을 담보할 수 있었습니다.



2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차

준비/계획단계

○ 학습데이터 구축 대상 및 범위 계획 수립

- 말뭉치 구축 대상 및 범위 선정
  - 두 사람이 특정주제 (10개 내외)로 자유롭게 대화
  - 대화 내용을 녹음하고 정제 (2,000명 이상의 화자가 발화한 총 3,000시간 이상 데이터셋, 대화당 15분 이하)
  - 해당 녹음자료에 대한 저작권 이용 허락 계약서 체결
  - 녹음된 내용 이중 전사 (발음전사 / 철자전사)
  - 구축된 전사자료에 대한 메타정보 구축 (녹음날짜, 화자명 및 정보, 대화주제 등)

세부 메타정보 항목	날짜, 대화명, 주제, 화자명, 화자정보 , 화자간 관계 등 ※ 단, 화자 개인정보는 비식별화
	예) 1. 녹음날짜 - 2020년 3월 26일 2. 참 여 자 - ① 김철수 (33세, 남, 직장인, 부산출생, 현 서울거주) ② 최영희 (31세, 여, 직장인, 서울출생, 현 서울거주) 3. 대화자관계 – 회사 거래처 담당자 4. 대화주제 – 주말 일상

- 학습데이터 주제 선정
  - 컨소시엄의 인공지능 전문가 그룹을 통해 방언 학습데이터 구축을 위한 상세 분석 후 주제 및 메타 정보를 검토
  - NIA(한국정보화진흥원) 검토 및 협의 후 최종 주제와 메타를 선정

구축/변환단계

○ 학습데이터 구축 및 변환 작업

- 녹음 환경 및 화자 모집

  녹음 품질 및 데이터 표본의 다양화를 위해 녹음지역을 전국에 있는 지사 사무실을 활용하여 진행하며, 두 명의 화자가 서로 자유롭고 편안한 상황에서 대화할 수 있도록 1 조 2 인 단위로 화자를 모집하여 자료 수집을 진행

- 음성 녹음 및 정제

  음성 녹음된 자료가 기준에 벗어나지 않도록 녹음 시 국립국어원과 협의된 기준에 부합하도록 작업하며, 녹음된 음성 자료를 전사 단위로 편집하고 개인 정보 및 불필요한 내용을 정제하는 과정도 효율적인 체계를 구축하여 업무상 인적, 물적 손실을 최소화 합니다.

  • 대화 전체 음성 파일(원본) / 억양구 단위로 분할된 파일(정제본)을 각각 제출
  - 단위 등 정제 기준은 주관기관과 사전 협의
  - 폴더 구조 및 파일명 등은 주관기관 제시자료를 따르되 상세 내용은 사전 협의

• 대화 주제와 무관한 내용(인사말 등)은 제외하여 정제

- 전사 및 학습데이터 구축

화자를 통해 녹음 및 정제가 된 음성 자료를 대상으로 작업 지침에 따라 전사 작업을 수행하고 학습 데이터의 정확성 확보를 위해 교차 검수를 진행하여 100% 정확도을 기함

• 전사 작업자 대상 작업 지침 교육
  - 전사 작업장 내 마련되어있는 교육장을 활용하여 대상자 교육 실시
  - 전사 관련 기본교육은 상시 실시하고 있으며, 본 사업에서 국립국어원이 제시한 규정을 재확인하여 해당 규정에 맞는 교육 실시
• 전사 지침
  - 발화된 그대로 전사하는 발음전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 함
  - 그 외 화자표시, 전사단위, 발화겹침, 불완전한 발화, 띄어쓰기 등 세부 내용은 국립국어원이 제시한 전사지침을 따름
• 학습데이터 지침
  - 전사 결과물에 대해 헤더정보 부착 등 표지 부착 작업 수행
  - 파일명 부여방식, 표지 부착방식, 형식 등은 협의하여 진행

절차	내 용	담당	산출물
음성 전사 학습데이터 교정	• 작업지침 교육 및 전사 프로그램 교육 • 음성 텍스트 자동 변환(STT) 후 전사 학습데이터 교정 • 발음 전사와 철자 전사가 병행 전사 • 영문과 숫자 한글로 표기되었는지 점검 • UTF-8로 저장	구축 팀	워크시트 작업지침서
메타 정보 점검	• 메타 정보가 일치하는지 확인 • 메타 정보의 오탈자 및 중복 확인 • 불일치 데이터 재작업	구축 팀	워크시트 작업지침서
전사 학습데이터 마크업	• 점검이 완료된 전사 학습데이터 마크업 • 표준 지침에 따라 변환 • 자동 변환된 마크업 점검	구축 팀	작업지침서 점검내역서
인계	• 점검이 완료된 음성 자료와 음성 전사 학습데이터/메타 정보는 품질 점검팀으로 인계	구축 팀	음성 자료 음성 전사 학습데이터

2.4.2 어노테이션/라벨링 기준

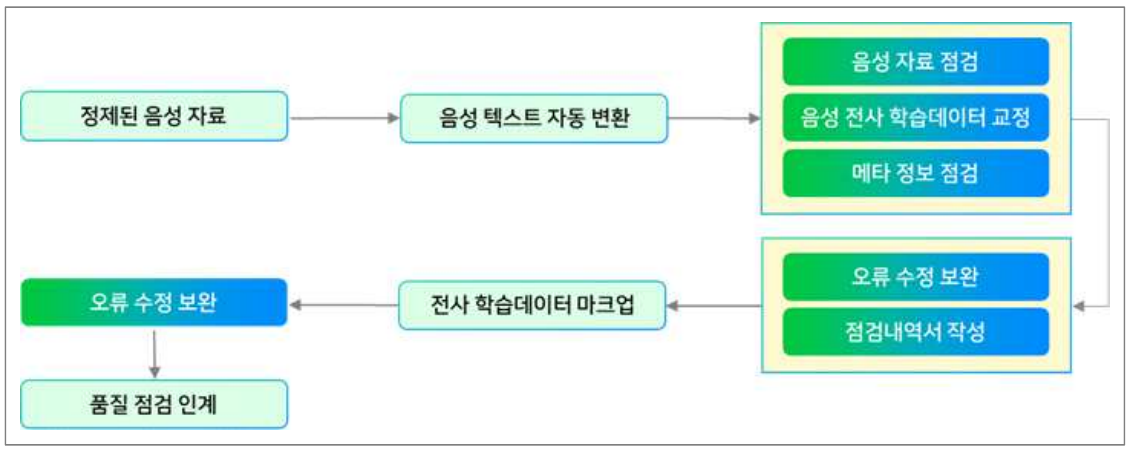
분류	전사규칙
개요	• 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 한다.
화자 표시	• 화자 아이디, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 '?'로 표시한다. • 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명

	하지 않을 경우에는 '?'로 표시한다.
전사 단위	• 기본 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 한다.
숫자/외래어/기호/단위	• 한글 맞춤법, 표준어 규정, 외래어 표기법 등 관련 어문 규정에 따라 한글로 적는다.
발음	• 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 전사를 한다.
발화 겹침	• 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. • 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다.
익명성 보장	• 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인 정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. - n : 사람 이름(단, 정치인, 연예인 등 유명인의 이름은 비식별화하지 않으며, 상호명은 부정적인 경우에만 비식별화) - social-security-num : 주민등록번호 - card-num : 신용카드 번호 - address : 주소(동 이하의 구체적인 주소만 비식별화) - tel-num : 전화 번호
기타 소리	• 기타 소리 중 웃음, 목청, 박수, 노래에 대한 4가지는 태그로만 전사한다. • 기침, 들숨, 날숨, 재채기, 코훌쩍, 하품 등은 전사하지 않는다.
축약형 표기	• 언어 경제성의 원칙에 의해 구어에서는 축약형이 많이 나타나며, 이는 모두 표기에 반영 • 모음 축약형은 ‘를 사용해서 두 음소를 연결해 준다.

2.5 검수

2.5.1 검수 절차

"방언 전사"와 "표준어 대응쌍"은 "(방언 전사 형태)/(표준어 대응쌍 형태)"의 꼴로 제시하고, 방언 전사한 내용이 표준어와 차이가 없을 경우는 방언 전사의 내용을 그대로 유지한다.



2.5.2 검수 기준

기본 원칙

1) 이 작업은 한국어 방언 AI데이터 구축을 위하여 5개의 지역으로 묶어서 수집한 방언을 일차적으로는 지역 방언의 특성을 살려 "방언 전사"하고, 표준어 규정에서 벗어나는 방언에 해당하는 부분에 대해 "표준어 대응쌍"을 제시하는 것이다.

2) "방언 전사"와 "표준어 대응쌍"은 "(방언 전사 형태)/(표준어 대응쌍 형태)"의 꼴로 제시하고, 방

언 전사한 내용이 표준어와 차이가 없을 경우는 방언 전사의 내용을 그대로 유지한다.

지역	보기
강원	이게 (다나?)/(다니?) 나도 이쪽 동네 (출신이래.)/(출신이야.) (이라)/(이렇게)
경상	어제 어디 (갔었노?)/(갔었니?) 미역 (쭐거리)/(줄기) (단디)/(단단히)
전라	혼자 다 (묵어)/(먹어) (분당께.)/(버린다니까.) 아 (실맹키로)/(실처럼) 가는 거 그거? (그랑께)/(그러니까)
제주	아까 (집드레)/(집으로) (가라.)/(가더라.) 너 (하*구정 한)/(하고 싶은) 대로 (하*라.)/(해라.) (아매나)/(아무렇게나)
충청	동네 사람들은 (워떡헌댜?)/(어떡한대?) 가만히 (두덜)/(두질) (못햐.)/(못해.) (그려.)/(그래.)

※ [참고] 제주 방언에 제시된 '하*'에서 '아*'는 아래아(ㆍ)를 나타내는 표기이다. '3.13. 지역별 방언 전사 주의 사항' 참조(17쪽).

3) 방언 전사하기

① 방언 전사: 각 지역에서 모은 사람들의 대화를 지역 언어의 특성이 드러나도록 소리 나는 대로 적는 것.

② 방법 및 유의 사항: 방언과 관련이 없는 표현은 표준어를 적는 방식으로 쓰되, 방언 표현은 방언의 특색을 드러나도록 표기한다. 이때 방언의 표기는 음성 그대로 소리나는 대로 쓰지 않고 방언의 형태가 드러나는 방식으로 쓴다. 예를 들어 "먹었지"의 경상도 방언은 [묵었지]로 쓰며, 소리나는 대로 연음한 [무걷찌]로 쓰지 않는다.

예) 올바른 전사 표기: (묵었지)/(먹었지), (갔었노?)/(갔었니?)
잘못된 전사 표기: (무걷찌)/(먹었지), (가썬노?)/(갔었니?)

지역	보기	
	올바른 전사 표기	잘못된 전사 표기
강원	그때는 (우뚱게)/(어떻게) 할 수가 없었어.	그때는 (우뜨케)/(어떻게) 할 수가 없었어.
경상	어제 어디 (갔었노?)/(갔었니?) 오늘 날씨가 너무 (추버서)/(추워서)	어제 어디 (같었노?)/(갔었니?) 어제 어디 (가썬노?)/(갔었니?) 오늘 날씨가 너무 (춥어서)/(추워서)
전라	니가 (멋이간디)/(뭣이관데) 큰 소리냐?	니가 (머시간디)/(뭣이관데) 큰 소리냐?
제주	하루에 같이 (검질멧주게.)/(김매었지).	하루에 같이 (검질멧쭈게.)/(김매었지). 하루에 같이 (검질멛주게.)/(김매었지). 하루에 같이 (검질멛쭈게.)/(김매었지).
충청	돈은 물어 주면 되지만 속상한 건 (위칙헌댜?)/(어떡한대?)	돈은 물어 주면 되지만 속상한 건 (위치컨댜?)/(어떡한대?) 돈은 물어 주면 되지만 속상한 건 (위치껀댜?)/(어떡한대?)

4) "표준어 대응쌍 전사"는 소리 나는 대로 적은 "방언 전사"가 표준어 규정에서 벗어난 경우에, 그에 대응하는 표준형을 함께 제시하는 것을 원칙으로 한다. 띄어쓰기를 기준으로 하여 방언과 표준어를 각각 괄호 안에 넣어서 전사하고 이들 사이에는 빗금(/)을 넣는다. 방언 전사를 먼저하고 표준어 대응쌍 전사를 그 뒤에 나란히 제시한다.

지역	보기
강원	(여서)/(여기서) 꾸물거리지 말고 (얼푼)/(얼른) 가라. 마을 사람들은 뭐든 (농가)/(나누어) 먹지요.
경상	근데 (지)/(자기) 생각이 옛날부터 그런 생각을 하더라고. 여기에 동그라미나 (꼽표)/(곱표) 치세요.
전라	아이들은 (훈지)/(그네) 뛰면서 놀고 있었다. 늦은 사람이 (댑대로)/(도리어) 큰소리친다.
제주	성격이 참 (요망지다.)/(야무지다.) (하르방)/(할아버지) 댁에 가는 길.
충청	여기 (졸)/(부추) 한 단에 얼마요? (고쿠락)/(아궁이) 불이 꺼졌나 좀 봐라.

지역	보기
강원	모처럼 해가 난 (날에느)/(날에는) 마실이나 (댕게오시우.)/(다녀오시오.) 애가 종일 (울민서)/(울면서) 쳐다보더라고. 돈이 (읎어도)/(없어도) 남한테 (아수운)/(아쉬운) 소리는 못하겠다.
경상	상담소에는 어떤 걸 기대하고 (왔으까?)/(왔을까?) 마음에 든다 싶으면은 그냥 다 하는 (스타일이래)/(스타일이어) 가지고. (글잖아)/(그렇잖아). (우리가잉)/(우리가) 그래도 둘은 됐으면 하는 생각이 있잖아.
전라	하루 종일 이엉만 (영끄고)/(엮고) 급히 약을 지었는데도 못 (나수고)/(낫고) (가 부렀어.)/(가 버렸어.) 그거 다 (이애기헐라면)/(이야기하려면) (미칠을)/(며칠을) 해도 안 돼.
제주	(게민)/(그러면) (모지레민)/(모자르면) (멋을)/(뭣을) 더 (가*라 주코?)/(말해 줄까?) 야 (무신)/(무슨) 그런 게 또 (시어.)/(있어.) 어떻든 (저디)/(저기) 다 (지내치민)/(지나치면) (되우다.)/(됩니다.)
충청	너 (또래)/(때문에) (여기꺼지)/(여기까지) 와야 (되겄어?)/(되겠어?) (오둥아를)/(오디를) 얼마나 (마이)/(많이) 먹었는지 입 안이 시커멓게 (물들었슈.)/(물들었어요.) 점심 때는 (밥얼)/(밥을) 먹구, (새이)/(새참) 때는 (국시를)/(국수를) 먹는(겨)/(거야).

5) 외래어, 외국어의 경우 들리는 대로 한글로 표기하고 대응쌍은 전사하지 않는다. 이 사업은 방언형의 표준형 대응쌍 이중전사를 정확하게 하는 것이 목적이다. 따라서 작업 효율을 고려하여 외래어, 외국어는 사전 표준형을 일일이 검색하기보다는 들리는 대로 적도록 한다. 예를 들어, 발화자가 '뻐쓰'로 발화한다면 '뻐쓰'만 전사하면 된다.

올바른 표기	잘못된 표기
아무래도 뻐쓰가 더 빠르지. ([뻐쓰]로 발음한 경우)	아무래도 버스가 더 빠르지. 아무래도 (뻐쓰가)/(버스가) 더 빠르지.
그린 뉴딜 시대에 맞는 그린 모빌리티 보급 확대	Green-New Deal 시대에 맞는 green mobility 보급 확대

6) 숫자, 외국어, 기호, 단위 등은 숫자나 기호가 아닌 한글로 표기한다. 규범 표기가 미확정된 외국어의 경우 <우리말샘>의 등재된 표기를 기준으로 삼는다.

올바른 표기	잘못된 표기
우리가 구 박 십 일 갔었나?	우리가 9박 10일 갔었나?
시급이 오천 원 육천 원짜리도 하는 데도	시급이 5천 원 6천 원짜리도 하는 데도 시급이 5000원 6000원짜리도 하는 데도
그거는 항상 백 프로 만족은 잘 없다라고 생각이 듭니다.	그거는 항상 100% 만족은 잘 없다라고 생각이 듭니다.
그린 뉴딜 시대에 맞는 그린 모빌리티 보급 확대	Green-New Deal 시대에 맞는 green mobility 보급 확대

화자 표시

- 화자 아이디, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 'NA'로 표시한다.
- 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명하지 않을 경우에는 'NA'로 표시한다.

방언 전사의 단위 구획과 문장 부호의 사용

1) 방언 전사의 단위 구분

- 방언 전사의 단위는 문장 단위를 기본 단위로 하되, 구어의 특성을 고려하여 너무 긴 문장은 쉼을 고려하여 단위를 구분한다.
- 글을 쓸 때에는 마침표로 문장이 끝났음을 알 수 있지만, 방언에서는 문장부호가 없으므로 문장과 유사한 단위를 기본 단위로 구획할 필요가 있다. 이에 대한 기준은 다음과 같다.
  ① 문장 단위를 기본 단위로 한다. 단 문장이 너무 길거나 중간에 쉼이 있는 경우는 아래 ②와 같이 단위를 구분한다.
  ② 한 단위로 전사하는 분량이 6초를 넘지 않도록 제한하며, 이때 띄어쓰기 단위는 10개 이내가 된다. 이 기준은 절대적인 기준은 아니며, 컴퓨터의 자동 의미 분석에서 단위가 지나치게 복잡해지도록 하지 않기 위함이다. 즉 말하는 사람의 나이나 말하는 속도에 따라 6초 이상이 될 수도 있다.
  ③ 단위의 구분은 문장 내용의 달라짐을 기준으로 한다. 가급적 하나의 문장이 완성될 때에 단위를 나눈다.

예) 전사 단위 구획 예(※음성파일 **sample01.pcm** 참고)
※ 총 7구간. 각 구간마다 5~6초로 나눔

하루 시간으로 따졌을 때 제일 길게 했던 아르바이트는 역시나 상하차 아르바이트가 제일 길었고요.
이제 기간으로 따졌을 때는
저번 주에 그만두게 되었는데 이제
중화요리 집에서 홀서빙을 했었습니다 한
그래도 그렇게까지 길진 않았어요 한 달 정도?
한 달 동안 일하고 그만둔 게 이제 저의 저한테 제일 긴 거였죠.
아까도 말씀드렸다시피 저는 이제 단기로 아르바이트를 많이 했어 가지고 단기로 간헐적으로

2) 문장 부호의 사용

① 마침표와 물음표를 사용하고, 느낌표나 쉼표는 사용하지 않는다.

② 마침표는 문장이 완전히 끝났을 때에만 사용하며, 문장을 끝맺지 못하였을 경우는 단위를 나눈 경우라도 마침표를 붙이지 않는다. 이때 문장이 완전히 끝났다는 것은 '-다', '-어요', '-어라' 등 종결어미로 끝났음을 뜻한다.

올바른 표기	잘못된 표기
뉴스룸의 앵커브리핑을 시작하겠습니다.	뉴스룸의 앵커브리핑을 시작하겠습니다!
중화요리 집에서 홀서빙을 했었습니다.	중화요리 집에서, 홀서빙을 했었습니다.
저번 주에 그만두게 되었는데 이제	저번 주에 그만두게 되었는데 이제,

③ 문장이 종결되었을 경우 마침표 또는 물음표를 반드시 붙이며, 마침표나 물음표를 붙인 경우는 반드시 단위를 나눈다.

올바른 표기	잘못된 표기
똑같이 했을 겁니다. 어~ 그게 이거는 습관이고 버릇이고	똑같이 했을 겁니다. 어~ 그게 이거는 습관이고 버릇이고

④ 말끝을 올리는 경우는 물음표를 붙인다. 특히 '-어', '-어요' 등 말끝을 올리거나 내리는 것에 따라 의미가 달라지는 경우, 반드시 마침표와 물음표를 사용하여 구분해 준다.

평서문	의문문
-	그냥 월급 루팡이 되는 듯한 기분?
밥은 먹었어.	밥은 먹었어?

대화 순서 겹침

- 대화의 순서가 겹쳐 동시에 소리가 들리는 경우는 따로 표시하지 않고 대화를 먼저 시작한 화자를 기준으로 시간 순서에 따라 적는다. 만약 상대방의 맞장구를 치는 표현(예: 네, 그렇죠, 맞아)이 중간에 나오면 앞선 대화를 완전히 적은 다음에 맞장구치는 표현을 줄을 바꾸어 적는다.

구분	보기
주 발화	1: 딸 하나 (나)/(낳아) 갖고
맞장구	2: 네.
주 발화	3: 세 살 (묵아)/(먹어) (잊아뿌고)/(잊어버리고)

끊어진 단어(단어가 불완전하게 발화된 경우)

- 단어가 완전히 발음되지 않고 끊어진 경우는 그대로 전사하고 아래 예와 같이 –를 해당 단어 앞뒤에 붙인다. 이러한 단어가 둘 이상인 경우에는 모두 –를 붙인다.

올바른 표기	잘못된 표기
-전- -전- 전통이라고 우리가 흔히 얘기할 때	전통이라고 우리가 흔히 얘기할 때 전 전 전통이라고 우리가 흔히 얘기할 때
-학- 학교 아니 유치원에	-학- -학교- 아니 유치원에

- 발화자가 실수로 원래 하려던 말이 아닌 다른 말을 한 뒤 올바르게 수정한 경우, 또는 같은 단어를 반복해서 말한 경우에는 -를 표시하지 않고 들리는 그대로 표기한다. 그리고 해당 음성 단위에는 반드시 코멘트를 달아 놓도록 한다.

	올바른 표기	잘못된 표기
1	-학- 학교 아니 유치원에	-학- -학교- 아니 유치원에
2	크라운 -베- 크라운 베이커리 생크림이 좀 맛있죠.	-크라운- -베- -크라운- 베이커리 생크림이 좀 맛있죠.

띄어쓰기

1) 띄어쓰기는 띄어쓰기 규정에 맞게 한다.
2) 의존명사는 띄어 쓰고, 수를 적을 때는 만 단위로 띄어 쓴다(예: 십이억 삼천백만 팔백구 불 등). 판단하기 어려운 경우에는 수시로 논의하여 결정한다(예: 오십대, 일 대 이).
3) 본용언과 보조용언은 띄어 쓴다(예: 먹어 버리다, 가고 싶다, 먹지 못하다).

※ [주의] 단어를 발음하는 중간에 쉼이 들어간 경우에는 띄어 쓰지 않는다.

올바른 표기	잘못된 표기
많이 먹는구나 그걸로 넘어지기야 하겠냐만은	많이 먹는ᵛ구나 그걸로 넘어ᵛ지기야 하겠냐만은

※ [주의] 방언이 축약되어 띄어쓰기 규정을 적용하기 어려운 경우는 붙여 쓴다. 예를 들어 경상방언 '뭐라카노'의 '-카-'는 '(뭐라)고 하(노)'의 축약형이다.

올바른 표기	잘못된 표기
뭐라카노	뭐라ᵛ카노

축약형의 표기

1) 구어에서는 발음의 축약 현상이 많이 나타나는데, 두 음절이 한 음절이 될 수도 있고 네 음절이 두 음절로 줄어들 수도 있다. 이때 발음된 음절수와 표기상의 음절수를 맞추는 것을 원칙으로 한다. 따라서 축약형의 경우 모두 전사 표기에 반영한다. 표준어 대응쌍에서는 원래 형태를 다시 밝혀 적는다.

축약 대상	보기
그냥	(걍)/(그냥)
그러니까	(그니까)/(그러니까)

※ [주의] <표준국어대사전>에 준말로 등재되어 있는 다음과 같은 단어들은 일일이 표준어 대응

쌍을 적지 않는다.
예) 근데, 얘기, 요새, 애, 담, 맘, 첨, 낼, 젤, 좀, 재밌다, 갖다, …

2) 사귀어, 바뀌어 등에서 모음 'ㅟ'와 'ㅓ'가 합쳐져 1음절로 축약되어 발음되는 경우는 다음과 같이 모음 'ㅕ'로 바꾸어 전사한다. 표준어 대응쌍에서는 원래 형태를 다시 밝혀 적는다.

축약 대상	보기
사귀어	(사겨)/(사귀어)
바뀌어	(바껴)/(바뀌어)
뛰어	(뗘)/(뛰어)

담화 표지

1) "담화 표지"는 대화 상황에서 말하는 이가 머뭇거림, 이야기를 계속 하고 싶어 하는 등의 의도나 심리적 태도를 전달하기 위해 사용하는 것이다. 여기서는 "이, 그, 저, 아, 어, 에, 음, 응, 뭐" 등 1음절 담화표지에 한해서만 본래의 품사와 구별하기 위해 물결표(~)를 붙여 전사한다.
2) 여기서 물결표(~)를 같이 전사하는 경우는 머뭇거림의 느낌을 주는 담화표지이다. 즉, "이 사람, 그 사람, 저 사람"처럼 가리키는 말로 쓰이는 "이, 그, 저"나 감탄의 "아, 어, 에, 음, 응, 뭐" 등이 원래의 의미로 쓰이지 않고, 말을 더듬거리거나 머뭇거릴 때 사용될 경우에만 물결표(~)를 붙여 표기한다. ("인제, 이제, 그냥, 무슨, 어떤" 등은 2음절이므로 물결표(~)를 붙이지 않음).

지시·감탄의 경우	담화 표지인 경우
그 돈 벌고 싶어서	그~ 돈 벌고 싶어서
그냥 저 통상적인 노하우인지	그냥 저~ 통상적인 노하우인지
응. 얼마만인지 모르겠네	응~ 얼마만인지 모르겠네

잘 들리지 않는 부분

1) 대화가 잘 들리지 않는 부분이나 잘 들리지 않지만 해당 부분을 추측할 수 있는 경우, 추측한 내용을 (**()**) 안에 적는다.
2) 잘 들리지 않는 부분 중 일부분만 들리거나 추측 가능한 경우, 추측 가능한 부분을 (**()**) 안에 적되, 들리지 않는 부분은 그 음절 수만큼 x로 나타낸다.

구분	보기
추정 불가능	**(())** 너무한 거 같더라.
추정 가능	그 전까지는 직장 생활 (하니라구)/(하느라고) **((더 힘들어))**
일부 추정 가능	그거 진짜 **((xx해야))** 되겠더라.

3) 잘 들리지 않는 부분을 전사하기 위해 반복 청취 등의 노력을 들이지 말고 가급적 (()) 처리하는 것이 바람직하다.

말소리를 제외한 기타 소리들

1) 웃음, 목청 가다듬는 소리, 박수, 노래 등 직접적인 말소리가 아닌 소리에는 @를 앞에 붙여서 다음과 같

이 적는다. 전사 후 최종 단계에서는 다음과 같이 마크업된다

구분	전사	마크업
웃음	@웃음	**{laughing}**
목청 가다듬는 소리	@목청	**{clearing}**
박수	@박수	**{applauding}**
노래	@노래	**{singing}**

2) 위의 4가지 외에 기침, 들숨, 날숨, 재채기, 코훌쩍임, 하품 등의 소리는 아예 전사하지 않는다.

개인 정보의 보호

1) 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화번호 등 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 표시를 해 둔다.

구분	전사	마크업
이름	@이름	**&name&**
상호명	@상호명	**&company-name&**
주민등록번호	@주민번호	**&social-security-num&**
카드번호	@카드번호	**&card-num&**
주소	@주소	**&address&**
전화번호	@전화번호	**&tel-num&**

※ [주의] 정치인, 연예인 등 유명인의 이름은 위와 같이 하지 않고 그대로 전사한다.
※ [주의] 공적 성격을 지닌 이름들의 경우도 비식별화하지 않고 그대로 전사한다. 예를 들어 학교, 기관명, 단체명, 영화 제목, 노래 제목, 책 제목, 방송 제목, 게임명, 상품명, 제품명 등이 있다.

2) 특정 상호가 발화되었을 경우 그대로 전사하지 않고 아래와 같이 적는다.

> 신촌에 @상호명은 진짜 맛없어.

※ [주의] 넷플릭스, 유튜브, 삼성, 엘지, 애플 등 널리 알려져 있는 상호에 대해서는 위와 같이 하지 않고 그대로 전사한다. 개인상호만 위와 같이 표시한다.
※ [주의] 개인 유튜브 채널명도 신분 보장을 위해 비식별화해야 하며, 이때 이름이 아닌 상호로 취급하여 '@상호명'으로 표시한다.

> 뭐~ @상호명1나 아니면 @상호명2 이런 거 자주 봐.

3) 주소는 동 이하의 구체적인 주소만 표시하며, 동 이상의 주소는 그대로 전사한다.

> 근데 너 연희동 살잖아.(o)

4) 여러 이름이 나올 때는 번호를 붙여 구별해야 한다. 이때 한 파일 내에서 해당 번호가 가리키는 대상이 일관성을 지녀야 한다.

그때 철수랑 민수랑 너랑 나랑 갔잖아. 철수도 알고 있지?	
올바른 표기	잘못된 표기
그때 @이름1이랑 @이름2이랑 너랑 나랑 갔잖아. @이름1도 알고 있지?	그때 @이름1이랑 @이름2이랑 너랑 나랑 갔잖아. @이름3도 알고 있지?

※ [주의] 담화 안에서 등장하는 사람 이름이 매우 많을 수 있으므로, 전사 작업 시 반드시 이름과 순번의 대응을 따로 기록해 두고 헷갈리지 않도록 주의한다.
※ [주의] 상호명의 경우도 여러 상호명이 나오면 상호명1, 상호명2, …로 구별하여 전사한다.

5) 받침이 없는 이름 뒤에 '-입니다', '-인데' 등이 결합해서 '-ㅂ니다', '-ㄴ데' 형태로 나타나더라도 '-입니다', '-인데' 등을 살려 쓴다.

저는 김철숩니다.	
나 김민순데 알지?	
올바른 표기	잘못된 표기
저는 @이름1입니다.	저는 @이름1ㅂ니다.
올바른 표기	잘못된 표기
나 @이름2인데 알지?	나 @이름2ㄴ데 알지?

기타 지침

1) 방언 전사를 위해 사용한 기호(예: -, {}, &, (()) 등)는 표준어 대응쌍 표기에는 사용하지 않는다.
2) 큰따옴표나 작은따옴표를 사용하지 않는다. 즉 발화자가 글을 읽을 때 따옴표로 표시하지 않고 내용 전사만 한다.
3) 말꼬리를 끌어 장음으로 발음한 경우에 이를 반영하지 않고 원래의 단어로 적는다.
예) 학스읍(x) => 학습, 소오름(x) => 소름

지역별 방언 전사 주의 사항

1) 경상방언

- 종결어미에 '-이'가 결합한 '-대이, -래이, -재이'은 소리대로 적는다.

집에 (갔대이.)/(갔다.) 전화 (해래이.)/(해라.) 다음에 (보재이.)/(보자.)

- 표준어의 '그러다'에 해당하는 '그카다, 그쿠다' 등은 소리대로 적는다.

(그카면)/(그러면) 저기 갔다 올 (끼가?)/(거가?) (그쿠면)/(그러면) 그 일은 (끝났나?)/(끝났니?) (그카고)/(그러고) 있지 말고 (일로)/(이리로) (온나)/(오너라)

(그쿠고)/(그러고) 잘난척은 (자가)/(쟤가) 잘한다.

- 받침 'ㅇ'이나 'ㄴ'이 나타나지 않으면 소리대로 적는다.

(주머이)/(주머니)
(어무이)/(어머니)
(사이)/(산이) 크다
(학새이)/(학생이)

2) 전라방언

- 표준어 '-으니까'의 방언형은 '-응께, -응게, -응께네, -으니께' 등으로 소리대로 적는다.

(인자는)/(인제는) 약이 (조응께)/(좋으니까)
그 공식적으로 다 (허니께)/(하니까)

- 둘째 음절 이하의 'ㅎ'이 나타나지 않는 말의 경우, 다음과 같이 소리대로 적는다.

(뭐다러)/(뭣 하러) 그러냐?
잘 하지도 (모다고)/(못하고)
(배과점에)/(백화점에) 가 (봉께)/(보니까)
벌써 (유강년이여?)/(육 학년이야?)
으메 (답다번거)/(답답한 것)
눈 앞이 (깝까버다.)/(갑갑하다.)

- 'ㄴ'이 나타나지 않으면 소리대로 적는다.

(가마이)/(가만히)
(마이씩)/(많이씩)
(아잉)/(아닌) 게 (아이라)/(아니라)

3) 제주방언

- 앞에 요소가 받침이 있는 음절이고 후행하는 요소가 모음으로 시작할 때 후행 단어의 첫 음절 자리에 받침 자음을 복사하여 발음한다. 이러한 경우는 소리대로 적는다.

(한국금식)/(한국음식)
(맏다덜)/(맏아들)
(비단놋)/(비단옷)
(감못)/(감옷)

- ᄋᆞ(아래아) 는 '아*'로 적는다.

(밟*아)/밟아
(나*말*)/(나물)
(아달*)/(아들)

※ [참고] 고동호(2008: 69-70),「제주방언 ·의 세대별 변화 양상」,『한국언어문학』65, 55-74., 김원보(2006: 134-135),「제주방언화자의 세대별(20대, 50대, 70대) 단모음의 음향분석과 모음체계」,『언어과학연구』39, 125-136. 등 최근 연구에서는 제주방언에서 70대 이상은 [ᄋᆞ] 발음을 유지하고 있고, 50대는 개인에

따라 어휘에 따라 있기도 없기도 하고, 20대는 [ㅇ] 발음이 거의 없다고 보고하였다.

4) 충청방언

- 종결어미 '-댜'는 소리대로 적는다.

> 그 낮에 꿈을 (꾸니께)/(꾸니까) (그라댜)/(그러더래)
> 우리 (야덜)/(아들) (잡어)/(잡아) (간댜.)/(간대.)

- 표준어 '어떻게'의 방언형은 '워떻게, 어티기' 등은 소리대로 적는다.

> 혹시 (워떻게)/(어떻게) 하는 건 줄 아세요?
> 장사 하려고 (어티기)/(어떻게) 집을 크게 (졌는디)/(졌는데)

5) 기타

- 어두 된소리화 현상

방언에서 흔히 나타나는 어두 된소리화의 경우, 방언의 특성으로 볼 수 있으므로 소리나는 대로 전사하고, 표준어 대응쌍 이중전사를 한다.

> 예) (쩌번에)/(저번에), (따르다)/(다르다), (꼐속)/(계속)

1.5.2 표준어 대응쌍 전사지침

기본 원칙

1) 현재 컴퓨터를 사용하여 한국어를 분석하는 도구는 기본적으로 표준어를 기반으로 개발되었다. 방언과 함께 표준어 대응쌍을 구축하는 것은 특수한 방언형과 표준어를 쌍으로 제시함으로써 표준어의 규칙을 벗어나는 구어 방언 데이터를 AI가 인식할 수 있도록 하기 위함이다.

2) '방언 전사'와 '표준어 대응쌍'은 "(방언 전사 형태)/(표준어 대응쌍 형태)"의 꼴로 제시한다. "방언 전사"를 먼저하고 "표준어 대응쌍 전사"를 그 뒤에 나란히 제시한다.

지역	보기
강원	어제는 막 아프다고 (그라)/(그렇게) 난리를 치더니 오늘은 좀 괜찮냐?
경상	고등학교 (댕길)/(다닐) 때 미역 (쭐거리)/(줄기) 반찬도 (마이)/(많이) (묵었지.)/(먹었지.)
전라	나는 (거까정은)/(거기까지는) 잘 (모릉께)/(모르니까) 이제 더 묻지 마시오.
제주	보리 (한*)/(한) 말이면 부인들 (비렁)/(빌려) (했주게.)/(했지).
충청	그럼 내가 (지비)/(집에) 갔다 올 (때꺼정)/(때까지) (지다리실텨?)/(기다리실 테요?)

3) 표준어 대응쌍의 작성은 국어 어문규범(한글 맞춤법, 표준어 규정, 외래어 표기법, 로마자 표기법)에 따른다.

4) 문장부호는 괄호 속에 넣어서 방언 전사형과 표준어 대응쌍 모두에 동일하게 제시한다.

지역	보기

강원	나도 이쪽 동네 (출신이래.)/(출신이야.) 니가 말한 건 이게 (다나?)/(다니?)
경상	떡을 (맹갈아)/(만들어) (묵었지.)/(먹었지.) 어제 어디 (갔었노?)/(갔었니?)
전라	혼자 다 (묵어 분당께.)/(먹어 버린다니까.) 난장을 쳐도 (쪼깨)/(조금) (늦것는디?)/(늦겠는데?)
제주	하루에 같이 (검질멧주게.)/(김매었지). (게민)/(그러면) (멋을)/(뭣을) 더 (가*라 주코?)/(말해 줄까?)
충청	(오둥아를)/(오디를) 먹었는지 입 안이 시커멓게 (물들었슈.)/(물들었어요.) 동네 사람들은 (워떡헌댜?)/(어떡한대?)

표준어 대응쌍 표기와 선별 기준

1) 방언형에 대응되는 표준어 대응쌍 표기는 국립국어원의 <우리말샘>의 정보를 기준으로 삼는다. 아래에서 방언 '맹갈다, 묵다'에 대해 '만들다, 먹다'를 표준어 대응쌍으로 전사한다.

※ <우리말샘> https://opendic.korean.go.kr/main

지역	보기
강원	돈이 (읎어도)/(없어도) 남한테 (아수운)/(아쉬운) 소리는 못하겠다. <우리말샘> 읎다 • 읎다 「001」「형용사」「방언」 '없다'의 방언(강원, 경기, 전북, 충청) 아숩다 • 아숩다 「001」「형용사」「방언」 '아쉽다'의 방언(강원, 경상, 전라, 충청)
경상	떡을 (맹갈아)/(만들어) (묵었지.)/(먹었지.) <우리말샘> 맹갈다 • 맹갈다 「001」「동사」「방언」 '만들다'의 방언(경남, 전남) 묵다 • 묵다 「004」「동사」「방언」 '먹다'의 방언(강원, 경상, 전라)
전라	그거 다 (이애기헐라면)/(이야기하려면) (미칠을)/(며칠을) 해도 안 돼. <우리말샘> 이애기 • 이애기 「001」「명사」「방언」 '이야기'의 방언(강원, 경북, 전라) 미칠 • 미칠 「001」「명사」「방언」 '며칠'의 방언(경남, 전남)
제주	성격이 (잘도)/(매우) (요망지다.)/(야무지다.) <우리말샘> 잘도 • 잘도 「001」「부사」「방언」 '매우'의 방언(제주) 요망-지다 • 요망-지다 「001」「형용사」「방언」 '야무지다'의 방언(제주)
충청	그럼 내가 갔다 올 (때꺼정)/(때까지) (지다리실텨?)/(기다리실 테요?) <우리말샘>

	꺼정 * **꺼정** 「001」「조사」「방언」 '까지'의 방언(강원, 경상, 충청, 함경). **지다리다** * **지다리다** 「001」「동사」「방언」 '기다리다'의 방언(충청)

다음의 '먼'도 <우리말샘>에 '무슨'의 방언형으로 기술되어 있으므로, 아래 예와 같이 '무슨'을 표준어 대응쌍으로 제시해야 한다.

(먼)/(무슨) 말이야?

먼

* **먼** 「002」「관형사」「방언」 '무슨'의 방언(강원, 경상, 전라).

※ '머(뭐)'의 처리('뭐'를 '머'로 발음하여 '머'로 전사하는 경우)
: '먼'과는 달리 '머'는 <우리말샘>에 방언형이 아닌 구어형으로 기술되어 있다. 그러나 예외적으로 표준어 대응쌍을 제시해야 한다.

올바른 표기	잘못된 표기
(머에)/(뭐에) 쫓기는 듯이 급하게	머에 쫓기는 듯이 급하게

※ 대화 상황에서 자주 사용되는 중앙방언형 어미 '-애', '-어', '-두', '-구' 등은 비표준이지만 예외적으로 표준어 대응쌍을 제시하지 않는다, 다만 어간에 방언형이 나타나 표준어 대응쌍을 제시할 경우에는 어미 '-애', '-어', '-두', '-구' 등도 함께 표준어를 제시한다.

올바른 표기	잘못된 표기
진짜 우리가 잘 되길 바래.	진짜 우리가 잘 되길 (바래.)/(바라.)
너무 많이 지난 것 같애.	너무 많이 지난 것 (같애.)/(같아.)
그런 생각은 아예 하지를 말어.	그런 생각은 아예 하지를 (말어.)/(말아.)
그 사람 말이 맞어.	그 사람 말이 (맞어.)/(맞아.)
이제 나는 하나두 모르겠다.	이제 나는 (하나두)/(하나도) 모르겠다.
손부터 씻구 밥을 먹어라.	손부터 (씻구)/(씻고) 밥을 먹어라.
아직도 안 돌아갔다구?	아직도 안 (돌아갔다구?)/(돌아갔다고?)
이제 밥을 먹으려구 한다.	이제 밥을 (먹으려구)/(먹으려고) 한다.
이래 (가주구)/(가지고)	이래 (가주구)/(가지구)

2) 여러 단어가 합쳐져서 만들어진 말의 경우 사전 표제어 형식을 확인한다. 두 단어가 -로 연결된 경우 전체를 한 덩어리로 간주한다. 두 단어가 ^로 연결된 경우나 연결되지 않은 경우는 각각의 단어에 대해 표준어 대응쌍을 표기한다.

천지-삐까리 (天地삐까리)

* **천지-삐까리** 「001」「명사」「방언」 매우 많음(경상).

서답 구덕

* **서답 구덕** 「001」「방언」 '빨래 바구니'의 방언(제주).

올바른 표기	잘못된 표기
할 게 (천지삐까리다.)/(매우 많다.)	할 게 (천지)/(매우) (삐까리다.)/(많다.)
(서답)/(빨래) (구덕에)/(바구니에) 담아	(서답 구덕에)/(빨래 바구니에) 담아

3) 방언형이 여러 개의 표준어와 대응되는 경우에는 다음과 같은 원칙에 따른다.

ㄱ. 형태적으로 가장 가까운 표준형을 선택한다.

올바른 표기	잘못된 표기
지금 (머라카노?)/(뭐라고 하니?)	지금 (머라카노?)/(뭐라는 거니?)

ㄴ. 대화 상황에서 어미는 대화 맥락이나 어감, 화자·청자의 관계 등에 따라 어울리는 표준형이 달라질 수 있다. 이러한 경우 상황에 맞게 표준어형을 선택하며, 이 부분에 대한 판단은 전사자에 따라 다소 주관적일 수 있다.

방언형	복수 대응 표준어형	사용 맥락
경상방언 '-노?'	어제 어디 (갔었노?)/(갔었냐?)	친구 등 동등한 관계에서 사용
	어제 어디 (갔었노?)/(갔었니?)	부자, 사제 등 상하 관계에서 사용
경상방언 '-꾸마'	내가 (하꾸마.)/(할게.)	친구 등 동등한 관계에서 사용
	내가 (하꾸마.)/(하마.)	부자, 사제 등 상하 관계에서 사용
충청방언 '기여(겨)', '그려'	그게 (기여?)/(맞아?) 그게 (기야?)/(맞아?) 그게 (기냐?)/(맞냐?)	내용을 확인하는 질문에서 사용
	(겨)/(그래) (아녀?)/(안 그래?) (기여)/(그래) (아니여?)/(안 그래?)	'겨 아녀?'의 맥락에서 사용
	(그려.)/(그래.) (겨.)/(그래.) (아녀.)/(아니야.)	질문에 대답하는 상황에서 사용

※ 이것은 맥락에 따라 어감이 달라지는 어미에만 해당하며, 명사나 동사 등 단어의 경우에는 맥락상 다소 어색하더라도 사전에 제시된 형태를 그대로 따른다.

머스마

* 머스마 「001」「명사」「방언」 '사내아이'의 방언(강원, 경상, 전북, 충청).

올바른 표기	잘못된 표기
그 직원이 (머스만데.)/(사내아인데.)	그 직원이 (머스만데.)/(남잔데.)

4) 방언형에 형태적으로 유사한 표준형이 없을 때는 의미적으로 가장 유사한 표준형을 선택한다. 예를 들어 경상방언의 '-매로', '-맨치로'는 형태적으로 비슷한 표준어 어휘가 없지만, 의미적으로는 '-처럼'과 거의 동일하므로 '-처럼'을 표준어 대응형으로 적는다.

(니매로)/(너처럼)
(니맨치로)/(너처럼)

5) 형태적으로도 의미적으로도 대응 표준어를 사전에서 발견할 수 없는 경우, 대응하는 표준형이 없는 것으로 간주한다. 이러한 경우 표준형을 입력할 자리에 '#방언형'을 입력하며, 방언형의 뜻을 풀이하여 표준형에 기술하지는 않는다.

올바른 표기	잘못된 표기
(뭉티기)/(#뭉티기) (무러)/(먹으러) 가자	(뭉티기)/(생고기를 엄지손가락만하게 썰어 낸 음식) (무러)/(먹으러) 가자
(하고재비가)/(#하고재비가)	(하고재비가)/(하고 싶어 하는 사람이)

6) 방언형에 대한 표준어 대응쌍은 가급적 음절 및 어절 수를 맞추어 제시한다. 그러나 음절 및 어절 수를 맞출 수 없는 경우도 있는데, 예로 아래 '경상, 충청'처럼 한 어절이 두 어절의 표준어 대응쌍을 가질 수도 있다.

지역	보기
강원	어제는 막 아프다고 (그라)/(그렇게) 난리를 치더니 오늘은 좀 괜찮냐?
경상	지금 (머라카노?)/(뭐라고 하니?)
전라	혼자 다 (묵어)/(먹어) (분당께.)/(버린다니까.)
제주	하루에 같이 (검질멧주게.)/(김매었지.)
충청	언제까지 (지다리실텨?)/(기다리실 테요?)

축약과 생략

1) 방언전사에서 탈락된 소리는 표준어 대응쌍에서 복원시켜 본딧말로 바꾸어 적는다.

전부 (수매르)/(수매를) (하꺼인디)/(할 것인데)

2) 방언에서 축약형으로 표기된 형태 역시 표준어 대응쌍에서 본딧말로 바꾸어 적는다. 축약형와 본딧말 모두 표준어로 사전에 등재되고, 문맥상 축약형이 본딧말보다 더 자연스러울 경우, 축약형도 허용한다. 예를 들어 '거'와 '것'은 모두 표준형 대응쌍에 사용할 수 있지만, '할 거인데(×)'와 같이 문맥 상 '거'로 바꾸는 것이 허용되지 않는 경우는 반드시 '것'으로 적어야 한다.

올바른 표기	잘못된 표기
(커능기)/(하는 것이) (커능기)/(하는 게)	· ·
(하꺼인디)/(할 것인데)	(하꺼인디)/(할 거인데)

3) <우리말샘>에 준말로 등재되어 있는 다음과 같은 단어들은 일일이 표준어 대응쌍을 적지 않는다.
예) 근데, 얘기, 요새, 애, 담, 맘, 첨, 낼, 젤, 좀, 재밌다, 갖다, …

4) 외래어, 외국어의 준말의 경우 다른 외래어, 외국어와 마찬가지로 표준어 대응쌍을 적지 않는다. 준말의 발화 그대로 전사한다.
예) 알바, 페북, …

띄어쓰기

- 방언형에서는 띄어쓰기가 무시되었더라도 표준형에서는 어문규범에 준하여 띄어쓴다.

(뭐라카노?)/(뭐라고 하니?) (뭐라캐)/(뭐라고 해) (쌓노)/(쌓니)

(이래)/(이렇게 해) (가주고)/(가지고) 농사를 (지야)/(지어) (노으면)/(놓으면)

방언권별 조사, 어미 표준어 대응쌍 목록(일부)

- <우리말샘>의 방언(조사, 어미, 품사없음)에 대한 표준어 대응쌍의 일부를 보이면 다음과 같다. 목록 전체는 별첨 자료(**우리말샘_방언_조사 어미 품사없음_5개권역)**를 참고하면 된다.

방언권	어휘	품사	표준어 대응쌍
강원	가	품사 없음	걔
	갠데	품사 없음	그런데
	-게르	어미	-게
	아무케	품사 없음	아무렇게
	-았댔-	어미	-았었-
	야	품사 없음	얘
	오러	품사 없음	요렇게
	우뜨케	품사 없음	어떻게
	우째	품사 없음	어째
	울-만큼	품사 없음	얼마큼
	이러	품사 없음	이렇게
	인데	조사	한테
	자	품사 없음	쟤
	처름	조사	처럼
전라	-간디	어미	-관데
	거따가	품사 없음	게다가
	거러치럼	품사 없음	그렇게
	고러치롬	품사 없음	그렇게
	-그레	어미	-기에
	까장	조사	까지
	-당께로	어미	-다니까
	땜시	품사 없음	때문에
	-라우	어미	-요
	매이로	조사	처럼
	-ㅂ디껴	어미	-ㅂ디까
	아무케	품사 없음	아무렇게
충청	-겄-	어미	-겠-
	겨	품사 없음	거야
	그랴도	품사 없음	그래도
	까장	조사	까지
	-ㄹ라구	어미	-려고
	-ㄹ텨	어미	-ㄹ래
	-어유	어미	-아요
	워떡-허다	품사 없음	어떡하다
	워째서	품사 없음	어찌하여서
	워척-허다	품사 없음	어떡하다
	워치게	품사 없음	어떻게
	이릏게	품사 없음	이렇게
	튜	품사 없음	테요

제주	-게겐	어미	-자꾸나
	게고-제고	품사 없음	그러고저러고
	드레	조사	으로
	-멍	어미	-면서
	-메서란	어미	-매
	-센	품사 없음	-라고
	신디	조사	한테
	아메-나	품사 없음	아무렇게나
	-앙근에	어미	-고서, -아서
	야이	품사 없음	얘
	-어근	어미	-어서
경상	까짐	조사	까지
	-꺼마	어미	-ㄹ게
	-꾸마	어미	-마
	끈	조사	까지
	-는기라	품사 없음	-는 거야
	-니껴	어미	-ㅂ니까
	-래이	어미	-라
	매추로	조사	처럼
	-시다	어미	-오
	-시소	어미	-십시오
	우짜모	품사 없음	어쩌면
	-응께	어미	-으니까
	이카면	품사 없음	이렇게 하면

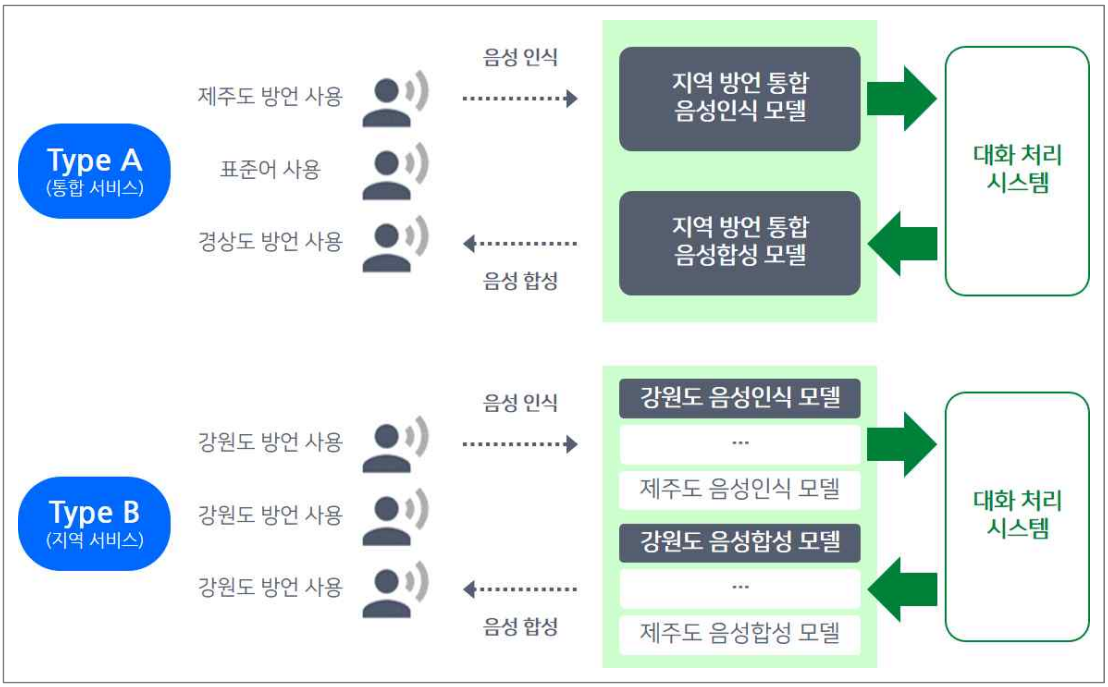
2.6 활용

2.6.1 활용 모델

2.6.1.1 모델 학습

과제명	AI 모델	모델 성능 지표
한국어 방언 발화 데이터 (강원도)	• 각 도별 음성인식 모델 (각 도별 5개, 5도 통합 1개) • 각 도별 음성합성 모델 (5개) • 각 도별 방언-표준어 기계번역 모델 (각 도별 5개, 5도 통합 1개) • GPT 기반 자연어 생성 기술을 이용한 일상 대화 모델 (각 도별 5개, 5도 통합 1개)	• 음성인식 모델 음절 인식율[1] • 음성합성 모델 음성 품질 수치 (MOS[2]: Mean opinion score) • 기계번역 모델 번역 품질 수치 (BLEU[3]: bilingual evaluation understudy) • 일상 대화 모델 대화 품질 수치 (Perplexity[4]; SSA[5]: Sensibleness and Specify Average, 구글에서 제안한 Human Evaluation Metric)
한국어 방언 발화 데이터 (경상도)		
한국어 방언 발화 데이터 (전라도)		
한국어 방언 발화 데이터 (제주도)		
한국어 방언 발화 데이터 (충청도)		

2.6.1.2 서비스 활용 시나리오



2.6.2 데이터 제공

1) AI Hub(www.aihub.or.kr) 활용

○ 구축된 데이터를 제공하는 AI Hub 페이지에서 한국어 방언 AI 데이터의 필요성과 구축 내용, 데이터셋 구조, 예시 등에 대해 자세한 내용을 제공하여 누구나 쉽게 데이터를 활용할 수 있는

환경 마련

○ 해당 데이터를 이용한 국민, 기관, 기업 등의 피드백을 수렴하고 결과를 분석하여 활용 방향 제고

2) 솔트룩스 AI Cloud(saltlux.ai) 활용

○ 솔트룩스의 3세대 AI 클라우드 서비스(saltlux.ai)는 현재 43개의 AI 모델을 사용할 수 있는 서비스를 제공하고 있음

○ 한국어 방언 AI 학습데이터의 활용한 AI 모델은 솔트룩스의 AI 클라우드 서비스 자원으로 무상으로 3년간 공개를 지원