

국립국어원 2019-01-08

발 간 등 록 번 호
11-1371028-000765-01

구어 자료 수집 및 원시 말뭉치 구축

사업 책임자
이 경 일

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘구어 자료 수집 및 원시 말뭉치 구축’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 6월 12일 ~ 2019년 12월 27일

2019년 12월 27일

사업 책임자: 이경일(주)솔트룩스

사업 수행자 (주)솔트룩스

사업 책임자 이경일
사업 참여자 박지혜, 주재현 외

<사업 수행자> (주)솔트룩스

사업 책임자	이경일((주)솔트룩스)
사업 참여자	박지혜((주)솔트룩스)
	주재현((주)솔트룩스)
	박선희((주)솔트룩스)
	김소정((주)솔트룩스)
	조성현((주)솔트룩스)
	배소영((주)솔트룩스)
	강수빈((주)솔트룩스)
	오지희((주)솔트룩스)
	김예하나((주)솔트룩스)
	김지영((주)솔트룩스)
	이혜련((주)솔트룩스)
	이선유((주)솔트룩스)
	안태성((주)솔트룩스)
	양승원((주)솔트룩스)
	김진우((주)솔트룩스)
	권기환((주)솔트룩스)
	이희철((주)솔트룩스)
	최창걸((주)솔트룩스)

구어 자료 수집 및 원시 말뭉치 구축 사업

본 사업은 구어 자료 수집 및 원시 말뭉치 구축 사업으로 자료의 수집 지침과 말뭉치 구축 지침에 따라 강연, 강의, 토론 등 구어 15,000시간, 드라마 대본 등 준구어 1,540만 어절 규모의 말뭉치를 새롭게 구축하는 데에 목적이 있다. 이에 따른 주요 과업과 사업의 성과는 다음과 같다.

구어/준구어 자료 수집: TV, 라디오, 인터넷 방송 등의 매체에서 구어 자료를 수집하고 드라마 대본 등에서 준구어 자료를 수집하여 민간 활용도 제고를 위한 저작권 이용허락 계약을 체결한다. 수집 자료가 산업계와 학계에 유효한 자료인지를 검증하고 유효하지 않는 자료는 수집 대상에서 제외한다.

원시 말뭉치 구축: 유효성이 검증된 수집 자료를 대상으로 말뭉치 구축 지침을 준수하여 전사한다. 전사 결과물을 대상으로 메타 정보를 포함한 구어 15,000시간, 준구어 1,540만 어절 이상의 원시 말뭉치를 구축한다.

원시 말뭉치 활용: 음성 인지 엔진과 언어 인지 엔진에 원시 말뭉치를 학습 데이터 형태로 활용하여, 구축된 원시 말뭉치의 활용 사례 및 방향성을 제시한다.

구어 자료 원시 말뭉치는 국가 주도의 말뭉치로서 인공 지능 기술 및 한국어 교육 연구에 활용됨으로써 인공 지능 산업의 국제 경쟁력 강화에 이바지할 것이다.

주요어: 원시 말뭉치, 구어 말뭉치, 준구어 말뭉치, 구어 말뭉치 수집, 준구어 말뭉치 수집, 원시 말뭉치 활용

차 례

제1장 사업 개요

1. 사업 목적	3
2. 사업 범위	4

제2장 사업 수행

1. 구어 자료 수집	7
2. 말뭉치 구축 대상 선정	9
3. 작업자 교육	16
4. 원시 말뭉치 구축 및 메타 정보 구축	21

제3장 사업 수행 결과

1. 원시 말뭉치 구축	33
2. 말뭉치 활용	40
3. 정책 제언	44

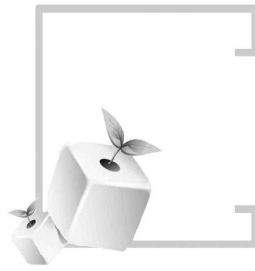
[붙임1] 현대 국어 구어 전사 말뭉치 지침	47
[붙임2] 전사 말뭉치 마크업 지침	53
[붙임3] 맞춤법 및 표기와 관련된 지침	55

표 차례

<표 1-1> 사업의 범위	4
<표 2-1> 방송사별 구어 자료 선정 시간	10
<표 2-2> 방송사별 준구어 자료 선정 시간	10
<표 2-3> 구어 자료 선정 대상 주제 분류	11
<표 2-4> 준구어 자료 선정 대상 주제 분류	11
<표 2-5> 구어 말뭉치 구축 대상 목록	13
<표 2-6> 준구어 말뭉치 구축 대상 목록	15
<표 2-7> 샘플 테스트 평가 지수	16
<표 2-8> 작업자 교육 지원 조직	18
<표 2-9> 작업자 교육 과정	18
<표 2-10> 보안 교육 및 관리	20
<표 2-11> 전사 지침(일부)	22
<표 2-12> 품질 점검 항목	27
<표 2-13> 파일명 부여 방식	29
<표 2-14> XML 변환 절차	29
<표 3-1> 구어 말뭉치 주제별 구축 결과	33
<표 3-2> 구어 말뭉치 방송 연도별 구축 결과	34
<표 3-3> 구어 말뭉치 주제×방송 연도별 구축 결과	36
<표 3-4> 준구어 말뭉치 주제별 구축 결과	37
<표 3-5> 준구어 말뭉치 방송 연도별 구축 결과	37
<표 3-6> 준구어 말뭉치 주제×방송 연도별 구축 결과	39

그림 차례

[그림 1-1] 인공 지능 시장 규모 추이.....	3
[그림 1-2] 사업의 핵심 성공 요소	4
[그림 2-1] 저작권 이용 허락 계약서 및 부속 합의서.....	8
[그림 2-2] 전문가 분석을 통한 구축 대상 선정	9
[그림 2-3] 샘플 테스트 절차	16
[그림 2-4] 샘플 테스트 품질 평가	17
[그림 2-5] 작업자 품질 평가표	17
[그림 2-6] 구축 지침 교육 자료(일부)	19
[그림 2-7] 말뭉치 구축 절차	21
[그림 2-8] 마크업 변환 절차	21
[그림 2-9] 전사 예시(발화자 표시)	22
[그림 2-10] 전사 예시(발음 전사)	23
[그림 2-11] 전사 예시(겹침 발화)	23
[그림 2-12] 음성 인식 자동 전사 도구	24
[그림 2-13] 수동 전사 도구	25
[그림 2-14] 작업자 실시간 피드백	26
[그림 2-15] 자체 품질 점검 절차	27
[그림 2-16] 품질 점검 후 수정 사항 반영 예시	28
[그림 2-17] 오류 사항 확인 및 수정 보완	30
[그림 3-1] 음성 인터페이스 기반 서비스	40
[그림 3-2] 음성 인식 개요	41
[그림 3-3] DNN-HMM 음향 모델의 구조	41
[그림 3-4] 언어 인지 엔진 구성도	42
[그림 3-5] 챗봇 서비스를 위한 음성 인식	43
[그림 3-6] 빅데이터 분석 시스템.....	43



제 1 장

사업 개요



1. 사업 목적

4차 산업혁명 대비 기반 기술 개발 및 인공 지능 기술 개발, 활용을 위해서는 대규모 말뭉치 구축을 통해 국어 자원의 활용도와 가치를 제고하는 것이 필요하다. 이러한 대규모 말뭉치는 민간 공유를 통해 언어 인공 지능 등 관련 산업 활용을 위한 기반을 마련하는 것과 동시에 학계에서의 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용이 된다.

인공 지능 시장은 해마다 증가하면서, 대규모 말뭉치의 확보는 대단히 중요시되고 있으며 해외 선도국 및 글로벌 기업에서는 많은 비용을 투자하여 언어 자료를 수집하고 말뭉치를 구축하여 인공 지능 기술 개발에 활용하고 있다.

본 사업은 TV, 인터넷 방송 등의 매체에서의 강연, 발표, 토론, 인터뷰, 대화 등의 구어 원자료와 드라마, 연극 대본 등의 준구어 원자료를 수집하고 음성 자료 전사를 통해 메타 정보가 있는 원시 말뭉치 구축하여 현재 확대되는 인공 지능 시장에 필요한 연구와 산업 발전에 이바지하는 데 목적이 있다.

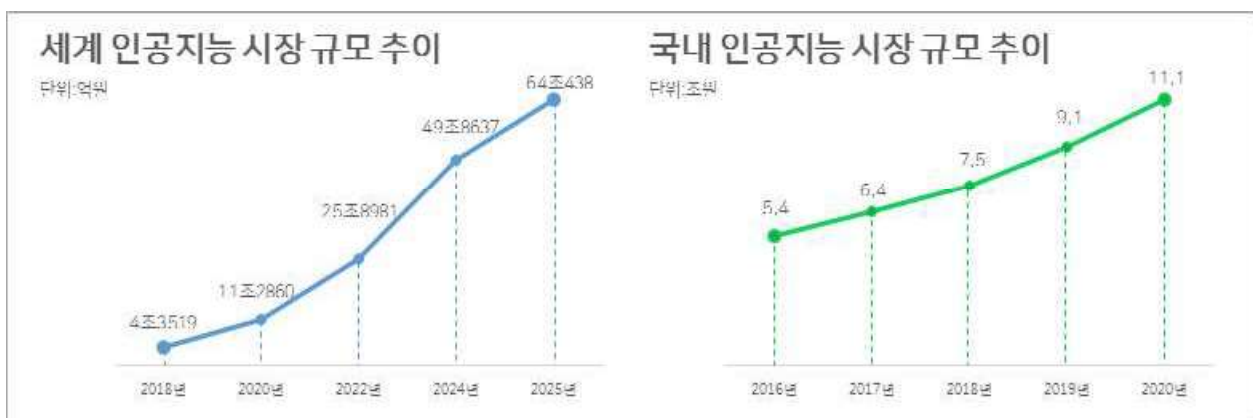


그림 1-1 인공 지능 시장 규모 추이

2. 사업 범위

본 사업에서 구축된 원시 말뭉치를 인공 지능 산업계와 학계에서 사용하기 위해서는 다양한 유형의 구어 자료 수집을 바탕으로 고품질의 원시 말뭉치를 구축하여 산업계와 학계의 인공 지능 기술 활용도를 제고시켜야 한다.

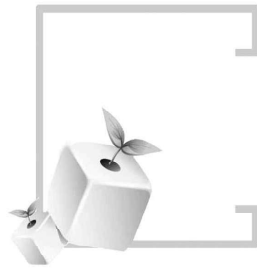


그림 1-2 사업의 핵심 성공 요소

이를 위한 본 사업의 수행 범위는 크게 세 부분으로 나눌 수 있다. 첫째는 다양한 유형의 구어 자료 수집을 위해 TV와 인터넷 방송 등의 원시 자료를 수집하는 것이다. 둘째는 산업계와 학계에서 활용이 가능한 유용한 말뭉치 확보를 위해 수집된 자료를 대상으로 유용성을 검증하는 것이다. 마지막으로 고품질 데이터 확보를 위해 원시 말뭉치 구축 시 전사 지침 및 품질 검증이 진행되어야 한다.

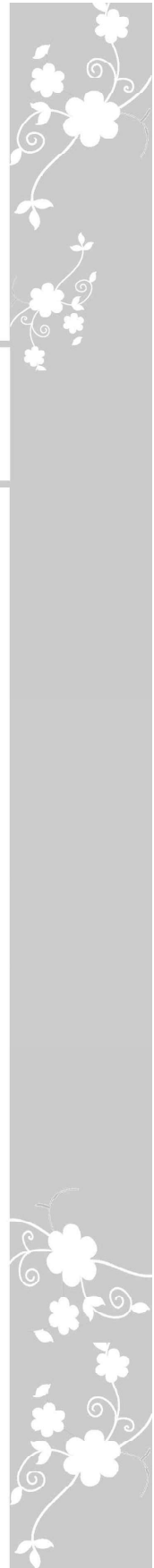
사업의 범위	사업 내용
구어 자료 수집	<ul style="list-style-type: none"> • TV, 인터넷 방송 등의 매체 자료 수집 • 강연, 토론, 인터뷰, 대화 등의 구어 원자료 수집 • 드라마, 연극 대본 등의 준구어 원자료 수집 • 수집 자료의 저작권 이용 허락 계약 체결
수집 자료 분류 및 유용성 검증	<ul style="list-style-type: none"> • 수집 대상 목록의 매체, 시기, 자료의 주제별 분류 • 산업계, 학계 전문가와 자문위원의 유용성 검증 • 원시 말뭉치의 메타 정보 선정 • 자료 및 메타 정보의 수정
원시 말뭉치 구축	<ul style="list-style-type: none"> • 음성 자료를 발화된 그대로 전사하여 전사 자료 구축 • 전사 자료에 대해 헤더 등 메타 정보 부착 • 능동적 기계 학습을 통한 4단계 품질 검증

표 1-1 사업의 범위



제 2 장

사업 수행



1. 구어 자료 수집

다양한 구어 자료를 수집하기 위해서는 구어 자료를 다량으로 확보하고 있는 방송사의 협조 및 공급 계약이 필요하다. 본 사업은 구어 자료 수집 이후에 수집된 자료를 토대로 원시 말뭉치를 구축해야 하므로 방송사와의 공급 계약뿐만 아니라 원시 말뭉치 구축을 위한 구어 자료의 저작권 이용 허락 계약이 필수적이다. 본 사업 진행 시 방송사별로 저작권 이용 허락 계약서에 대한 법적 검토가 있었고, 공급 계약은 가능하나 저작권 이용 허락 계약이 불가능한 경우가 발생하여 최종적으로 계약이 결렬된 사례가 있었다. 방송사별로 저작권 이용 허락 계약서 승인을 위해 계약서 검토를 진행하였고, 최종적으로 저작권 이용 허락 계약서를 승인한 방송사를 대상으로 구어 자료를 수집하였다.

1.1. 방송사 협의

대량으로 방송 대화를 수집하기 위해 MBC 등 약 10곳 이상의 대형 방송사를 접촉하였고, 대형 방송사로부터 제공받기 어려운 과학, 건강 등 특정 주제의 구어 자료 확보를 위해 5곳 이상의 인터넷 방송사를 접촉하였다.

방송사별로 최소 5차례 이상의 협의를 통해 본 사업의 목적 및 저작권 이용 허락 계약의 필요성을 설명하였고, 최종적으로 MBC, EBS, TV조선, TBS교통방송 이상 4곳의 대형 방송사와 과학과사람들, 마인드코칭연구소, 펄스교양 등 인터넷 방송사와 구어 자료 수집 계약을 체결하였다.

준구어 말뭉치 구축을 위한 방송 대본 수집의 경우, MBC가 한국방송작가협회와 본 사업을 위한 계약을 체결하여 드라마 대본 공급 및 저작권 이용 허락 계약을 충족하였다.

1.2. 저작권 이용 허락 계약 체결

구어 자료 공급 계약을 체결한 방송사별로 원시 말뭉치 구축을 위한 저작권 이용 허락 계약을 체결하였고, 공급 계약 비용을 지급을 위해 계약서와 별도로 부속 합의를 체결하였다. 부속 합의서에는 프로그램명, 수집 기간, 수집 편수, 회당 시간, 지급 금액 등을 명시하였고, 본 사업 진행 시 프로그램별로 구어 자료 추가 수집이 필요할 경우 부속 합의를 갱신하거나 차수를 추가하여 별도 계약을 체결하였다.

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

저작자 및 저작권 이용허락자 ㈜조선방송(이하 "권리자"이라 함)과 저작권 이용자 국립국어원(이하 "이용자"이라 함)은 아래 저작물에 관한 저작권산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

제1조 (계약의 목적)
본 계약은 저작권산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)
본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상저작물")에 대한 저작권 재산권 중 당사자가 합의한 권리로 한다.

저작물:
<신동방송> <보도본부 핫라인> <새한비밀24> <강적들> <내 몸 사용설명서> <남한는개>
저작자: ㈜조선방송
종별: ☒ 영상저작물
권리: 이용허락 대상 권리는 아래와 같다.

※ 저작권 이용허락 대상 권리의 내용

- 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 대상 저작물의 음성을 청취·전사한 텍스트를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
- 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자음, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물의 음성을 청취·전사하여 텍스트로 변환하고 그 텍스트를 복제·변형(독자·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어회·비언어회 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

부속 합의서

일 자 2019년 08월 09일 장 소 (주) 숭례방송 주 소 서울 중구 세종대로 21길 40 조선일보 씨스퀘어빌딩	장 소 (주) 숭례방송 주 소 서울 강남구 테헤란로 538 대우빌딩 3~5층
--	---

국립국어원의 '국어 언어 자원(말뭉치) 구축 및 활용 저작권 계약서'를 위해 숭례방송은 TV조선 측에서 제공하는 프로그램에 대해 아래와 같이 비용을 지급한다.

합계금액	시간당 25,000원 × 2,109.3시간 = 52,732,500원
------	---------------------------------------

프로그램명	수입기간	카테고리	수입원수	회당시간	총시간	비고
신동방송	2017. 11~2019. 05	시사 및 정치	370	90	555.0	TV
보도본부 핫라인	2017. 11~2019. 05	시사 및 정치	370	90	555.0	TV
사건파워24	2017. 07~2019. 05	시사 및 정치	470	80	626.7	TV
강적들	2017. 01~2019. 05	시사 및 교양	110	100	183.3	TV
내몸사용설명서	2018. 01~2019. 05	교양/다큐	67	80	89.3	TV
남한는개	2014. 01~2016.02	교양	100	60	100.0	TV
TV 방송 시간 소계					2,109.3	

* 프로그램 편성은 계약 후 90일 내에 TV조선과 숭례방송 협의한 일정에 따라 전달한다.
* 해당 부속합의서에 추가 프로그램 사용 시 별도 부속합의서를 제출한다.

그림 2-1 저작권 이용 허락 계약서 및 부속 합의서

1.3. 방송 구어 자료 수령

구어 자료 공급 계약 체결 후 음성에 대한 품질 확인을 위해 방송사별로 샘플 확인을 진행하였다. 구어 자료 공급 계약을 체결한 모든 방송사의 프로그램을 대상으로 원시 말뭉치 구축을 위해 음성 품질이 이상이 없는지 적정성 여부를 판단하였다. 방송사별로 음성 파일을 준비하는 시간이 필요하여 총 4개월간 순차적으로 음성 파일을 수령하였다.

2. 말뭉치 구축 대상 선정

2.1. 수집 목록 취합

본 사업에서 구축한 원시 말뭉치가 산업계 및 학계에서 유용하게 활용되도록 하기 위해서는 산업계와 학계의 전문가가 실제로 수집된 자료를 검토하는 것이 필요하다. 본 사업의 원시 말뭉치 구축 목표는 구어 15,000시간, 준구어 15,400,000어절이지만 전문가 검토 시 원시 말뭉치 구축 대상에서 제외될 자료가 있을 것으로 판단하여 본 사업의 원시 말뭉치 구축 목표에 비해 약 1.5~2.0배 이상의 방송 자료를 수집하였다.

2.2. 전문가 분석

취합된 수집 목록은 산업계와 학계의 전문가가 구어와 준구어 자료의 유용성 여부를 검토한 후 최종적으로 원시 말뭉치 구축 대상으로 선정하였다. 구어 자료 분석 시에는 자료의 유형과 주제, 발화 인원에 따른 분류, 발화자의 연령 및 지역에 따른 분류 등이 고려되어야 한다. 또한 구어 자료 내용 확인 시 중복되는 어절의 반영 방안, 잡음 발생 시 처리 방안 등 원시 말뭉치 구축 시 지침 사항을 확인하고 메타 정보를 확정하는 것이 필요하다. 준구어 자료 분석 시에는 산업계와 학계에서 필요한 주제 선정 여부와 그에 맞는 다양한 대본 대상 선정이 필요하다. 자연어 처리, 빅데이터 분야의 언어 인공 지능 기술 전문가, 국어국문학과 등 언어학 전공 교수진, 말뭉치 구축과 연관성이 높은 산업계 연구원 등의 전문가가 유용성을 분석하여 최종 구축 대상을 선정하고, 원시 말뭉치 구축 시 메타 정보 입력 지침을 최종 확정하였다.

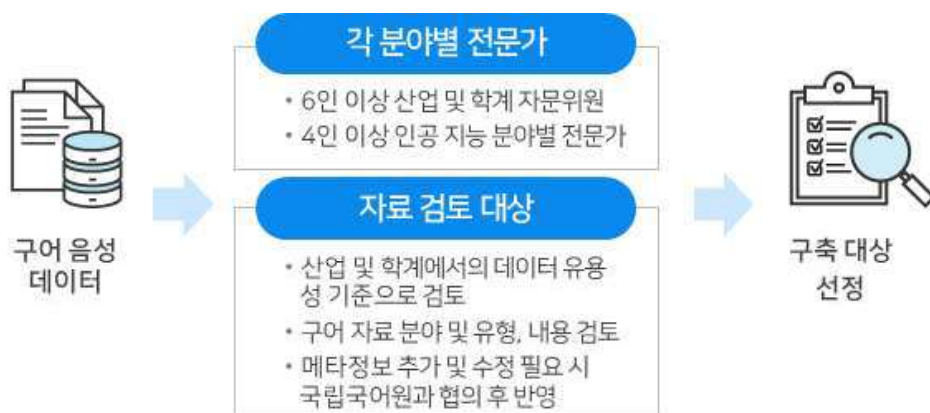


그림 2-2 전문가 분석을 통한 구축 대상 선정

구어의 경우 방송사별로 수집 가능한 대상을 분석하였고 발화 유형 및 주제 분류, 수집 편수 등을 확인하였다. 방송 자료를 대상으로 산업계, 학계 자문 위원에게 유용성 여부를 검토를 진행하였다. 원시 말뭉치 구축 시 최종 대상에서 제외될 가능성을 생각하여 구어 자료는 구축 목표에서 약 50시간을 추가하여 TBS 4,994시간, TV조선 4,098시간, MBC 3,605시간, EBS 1,307시간 등 총 15,052시간 분량을 선정하였다.

방송사	방송 시간	비율
TBS	4,994	33.18%
TV조선	4,098	27.23%
MBC	3,605	23.95%
EBS	1,307	8.68%
과학하고 앉아있네	422	2.80%
마인드 코칭 연구소	315	2.09%
펄스교양	311	2.07%
합계	15,052	100%

표 2-1 방송사별 구어 자료 선정 시간

준구어의 경우 15,400,000어절 구축을 위해서 시간당 7,000어절을 예상하여 방송 시간 기준으로 약 2,200시간 분량의 대본 수집이 필요하지만, 추가 필요 가능성을 생각하여 총 3,064시간 분량의 드라마 대본을 선정하였다.

방송사	방송 시간	비율
MBC	3,064	100%
합계	3,064	100%

표 2-2 방송사별 준구어 자료 선정 시간

구축 대상 선정 목록 중 구어의 경우 주제의 편중을 방지하고, 다양성을 확보하기 위해 과학, 교양, 교육, 문화, 법률, 스포츠, 예능, 건강, 시사 등 총 18가지 분야로 선정 대상을 분류하였다.

주제	방송 시간	비율
시사정치	5,357	35.6%
시사교양	2,179	14.5%
생활정보	1,525	10.1%
문화예술	911	6.1%
교육	739	4.9%
교양	708	4.7%
시사문화	564	3.7%
예능	437	2.9%
과학	422	2.8%
법률	344	2.3%
토론	321	2.1%
상담	315	2.1%
토크쇼	285	1.9%
보도토론	249	1.7%
도서	242	1.6%
문화	235	1.6%
스포츠	137	0.9%
건강	82	0.5%

표 2-3 구어 자료 선정 대상 주제 분류

준구어의 경우 대본 수집 시 주제의 편중을 방지하기 위해 가족, 로맨스, 범죄 등 16가지 분야로 선정 대상의 주제를 분류하였다.

주제	방송 시간	비율
가족	977	31.9%
로맨스	680	22.2%
사극	414	13.5%
직업	297	9.7%
시대	184	6.0%
법률	120	3.9%
범죄	84	2.7%
판타지	68	2.2%
요리	54	1.8%
의학	50	1.6%
코믹	39	1.3%
첩보	34	1.1%
음악	18	0.6%
스포츠	17	0.6%
방송국	16	0.5%
호러	12	0.4%

표 2-4 준구어 자료 선정 대상 주제 분류

2.3. 말뭉치 구축 대상

최종적으로 구어 말뭉치 구축 대상으로 총 67종의 방송 프로그램을 확정하였다.

번호	방송사	프로그램명	시간(분)	파일 수	방송 연도	주제
1	EBS	공감시대	34,953	745	2018~2019	시사교양
2	EBS	교육대토론	5,628	72	2013~2018	교양
3	EBS	미래 교육 플러스	1,301	28	2019	교육
4	EBS	생각하는 콘서트	1,998	41	2018~2019	교양
5	EBS	스타특강	6,484	137	2015~2016	교육
6	EBS	지식의 기쁨	3,257	84	2019	교양
7	EBS	초대석	8,216	172	2013~2019	교양
8	EBS	최고의 요리비결	16,590	668	2017~2019	교양
9	과학하고 앉아있네	과학하고 앉아있네	25,322	241	2013~2019	과학
10	마인드코칭연구소	참나원	18,918	586	2017~2019	상담
11	필스교양	필스교양	18,645	287	2018	교육
12	MBC	100분 토론	19,280	244	2013~2019	토론
13	MBC	2시 뉴스 외전	14,944	174	2018~2019	보도토론
14	MBC	기분 좋은 날	45,006	736	2016~2019	생활정보
15	MBC	김신영의 TMI	181	11	2018~2019	토크쇼
16	MBC	능력자들	2,848	37	2015~2016	생활정보
17	MBC	도전 발명왕	1,534	29	2013~2014	생활정보
18	MBC	라디오 스타	21,057	297	2013~2019	예능
19	MBC	문화사색	14,073	318	2013~2019	문화
20	MBC	베란다쇼	4,974	216	2013~2014	토크쇼
21	MBC	블라인드 테스트	1,942	28	2013	예능
22	MBC	생방송 오늘 아침	41,656	803	2016~2019	생활정보
23	MBC	세바퀴	8,782	115	2013~2015	토크쇼
24	MBC	스트레이트	2,459	51	2018~2019	시사 문화
25	MBC	스포츠	6,038	37	2013~2018	스포츠
26	MBC	여성토론	3,892	68	2013~2014	시사 문화
27	MBC	오빠생각	1,245	16	2017	예능
28	MBC	이슈를 말한다	8,314	182	2014~2018	시사문화
29	MBC	침착한 주말	100	6	2019	생활정보
30	MBC	탐나는 TV	2,223	41	2018~2019	교양
31	MBC	통일전망대	11,238	273	2013~2019	시사 문화
32	MBC	판결의 온도	365	8	2018	생활정보
33	MBC	황금어장 무릎팍도사	2,002	28	2013	예능
34	MBC	MLB 핫토크	2,158	38	2015~2019	스포츠
35	TBS	기분좋은 토요일 조현아입니다	4,266	168	2017~2019	문화예술
36	TBS	김규리의 풍당풍당	4,641	153	2019	문화예술
37	TBS	김어준의 뉴스공장	41,979	896	2018~2019	시사정치
38	TBS	김필수의 교통시대	4,349	96	2017~2019	법률
39	TBS	더 룸	3,178	53	2019	토크쇼
40	TBS	라디오 와이파이	11,976	664	2016~2019	문화예술
41	TBS	마음 산책	5,722	219	2017~2018	문화예술
42	TBS	민생연구소	7,922	166	2019	시사문화
43	TBS	상담받고 대학가자	17,889	361	2017~2019	교육
44	TBS	색다른 시선	10,200	245	2019~	시사정치

번호	방송사	프로그램명	시간(분)	파일 수	방송 연도	주제
		이숙이입니다				
45	TBS	서울속으로 항원찬입니다	16,289	874	2016~2019	법률
46	TBS	송정애의 좋은사람들	20,269	1459	2017~2019	문화예술
47	TBS	시시각각	17,028	365	2017~2019	시사정치
48	TBS	열린 아침 김만흠입니다	16,709	451	2016~2019	시사정치
49	TBS	이슈파이터	25,032	333	2017~2019	시사정치
50	TBS	최일구의 허리케인 라디오	46,439	1370	2017~2019	시사교양
51	TBS	팩트인스타	7,780	147	2016~2019	문화예술
52	TBS	품격시대	23,443	311	2016~2018	시사정치
53	TBS	TV책방 북소리	14,492	286	2014~2019	도서
54	TV조선	강적들	10,343	259	2017~2019	시사정치
55	TV조선	낭만논객	4,544	100	2014~2016	교양
56	TV조선	내 몸 사용 설명서	4,944	109	2017~2019	건강
57	TV조선	박종진 라이브쇼	10,627	143	2016~2017	시사정치
58	TV조선	보도본부 핫라인	33,818	475	2017~2019	시사정치
59	TV조선	사건파일 24	36,550	546	2017~2019	시사교양
60	TV조선	시사탱크	59,871	1229	2012~2016	시사정치
61	TV조선	시사토크 판	11,727	445	2011~2013	시사교양
62	TV조선	신율의 시사열차	8,094	194	2012~2013	시사정치
63	TV조선	신통방통	36,723	551	2017~2019	시사정치
64	TV조선	이슈본색	4,359	71	2016	시사정치
65	TV조선	이슈 해결사 박대장	11,798	194	2015~2016	시사정치
66	TV조선	이하원의 시사큐	11,403	176	2015~2016	시사정치
67	TV조선	최희준의 왜	1,101	24	2016~2017	시사교양

표 2-5 구어 말뭉치 구축 대상 목록

준구어 말뭉치 구축 대상으로 총 75종의 방송 프로그램을 확정하였다. 말뭉치 구축 대상으로 선정된 자료는 2004년부터 2019년까지 방송된 드라마 대본이다.

번호	프로그램명	작가	시간(분)	방송 연도	주제
1	그대없인못살아	김선영	4,180	2012	가족
2	그분이오신다	신정구	2,522	2008	가족
3	금나와라뚝뚝	하청옥	3,600	2013	가족
4	내생애봄날	박지숙	1,040	2014	가족
5	내생애마지막스캔들	문희정	1,040	2008	가족
6	당신참예쁘다	오상희	4,995	2011	가족
7	민들레가족	김정수	2,850	2010	가족
8	밥집	서영명	3,885	2009	가족
9	보석비빔밥	임성한	3,250	2009	가족
10	사랑해서남주나	최현경	3,550	2013	가족
11	살맛납니다	박현주	4,921	2009	가족
12	양큼한돌싱녀	이하나, 최수영	1,056	2014	가족
13	엄마의정원	박정란	4,410	2014	가족
14	오늘만갈아라	최현경	4,864	2011	가족
15	왔다장보리	김순옥	3,692	2014	가족

번호	프로그램명	작가	시간(분)	방송 연도	주제
16	천하일색박정금	하청옥	2,860	2008	가족
17	하얀거짓말	조은정	5,883	2008	가족
18	개인의취향	이새인, 김희주	1,168	2010	로맨스
19	남자가사랑할때 (동명드라마유익)	김인영	1,278	2013	로맨스
20	넌어느별에서왔니	정유경	1,120	2006	로맨스
21	닥터깽 (Dr.깽)	김규완	1,040	2006	로맨스
22	대장금이보고있다	박은정, 최우주	990	2018	로맨스
23	데릴남편오작두	유윤경	1,440	2018	로맨스
24	마이프린세스	장영실	1,168	2011	로맨스
25	멈출수없어	김홍주	4,902	2009	로맨스
26	메리대구공방전	김인영	1,040	2007	로맨스
27	빛나는로맨스	서현주	4,636	2013	로맨스
28	쇼핑왕루이	오지영	1,056	2016	로맨스
29	여자를올려	하청옥	2,698	2015	로맨스
30	운명처럼널사랑해	주찬옥, 조진국	1,188	2014	로맨스
31	잘했군잘했어	박지현	2,280	2009	로맨스
32	최고의사랑	홍정은, 홍미란	1,152	2011	로맨스
33	최고의연인	서현주	4,370	2015	로맨스
34	한번더해피엔딩	허성희	1,056	2016	로맨스
35	흔들리지마	이홍구	5,740	2008	로맨스
36	폭풍의연인	나연숙	2,484	2010	로맨스
37	스포츠라이트	황주하, 최윤정	960	2008	방송국
38	검법남녀	민지은, 원영실	870	2018	범죄
39	라이프특별조사팀	김수진, 여은희, 최윤정	780	2008	범죄
40	아이템(ITEM)	정이도	496	2019	범죄
41	전설의마녀	구현숙	2,880	2014	범죄
42	개과천선	최희라	1,056	2014	법률
43	아현동마님	임성한	6,120	2007	법률
44	구암허준	최완규	4,860	2013	사극
45	기황후	장영철, 정경순	3,366	2013	사극
46	불의여신정이	권순규, 이서윤	2,272	2013	사극
47	빛나거나미치거나	권인찬, 김선미	1,560	2015	사극
48	왕은사랑한다	송지나	640	2017	사극
49	이산	김이영	5,005	2007	사극
50	제왕의딸수백향	황진영	3,852	2013	사극
51	화정	김이영	3,300	2015	사극
52	9회말2아웃	여지나	1,040	2007	스포츠
53	글로리아	정지우	3,300	2010	시대
54	메이퀸	손영목	2,812	2012	시대
55	영웅시대	이환경	4,900	2004	시대
56	내이름은김삼순	김도우	960	2005	요리
57	신들의만찬	조은정	2,272	2012	요리
58	베토벤바이러스	홍진아, 홍자람	1,105	2008	음악
59	뉴하트	황은경	1,380	2007	의학
60	닥터진	한지훈, 전현진	1,606	2012	의학
61	다시시작해	원영옥	4,598	2016	직업
62	더뱅크	서은정, 오혜란, 배상옥	512	2019	직업
63	밤이면 밤마다	윤은경, 김은희	1,105	2008	직업
64	스탠바이	박민정, 김윤희, 박재현, 이은영, 양서윤	2,886	2012	직업
65	압구정백야	임성한	4,884	2014	직업

번호	프로그램명	작가	시간(분)	방송 연도	주제
66	옥션하우스	김남경, 권기경, 진현수	780	2007	직업
67	자체발광오피스	정희현	792	2017	직업
68	호텔킹	조은정	2,272	2014	직업
69	나도꽃	김도우	1,080	2011	첩보
70	에어시티	이선희, 이서윤	960	2007	첩보
71	안녕프란체스카시즌3	김현희	2,350	2005	코믹
72	궁	인은아	1,344	2006	판타지
73	밤을걷는선비	류용재, 장현주	1,320	2015	판타지
74	아랑사또전	정윤정	1,440	2012	판타지
75	혼(납량특집혼)	고은님, 인은아	700	2009	호러

표 2-6 준구어 말뭉치 구축 대상 목록

3. 작업자 교육

본 사업에서는 구어 15,000시간 분량과 준구어 15,400,000어절의 대량의 원시 말뭉치를 구축해야 하므로 월별 최소 200명의 말뭉치 작업자가 투입되어야 한다. 사업 책임자 및 품질 부문 관리자는 적정 자격 조건을 통과한 작업자를 대상으로 집체 교육 및 단계별 맞춤형 실무 중심의 교육을 진행하였고, 말뭉치 품질을 관리하고 오류 발생 시 올바른 해결안을 도출하기 위해 주 1회 사례를 공유하는 회의를 개최하였다.

3.1. 말뭉치 작업자 선발

원시 말뭉치 구축을 위해 사업 기간 동안 월별 최소 200명의 작업자 투입이 필요하였다. 말뭉치 구축 작업 지원자를 대상으로 작업자 교육 및 샘플 테스트를 진행하였고, 품질 점검팀이 작업 결과물의 품질을 확인하여, 샘플 테스트 품질 점수가 99.5% 이상인 지원자를 실제 말뭉치 구축 작업에 투입하였다. 월별 최소 200여명의 인원이 필요하였기 때문에 사업 기간 내내 작업자 선발을 상시 진행하였고, 월별 200~220명의 인원을 유지하면서 원시 말뭉치 구축을 완료하였다.

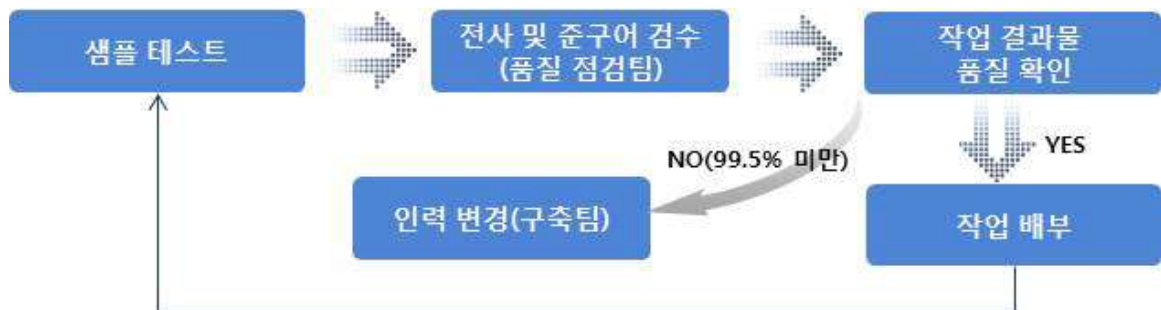


그림 2-3 샘플 테스트 절차

작업자의 작업 결과물 품질 확인을 위한 샘플 테스트는 치명적인 오류, 맞춤법 오류, 태그 오류, 문장 부호 오류 여부 등의 검사 내용을 기준으로 각각 -0.1~-1.0의 평가 지수를 적용하여 작업자별 품질 지수를 확인하였다.

검사 항목	검사 내용	평가 지수
치명적인 오류	문장 누락, 발화자 정보 누락, 발화 누락	-1.0
맞춤법 오류	오타자, 의미를 해치는 띄어쓰기 오류	-0.3
태그 오류	웃음, 박수, 노래 등 태그 처리, 비식별화 태그 처리	-0.2
표기 오류	숫자나 기호, 영문 등의 발음에 따른 한글 표기	-0.3
문장 부호 오류	마침표를 제외한 부호의 사용 여부	-0.1

표 2-7 샘플 테스트 평가 지수

[390회] EBS_조대석_박선희 * x

1 1: 정관용, 50대, 남자. ①
 2 2: 김산하, 40대, 남자, 생명다양성재단, 사무국장
 3
 4 1: 여러분 안녕하십니까.
 5 EBS, ② 씨, 정관용입니다.
 6 지구온난화와 기상이변, 어거, 이제, 이상한게, ③ 이다.
 7 거의, 일상이죠.
 8 그리고, 이걸, 모든, 것이, 인간, 잘못이라는, 것, 다, 알고, 있습니다. ④
 9 그럼에도, 불구하고, 우리는, 매일, 자동차를, 타고, 무심코, 쓰레기를, 버리고, 있죠. 하하. ⑤
 10 자, 이걸, 한, 마디로, 어~, 우리, 사람들이, 생태감수성을, 잃어버렸기, 때문이다.
 11 이렇게, 주장하시는, 분이, 있어서, 오늘, 초대석에, 모셔, 왔습니다.
 12 우리나라, 최초의, 야생, 영장류, 학자고, 생태학자인, 김산하, 박사를, 초대했습니다.
 13 어서, 오십시오, 김박사님.
 14 2: 안녕하세요.
 15 1: 네, 생태감수성이란, 단어가, 있나요?
 16 2: 예, 뭐, 사실, 그, 학계에서, 공식적으로, 인정하는, 단어라고, 할, 수는, 없지요.
 17 어떻게, 보면, 이제, 일상언어랑, 지금, 현재, 지구가, 겪고, 있는, 위기를, 생각해서, 많은, 사람들, 회자하는, 걸, 제가, 그냥, 골라서, 사용하는, 단
 18 1: 어떤, 뜻인가, 생태감수성?
 19 2: 예, 제가, 생각하기에는, 세상에, 대한, 나의, 상관. ⑥
 20 2: 어떻게, 표현하고, 싶다면요.
 21 좀, 나와, 세상이, 어떻게, 연결되어, 있는,지에, 대한, 감.
 22 1: 음.
 23 2: 그런, 거에, 대한, 생각과, 어떤, 지식, 이런, 것들을, 총체를, 해서, 생태감수성이라고, 저는, 부릅니다.
 24 예를, 들어서.
 25 뭐, 얼마, 전에, 비가, 굉장히, 많이, 왔는데요.
 26 1: 예.
 27 2: 비가, 많이, 오면, 일상생활이, 불편하잖아요.
 28 1: 음.
 29 2: 근데, 어~, 사실, 그, 비가, 우리나라에는, 이제, 여름에, 집중이, 돼서, 나중엔, 가을이나, 겨울, 특히, 그, 가을이, 심할, 때, 우리가, 바로, 마시.

1 발화자 정보 누락

정관용, 50대, 남자,
↓
정관용, 50대, 남자, **앵커**

2 영문 표기 오류

EBS
↓
이비에스

3 띄어쓰기

이상한게
↓
이상한 게

4 오타 및 문장부호

있습니다
↓
있습니.

5 태그 오류

하하
↓
@웃음

6 발화 누락

※동시 발화로 누락됨
↓
1: 그렇군요.

그림 2-4 샘플 테스트 품질 평가

자료명	작업자	배부	완료	어절	치명적인 오류(1.0)	맞춤법 오류(0.3)	태그오류 (0.2)	표기오류 (0.3)	문장부호 오류(0.1)	순도	평가
373회_EBS_조대석	김*숙	11/20	11/21	5,784	5	19	9	8	6	99.71%	합격
374회_EBS_조대석	박*현	11/20	11/21	5,911	4	10	5	6	4	99.82%	합격
375회_EBS_조대석	이*진	11/20	11/21	5,852	10	40	15	11	15	99.42%	대체
376회_EBS_조대석	최*용	11/20	11/21	5,633	3	14	9	8	6	99.78%	합격
377회_EBS_조대석	김*숙	11/22									
378회_EBS_조대석	박*현	11/22									

그림 2-5 작업자 품질 평가표

3.2. 작업자 교육

본 사업 기간 동안 각 단계별 작업자 교육 지원 조직을 통해 말뚝치 작업자를 대상으로 사업의 이해를 도울 수 있는 기본 교육과 실무 중심의 작업자 교육을 진행하였다. 구축 및 품질 검수 단계 중 난도 높은 자료를 담당하는 작업자에게는 반복 교육을 통해 이해도를 높이고, 품질 위험 요소를 사전에 방지하였다.

작업자 교육 지원 조직	역할
사업 책임자	교육 지원(교재, 교육 환경 등) 보안 및 사업 관련 기본 교육
구축 부문	말뚝치 구축 지침서 작성 지침서와 사례와 실습을 통한 작업자 대상 교육
품질 부문	수집 및 구축 부문 품질 지침서 작성 품질 관리 담당자 대상 교육
기술 부문	소프트웨어 사용법 및 발생 가능한 오류와 대응 방법 교육

표 2-8 작업자 교육 지원 조직

매주 1회 작업자의 구축 지침 교육을 통해 자료 구축 공정 지침 및 사례를 통한 오류 요소를 공유하였고, 품질 지침 교육을 통해 자료의 검사 방법 등을 전달하였다. 작업자 교육을 위한 구축 지침서와 품질 지침서를 작성하였고, 작업 도구에 익숙하지 않은 작업자를 대상으로 별도의 작업 도구 교육을 진행하였다.

구분	구축 지침 교육	품질 지침 교육
시기	매주 1회	매주 1회
대상	구축 작업자	구축 작업자, 품질 작업자
교육 내용	자료 구축 공정 지침, 사례를 통한 오류 요소 공유	품질 검수 공정 지침, 자료 검사 방법 및 작업 도구 사용법 교육
교육 자료	구축 지침서	품질 지침서

표 2-9 작업자 교육 과정

현대 국어 구어 전사 말뭉치 지침
2019.07. -Ver.1.2

1. 사용도구

- 가) EmEditor를 다운로드 사용한다.(<https://enemeditor.com/>)
나) 배운된 음성 파일을 듣고 EmEditor에 전사하는 방식으로 작업한다.
다) 작업 후 파일저장증 UTF-8(서명 없음)로 저장한다.

2. 전사지침

가) 발화자 표시

- 가) 모든 발화자에 관한 정보는 파일 맨 상단에 [예시]와 같이 붙인다.
• 가급적 후발화자를 1로 표시하고, 나머지는 통장 순서대로 한다.
• 발화자 정보는 이름,성별,연령대,직업을 간단하게 표시한다.
• 발화자에 대한 정보를 모를 경우에는 'X'로 표시한다.

[예시]

1. 알권도남35.0세,지자명본기
2. 이나미여자35정신과 전문의
3. 9.9.9.9 <=> 청중

- 나) 본문 전사에서 발화자 정보와 발화자 표시는 반드시 일치해야 한다.
• 발화자가 분명하지 않을 경우에는 'X'로 표시한다.
• 필요에 따라서는 '모두'나 '나머지' '청중' 등의 지칭을 사용할 수 있다.
• 화자 2와 화자 3이 동시에 말하는 경우는 '2,3'으로 표시하기도 한다.

[예시]

1. 이 프레임을 진행하신 이유가 궁금하네요.
2. 저 같은 경우에는
3.4 예
모두 아-

- 다) 발화자 표시는 나중에 T로 태그로 일괄 전환되어야 하므로 일관성 있게 표시한다.

나) 전사 단위 기준

- 가) 기본 전사 단위는 줄이며 기호로 한다.

[예시]

1. 그리고 일본 정부도 더 이상 도쿄전력에만 사고처리를 맡일 수 없고 우리
가 이제 직접 통제를 하겠다고 전 세계에 공언을 하면서 이제 이 문제가
다시 불거졌는데 사실은 일본 정부가 인정을 한 거죠.
2. 네

단, 부사나 형용사가 종결 어미 뒤에 나오는 도치된 발화된 경우는 한 문장으로 본다.
(발화 내용이 한 문장으로 말하는 내용이면 한 문장으로 본다)

[예시]

1. 어제 몇 갔다고?
2. 제가 사과를 샀어요. 어제
1. 커는 정말 좋아해요. 사과를

공정의문, 부정의문의 경우도 동일하다.

[예시]

1. 할까요? 안 할까요?
1. 맞아요? 틀려요?

- 나) **간접에 의해** 문장이 바뀌거나 발음이 호러진 채로 발화가 종료되면 한 문장으로 보고 전
사한다.

[예시]

2. 당직은 마담이 아니다.
그러니까 이제 비주류 사이티라고
1. 음

- 다) 느낌표나 쉼표는 사용하지 않는다. 문장이 완전히 종결이 되었을 때에 마침표를 사용한다.

[예시]

1. 저 이제 유럽에 처음 왔는데 음~ 이전엔의 모습과 많이 달라졌습니
다. 어서 오십시오.
2. 네. 안녕하세요.
1. 날씨 변화가 심하더라면 할 말 말이 많이 줄었을 거 같은 ㅎㅎ음 생각이
2. 그렇죠.
1. 네.
2. 네.
날씨는 정말 영악한 화제인 거 같아요.

- 예) 악양에 의해 의미가 달라지는 경우 마침표와 줄임표를 사용하여 구분에 준다.(-어, -어요
등)

[예시]

1. 어제 새로 기록한 영화 줄어
1. 어제 새로 기록한 영화 줄어?

1

2

다) 발화자 생치는 경우

- 가) 결점 발화는 시간 순서에 따라 적는다.

만약 발화자 발화가 있을 경우 순서에 맞춰 뒷말과 발화자 사이에 넣어 조발화를 나눈다.
네, 예 등의 뒷말과 발화는 한 문장으로 보고 전사한다.

[예시]

1. 국회는 열심히 해왔어 갯에서
2. 어
1. 막 분위기에서 경매인 요일에 갔다 버렸는데
2. 네.
1. 생각 수기해한 사람은 항상 그냥 비리드라 ㅎㅎ음
2. 네.
1. 소감장으로 다 들어가 버리더라
2. 네.

- 나) 상대방 앞에 뒷말과 발화에서 네네, 예예, 음음 등의 경우는 붙여서 전사한다.

[예시]

1. 다시 올리겠다는 건데
2. 올리겠다는 거 아니에요?
1. 네네. 대단합니다

라) 전사 기준

- 가) 발화 내용은 기본적으로 맞춤법을 기준으로 전사한다.

다만, 구어의 발음 특성, 개인의 발음 특성, 지역적인 특성 등에 의해
맞춤법대로 소리 나지 않는 발음(표준 발음이 아닌 경우)은 발음 나는 대로 적는다.

[예시]

1. 어~ 기호가 몇 줄지 않아요 그런거요?
2. 그분 거 기호요

- 나) 모음의 변화, 후의적 경음화 등을 반영하여 소리 나는 대로 전사한다.

[예시]

1. 요즘 한자 먹느라 오늘 죄주는 못 마시겠네요.
2. 그럼 어쨌게 회람도 못 마시요?
1. 한잔도 못 마시요. 이제

- 다) 약화 현상에 의한 이형대는 반영하지 않는다.

예를 들어 의문사 '뭐'가 '마', '모'로 모음이 약화되어 플리도 '뭐'로 전사한다.

- 예) 숫자나 기호, 영문 등도 발음에 따라 한글로 전사한다

[예시]

1. 네.
아이템에서 사해 직후에 그 논쟁의 실업자를 비디오 다들 보고 말
죠.
2. 예
1. 어~ 음 모기가 운동도 하고
2. 맞습니다.
1. 그리고 불과 된 한 일 년 남짓 후에?

- 나) 불완전 발화나 수정 발화의 경우에 구별하는 표시 's'를 한다.

• 불완전 발화는 단어가 끊어지거나 불완전하게 발화된 경우를 말한다.

• 수정 발화는 발화된 내용을 전체적으로 수정한 발화를 말한다.

• 발화된 그대로 전사하며, 's'를 붙여 정상적인 단어와 구별한다.

• 불완전하게 발화된 단어(어절)가 줄 이상인 경우 어절마다 's'를 붙인다.

• 단, 반복 발화에는 표시하지 않는다.

[예시]

1. 전제가 뒤배를 기쁘게는 구조를 s를 덧붙여는 거고 있고 있거든요(불완전 발
화)
2. 그럼요.
1. 그래서 하부트파 한~ 학생들끼리 이 공부하는 건데(불완전 발화)
2. 우리가 하는 생각할 수 있는 모든 날씨 얘기가 다 나옵니다(수정 발화)

예외) 그 순지 순지 나기때다가 줄 중절을 둔 듯한 그런 느낌이에요(반복 발화)

- 나) 띄어쓰기의 경우 맞춤법에 맞게 한다.

• 합성명사의 경우 맞춤법에 맞게 쓴다.(예, 발화 단위 → 발화단위)

• 의존명사는 띄어 쓴다.

• 수를 적을 때는 만 단위로 띄어 쓴다.(예, 십이억 삼천백만 팔백구 열러 등)

• 수와 단위가 함께 쓰이는 경우 띄어 쓴다.(예, 구십구 점 오 프로, 천구백삼십 년대 등)

• 판단하기 어려운 경우에는 수시로 논의하여 결정한다.(예, 오십대, 일 대 이 등)

• 본 음원과 비교 용도 띄어 쓴다.(예, 구입하기도 했다, 드라마가 보고 싶다 등)

- 나) 속삭임의 표기(정확한 발음이 나타난 경우에만 반영한다.)

• 두 음절이 한 음절의 사잇소리가 되거나, 두 음절이 한 음절 겹침소리가 되는 등의 경우

• 발음되는 음절소리와 무개상의 음절소리를 맞추어야 하므로 속삭임의 경우 무도 무기에 반

영한다.

[예시]

1. 그나란 그거 말에 일부 줄 수가 없는 거요.
2. 아유 그런 어때요?

- 받모음 /r/, /l/, /r/의 경우 /r/, /r/와 속삭임은 현상이 구에서 자주 나타나는데,

한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로

전사 시 {작은따옴표, under}를 사용해서 두 음소를 연결한다.

[예시]

그림 2-6 구축 지침 교육 자료(일부)

3.3. 보안 교육 및 관리

본 사업에 참여하는 관리자, 작업자 등 모든 사업 참여자를 대상으로 보안 교육을 실시하였다. 보안 업무 취급 규정 및 보안 대책을 철저히 준수하며 물리적, 관리적, 기술적 측면의 구체적인 보안 대책을 강구하고, 저작권을 존중하는 총체적인 보안 대책을 수립하였다. 기밀 보안을 위해서 정보 보안 담당자가 사업 착수, 사업 진행, 사업 종료 시에 절차에 따라 보안 점검을 진행하였다.

역할	내용
정보 보안 담당자 보안 착수 지원	<ul style="list-style-type: none"> - 사업 참여자 보안 교육 실시 - 사업 참여자별 보안 서약서 작성 - PC 및 네트워크 보안 - 사업을 위한 별도의 작업장 구축
정보 보안 담당자 보안 점검	<ul style="list-style-type: none"> - 물리적 보안 <ul style="list-style-type: none"> 별도의 자료 보관함 비치 중요도에 따른 자료 분류 및 보관 - 기술적 보안 <ul style="list-style-type: none"> 보안키를 통한 작업장 출입 작업장 내 CCTV 설치 사내 서버실 설치 - 관리적 보안 <ul style="list-style-type: none"> 보안 책임자 지정 작업자에 대한 보안 교육 실시
보안 관리자 보안 점검 확인	<ul style="list-style-type: none"> - 사업 완료 후 작업자 PC 내 중요 정보 삭제 및 초기화 - 사업 완료 후 작업자별 서버 접속 계정 삭제

표 2-10 보안 교육 및 관리

4. 원시 말뭉치 구축 및 메타 정보 구축

4.1. 구축 절차

말뭉치 구축 대상으로 선정된 음성 및 영상, 대본 등의 자료를 대상으로 헤더를 부착하고 목록을 작성하였다. 구어와 준구어 원시 말뭉치 구축을 위해 지침에 따라 구축 작업을 수행하고 말뭉치의 정확성 확보를 위한 검수를 진행하였다.



그림 2-7 말뭉치 구축 절차

구축이 완료된 구어, 준구어 말뭉치를 대상으로 XML 변환 프로그램을 이용하여 마크업 변환을 진행하였다. 검수 완료된 자료는 마크업 변환 프로그램을 이용하여 XML 변환을 검증하였고, 검증이 완료된 말뭉치는 DB 서버에 등록하여 저장하였다.

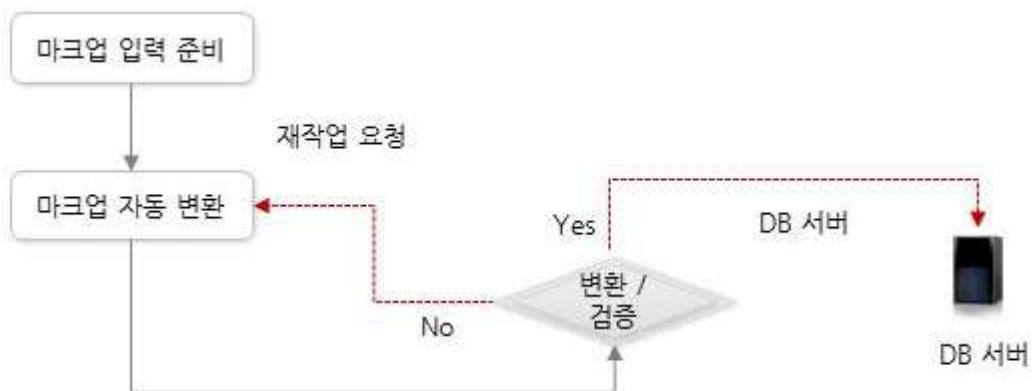


그림 2-8 마크업 변환 절차

4.2. 전사 지침

전사는 기본적으로 한글 맞춤법에 따르는 것이 원칙이나, 국립국어원과 협의한 현대국어 구어 전사 말뭉치 지침을 기준으로 구어의 발음 특성, 개인의 발음 특성, 지역적인 특성 등에 의해 표준 발음법에 따라 소리 나지 않는 경우에 모음의 변화, 수의적 경음화 등을 반영하여 소리 나는 대로 전사하였다.

분류	전사 지침(일부)
개요	자료의 전사는 발화된 그대로 전사하는 것을 원칙으로 함.
발화자 표시	모든 발화자에 관한 정보는 파일 맨 처음에 아래와 같이 붙임. 1: 홍민영, 20대, 여자, 교사. 2: ?, 10대, 남자, 학생. 3: ?, 30대, 남자, ?
단위	기본 전사 단위는 문장 단위로 함.
영문	발음에 따라 한글로 입력
숫자	숫자는 발화한 대로 한글로 표기함.
문장 부호	느낌표나 쉼표는 사용하지 않음. 문장이 완전히 종결이 되었을 때는 마침표를 사용
기타 소리	기타 소리 중 웃음, 목청, 박수, 노래에 대한 4가지는 태그로만 전사 기침, 들숨, 날숨, 재채기, 코흘쩍, 하품 등은 전사하지 않음.
축약형 표기	언어 경제성의 원칙에 의해 구어에서는 축약형이 많이 나타나며, 이는 모두 표기에 반영

표 2-11 전사 지침(일부)

발화자의 표시는 전사 지침에 따라 파일의 상단에 배치하였고, 본문 전사 시에 발화자 정보와 발화자 표시가 일치하도록 헤더와 전사 내용을 확인하였다. 가급적 주 발화자를 1로 표시하였고, 나머지 발화자는 등장 순서대로 표시를 하였다. 발화자에 대한 정보는 이름, 성별, 연령대, 직업으로 표기하였으나 불분명할 경우 '?'로 표시하였다.

1: 김혜지, 여자, 30대, 진행자
2: 이창현, 남자, 60대, 교수
3: 허희, 남자, 30대, 평론가
4: 이가희, 여자, 20대, 진행자

1: 책으로 세상을 바라고 책으로 삶을 살피우는 시간 티비 책방 북소리 북마스터 김혜지입니다.
2: 네, 안녕하세요.
국민대 언론정보학부 이창현 교수이구요,
여기서 북마스터하고 있습니다.
3: 네, 안녕하세요.
책읽고 들숨 쓰는 문학 평론가 허희입니다.
1: 네, 오늘은 영화 이야기로 먼저 시작을 해볼까 하는데요.
앤 헤서웨이 그리고 로버트 드 니로
2: 음...
1: 이 두 사람이 호흡을 맞췄던 영화 인턴 보셨나요?
2: 네, 봤습니다.
1: 하~
3: 거기서 보면 앤 헤서웨이가 성공한 청년 사업가잖아요.
그런데 자기가 사업을 해나가는데 그때 뭔가 좀 도움을 줄 사람이 있으면 좋겠다 그렇게 바라던 차에 한 회사의 중역이었던 로버트 드 니로가 인턴으로 들어오게 되잖아요.

그림 2-9 전사 예시(발화자 표시)

발음 전사 시 “이, 그, 저, 아, 어” 등 기존 품사의 의미, 기능을 가지지 않는 것은 담화 표지로 보고 ‘~’로 표시를 하며, 웃음소리나 박수 소리, 노래 등의 기타 소리는 ‘@’ 표시로 표시하여 전사하였다. ‘@’로 표기하는 요소는 @웃음, @목청, @박수, @노래 이상 4가지로 국한하였다.

1: 원종우, 남자, 50대, 과학과사람들대표
 2: 이강환, 남자, 40대, 천문학자
 3: 최진영, 여자, ?, 과학과사람들팀장
 4: ?, 남자, ?, ?
 1: 요즘 감기가 이 묘한 그 목감기 코감= 코감기 섞여서
 2: 네, 기침만 나오죠.
 1: 네 아주 막 아파서 막 쉬고 싶다거나,
 2: 네,
 1: 회사를 안 간다거나 그런 건 또 아니고
 3: @웃음
 1: @목청
 4: 이 좁은 데 있으면 씨 나까지 걸리는 거 아닌지 모르겠다.
 1: 여기 일하시는 분들까지 다.
 모두: @웃음
 2: 마이크를 통해서 다 전국에
 1: 그러게요.
 2: 마이크 소독하세요 끝나고
 1: 아 그럼 술을 시작을 해 볼까요.
 전문가와 비전문가가 함께하는 과학전문 팟캐스트 방송 과학하고 앉아있네.

그림 2-10 전사 예시(발음 전사)

발화자의 발화가 겹치는 경우 시간 순서에 따라 적으며, 만약 맞장구 발화가 있을 경우 순서에 맞춰 맞장구 발화를 사이에 넣어 주 발화를 나누었다. 또한 ‘네’, ‘예’ 등의 맞장구 발화는 한 문장으로 보고 전사하였다.

1: 원종우, 남자, 50대, 과학과사람들대표
 2: 이강환, 남자, 50대, 천문학자
 3: 최진영, 여자, ?, 과학과사람들팀장
 4: ?, 남자, ?, ?
 5: ?, 여자, ?, 질문자
 6: ?, 여자, ?, ?
 1: 제가 오프닝감으로 뭐를 좀 찾다가 보니까 우리 저 다저스 홈...
 2: 네.
 1: 그 다저스의 에이스가
 2: 네.
 1: 이 커쇼라는 선수거든요.
 2: 네.
 1: 클레이튼 커쇼가.
 2: 네, 클레이튼 커쇼 잘 생겼더라고요.
 1: 네.
 1: 그러게요.
 1: 잘 생겼더라고요.
 1: 네, 절고 잘 생기고
 2: 네.
 1: 돈 많고
 2,3: @웃음
 3: 갑자기
 1: 애가 보통 잘 하는 애가 아니에요.
 애가 재작년에 사이언상도 받았고
 2: 음
 1: 그 나이에 앞으로 소위 말하면 그 탑에이스들의 계보를 이어갈 그런 투수라고 하는데 이 친구가 그 명왕성 퇴출과 관련된 무슨 인터뷰를 했어요.
 2: 네, 저도 봤어요.
 1: 예, 이게 무슨 소리가 했더니
 1: 종조부?
 2: 예, 종조부 할아버지의 형제
 1: 형제죠.
 2: 네.

그림 2-11 전사 예시(겹침 발화)

4.3. 음성 인식 자동 전사

본 사업 진행 시 고품질의 원시 말뭉치 구축을 위해 음성 인식 기술을 테스트하였다. 방송국에서 수집한 음성 파일을 기준으로 사업 수행자의 음성 인식 기술을 적용하여 자동 전사를 수행하였고, 사용 가능 여부를 판단하고 전사의 품질을 검증하였다. 음성 인식 테스트 결과, 겹치는 발화가 많은 경우와 잡음이 많은 음성 파일의 경우에는 정확도가 낮아져서 수동 전사 절차를 진행하였고, 스튜디오에서 녹음한 1인 발화 혹은 2인 발화와 같이 잡음이 적고 겹침 발화가 적은 경우 음성 인식 정확도가 85% 이상으로 판단되어 음성 인식에 대한 자동 전사를 선택적으로 적용하고 품질을 검증하였다. 단, 음성 인식 자동 전사의 정확도가 높더라도 본 사업의 원시 말뭉치 구축 전사 지침에 따라 소리 나는 대로 전사를 해야 하였으므로 품질 검증 시 자동 전사가 적합하지 않은 파일의 경우, 아무리 자동 전사의 정확도가 높더라도 자동 전사를 하지 않고 100% 수동 전사로 전환하였다.

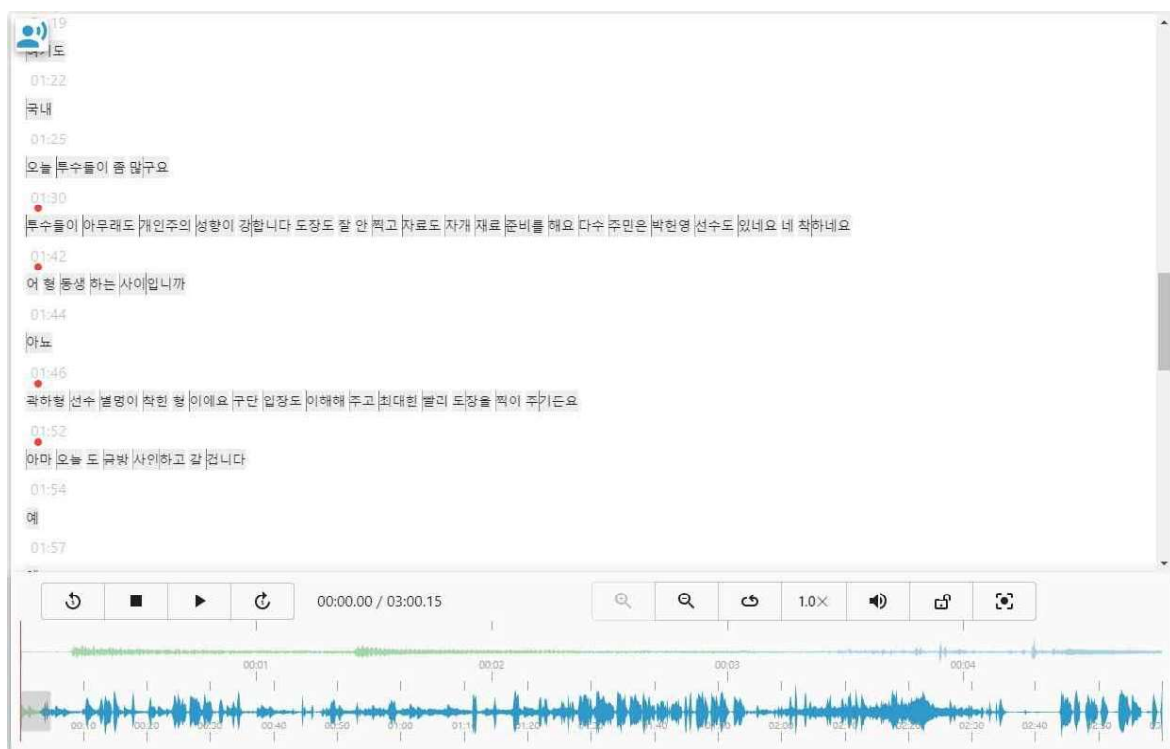


그림 2-12 음성 인식 자동 전사 도구

음성 인식 자동 전사 도구에서 오류 가능성이 있는 부분을 자동으로 확인하여 오류 여부를 빨리 파악할 수 있었고, 품질 검증의 도구로 사용할 수 있었다. 또한 구간 반복 청취를 통해 작업자가 발화자의 발음을 손쉽게 확인하여 전사를 진행할 수 있었다.

4.4. 수동 전사

음성 인식 자동 전사 대상이 아니거나 자동 전사 이후 추가 작업이 필요한 목록을 대상으로 전사 지침 항목에 맞춰 수동으로 전사를 하였다. 수동 전사 시 작업자별 업무 효율을 높이기 위해 도구를 사용하여 전사 지침에 맞춰 원시 말뭉치 구축 작업을 수행하였다.

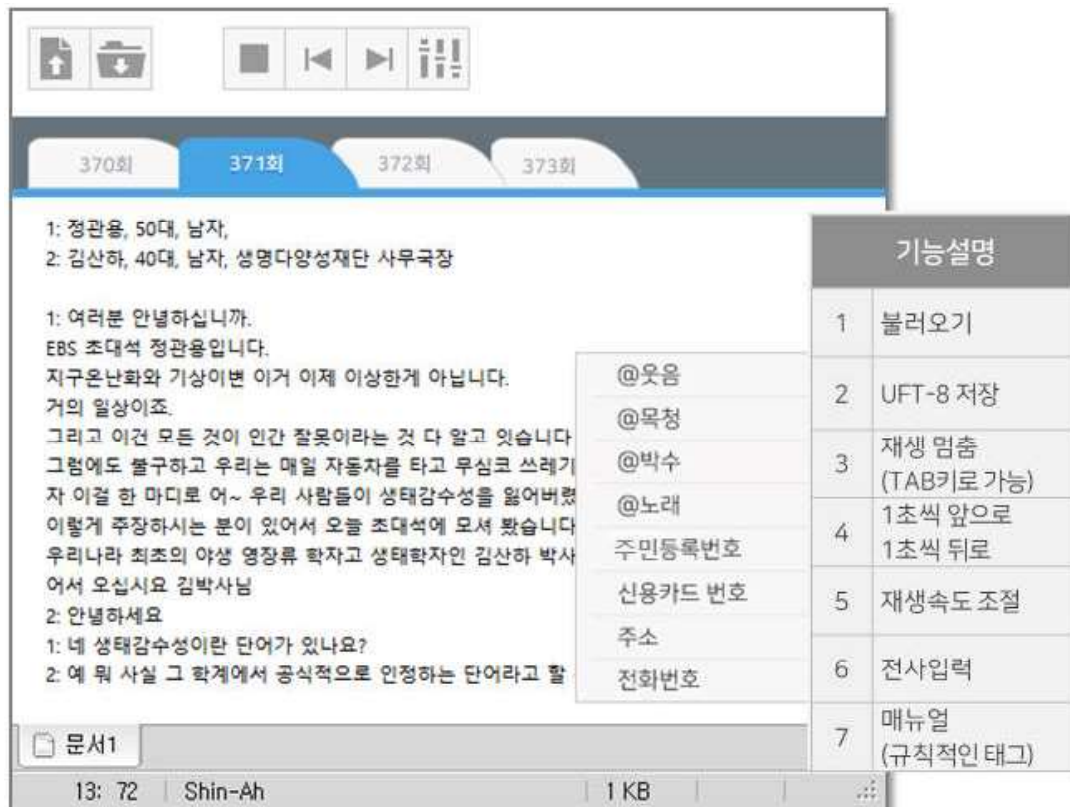


그림 2-13 수동 전사 도구

수동 전사를 위해 작업자의 작업 시간을 산출해 본 결과, 1인당 1일 8시간의 작업 시간 기준으로 평균 40분~1시간의 구어 원시 말뭉치 구축이 가능하였다. 약 5개월 동안의 수동 전사 작업 기간에 월별 약 200명의 인원이 투입되었고, 약 30~50명 이상의 인원이 품질 관리를 위해 투입되었다.

수동 전사 작업자는 구축 지침서 등 교육 자료를 토대로 샘플 테스트를 통해 선발하였고, 단계별 맞춤형 실무 중심의 교육과 주 1회 사례 공유 회의를 진행하여 작업자별 실시간 피드백을 주어 자료의 품질을 관리하였다. 피드백 이후에도 동일한 품질 오류가 반복적으로 발생하는 작업자에 대해서는 작업 중단과 신규 작업자 선발을 동시에 진행하여 사업 일정에 이상이 없도록 관리하였다.

파일명	전사자	검수자	검수완료	검수내용	평가
라디오_와이파이_190130	양*오	함*운	10/08	1. 가태요/가타요 등 맞춤법 수정 2. 불완전/수정발화 표시(=) 뒤 띄어쓰기 주의 3. 발화자 혼동 주의 4. 그렇게/이제/또 등 문장 중간 누락 주의 5. 음악 노트 후에 발화자 표시	
라디오_와이파이_190131	양*오	함*운	10/08	1. @웃음 누락 2. 발화자 정보 누락 3. 음/네 뒤에 마침표 누락 4. 불완전/수정발화 표시(=) 뒤 띄어쓰기 주의	
라디오_와이파이_190301	김*혜	피*라	10/07	양호	
라디오_와이파이_190305	류*영	김*희	08/09	양호	
pong20190225	김*은	김*희	08/13	한 음절만 물결 표시 사용	
pong20190227	김*은	김*희	09/16	재작업 9-16 1. 한 음절에만 물결 2. 어절 누락 3. 마침표 누락 4. 맞장구 누락 5. 전사 정확하지 않음	
pong20190308	김*린	나*하	09/16	1. 노트 처리 후 발화자 표시 누락 주의 2. =표시는 비정상적인 발화에 붙이고 한 칸 띄어쓰기 3. 대답에 마침표 누락 주의 4. 띄어쓰기 주의(님)	
nf20180606001	김*현	박*희	10/25	재작업 9-10(2) 1. 단어 누락 및 수정 다수 2. 전사 내용 확인하며 전사바람 3. 발화자 누락 4. 숫자는 한글로 전사	중단
nf20180606002	서*번	나*하	09/02	1. 노트 처리 형식 준수할 것 2. 맞장구 발화, 대답 누락 주의 3. 머뭇거리는 발에는 '=' 표시 안 함 4. 전사 누락 주의	
nf20180607002	곽*람	조*스더	08/28	오타와 누락 많음	중단
nf20180608001	김*정	조*스더	08/27	양호	
nf20181107001	이*웅	박*희	10/24	9-17 재작업 1. 전사 내용 정확하게 듣기 2. 오타 및 단어 수정 다수 3. 선체적 단락 수정	중단
nf20181116001	이*성	김*희	09/18	1. 이음 오타 2. 음악 노트 처리 누락 3. 종결어미 마침표 4. 동일 발화자 번호 연달아 나옴 5. 어절 및 문장 누락 6. 정확히 전사할 것	중단

그림 2-14 작업자 실시간 피드백

전사가 완료된 구어와 준구어 원시 말뭉치는 품질 검수 담당자가 품질을 점검하였다. 자체 품질 점검을 통해 오류 여부를 확인하여 99.9% 미만의 품질일 경우 수정 조치를 하였고, 재작업이 필요한 말뭉치 파일의 경우에는 전사 작업자에게 재작업을 요청하여 진행하였다. 품질 점검 시 음성 파일과 전사 파일의 일치 여부, 텍스트 전사 점검, 지침 사항 점검, 오류 수정 여부 점검과 전사 작업자의 평가를 진행하여 교체 여부를 결정하였다.

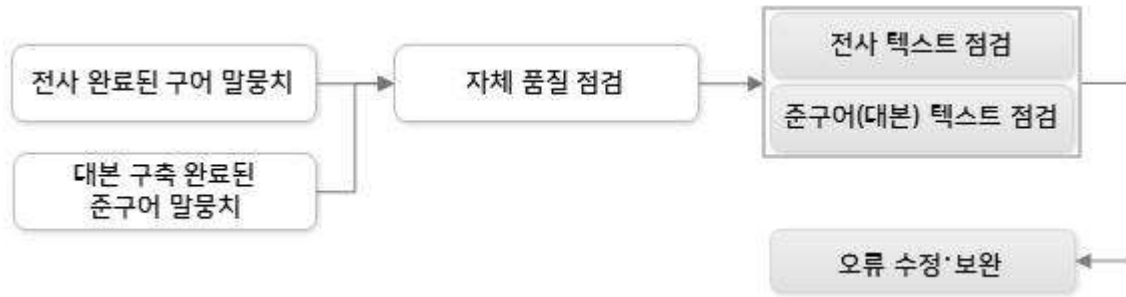


그림 2-15 자체 품질 점검 절차

점검 항목	세부 내용
음성 파일	음성 파일과 전사 파일 일치 여부 점검
텍스트 전사 점검	전사 텍스트 오타자 및 띄어쓰기 등의 점검
지침 사항 점검	전사 지침 준수 여부 점검
오류 수정 여부 점검	오류 수정 요청에 대한 반영 여부 점검
작업자 평가	품질 점검 후 작업자 교체 여부 평가

표 2-12 품질 점검 항목

품질 점검 담당자별로 작업자의 전사 파일을 점검하였으며, 자체 품질 점검 후 전사 작업자에게 보완 요청하여 수정된 사례는 다음과 같다. 그림 2-16에서 왼쪽은 품질 점검 전 파일이며, 오른쪽은 품질 점검 후 전사 작업자가 재작업하여 품질 보완한 파일이다. 또한 반복적으로 나오는 오류 유형의 경우 전사 파일 전체를 분석하여 오류가 있는 부분을 파악하여 수정하였다.

1: 네.
안녕하십니까?
엄밀정입니다.
창조경제에 대한 새로운 그 해석들이 나타나고 있는 가운데 사실은 우리 한국 한류문화에 대해서 어 많은 외국사람들은 그 깊이 있고
또한 다양한 전통에서 오는 것이 아닌가 그런 평가를 내리게 됩니다.
어 사실 우리 문화가 다른 어떤 나라보다 우월성을 갖는 것은 우리가 문화를 숭상하고 또 온
종종 가운데 다양한
삶의 형태를 가져온 전통에서 오는 것이 아니라 이런 우리 나름대로의 자평도 해 보게 됩니다.
전 세계가 지금 창조경제를 향하여 달려가고 있습니다.
우리가 갖는 우월성과 고유성이 있다면 바로 그 사이에
내려주시는 점층해온 에 우리에게 그 가치하고도 자신감 넘치는
전통 문화에 있지 않나 생각을 합니다.
이번 인천십사 년 한국 전승공예대전에서 소목장 양석중 씨가 대통령상을 받았습니다.
자 가구장인 양석중씨의 삶을 보면 우리 목가구의 전통은 물론이고 더
전통 전승 문화인들의 삶을 볼 수 있는 시사하는 바가 큰 대목이 어 아닐 수가 없습니다.
네, 오늘 초대했습니다.
여러분 양석중 소목장을 소개하겠습니다.
어서 오십시오.
2: 네, ...
1: 먼저 축하드리겠습니다.
2: 아 예 감사합니다.
1: 대통령상은 그 뭐 어마어마한 작가들이 받으시는 건데 소감이 어떠셨어요?
2: 전승공예대전에 아 그간 여섯 번 정도 도전했었고요.
1: 예
2: 중간에 특선 정도를 받은 적도 있었지만 세 번은 낙방했었고
1: 예.
2: 이번에 여섯 번 만에 대통령상을 받게 되었습니다.

22 1: @웃음
23 2: @웃음
24 이제 이제 나무를 다뤄서 일하는 목수를
송나라 때부터요 뭐 그런 이름을 지었다 하는 얘기가 있던 하던데
25 1: 네.
26 2: 어 집을 짓는 기둥을 세우고 서까래를 깔고 하는 그런 목수를 대목이라고 불렀고 그다음에
이제 문을 만들고 가구를 만들고 뭐 이런 이제 목수를
27 1: 예.
28 2: 또 소목장이라고 불렀고요.
29 1: 예.
30 2: 근대 그런 명칭은
조선시대를 이어서 지금 현재도 문화재청에서는 이제 중요문화재 무형문화재 종목 중에
31 1: 예
32 2: 대목장이 있고 소목장이 또 따로 있습니다.
33 1: 아 그럼 지금 선생님이 지금 활동하시는 소목장 분야의 중요무형문화재는 어느
분야인가요?

1: 그런 분들은 우리 민간에서 어떻게 이렇게 맥을 이어 왔을까요?
지금까지
2: 어 대부분 **오는데** 그 가계를 이어거나 또는
승승으로부터 배워서 그 분야 또 그 전통방식을 본인이 갖고 계신 거죠.
1: 예.
2: 그래서
중요무형문화재 중에는 그 전통적으로 배워온 기수 기술과 기능을 가지고 있느냐 하는 부분들
은 이제 전문가들이 심사를 하셔서
1: 네.
2: **그분들** 인정하는 과정이 **있지요**.
별도로
1: 그런데 우리가 이제 이 가구가 현대화 되면서 대량생산 또 기계와 심지어 뭐 대기업이 진출하고
2: 네.
1: 이런 역사를 가지고 있는데 **그런** 동안 소목장들께서 어떻게 맥을 이어 오셨을까요?
2: 그래서 넉넉하게 사는 소목장들은 몇 분 안 계시고요.
1: 예.
2: 어떻게 보면 이제 전통적인 방식으로 어 소목장을 직업으로 가지신 분들이 현대적인 어떤
시장경제의 논리라면 빠시면
1: 네.
2: 결코 잘 살기는 어려운 환경입니다.
그렇지만 **이제** 나름 갖고 있는 사명감
1: 네.
2: 또 그 그런 고생 끝에 어쨌든 몇 작품을 만들었을 때 그것을 인정받고
1: **응**
2: 또 그거를 사주시는 분들 이런 분들을 보람으로 해서 사명감과 보람으로 또 어느 정도의
이제
저 선생님처럼 경지에 이르시면 저처럼 가서 가르쳐 달라고 뭐 애걸복걸 하는 사람들도 많이
생겨나거든요.
그런 보람으로 사시는 게 아닌가 하는 생각이 듭니다.
1: 네.
자 이번에 그 이 영예의 상을 타신 삼층상 이야기를 좀 하겠습니까?
심사위원들이 어떤 점을 높이 사시던가요?
2: 전통 목가구 중에서 삼층상이라고 하면 뭐 문갑도 있고 책상도 있고 뭐 여러 가지 합 종류
두 있지만 가장 대표적인 가구의 종류라고 보시면 됩니다.
그런데 **이제**
그만큼 많은 소목장들이 만들고 싶어라 하기도 하고 또 도전해 보기 위한 과제 얘기도 하고
뭐 그런

1: 네.
2: 안녕하십니까?
3: 엄밀정입니다.
창조경제에 대한 새로운 그 해석들이 나타나고 있는 가운데 사실은 우리 한국 한류문화에 대해서 어 많은 외국사람들은 그 깊이 있고
또 아주 다양한 전통에서 오는 것이 아닌가 그런 평가를 내리게 됩니다.
어 사실 우리 문화가 다른 어떤 나라보다 우월성을 갖는 것은 우리가 문화를 숭상하고 또 온
종종 가운데 다양한
삶의 형태를 가져온 전통에서 오는 것이 아니라 이런 우리 나름대로의 자평도 해 보게 됩니다.
전 세계가 지금 창조경제를 향하여 달려가고 있습니다.
우리가 갖는 우월성과 고유성이 있다면 바로 **조상**이
내려주시는 점층해온 에 우리에게 그 가치하고도 자신감 넘치는 **에**
전통 문화에 있지 않나 생각을 합니다.
8: **이** 이번 인천십사 년 한국 전승공예대전에서 소목장 양석중 씨가 대통령상을 받았습니다.
9: 자 가구장인 **양석중**씨의 삶을 보면 우리 목가구의 전통은 물론이고 또
전통 전승 문화인들의 삶을 볼 수 있는 시사하는 바가 큰 **이** 대목이 어 아닐 수가 없습니다.
10: 네, 오늘 초대했습니다.
11: 여러분 양석중 소목장을 소개하겠습니다.
12: 어서 오십시오.
13: 2: 네, ...
14: 1: 먼저 축하드리겠습니다.
15: 2: 아 예 감사합니다.
16: 1: 대통령상은 그 뭐 어마어마한 작가들이 받으시는 건데 소감이 어떠셨어요?
17: 2: 전승공예대전에 아 그간 여섯 번 정도 도전했었고요.
18: 1: 예
19: 2: 중간에 특선 정도를 받은 적도 있었지만 세 번은 낙방했었고
20: 1: 예.
21: 2: 이번에 여섯 번 만에 대통령상을 받게 되었습니다.

20 1,2: @웃음
21 2: 이제 이제 나무를 다뤄서 일하는 목수를 뭐
송나라 때부터요 뭐 그런 이름을 지었다 하는 얘기가 있던 하던데
22 1: 네.
23 2: 어 집을 짓는 기둥을 세우고 서까래를 깔고 하는 그런 목수를 대목이라고 불렀고 그다음에
이제 문을 만들고 가구를 만들고 뭐 이런 이제 목수를
24 1: 예.
25 2: 소목장이라고 또 불렀고요.
26 1: 예.
27 2: 근대 그런 명칭은 뭐
조선시대를 이어서 지금 현재도 문화재청에서는 이제 중요문화재 무형문화재 종목 중에
28 1: 예
29 2: 대목장이 있고 소목장이 또 따로 있습니다.
30 1: 아~
31 2: 예.
32 1: 그럼 지금 선생님 **작** 지금 활동하시는 소목장 분야의 중요무형문화재는 어느 분야인가요?

2 1: 그런 분들은 우리 민간에서 어떻게 이렇게 맥을 이어 왔을까요? 지금까지
3 2: 어 대부분 **이제** **그** 그 가계를 이어거나 또는 **뭐**
승승으로부터 배워서 그 분야 또 그 전통방식을 본인이 갖고 계신 거죠.
4 1: 예.
5 2: 그래서 **이제**
중요무형문화재 중에는 그 전통적으로 배워온 기수 기술과 기능을 가지고 있느냐 하는 부분들
은 이제 전문가들이 심사를 하셔서
6 1: **네**
7 2: **그분** 인정하는 과정이 **있지요**. 별도로
8 1: 그런데 우리가 이제 이 가구가 현대화 되면서 대량생산 또 기계와 심지어 뭐 대기업이 진출하고
9 2: 네.
10 1: 이런 역사를 가지고 있는데 **그러는** 동안 소목장들께서 어떻게 맥을 이어 오셨을까요?
11 2: 그래서 **이제** 넉넉하게 사는 소목장들은 몇 분 안 계시고요.
12 1: 예.
13 2: 어떻게 보면 이제 전통적인 방식으로 어 소목장을 직업으로 가지신 분들이 현대적인 어떤
시장경제의 논리라면 빠시면
14 1: 네.
15 2: **아** 결코 잘 살기는 어려운 환경입니다.
16 **그렇지만 이제** 나름 갖고 있는 사명감
17 1: 네.
18 2: 또 그 그런 고생 끝에 어쨌든 **하** 몇 작품을 만들었을 때 그것을 인정받고
19 1: **응**
20 2: 또 그거를 사주시는 분들 이런 분들을 **뭐**
보람으로 해서 사명감과 보람으로 또 어느 정도의 **이제**
저 선생님처럼 경지에 이르시면 저처럼 가서 가르쳐 달라고 뭐 애걸복걸 하는 사람들도 많이
생겨나거든요.
21 그런 보람으로 사시는 게 아닌가 하는 생각이 듭니다.
22 1: 네.
23 **자** 이번에 그 이 영예의 상을 타신 삼층상 이야기를 좀 해하겠습니다.
24 심사위원들이 어떤 점을 높이 사시던가요?
25 2: 전통 목가구 중에서 삼층상이라고 하면 뭐 문갑도 있고 책상도 있고 뭐 여러 가지 합 종류
두 있지만 가장 대표적인 가구의 종류라고 보시면 됩니다.
26 **그런데 이제**
그만큼 많은 소목장들이 만들고 싶어라 하기도 하고 또 도전해 보기 위한 과제 얘기도 하고
뭐 그런

그림 2-16 품질 점검 후 수정 사항 반영 예시

4.5. 메타 정보 구축

전사가 완료된 말뭉치를 대상으로 국립국어원과 협의한 기준으로 파일명을 부여하고 헤더와 마크업을 부여하여 최종적으로 원시 말뭉치를 구축하였다. 구축 완료 후 태그 오류 여부를 확인하였고, 오류가 있는 부분을 재수정하였다.

파일명은 공적 독백, 공적 대화, 준구어 대본 등의 분류와 구축 연도, 8자리 일련번호를 부여하였다.

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	8자리 일련번호
S: 구어 말뭉치	A: 공적 독백 B: 공적 대화 E: 준구어-대본	RW: 원시 말뭉치	19	#####

표 2-13 파일명 부여 방식

- 예시

- SARW1900000001.sjml 구어(공적 독백) 원시 말뭉치 첫 번째 파일
- SBRW1900000001.sjml 구어(공적 대화) 원시 말뭉치 첫 번째 파일
- SERW1900000001.sjml 준구어(대본) 원시 말뭉치 첫 번째 파일

파일명 부여 후 기술 지원팀에서 XML 변환 프로그램을 통해 자료를 변환하고, 오류 여부 및 오류의 발생 원인을 검증하였다.

절차	내용
XML 입력 준비	사용 태그 유형 정의 태깅 지침, 외부 파일 연결 지침 수립
마크업	자료의 마크업
XML 자동 변환	XML 변환 프로그램 이용
테스트/검증	오류 검증 및 수정

표 2-14 XML 변환 절차

XML 변환 후 오류 사항을 확인하였고, 반복적인 테스트와 검증을 통해 오류 보완 및 수동 확인이 필요한 부분은 전체 파일을 대상으로 확인하여 자동 변환을 보완하는데 사용하였다.

<u who="P3" n="567">반 사회적 그 얘기하시는 거예요?</u>
 <u who="P2" n="568">어떤 어떤 그 지독한 경험들을 통해서</u>
 <u who="P모두" n="569">외상 후 스트레스 장애</u>
 <u who="P2" n="570">예. 그게 제 기억이 외상후 스트레스 증후군입니다.</u>
 <u who="P1" n="571">네네.</u>

<u who="P3" n="567">반 사회적 그 얘기하시는 거예요?</u>

<u who="P6" n="986">글쎄요.</u>
 <u who="P6" n="987">그게 뭐 어떠한 남의 남의 무슨 인권을 침해한다거나 남의 법 이익을 침해하는 건
 아니지 않습니까? <vocal desc="목청가다듬는소리"/></u>
 <u who="P7" n="988">근데 이제 그와 관련해서 여러 가지 인제 금방 또 이천팔년 그 쫓돌 말씀하셨는데
 실질적으로 그러한 어떤 폭력적 행위에 대해서 계속 인제 그 강조를 하시는 것 가태요.</u>



<u who="P6" n="986">글쎄요.</u>
 <u who="P6" n="987">그게 뭐 어떠한 남의 남의 무슨 인권을 침해한다거나 남의 법 이익을 침해하는 건
 아니지 않습니까? <vocal desc="목청가다듬는소리"/></u>
 <u who="P7" n="988">근데 이제 그와 관련해서 여러 가지 인제 금방 또 이천팔년 그 쫓돌 말씀하셨는데
 실질적으로 그러한 어떤 폭력적 행위에 대해서 계속 인제 그 강조를 하시는 것 가태요.</u>

<u who="P3" n="389">잘생겼어요.</u>
 <u who="P3" n="390">xx해요.</u>
 <u who="unknown" n="391">어</u>



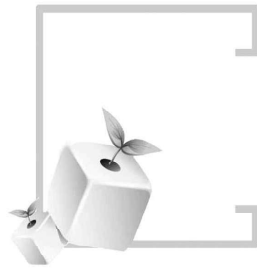
<u who="P3" n="389">잘생겼어요.</u>
 <u who="P3" n="390"><unclear>xx해요</unclear>.</u>
 <u who="unknown" n="391">어</u>

<sp>
 <stage>(플래시백) (71회 씬4에서)</stage>
 <speaker>빛나</speaker>
 <p>도무지 날 낳아준 사람 같지 않게.. 저주를 퍼붓는 거 같았어요.</p>
 </sp>



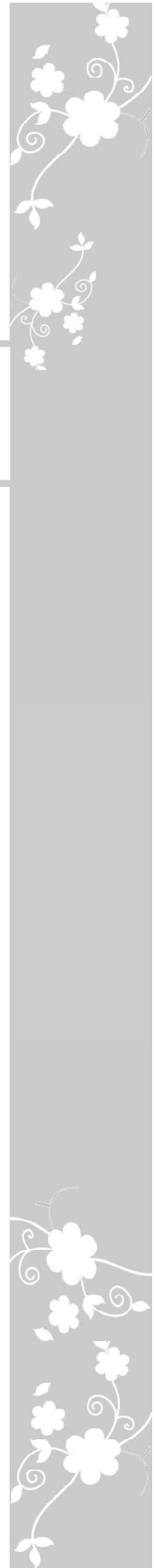
<stage>(플래시백) (71회 씬4에서)</stage>
 <sp>
 <speaker>빛나</speaker>
 <p>도무지 날 낳아준 사람 같지 않게.. 저주를 퍼붓는 거 같았어요.</p>
 </sp>

그림 2-17 오류 사항 확인 및 수정 보완



제 3 장

사업 수행 결과



1. 원시 말뭉치 구축

1.1. 구어 말뭉치 주제별 구축 결과

구어 말뭉치는 18개 주제의 총 67개의 프로그램을 대상으로 구축하였다. 주제별 구축 결과를 보면 시사정치 분야가 35.59%로 가장 높은 비율을 차지하였으며, 다음으로 시사교양 분야 14.48%, 생활정보 분야 10.13%였다.

주제	분량(시간)	비율	프로그램 수
시사정치	5,357.1	35.59%	15
시사교양	2,179.5	14.48%	5
생활정보	1,525.2	10.13%	6
문화예술	910.9	6.05%	6
교육	738.6	4.91%	4
교양	707.6	4.70%	7
시사문화	563.8	3.75%	5
예능	437.4	2.91%	4
과학	422	2.80%	1
법률	344	2.29%	2
토론	321.3	2.13%	1
상담	315.3	2.09%	1
토크쇼	285.2	1.89%	4
보도토론	249.1	1.65%	1
도서	241.5	1.60%	1
문화	234.6	1.56%	1
스포츠	136.6	0.91%	2
건강	82.4	0.55%	1
합계	15,052.1	100%	67

표 3-1 구어 말뭉치 주제별 구축 결과

1.2. 구어 말뭉치 방송 연도별 구축 결과

구어 말뭉치 구축 결과를 방송 연도별로 살펴보면 최근 3개년도(2017~2019년)가 69.2%이었으며, 최근 5개년도(2015~2019년)가 87.7%였다.

방송 연도	분량(시간)	비율
2011	12.0	0.1%
2012	171.0	1.1%
2013	721.2	4.8%
2014	942.6	6.3%
2015	1,199.5	8.0%
2016	1,590.1	10.6%
2017	2,149.7	14.3%
2018	5,222.8	34.7%
2019	3,043.1	20.2%
합계	15,052.1	100%

표 3-2 구어 말뭉치 방송 연도별 구축 결과

1.3. 구어 말뭉치 주제×방송 연도별 구축 결과

구어 말뭉치 구축 결과를 주제와 방송 연도별로 살펴보면 주제별 분류에서 가장 높은 비율을 차지한 시사정치의 경우 2012년부터 2019년까지 약 5,357시간 분량의 자료를 구축하였다. 이 가운데 최근 3개년도 자료가 약 3,402시간 분량이었다.

주제	방송 연도	분량(시간)	비율	비고
건강	2017	16.9	0.1%	TV조선
	2018	35.4	0.2%	
	2019	30.2	0.2%	
과학	2013	30.8	0.2%	과학하고 아아있네
	2014	42.4	0.3%	
	2015	47.0	0.3%	
	2016	79.8	0.5%	
	2017	79.5	0.5%	
	2018	87.1	0.6%	
	2019	55.5	0.4%	
교양	2013	30.9	0.2%	EBS MBC TV조선
	2014	70.1	0.5%	
	2015	77.4	0.5%	
	2016	53.9	0.4%	
	2017	111.3	0.7%	
	2018	246.2	1.6%	
	2019	117.7	0.8%	
교육	2015	107.26	0.7%	EBS TBS 필스교양
	2016	0.82	0.0%	
	2017	140.37	0.9%	
	2018	432.36	2.9%	
	2019	57.84	0.4%	

주제	방송 연도	분량(시간)	비율	비고
도서	2014	23.75	0.2%	TBS
	2015	41.88	0.3%	
	2016	46.14	0.3%	
	2017	45.46	0.3%	
	2018	46.75	0.3%	
	2019	37.57	0.2%	
문화	2013	33.19	0.2%	MBC
	2014	31.79	0.2%	
	2015	33.52	0.2%	
	2016	34.57	0.2%	
	2017	38.07	0.3%	
	2018	36.02	0.2%	
	2019	27.4	0.2%	
문화예술	2016	37.87	0.3%	TBS
	2017	184.79	1.2%	
	2018	384.59	2.6%	
	2019	303.67	2.0%	
법률	2016	31.85	0.2%	TBS
	2017	54.2	0.4%	
	2018	206.24	1.4%	
	2019	51.67	0.3%	
보도토론	2018	92.45	0.6%	MBC
	2019	156.62	1.0%	
상담	2017	81.78	0.5%	마인드코칭 연구소
	2018	165.21	1.1%	
	2019	68.32	0.5%	
생활정보	2013	25.6	0.2%	MBC
	2015	9.4	0.1%	
	2016	506.1	3.4%	
	2017	411.4	2.7%	
	2018	378.8	2.5%	
	2019	193.9	1.3%	
스포츠	2013	19.07	0.1%	MBC
	2014	22.22	0.1%	
	2015	10.56	0.1%	
	2016	24.02	0.2%	
	2017	3.75	0.0%	
	2018	50.47	0.3%	
	2019	6.5	0.0%	
시사교양	2011	11.97	0.1%	EBS MBC TV조선
	2012	159.37	1.1%	
	2013	24.1	0.2%	
	2016	15.28	0.1%	
	2017	190.99	1.3%	
	2018	1202.71	8.0%	
	2019	575.06	3.8%	

주제	방송 연도	분량(시간)	비율	비고
시사문화	2013	49.68	0.3%	MBC TBS
	2014	73.94	0.5%	
	2015	73.08	0.5%	
	2016	71.96	0.5%	
	2017	69.09	0.5%	
	2018	66.23	0.4%	
	2019	159.78	1.1%	
시사정치	2012	11.7	0.1%	TBS TV조선
	2013	216.9	1.4%	
	2014	514.4	3.4%	
	2015	636.0	4.2%	
	2016	575.8	3.8%	
	2017	610.7	4.1%	
	2018	1,693.2	11.2%	
예능	2019	1,098.4	7.3%	MBC
	2013	118.0	0.8%	
	2014	50.3	0.3%	
	2015	55.7	0.4%	
	2016	57.6	0.4%	
	2017	71.1	0.5%	
	2018	62.4	0.4%	
토론	2019	22.3	0.1%	MBC
	2013	62.7	0.4%	
	2014	44.6	0.3%	
	2015	58.0	0.4%	
	2016	54.3	0.4%	
	2017	40.3	0.3%	
	2018	34.5	0.2%	
토크쇼	2019	27.0	0.2%	MBC TBS
	2013	110.3	0.7%	
	2014	69.2	0.5%	
	2015	49.7	0.3%	
	2018	2.2	0.0%	
	2019	53.8	0.4%	

표 3-3 구어 말뭉치 주제×방송 연도별 구축 결과

1.4. 준구어 말뭉치 주제별 구축 결과

준구어 말뭉치 구축 결과, 총 15,591,197어절의 준구어 말뭉치를 구축하였다. 주제별로 구축 결과를 살펴보면 가족 분야가 총 4,894,707어절(31.39%)이었으며, 로맨스 분야는 3,514,034어절(22.54%), 사극 분야는 2,064,625어절(13.24%)이었다.

주제	분량(어절 수)	비율	프로그램 수
가족	4,894,707	31.39%	17
로맨스	3,514,034	22.54%	19

사극	2,064,625	13.24%	8
직업	1,754,580	11.25%	8
시대	759,868	4.87%	3
범죄	594,379	3.81%	4
법률	421,493	2.70%	2
판타지	309,463	1.98%	3
의학	275,170	1.76%	2
요리	265,335	1.70%	2
코믹	191,699	1.23%	1
첩보	171,408	1.10%	2
음악	130,507	0.84%	1
방송국	93,913	0.60%	1
스포츠	89,394	0.57%	1
호러	60,622	0.39%	1
합계	15,591,197	100%	75

표 3-4 준구어 말뭉치 주제별 구축 결과

1.5. 준구어 말뭉치 방송 연도별 구축 결과

준구어 말뭉치 구축 결과를 방송 연도별로 살펴보면, 다양한 준구어 말뭉치 구축을 위해 2004년부터 2019년까지의 방송을 대상으로 16년간의 말뭉치를 구축하였고, 방송 연도별 최대 15%는 넘지 않도록 조절하였다. 가장 어절 수가 많은 방송 연도는 2013년 준구어 말뭉치로 2,166,560어절(13.9%)이었다.

방송 연도	분량(어절 수)	비율
2004	273,297	1.8%
2005	270,693	1.7%
2006	237,242	1.5%
2007	1,248,134	8.0%
2008	1,962,639	12.6%
2009	1,626,007	10.4%
2010	751,688	4.8%
2011	1,072,184	6.9%
2012	1,338,165	8.6%
2013	2,166,560	13.9%
2014	1,999,102	12.8%
2015	1,124,915	7.2%
2016	642,904	4.1%
2017	212,514	1.4%
2018	484,524	3.1%
2019	180,629	1.2%
합계	15,591,197	100%

표 3-5 준구어 말뭉치 방송 연도별 구축 결과

1.6. 준구어 말뭉치 주제×방송 연도별 구축 결과

준구어 말뭉치 구축 결과를 주제와 방송 연도별로 살펴보면 주제별 분류에서 가장 높은 비율을 차지한 가족의 경우 2008년부터 2014년까지 7년간의 방송 자료를 통해 4,894,707어절의 말뭉치를 구축하였고 이 가운데 가장 높은 비율을 차지한 방송 연도는 2008년 1,079,709어절(6.9%), 2009년 1,005,315어절(6.4%), 2014년 932,866어절(6.0%)어절이었다.

주제	방송 연도	분량(어절 수)	비율
가족	2008	1,079,709	6.9%
	2009	1,005,315	6.4%
	2010	226,751	1.4%
	2011	781,268	5.0%
	2012	317,991	2.0%
	2013	550,807	3.5%
	2014	932,866	6.0%
로맨스	2006	163,924	1.1%
	2007	87,481	0.6%
	2008	471,855	3.0%
	2009	560,070	3.6%
	2010	270,887	1.7%
	2011	196,104	1.3%
	2013	554,126	3.6%
	2014	102,230	0.7%
	2015	604,161	3.9%
	2016	202,960	1.3%
방송국	2008	300,236	1.9%
	2008	93,913	0.6%
범죄	2008	73,824	0.5%
	2014	247,048	1.6%
	2018	184,288	1.2%
	2019	89,219	0.6%
법률	2007	327,389	2.1%
	2014	94,104	0.6%
사극	2007	463,309	3.0%
	2013	1,061,627	6.8%
	2015	406,737	2.6%
	2017	132,952	0.9%
스포츠	2007	89,394	0.6%
시대	2004	273,297	1.7%
	2010	254,050	1.6%
	2012	232,521	1.5%
요리	2005	78,994	0.5%
	2012	186,341	1.2%
음악	2008	130,507	0.8%
의학	2007	137,362	0.9%
	2012	137,808	0.9%

주제	방송 연도	분량(어절 수)	비율
직업	2007	66,603	0.4%
	2008	112,831	0.7%
	2012	341,376	2.2%
	2014	622,854	4.0%
	2016	439,944	2.8%
	2017	79,562	0.5%
	2019	91,410	0.6%
첩보	2007	76,596	0.5%
	2011	94,812	0.6%
코믹	2005	191,699	1.2%
호러	2009	73,318	0.5%
판타지	2006	122,128	0.8%
	2012	114,017	0.7%
	2015	60,622	0.4%

표 3-6 준구어 말뭉치 주제×방송 연도별 구축 결과

2. 말뭉치 활용

말뭉치 활용 검증을 위해 본 사업으로 구축된 말뭉치를 음성 인식을 위한 음향 모델링과 언어 모델링에 적용하였으며, 2가지 모델을 국립국어원에 제공하였다. 음향 모델의 경우 원시 말뭉치 1,000시간을 학습 데이터로 활용하였고, 언어 모델의 경우 원시 말뭉치 3,000시간을 학습 데이터 형태로 변환하여 학습하였다.

2.1. 음성 인식을 위한 음향 모델

음성 인식(Speech Recognition)이란 사람이 말하는 음성 언어를 컴퓨터가 해석해 그 내용을 문자 언어로 전환하는 처리를 말하며 STT(Speech-to-Text)라고도 한다. 음성 인식은 음성 인터페이스를 기반으로 하는 다양한 서비스에 활용되고 있다.

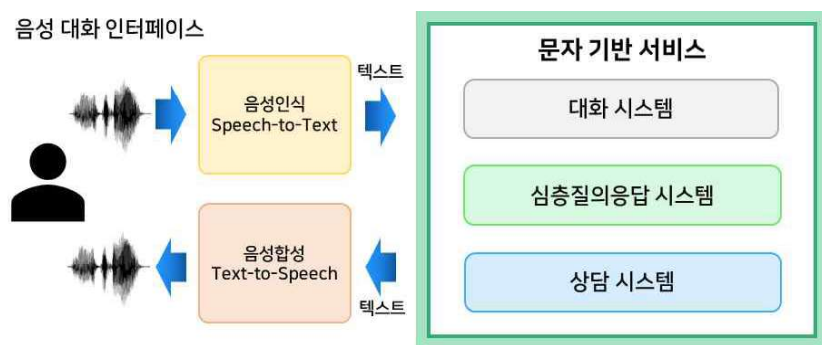


그림 3-1 음성 인터페이스 기반 서비스

음성 인식은 크게 모델을 학습하는 단계, 학습된 모델을 통합하는 단계, 이를 기반으로 음성을 인식하는 세 단계로 구성되며 이 가운데 음향 모델과 언어 모델을 학습하는 부분이 매우 중요하다.

음성 인식 엔진은 음향과 언어 정보라는 중요한 지식을 사용하여 음성 신호로부터 문자 정보를 생성하게 되는데, 이때 개념적으로 음성 신호를 문자 기호로 해석한다는 차원에서 음성 인식 알고리즘을 디코더(decoder)라고 부르기도 한다. 그림 3-2와 같이 학습 단계에서는 음성 데이터를 활용해 음향 모델을 생성하고, 말뭉치와 같은 텍스트 데이터를 활용해 음소 변환(G2P: Grapheme-to-Phoneme)을 수행하여 언어 모델과 발음 사전을 생성한다. 이렇게 생성된 음향 모델, 언어 모델, 발음 사전을 통합한 후 인식 단계에서 문장 단위의 학습에 최적화되어 속도와 인식률을 향상시키며, 대용량 연속 어휘에 대한 자연어 음성 인식을 가능하도록 한다.

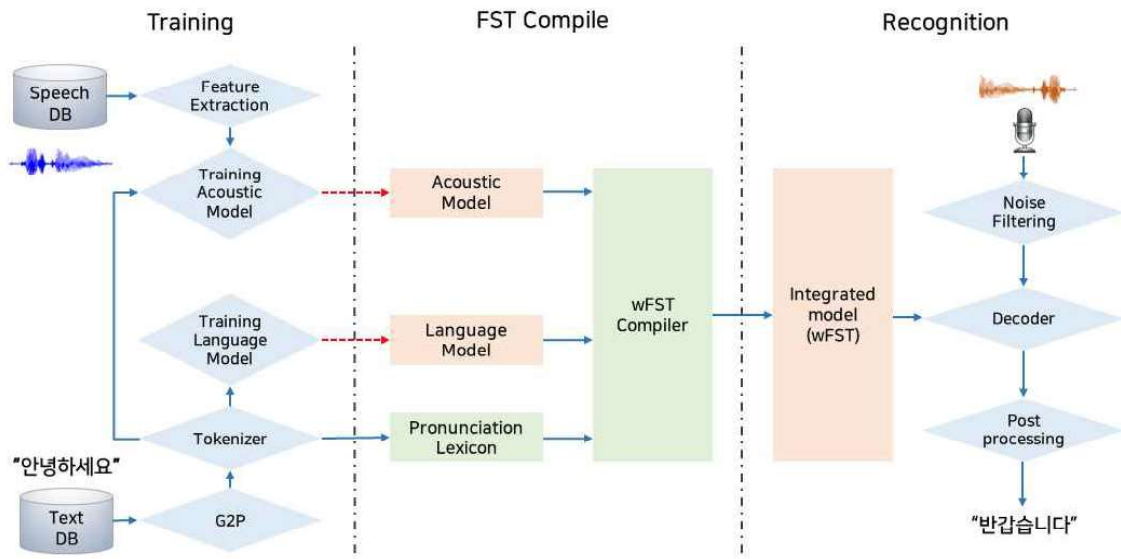


그림 3-2 음성 인식 개요

음향 모델은 음성 데이터에서의 음향적 특성을 통계적으로 모델링하여 구성하게 된다. 음향 모델 구성을 위하여 기존에는 GMM(Gaussian Mixture Model)으로 음소를 모델링하고 해당 음소들의 변화를 HMM(Hidden Markov Model)으로 예측하는 GMM-HMM 방식을 사용하였는데, 최근에는 음소 모델링을 위한 GMM 확률 모델링 부분을 DNN(Deep Neural Network)으로 대체한 DNN-HMM에 기반한 음성 인식이 좋은 성능을 보이고 있다.

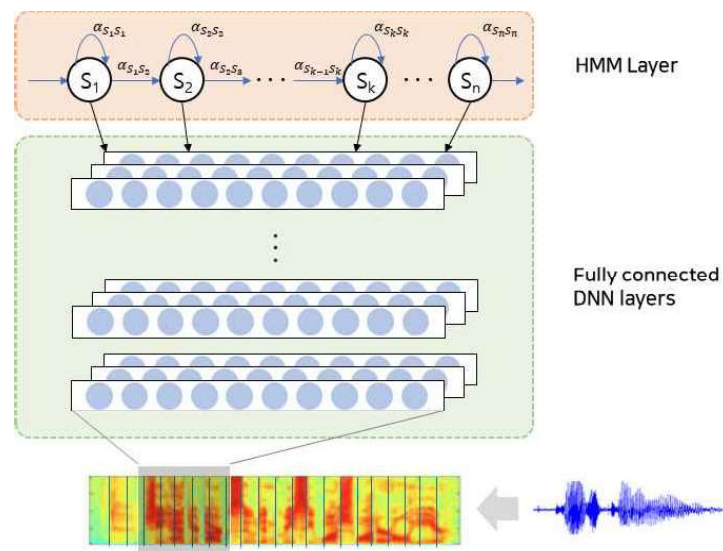


그림 3-3 DNN-HMM 음향 모델의 구조

기본 음향 모델을 기반으로 실제 음성 인식 적용이 필요한 환경 및 분야의 음성 특성을 추가하는 적응 학습이 가능하다. 콜센터 등 특정 분야에서 수집된 음성 데이터를 학습 데이터로 구성하여 기존 음향 모델에 적응 학습을 수행할 수 있다. LSTM(Long

Short-Term Memory) 기반으로 학습된 음향 모델은 기존의 모델에 비해 높은 음성 인식 성능을 보이고 있다.

2.2. 언어 이해를 위한 언어 모델

자연어 이해 기술은 컴퓨터가 사람의 말과 글을 이해하고 사람과 기계가 상호 소통할 수 있도록 하는 대화형 인공지능의 핵심 기반 기술이다. 자연어 이해를 위해서는 형태소 분석, 개체명 인식, 구문 분석, 감성 분석, 의미 분석 등의 자연어 처리 결과를 바탕으로 문장에 숨겨진 의도를 이해하거나 질문의 유형을 파악하는 등의 한 단계 높은 수준의 분석이 수행되어야 한다.

자연어 이해 기술이 적용된 언어 인지 엔진에는 기계 학습과 심층 학습(인공 신경망) 기술이 적용되어 있으며, 말뭉치와 같은 대규모 학습 데이터와 사전과 규칙 등의 언어 자원을 통해 구성한다.



그림 3-4 언어 인지 엔진 구성도

최근에는 BERT(Bidirectional Encoder Representations from Transformers)와 같은 사전 학습된 언어 모델(Pre-Trained Language Model)을 통해 인간 수준의 언어 인지가 가능해지고 있다. BERT는 대용량의 말뭉치로부터 비지도 학습을 통해 일반 목적의 언어 모델을 구축하고 지도 학습을 통해 모델을 최적화함으로써 질의응답, 문장 유사도 비교 등의 다양한 응용 분야에 적용하는 준지도 학습(Semi-supervised Learning) 모델이다.

2.3. 적용 사례

위와 같이 말뭉치를 학습하여 생성된 음향 모델, 언어 모델 등은 음성 인식 엔진, 언어 처리 엔진 등에 적용된다. 인공 지능 기술을 활용한 상담 시스템과 챗봇 등 질의응답 시스템 등에 음성 인식 엔진이 사용되고 있다. 또한 형태소 분석, 개체명 및 의미 분석 등을 수행하는 언어 인지 엔진은 빅데이터 분석 시스템 등 다양한 데이터 수집 및 분석 분야에 적용되고 있다.

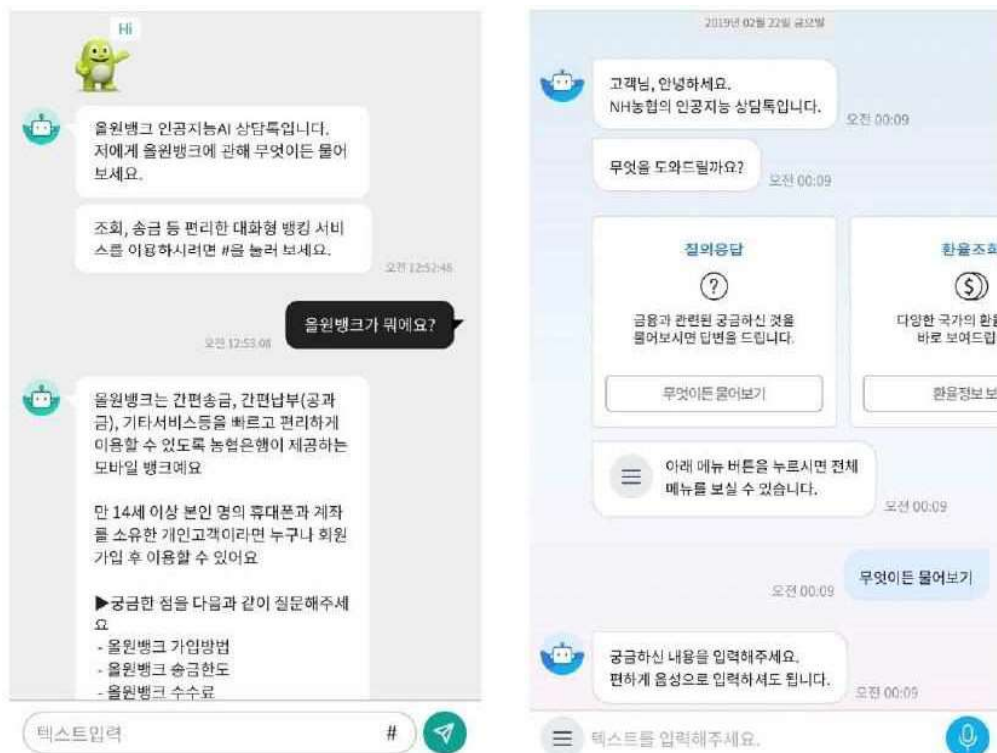


그림 3-5 챗봇 서비스를 위한 음성 인식



그림 3-6 빅데이터 분석 시스템

3. 정책 제언

3.1. 사업 수행 과정의 현안 및 해결 결과

구어 자료 수집 및 원시 말뭉치 구축 사업을 수행하면서 여러 가지 사업 수행상 현안과 문제점을 확인하였으며, 이에 대한 해결안을 도출하면서 사업 수행을 완료하였다. 향후 유사한 사업 수행 추진을 위해 사업 수행 과정의 주요 현안 및 이에 대한 해결 결과를 아래와 같이 제시하였다.

- 구어 수집 시 저작권 해결
사업 초기에 여러 방송사에 구어 자료 수집을 제안하였으나, 방송사별로 저작권 이용 허락 계약서에 대한 법률적인 검토 후 방송 대화 제공이 어렵다는 의사를 표명하였다. 이후 방송사별로 추가적인 협의를 통해 사업 목적의 취지에 벗어나지 않는 범위 내에서 저작권 이용 허락 계약서의 일부 내용을 변경하여 최종적으로 저작권 이용 허락 계약을 체결하고 자료를 수집하였다.
- 구어 자료 수집 시 개인 방송 자료의 품질
팟캐스트 등 인터넷 방송 자료의 경우 저작권 이용 허락 계약은 대형 방송사에 비해 쉽게 체결할 수 있었으나, 대형 방송사의 자료와 달리 음성이 겹치는 부분이 많이 발생하는 등 품질 문제가 있는 경우가 있었다. 다른 자료로 대체하는 것이 좋겠다는 전문가 검토 결과에 따라 일부 수집 목록은 다른 자료로 대체하여 구축하였다.
- 구어 자료의 품질
일부 자료의 경우 발화자가 많거나, 음성이 겹치는 부분 등이 많아 말뭉치 구축에 소요되는 시간이 사업 초기 예상한 시간보다 많이 소요되었다. 또한 외부 인터뷰나 음악 등 수집이 제외되는 부분이나 음성에 오류가 있는 부분이 발견될 경우, 구축 시 예상한 분량과 실제 원시 말뭉치 분량의 차이가 발생하여 추가 자료를 수집하여 부족한 분량을 보완하는 업무가 필요하였다.
- 구어 자료의 주제 분류 다양화
원시 말뭉치의 주제가 한쪽으로 치우치지 않도록 다양한 주제의 자료 수집을 위해 방송사별로 3~4차례 이상 수집 목록에 대한 의견을 교환하고 목록별 수집 시간을 수정하였다. 저작권 이용 허락 계약 체결 후에도 예능, 스포츠 등 저작권 문제로 수집이 어려운 자료가 있어 이를 제외하고 다양한 주제의 자료 수집

을 위한 목록을 수정하고 수집하는 데에 많은 시간이 소요되었다.

- 준구어 자료 수집 시 저작권 문제 및 대본 품질

준구어 대본의 경우 유명 작가의 대본은 저작권 문제로 수집할 수 없었고, 저작권 문제가 발생하지 않는 자료 위주로 방송사와 한국방송작가협회의 동의하에 준구어 대본을 수집하였다. 대부분의 드라마 대본이 파일로 되어 있지 않아 수집에 어려움이 있었으며, 파일로 되어 있는 대본의 경우에도 보관을 위한 최종본이 아닌 방송사 소장용에 불과하여 철자나 맞춤법 등 오류가 상당히 많아 말뭉치 구축 시 품질을 높이는 데 어려움이 있었고, 많은 인력이 투입되어 대본을 수정하였다.

- 구축 자료 확정을 위한 수집 자료의 양

원시 말뭉치 구축 시 수집 자료의 품질 등을 이유로 제외되는 자료가 많을 것을 사업 초기에 예상하였다. 사업 초기 방송사와 협의하여 수집 대상을 사전 검토하여 선정하였고, 구축 목록에서 제외되는 자료가 있을 경우 바로 대체 자료를 방송사에 요청하여 사업 일정에 영향이 없도록 관리하였다.

- 구축 인력의 지침서 이해도 차이

원시 말뭉치 구축 인력만 월별로 약 200명 이상이 투입이 되므로 사전 교육과 실시간 품질 점검을 주기적으로 하여도 지침서에 대한 개인별 이해도 차이로 인한 오류가 발견되었다. 이를 개선하기 위해 구축 인력에게 피드백을 주기 위한 전담 인력과 품질 점검 인력을 별도로 조직하였고, 관리 인력을 예상보다 추가로 투입하여 품질에 이상이 없도록 관리하였다.

- 구축 목록에 따른 구축 인력의 세분화 필요

방송 내용 중 정치, 경제, 과학, 스포츠 분야의 경우 해당 분야 관련 지식이 없는 구축 인력이 작업할 경우 방송 내용에 대한 이해도 부족으로 용어 등에 내용 오류가 발생하는 것을 확인하였다. 이후 방송 내용에 대한 이력에 맞는 구축 인력을 투입하여 내용에 대한 오류가 없도록 개선하였다.

3.2. 향후 사업 진행을 위한 시사점

위와 같이 사업 수행 과정 중에 다양한 현안이 발생하였으나, 이를 해결하면서 사업 수행을 완료하였다. 향후 유사한 사업 진행 시 고려할 사항은 아래와 같다.

- 구어 및 준구어 자료 수집 부문

방송사별로 저작권에 대한 인식 및 법률적인 검토 의견이 다르므로 구어 및 준구어 자료 수집을 위해서는 방송사별로 여러 차례의 회의를 통해 저작권 이용 허락 계약에 대한 협의를 하는 것이 필요하다. 또한 방송별 혹은 회차별로 자료를 대량으로 받을 시 내용에 대한 검토를 사전에 하기 어려우므로 목표량의 배가 되는 분량을 수집하여 내용을 검토하는 공정이 반드시 필요하다. 또한 준구어 자료의 경우 방송사와 한국방송작가협회에서 저작권 이용 허락 동의는 가능하나 대본 자체가 국내에 파일로 되어 있는 경우가 거의 없거나 소량이므로 향후 준구어 자료 수집 시 방송사의 준구어 대본 보유 여부를 파악하는 것이 필요하다.

- 원시 말뭉치 구축 부문

일정 기간 내에 대량의 원시 말뭉치를 구축하므로 수백 명의 구축 인원 동원이 필수이다. 구축 인원별 사업 이해도와 수준의 차이가 있으므로 이를 교육하고 품질을 표준화할 수 있는 공정이 필요하다. 또한 공정별 자동화 도구를 사용하여 품질 오류가 발생할 가능성을 줄여야 사업 기간 내에 말뭉치 구축 목표를 달성할 수 있다.

현대 국어 구어 전사 말뭉치 지침

1. 사용도구

- 가) EmEditor를 다운받아 사용한다.(<https://ko.emeditor.com/>)
- 나) 배분된 음성 파일을 듣고 EmEditor에 전사하는 방식으로 작업한다.
- 다) 작업 후 파일저장은 UTF-8(서명없음)로 저장한다.

2. 전사지침

가) 발화자 표시

- ㄱ) 모든 발화자에 관한 정보는 파일 맨 상단에 [예시]와 같이 붙인다.
 - ◆ 가급적 주발화자를 1로 표시하고, 나머지는 등장 순서대로 한다.
 - ◆ 발화자 정보는 이름,성별,연령대,직업을 간단하게 표시한다.
 - ◆ 발화자에 대한 정보를 모를 경우에는 '?'로 표시한다.

[예시]

1: 정관용,남자,50대,시사평론가
2: 이나미,여자,?,정신과 전문의
3: ?,?,?,? ==> 청중

- ㄴ) 본문 전사에서 발화자 정보와 발화자 표시는 반드시 일치해야 한다.
 - ◆ 발화자가 분명하지 않을 경우에는 '?'로 표시한다.
 - ◆ 필요에 따라서는 '모두'나 '나머지' '청중' 등의 지칭을 사용할 수 있다.
 - ◆ 화자 2와 화자 3이 동시에 말하는 경우는 '2,3'으로 표시하기도 한다.

[예시]

1: 이 주제를 선택하신 이유가 궁금하네요.
?: 저 같은 경우에는
3,4: 예.
모두: 아~

- ㄷ) 발화자 표시는 나중에 TEI 태그로 일괄 전환되어야 하므로 일관성 있게 표시한다.

나) 전사 단위 기준

- ㄱ) 기본 전사 단위는 종결어미 기준으로 한다.

[예시]

1: 그리고 일본 정부도 더 이상 도쿄전력에만 사고처리를 맡길 수 없고 우리가 이제 직접 통제를 하겠다고 전 세계에 공언을 하면서 이제 이 문제가 다시 불거졌는데 사실은 일본 정부가 인정을 한 거죠.
2: 네.

단, 부사나 형용사가 종결 어미 뒤에 나오는 도치된 발화된 경우는 한 문장으로 본다.
(발화 내용이 한 문장으로 말하는 내용이면 한 문장으로 본다)

[예시]

1: 어제 뭘 샀다고?
2: 제가 사과를 샀어요. 어제
1: 저는 정말 좋아해요. 사과를

긍정의문, 부정의문의 경우도 동일하다.

[예시]

1: 할까요? 안 할까요?
1: 맞아요? 틀려요?

ㄴ) 긴 쉼에 의해 문장이 나뉘거나 말끝이 흐려진 채로 발화가 종료되면 한 문장으로 보고 전사한다.

[예시]

2: 당신은 마임이 아니다.
그러니까 이제 비주얼 시이터라고
1: 음

ㄷ) 느낌표나 쉼표는 사용하지 않는다. 문장이 완전히 종결이 되었을 때에 마침표를 사용한다.

[예시]

1: 자 이제 유월에 처음 맞는 음~ 이준형의 보물상자 이준형 씨 모셨습니다.
어서 오십시오.
2: 네, 안녕하세요.
1: 날씨 변화가 없었더라면 참 할 말이 많이 줄었을 거 같은 @웃음 생각이
2: 그럼요.
1: 네.
2: 네.
날씨는 정말 영원한 화제인 거 같아요.

ㄹ) 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분해 준다.(-어, -어요 등)

[예시]

1: 어제 새로 개봉한 영화 봤어.
1: 어제 새로 개봉한 영화 봤어?

다) 발화가 겹치는 경우

ㄱ) 겹침 발화는 시간 순서에 따라 적는다.

만약 맞장구 발화가 있을 경우 순서에 맞춰 맞장구 발화를 사이에 넣어 주발화를 나눈다.

네, 예 등의 맞장구 발화는 한 문장으로 보고 전사한다.

[예시]

1: 우리는 열심히 깨끗이 씻어서
 2: 어
 1: 막 분리해서 정해진 요일에 갖다 버렸는데
 2: 네.
 1: 정작 수거해간 사람은 왕창 그냥 버리더라. @웃음
 2: 네.
 1: 소각장으로 다 들어가 버리더라.
 2: 네.

ㄴ) 상대방 말에 맞장구 발화에서 네네, 예예, 음음 등의 경우는 붙여서 전사한다.
 [예시]

1: 다시 올라갔다는 건데
 2: 올라갔다는 거 아니에요?
 1: 네네. 대단합니다.

라) 전사 기준

- ㄱ) 발화 내용은 기본적으로 맞춤법을 기준으로 전사한다.
 다만, 구어의 발음 특성, 개인의 발음 특성, 지역적인 특성 등에 의해 맞춤법대로 소리 나지 않는 발음(표준 발음이 아닌 경우)은 발음 나는 대로 적는다.
 [예시]

1: 어~ 기후가 썩 좋지 않아서 그런가요?
 2: 그런 거 가태요.

ㄴ) 모음의 변화, 수의적 경음화 등을 반영하여 소리 나는 대로 전사한다.
 [예시]

1: 요즘 한약 먹느라 오늘 썩주는 못 마시겠네요.
 2: 그럼 어뜨케 썩끔도 못 마셔요?
 1: 한잔도 못 마셔요. 이제

ㄷ) 약화 현상에 의한 이형태는 반영하지 않는다.
 예를 들어 의문사 '뭐'가 '머', '모'로 모음이 약화되어 들려도 '뭐'로 전사한다.

ㄹ) 숫자나 기호, 영문 등도 발음에 따라 한글로 전사한다.
 [예시]

1: 네.
 아이엠에프 사태 직후에 그 눈물의 실업자들 비디오 다들 보고 말이죠.
 2: 예.
 1: 어~ 금 모으기 운동도 하고
 2: 맞습니다.
 1: 그리고 불과 뭐 한 일 년 남짓 후에?

- ㄱ) 불완전 발화나 수정 발화의 경우에 구별하는 표시 '='를 한다.
- 불완전 발화는 단어가 끊어지거나 불완전하게 발화된 경우를 말한다.
 - 수정 발화는 발화된 내용을 반복하여 수정한 발화를 말한다.
 - 발화된 그대로 전사하며, '='를 붙여 정상적인 단어와 구별한다.
 - 불완전하게 발화된 단어(어절)가 둘 이상인 경우 어절마다 '='를 붙인다.

- ◆ 단, 반복 발화에는 표시하지 않는다.

[예시]

1: 선배가 후배를 가르치는 구조를 =를 옛날에는 가지고 있었거든요.(불완전 발화)
2: 그랬죠.
1: 그래서 하부르따 학= 학생들끼리 이 공부하는 건데(불완전 발화)
2: 우리가 하= 생각할 수 있는 모든 날씨 얘기가 다 나옵니다.(수정 발화)
예외) 그 속지 속지 내지에다가 좀 중점을 둔 듯한 그런 느낌이에요.(반복 발화)

ㄴ) 띄어쓰기의 경우 맞춤법에 맞게 한다.

- ◆ 합성명사의 경우 대부분 붙여 쓴다.(*예, 발화 단위 → 발화단위)
- ◆ 의존명사는 띄어 쓴다.
- ◆ 수를 적을 때는 만 단위로 띄어 쓴다.(*예, 십이억 삼천백만 팔백구 달러 등)
- ◆ 수와 단위가 함께 쓰이는 경우 띄어 쓴다.(*예, 구십구 점 오 프로, 천구백삼십 년대 등)
- ◆ 판단하기 어려운 경우에는 수시로 논의하여 결정한다.(*예, 오십대, 일 대 이 등)
- ◆ 본 용언과 보조 용언도 띄어 쓴다.(*예, 구입하기로 했다, 드라마가 보고 싶다 등)

ㄴ) 축약형의 표기(정확한 발음이 나타난 경우에만 반영한다.)

- ◆ 두 음절이 한 음절의 사잇소리가 되거나, 두 음절이 한 음절 겹핥소리가 되는 등의 경우
- ◆ 발음되는 음절수와 표기상의 음절수를 맞추어야 하므로 축약형의 경우 모두 표기에 반영한다.

[예시]

1: 그니깐 그거 땀에 일루 올 수가 없는 거죠.
2: 아휴 그럼 어쨌요?

- ◆ 반모음 /ㄱ/, /ㄴ/의 경우 /ㄱ/, /ㄴ/와 축약되는 현상이 구어에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 전사 시 ' (작은따옴표, enter키 옆)를 사용해서 두 음소를 연결한다.

[예시]

1: 입학해서 첫 친구를 사꺀어요.(X) -> 입학해서 첫 친구를 사꺀었어요.(O)
그 친구 덕에 성격이 많이 바꺀죠.(X) -> 그 친구 덕에 성격이 많이 바뀌꺀쥬.(O)

ㄴ) 담화표지는 '~'로 표시한다.

“이, 그, 저, 아, 어” 등 기존 품사의 의미, 기능을 가지지 않는 것은 담화표지로 보고, '~'(물결표, 숫자1 key 옆)을 이용하여 표시한다.

(주로 머뭇거림의 표지로 사용되는 이~, 그~, 저~, 어~, 아~, 에~ 등이 해당됨.
인제, 이제, 그냥, 무슨, 어떤, 그게 등은 붙이지 않음)

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

[예시]

1: 그 이후에 변화하는 세상에 대응하는 어~ 우리의 능력은 조금 달랐어야 되는데
2: 아~
1: 그 변화를 만들지 못하다 보니까 어~ 우리가 생각했던 것만큼 어~ 큰 변화를 만들지 못하구 어떻게 보면은 조금 어~

ㄴ) 잘 들리지 않는 부분

- ◆ 잘 들리지 않는 부분은 들리는 대로 괄호 '()' 안에 전사한다.

[예시]

1: 집에 있는 히키코모리 대신 나왔(네요.) @웃음

2: 네.

집에 있는 히키코모리 대신 나오니라구

- ◆ 발화 내용이 전혀 들리지 않는 부분은 ‘...(마침표 3개)’로 전사한다.

[예시]

1: 더위가 요즘 심하다가 지금 장마도 엄청 심하고 굉장히 힘들잖아요.

어 또 배포 ...

네. @웃음

- ◆ 들리지 않는 음절은 그 음절의 수만큼 x(영문 소문자)를 붙인다.

[예시]

1: 입을 썸 풀어보도록 하겠습니다.

2: 아 네.

많은 고민들을 안고 계시는데 우리 부모님들께서도 이 시간 귀
xxxx 것 같습니다.

그러면 미술교육을 전공하셨어요?

㉞) 노트 처리(전사자의 설명)

- ◆ 전사가 어려운 사항이나 특별한 설명을 붙일 필요가 있을 때는 발화자 경계에서 (엔터를 넣고) 노트로 내용을 적고 발화를 전사하지 않는다. (노트 처리는 정해진 입력 메시지는 없으며, 간단하게 노트 처리 사유를 쓴다.)

[예시]

1: 한 범조인을 만나보겠습니다.

<note>배경 화면 잠깐 나옴</note>

이 자리에 기독교반성폭력센터의 이사장이시죠.

박종운 변호사님을 모셨습니다.

- ◆ 주 발화자의 발화 외 관련 화면이 삽입된 경우, 전사하지 않고 다음과 같이 노트로 적는다.

[예시]

1: 바로 세입자들의 이야깁니다.

<note>관련 인터뷰 화면 나옴</note>

이십년 동안 무려 스무 번.

- ◆ 외국인 발화자(패널)이 나오는 경우, 이 부분의 대화는 전사하지 않고 역시 노트 처리한다. (이 경우 작업 파일명에 ‘외국인 발화’를 표시한다.)

[예시]

<note>외국인 발화자 나옴</note>

- ◆ 수화가 나오는 경우, 이 부분의 자막을 전사하지 않으며 노트 처리한다.

[예시]

<note>수화 나옴</note>

- ◆ 정식 방송 시작 전의 예고편이 앞부분에 포함되어 있는 경우, 전사하지 않고 노트 처리한다.

[예시]

<note>앞부분에 예고편이 편집되어 있음</note>

- ◆ 라디오, 방송에서 프로그램에 대한 오프닝 멘트에 대해 전사하지 않고 노트 처리한다.

[예시]

<note>오프닝 멘트 있음</note>

- ◆ 라디오, 방송에서 프로그램에서 협찬 및 광고 멘트는 전사하지 않고 노트 처리한다.

[예시]

2: 여성시대 사연이 소개되신 분들께는 이런 선물을 드립니다.

<note>광고 멘트 있음</note>

ㄱ) 준음성 및 기타 소리 전사

- ◆ 웃음소리나 박수소리, 노래를 부르는 등의 기타 소리는 '@'표시로 태그하여 전사한다.
- ◆ @로 표기하는 요소는 "@웃음, @목청, @박수, @노래"의 4가지로 국한한다.
이 4가지는 음성 상징으로 들리는 대로 적지 않고, 태그로만 전사한다.
('@목청'은 목청 가다듬는 소리/크흠 등/를 의미함)

[예시]

1: 해가 좋다 비가 온다 뭐 다음 주는 좋을 거라더라 뭐 등등

2: @웃음 뉴스에서 봤는데 뭐

- ◆ 감탄, 놀람 등은 "오오, 앓, 어머" 등으로 음성상징어에 따라(들리는 대로) 전사하며, '@'표시로 태그하지 않는다.

1: 어머 그런 일이 있었어?

- ◆ 노래는 가사를 적지 않고 해당 부분에 "@노래"로만 표기한다.
- ◆ 시를 읊는 경우는 전사한다.

전사 말뭉치 마크업 지침

1. 전체구조

(전사)

1: 그래서 그랬는데 이번에 여의도에 갔었는데
여의도 거기 벚꽃 했잖아요

2: 윤중로

1: 예.

(마크업)

<u who="P1" n="1">그래서 그랬는데 이번에 여의도에 갔었는데</u>

<u who="P1" n="2">여의도 거기 벚꽃 했잖아요</u>

<u who="P2" n="3">윤중로</u>

<u who="P1" n="4">예.</u>

(참고)

1) 콜론 다음에 스페이스가 있기도 하고 없기도 한다.

2) 발화자에 대한 정보를 모를 경우에는 다음과 같이 표시한다.

(전사)

?: 예.

(마크업)

<u who="unknown" n="4">예.</u>

2. 발화자가 2명 이상의 동시 발화인 경우와 기타

<u who="P3,P4" n="5">그지.</u>

‘모두’나 ‘나머지’는 all, others 등으로 표시한다.

3. 끊어진 단어

(전사)

전= 전= 전통이라고 우리가 흔히 얘기할 때

(마크업)

<u who="P1" n="4"> <trunc>전</trunc> <trunc>전</trunc> 전통이라고 우리가 흔히 얘기할 때</u>

4. 잘 들리지 않는 부분(추정 전사)

(전사)

그 전까지는 직장 생활 하니라구 (더 힘들어)

(마크업)

<u who="P1" n="4">그 전까지는 직장 생활 하니라구 <unclear>더 힘들어</unclear></u>

5. 잘 들리지 않는 부분(전사 못한 경우)

(전사)

2: ... 너무한 거 같더라. => 마침표 3개

(마크업)

<u who="P1" n="4"> <unclear/> 너무한 거 같더라.</u>

잘 들리지 않는 음절

들리지 않는 음절은 그 음절의 수만큼 x(영문 소문자)를 붙인다.

예.

2:근데 그거 진짜 xx해야 되겠더라.

<u who="P1" n="4"> 근데 그거 진짜 <unclear>xx해야</unclear> 되겠더라.</u>

6. 전사자의 설명

(전사)

1: 뉴스룸의 앵커브리핑을 시작하겠습니다.

삼김퀴즈.

<note>배경 화면 잠깐 나옴</note>

최양락 배철수 콤비가 진행한 라디오 시사 토크였습니다.

(마크업)

<u who="P1" n="1">뉴스룸의 앵커브리핑을 시작하겠습니다.</u>

<u who="P1" n="2">삼김퀴즈.</u>

<note>배경 화면 잠깐 나옴</note>

<u who="P1" n="4">최양락 배철수 콤비가 진행한 라디오 시사 토크였습니다.</u>

7. 준음성과 기타 소리들

(전사와 마크업 매칭)

@웃음 <vocal desc="laughing"/>

@목청 <vocal desc="목청가다듬는소리"/>

@박수 <vocal desc="applauding"/>

@노래 <vocal desc="singing"/>

8. 어절수에 제외되어야 할 태그

- 발화자 태그
- note 전체
- <unclear/>
- <vocal desc="applauding"/>
- <vocal desc="singing"/>
- <anon type="name1"/>는 1어절로 계산

맞춤법 및 표기와 관련된 지침

1. 축약형 표기

언어경제성의 원칙에 의해 구어에서는 아래와 같은 축약형이 많이 나타나며, 이는 모두 표기에 반영한다.

(1) 준말형

○ 뒤홀소리로 바뀐 후 축약

일부 어휘 내에서 ‘ㄷ, ㅌ, ㄱ, ㅋ’ 모음에 앞홀소리 ‘ㅣ’가 후행하면 ‘ㄷ, ㅌ, ㄱ, ㅋ’가 뒤홀소리 ‘ㅌ, ㅍ, ㄲ, ㅋ’로 바뀌고 축약 현상이 일어난다. ‘여기, 거기’가 ‘예, 게’가 되는 예에서처럼 그 사이 자음이 있으면 그 자음은 탈락되고 축약된다.

예 아이→예, 사이→새, 요사이→요새, 이야기→얘기, 여기→예, 거기→게

○ ‘ㅡ’ 탈락 후 축약

‘다음, 마음, 처음’과 같은 어휘 내에서 ‘음’에서의 ‘ㅡ’가 탈락하고 한 음절로 축약된 형태가 나타난다.

예 다음→담, 마음→맘, 처음→참

○ ‘ㅣ’ 탈락 후 축약

아래와 같은 일부 어휘 내에서 ‘ㅣ’모음이 탈락하고 축약된 형태가 나타난다.

예 제일→젤, 내일→넬, 재미있다→재밌다, 가지다→갓다

○ ‘그’ 탈락 후 축약

‘조금, 지금’ 등의 단어에서 ‘그’가 탈락하고 ‘죵, 짐’ 등의 형태가 된다.

예 조금→죵, 지금→짐

○ ‘러’ 탈락 후 축약

지시부사 ‘이렇게, 그렇게, 저렇게’에서 ‘러’가 탈락되고 축약되어 ‘이케, 그케, 저케’의 형태로 나타난다.

예 이렇게→이케, 그렇게→그케, 저렇게→저케

(2) 준꼴형

○ ‘ㅣ’ 탈락 후 축약

‘지’와 높임을 나타내는 보조사 ‘요’가 결합한 꼴인 ‘지요’에서 ‘ㅣ’가 탈락하고 축약되어 ‘쵸’의 형태가 된다. 그리고 ‘이다’의 활용형에서 앞의 체언이 받침이 없는 경우 ‘이다’의 ‘ㅣ’가 탈락하고 축약된 형태가 많이 나타난다. 특히 ‘것’과의 결합에서 ‘것’의 ‘ㅅ’과 ‘이다’의 ‘ㅣ’가 함께 탈락하고 줄어드는 형태가 많다.

▶ 어미 지요→ 쵸

예 말하지요→말하쵸, 거지요→거쵸, 드시지요→드시쵸, 있겠지요→있겠쵸

▶ ‘이다’의 여러 활용형에서

입니다→ㅂ니다, 이다→다, 인데→ㄴ데, 인 줄→ㄴ 줄,

인지→ㄴ지, 이거든요→거든요, 이네→네

예 상태입니다→상탭니다, 것이거든요→거거든요, 것인지→건지

▶ 일부 어휘에서

예 어디다→어따

○ ‘ㅎ’ 탈락 후 축약

‘놓다, 넣다, 날다, 달다, 땀다, 뺏다, 쌓다’ 등 ‘ㅎ’으로 끝나는 용언의 활용형에서 ‘ㅎ’이 탈락하고 축약된 형태가 나타난다.

예) 놓아→놔, 놓아야→놔야, 놓은→논
 넣었습니다→넉습니다, 넣었더니→넉더니
 낳았으면→났으면

○ ‘ㅅ’ 탈락 후 축약

지시대명사 ‘이것, 그것, 저것, 요것, 조것’이나 의존명사 ‘것’, 의문대명사 ‘무엇’이 ‘이’, ‘은’, ‘을’, ‘으로’ 등의 조사와 결합할 때 ‘ㅅ’이 탈락하고 축약된 형태가 나타난다. ‘은’, ‘을’과의 결합시에는 ‘ㅡ’도 같이 탈락하며, ‘으로’와의 결합시에는 ‘ㄹ’이 첨가되는 경우도 있다.

▶ 지시대명사와 조사와의 결합형에서

예) 이것이→이게, 이것은→이건, 이것을→이걸, 이것으로→이걸로
 그것이→그게, 그것은→그건, 그것을→그걸, 그것으로→그걸로

▶ 의존명사 ‘것’과 조사와의 결합형에서

예) 것이→게, 것은→건, 것을→걸, 것으로→걸로(거로)

▶ 의문대명사 ‘무엇’과 조사와의 결합형에서

예) 무엇→뭐, 무엇이→뭐가, 무엇을→뭐를→뭇→뭇, 무엇으로→뭇로

○ ‘ㄴ’ 탈락 후 축약

체언이나 조사, 어미 등이 조사 ‘는’과 결합할 때 ‘ㄴ’이 탈락하고 축약된 형태가 나타난다.

▶ 체언+는

예) 호랑이는→호랑인, 볼 수는→볼 순, 익히는 데는→익히는 덴

▶ 어미/조사+는

에는→엔, 에서는→에선, 으로는→으론, 한테는→한텐
 고는→곤, 기는→긴, 까는→깐, 다가는→다간
 라는/ 다는/ 자는→란/ 단/ 잔, 서는→선, 지는→진

예) 옛날에는→옛날엔, 아가씨한테는→아가씨한텐, 오래됐다고는→오래됐다라고곤

○ ‘ㄹ’ 탈락 후 축약

체언이나 조사, 어미 등이 조사 ‘를’과 결합할 때 ‘ㄹ’이 탈락하고 축약된 형태가 나타난다.

▶ 체언+를

예) 너를→넌, 영화를→영활

▶ 어미/조사+를

예를→엘, 게를→겔, 기를→길, 지를→질

예) 학교예를→학교엘, 풀지를→풀질, 먹기를→먹길, 하지를→하질

○ ‘ㅈ’ 탈락 후 축약

‘하다’로 끝난 일부 용언이 ‘기, 지, 도록’ 등의 어미와 결합할 때 ‘ㅈ’이 탈락되고 축약된 형태가 나타난다.

예) 예상하기로는→예상키로는, 심심하지→심심치, 조사하도록→조사토록

2. 그 외의 이형태 표기

(1) 탈락 현상에 의한 이형태

시간적인 제약을 받는 구어에서는 아래와 같은 탈락 현상이 많이 나타난다.

○ ‘ㄹ’탈락

어미 ‘(으)ㄹ게(요), (으)ㄹ까(요)’에서 ‘ㄹ’이 탈락하고 ‘(으)께(요), (으)까(요)’가 되는데 ‘(으)ㄹ게(요)→(으)께(요)’의 경우 ‘게’가 ‘께’로 되는 된소리화 현상도 같이 일어난다.

▶ 어미에서

(으)ㄹ게(요)→(으)께(요)

(으)ㄴ까(요)→ (으)까(요)

예 드릴게요→드리게요, 줄게→주게, 뻔까요→뻔까요, 좋을까요→좋으까요

*[드릴게요]라고 제대로 발음한 경우는 '드릴게요'로 적음

▶ 일부 어휘에서

예 둘째→두째

○ 'ㅅ' 탈락

'것'이 포함된 지시대명사와 의존명사 그리고 어미형태류 ' 것 같아요'에서 '것'의 'ㅅ'이 탈락한 형태가 많이 나타난다.

▶ 지시대명사와 의존명사에서

예 이것→이거, 것은→거는, 그것→그거, 것에→거에, 저것→저거, 것만→거만

▶ 것 같아요→ 거 같아요.

예 했었던 것 같아요→했었던 거 같아요

○ '가' 탈락

' 다가'가 포함된 어미와 조사 그리고 일부 어휘에서 '가'가 탈락하는 경우가 있다.

▶ 어미와 조사에서

에다가→에다, 다가→ 다, 어다가→ 어다

예 손에다가→손에다, 넣다가→넣다, 꺾다가→꺾다

▶ 일부 어휘에서

예 깨다가→깨다

○ '고', '구' 탈락

어미 ' (으)ㄴ려구(라구)'나 ' 라고/ 다고/ 자고'에서 '고'나 '구'가 탈락하는 경우가 있다.

(으)ㄴ려구(라구)→ (으)ㄴ려(라)

라고/ 다고/ 자고→ 라/ 다/ 자

예 할려구(할라구)→할려/할라

결혼했다고→결혼했다, 빼라고→빼라, 있으라고→있으라

○ '기' 탈락

'여기, 거기'에서 '기'가 탈락하고 '여'나 '거'의 형태로 나타난다.

예 여기→여, 거기→거

○ '그' 탈락

지시대명사 '그것'이나 'ㅅ'이 탈락한 형태인 '그거'에서 '그'가 탈락하고 '것'이나 '거'로 나타나는 경우가 있다.

예 그것→것, 그거→거

○ 조사 '에' 탈락

체언과 조사 '에', '에서'와의 결합에서 '에'가 탈락되는 경우가 많다.

예 여기에다가→여기다가, 때에도→때도, 거기에서→거기서, 주는 데에서→주는 데서

(2) 모음의 단순화 현상에 의한 이형태 (반영 X)

시간상의 제약과 편의성을 추구하는 구어에서는 “봤거던→봤거던, 뇌→나, 관계→간계, 최근→채근, 되다→대다, 되게→대게, 뭐→머, 거예요→거예요” 등 이중모음을 단순화시켜서 발음하는 일이 종종 있다. 경향성은 인정하되 실제 전사에서는 반영하지 않기로 한다.

(3) 교체 현상에 의한 이형태

○ 조사 '의'의 발음

조사 '의'의 경우 구어에서 '에'로 발음하는 경우가 많기 때문에 이 정보를 반영해야 할 필요가 있다. 따라서 교체가 되어 발음이 [에]로 발음되면 '에'로 적고, [의]로 발음되는 경우만 '의'로 적는다.

○ ‘ㄴ’가 ‘ㄷ’로 바뀔

‘ㄴ’가 ‘ㄷ’로 바뀌는 현상은 구어에서 아주 광범위하게 나타나는 현상이다. 같은 등근 홀소리이고 뒤홀소리이지만 ‘ㄷ’가 ‘ㄴ’보다 발음이 쉽기 때문에 이러한 바뀔 현상이 생긴다.

▶ 조사에서: 도→두, 으로→으루

예) 오늘도→오늘두, 애도→애두, 구체적으로→구체적으루, 손으로→손으루

▶ 어미에서

-고→ 구, 고서→ 구서, 고요→ 구요, 더라고요→ 더라구요
라고/ 다고/ 자고→ 라구/ 다구/ 자구
아/어/여도→ 아/어/여두

예) 얘기하고→얘기하구, 가지고서→가지구서, 뜨고서는→뜨구서는, 언어고요→언어구요
뭐라고→뭐라구, 안다고→안다구, 몰라도→몰라두

▶ 일부 어휘에서

예) 하도→하두, 그래도→그래두, 별로→별루, 바로→바루,
서로→서루 그대로→그대루, 함부로→함부루

○ ‘ㄷ, ㄱ’가 ‘ㄱ, ㄴ’로 바뀔

구어에서는, 후설모음(뒤혀홀소리) ‘ㄷ, ㄱ, ㄴ, ㄷ, ㅡ’ 뒤에 전설모음(앞혀홀소리) ‘ㄱ’이 후행하면 역행동화 현상으로 후설모음 ‘ㄷ, ㄱ, ㄴ, ㄷ, ㅡ’가 전설모음 ‘ㄱ, ㄴ, ㄱ, ㄱ, ㅏ’로 변동하는 전설모음화 현상이 나타난다. 이러한 전설모음화 현상은 보편적, 필연적 변동 현상에 의한 것이 아니라 수의적인 이형태이므로 표준 표기법에는 반영되지 않지만, 구어 전사시에는 이를 반영하기로 한다.

전설모음화 현상 뿐 아니라 ‘만들다→맨들다, 놀라다→놀래다’와 같은 일부 용언이나 ‘라서→래서, 라던데→래던데’ 등 일부 어미 및 어미형태류에서 이러한 바뀔 현상을 확인할 수 있다. 특히 간접 인용형이 포함된 어미형태류들에 그러한 예가 많다.

▶ 일부 어휘에서

예) 창피하다→챙피하다, 먹이다→멕이다, 벗기다→벧기다
만들다→맨들다, 놀라다→놀래다, 같아요→갈해요, 간덩이→간텅이
꿀보기 싫다→꿀뵈기 싫다

▶ 어미 및 어미형태류에서

더라도(두)→ 더래도(두), 라서→ 래서
라는/ 다는/ 자는→ 래는/ 대는/ 재는
라면/ 다면/ 자면→ 래면/ 대면/ 재면
라니까/ 다니까/ 자니까→ 래니까/ 대니까/ 재니까
라던데/ 다던데/ 자던데→ 래던데/ 대던데/ 재던데
라더라/ 다더라/ 자더라→ 래더라/ 대더라/ 재더라

예) 하더라도→하더래도, 한다는→한대는, 심하다니까요→심하대니까요
그랬다던데→그랬대던데

○ ‘ㄱ’가 ‘ㄷ’로 바뀔

선어말어미 ‘더’가 ‘드’로 바뀐다든지 어미 ‘거든요, 거든’에서 ‘거’가 ‘그’로 바뀐다든지 혹은 일부 어휘에서 ‘ㄱ’가 ‘ㄷ’로 바뀌는 경우가 있다.

▶ 선어말어미 더 → 드

더라구(요)→ 드라구(요), 던가요→ 든가요, 더라는→ 드라는

예) 가더라구요→가드라구요, 평범하더라는→평범하드라는, 좋던데→ 좋든데

▶ 어미에서

거든요→ 그든요, 거든→ 그든

예) 당사국이거든요→당사국이그든요, 있거든→있그든

▶ 일부 어휘에서

예 정말→증말, 그런→그른, 이런→이른, 어른→으른, 넣다→눔다
이렇게/그렇게/저렇게/어떻게→이릉게/그릉게/저릉게/어똥게

○ ‘ㄴ’가 ‘ㄷ’로 바뀜

‘든요’가 ‘던요’로 ‘르든지’가 잘못 쓰인 꼴인 ‘르른지’가 ‘르런지’로 조사 ‘라든가’가 ‘라던가’로 일부 어미 및 어미 형태류와 조사에서 ‘ㄴ’가 ‘ㄷ’로 바뀌는 경우가 있다.

▶ 어미에서

거든요→ 거던요(덩요)

르든지→ 르런지

예 그랬거든→그랬거덜, 따르거든→따르거덜, 예쁘거든→예쁘거덜, 하거든→하거덜
떨든지도→떨런지도

▶ 조사에서

라든가→라던가

예 자음들이라든가→자음들이라던가, 교재라든가→교재라던가

○ ‘ㄴ’가 ‘ㄷ’로 바뀜

‘잠그다, 담그다’ 등의 용언에서 ‘ㄴ’가 ‘ㄷ’로 교체된 어간 이형태가 나타난다.

예 잠그다→잠구다(잠가→잠귀), 담그다→담구다(담가→담귀)

○ ‘ㄷ’가 ‘ㄷ’로 바뀜

‘연거푸’ 등의 단어에서 ‘ㄷ’가 ‘ㄷ’로 교체된 이형태가 나타난다.

예 연거푸→연거퍼

○ ‘ㄷ’가 ‘ㄷ’나 ‘ㄴ’로 바뀜

‘-면, -라면서/다면서/자면서(요)’ 등의 어미형태류에서 ‘면’이 ‘문’으로 되거나 ‘라면서/다면서/자면서’의 준꼴인 ‘-라며/다며/자며’에서 ‘며’가 ‘메’로 바뀌어서 발음 나는 경우가 있다. 그리고 ‘며칠→메칠, 몇→멧’ 등의 일부 어휘에서도 이러한 바뀔 현상을 확인할 수 있다.

▶ 어미 및 어미형태류에서

라며/다며/자며→ 라메/다메/자메, 면→ 문, 면서→ 문서

라면서/다면서/자면서(요)→ 라문서/다문서/자문서(요)

예 좋아한다며→좋아한다메, 나오면→나오문, 먹으면서→먹으문서, 한다면서→한다문서

▶ 일부 어휘에서

예 별안간에→베란간에, 몇→멧, 며칠→메칠, 별의별→벨에벨

○ ‘ㅏ’가 ‘ㅏ’로 바뀜

일부 어휘에서 ‘ㅏ’가 ‘ㅏ’로 바뀐다.

예 한테→헌테, 같고→겉고, 납파하고→납파허고

○ ‘ㅓ’가 ‘ㅓ’로 바뀜

일부 어휘에서 ‘ㅓ’가 ‘ㅓ’로 바뀐다.

예 예쁘다→이쁘다, 계집애→기집애

○ ‘ㅓ’가 ‘ㅓ’로 바뀜

‘줍다’의 경우 사람들이 기본형을 ‘줍다’가 아니라 ‘줏다’로 인식해서, 활용형에서도 ‘주워서, 주운’이 아닌 ‘주서서, 주슨’ 등으로 나타난다.

예 줍다→줏다(주워서→줏어서, 주운→줏은)

(4) 된소리화 현상에 의한 이형태

근대국어, 현대국어를 거치면서 개별적인 단어들에서 평음이 된소리나 거센소리로 바뀌는 현상이 계속 되고 있으면, 이러한 현상은 말의 변화가 가장 첨예하게 나타나는 구어에서 많이 나타난다.

○ 어두에서의 된소리화 현상에 의한 이형태

어두에서의 된소리화 현상은 순우리말 단어에서 많이 나타나며, 방언의 영향도 적지 않은 것으로 보인다. 구체적인 예를 품사별로 나누어서 살펴보면 다음과 같다.

▶ 명사에서

예 고추장→꼬추장, 고춧가루→꼬춧가루, 소주→쏘주, 집게→쩍게, 꽃감→꽃감
조각→쪼각, 생맥주→쌩맥주, 파사무실→파사무실, 중국→쥬국, 자투리→짜투리
숙맥→쑥맥, 쯔뽕이→쑤뽕이, 껌동이→껌동이, 고물→꼬물

▶ 동사, 형용사에서

예 (힘이)달린다→딸린다, 당기다→땅기다, 던지다→뎌지다, 부러지다→뿌러지다
(체중이)붙다→뽏다, 베끼다→뻬끼다, 자르다→짜르다, 절다→쩔다, 줄다→쥘다
세다→썰다, 동그랗다→똥그랗다, 작다→ 짹다, 조그맣다→쪼그(끄)맣다

▶ 부사, 관형사에서

예 거꾸로→꺼꾸로, 좀→쑤, 조금→쪼금(쪼끔/쪼뽀)
다른→따른

▶ 조사에서

예 밖애→뽓애

(5) 첨가 현상에 의한 이형태

발음의 편의성을 추구하는 구어에서는 유음 ‘ㄹ’이나 비음 ‘ㄴ, ㄹ, ㅇ’, 반모음 ‘j’ 등이 첨가되는 일이 있다.

○ ‘ㄹ’첨가

▶ 일부 용언에서

예 누르다→눌르다, 다르다→달르다, 모르다→몰르다, 바르다→발르다, 부르다→불르다
어르다→얼르다, 오르다→올르다, 지르다→질르다, 흐르다→흘르다

▶ 어미에서

(으)려면→ (으)르려면
(으)려고(구)→ (으)르려고(구)

예 올라가려고→올라갈려고, 오려면→올려면

▶ 일부 어휘에서

예 여기로는→여길로는, 날아→날라

○ ‘ㄴ’ 첨가

‘이제’가 ‘인제’로 되는 예에서 ‘ㄴ’ 첨가의 예를 찾을 수 있다. 그리고 ‘매일’의 구어형인 ‘맨날’에서도 좀 더 복합적인 현상이긴 하지만 ‘ㄴ’ 첨가 현상을 볼 수 있다.

예 이제→인제, 균열→균널

○ ‘ㅁ’ 첨가

조사 ‘보다, 부터’가 구어에서는 ‘보담, 부터’라는 이형태를 가진다.

예 과거보다도→과거보담도, 여기부터→여기부터

○ ‘ㅇ’ 첨가

조사 ‘까지’가 구어에서는 ‘까징’이라는 이형태로 실현되기도 한다.

예 상황까지→상황까징

(6) 모음조화파괴 현상에 의한 이형태

모음 조화의 파괴 현상은 한국어에서 통시적으로 진행되고 있으며, 문어와 구어에서 모두 나타나지만 구어에서 그 현상이 더욱 두드러진다.

예) 달라→달러, 따라→따러, 알아→알어, 많아→많어
나빠→나빠, 갈라→갈러, 맞아→맞어, 아파→아퍼
앉아→앉어, 팔아→팔어, 앉아서→앉어서

(7) 부사화 접미사와 관련된 이형태

한글맞춤법 제 51항에서는 부사의 끝 음절이 분명히 ‘이’로 나는 것은 ‘-이’로 적고, ‘히’로만 나거나 ‘이’나 ‘히’로 나는 것은 ‘-히’로 적는다고 규정되어 있다. 일반적으로 구별하는 방법은 ‘-하다’가 붙는 말은 ‘-히’를 그렇지 않은 말은 ‘-이’를 붙이면 된다는 것이다. 그러나 ‘-하다’가 붙는 말 중에서 어근의 끝소리가 ‘ㅅ’, ‘ㄱ’일 경우에는 ‘-이’가 붙는다. ‘깨끗이, 나긋나긋이, 기웃이, 남짓이, 느긋이, 따뜻이, 뜨뜻이, 반듯이, 빠듯이, 산뜻이, 의젓이, 지긋이, 너부죽이, 큼직이’ 등이 그 예이다. 이 중 ‘깨끗이, 땃땃이, 뜨뜻이, 따뜻이, 산뜻이, 큼직이, 너부죽이’ 등은 실제 구어에서 ‘-이’대신에 ‘-히’가 붙기도 한다. 이 경우 아래와 같이 표기한다.

예) 깨끗이→깨끗히, 땃땃이→땃땃히, 뜨뜻이→뜨뜻히, 따뜻이→따뜻히, 산뜻이→산뜻히

(8) 사이시옷과 관련된 이형태

구어에서는, 사이시옷 또한 표준 규정(한글맞춤법 제30항)과는 다른 형태로 나타나는 경우가 많다. 표준어에는 있으나 없는 형태로 발화되기도 하고, 표준어에는 없지만 실제 발화에서는 사이시옷이 나타나기도 한다.

예) 노랫말→노래말, 머리말→머릿말, 인사말→인삿말, 기와집→기왓집, 나이대→나잇대

(9) 복합형 및 기타 이형태

○ 접속부사

‘그’ 류의 접속부사는 아래와 같이 구어에서 다양한 형태로 나타난다.

예) 그리고→그르구/그르고/그리구/그려구/글구, 그런데→그른테/근테
그러면→그르면/그러면/그면, 그러면→그르면/그럼
그러니까→그르니까/그니까/그까/근까, 그러니깐→그르니깐/그니깐/그깐/근깐
그렇지만→그룽지만/그치만, 그렇다면→그룽다면/그타면
그런데도(두)→그른데도(두)/근데도(도), 그럼→금, 그래서→그서

○ 기타

이 외에도 ‘(으)려고(구)’가 ‘(으)라고(구)’가 되는 것 등의 여러 가지 복합적인 변이형이 나타나고 있다.

▶ 어미에서

(으)려고(구)→(으)라고(구)

예) 하려구→할라구, 말씀하시려고→말씀하실라고

▶ 지시관형사 ‘이, 그, 저’+아이(는)

예) 이 아이→애 이 아이는→앤
 그 아이→개 그 아이는→겐
 저 아이→재 저 아이는→젠

▶ 특이형

예) 그렇죠→그룽쵸/그쵸/그쵸, 그렇지→그룽지/그치/그지
 큰일났다→쿨났다, 때문에→땀에, 내버려두다→내비두다/냅두다, 그만두다→관두다
 조그마하다→쭈만하다
 너, 네→니, 너희→너네/느네/늬 자기→지, 그냥→기냥
 매일→맨날, 무슨→뮌/먼

(10) 표기에 반영하지 않는 이형태

○ 약화 현상에 의한 이형태

‘은행, 전화, 팬히, 지하철, 그니까, 아니, 그것두’ 등 구어에서는 ‘ㅎ, ㄴ, ㄱ’이 ‘비음/모음+모음’ 환경에서 약화되는 예들이 나타나거나, 경우에 따라서는 완전히 탈락되는 것처럼 들리기도 하지만, 이를 표기에 반영하지 않는다.

○ 조음위치동화(변자음화)에 의한 이형태

‘신문/심문/, 엿보다/엮보다/, 받고/박고/, 손가락/송가락/, 감기/강기/, 밥그릇/박그릇/’ 등 구어에서는 조음위치동화(변자음화)에 의해 양순음 ‘ㅂ, ㅁ’ 앞에서 치조음 ‘ㄷ, ㄴ’이 양순음으로 되거나 연구개음 ‘ㄱ, ㅇ’ 앞에서 치조음 ‘ㄷ, ㄴ’이나 양순음 ‘ㅂ, ㅁ’이 연구개음으로 변동하는 현상이 나타난다. 이 현상은 종성(받침)규칙을 거쳐서 적용되므로, ‘ㄷ, ㄴ’, ‘ㅂ, ㅁ’ 외에도 종성규칙을 거쳐서 ‘ㄷ, ㄴ’, ‘ㅂ, ㅁ’이 된 ‘ㄷ, ㅌ, ㅊ, ㅊ, ㅎ’과 ‘ㅍ’ 등의 음절말 자음에도 적용된다. 그렇지만, 실제 전사시 이 현상을 구별하여 표기하기가 현실적으로 어려우므로 반영하지 않는 것을 원칙으로 한다.

3. 방언형 표기

○ 방언형은 다른 이형태와 마찬가지로 발음 나는 대로 표기를 한다.

○ 방언형인지 구어 변이형인지 분명하지 않은 경우에는 목록화해 두기로 한다. 예를 들어 ‘쑥맥’같은 경우 연세한국어사전(1998)에서는 ‘숙맥’의 비표준어로 보고 있으나, 표준 화자들의 발음에서 보편적으로 나타나는 현상이라는 점을 고려하면 구어 변이형으로 보는 것이 바람직하다.

예) 억수로

4. 약어 표기

○ ‘IPA(International Phonetic Alphabet)’, ‘for example→e.g’ 등 영어에서의 약어는 음소 차원으로 표시되는 경우가 많지만, ‘비냉, 불낙, 국영수’ 등 한국어의 약어(abbreviation)는 음절 차원에서 머리글자로 표시되는 경우가 대부분이다. 따라서, 원래의 형태로 복원하지 않고 줄인 꼴 그대로 표기하는 것을 원칙으로 한다.

예) 비냉, 불낙, 국영수, 중고등학교, 오공, 육공

5. 숫자 표기

○ 숫자는 아래와 같이 표기 원칙에 맞춰 한글로 적는다.

예) 2002년→이천이년, 5개→다섯 개

6. 외래어.외국어 표기

○ 한국어의 경우 표준맞춤법에서 정한 외래어, 외국어 표기와 실제 발음이 다른 경우가 많다. 따라서, 이를 현행 맞춤법 규정에 따라 표기할 경우 실제 구어 발화의 특징이 반영되지 않는다는 문제가 발생한다. 따라서 외래어나 외국어 또한 가능한 한 실제 발화에 가깝게 전사하는 것을 원칙으로 한다. 물론 이 경우에도 표기 원칙에 맞춰 한글로 적는다.

○ 외래어에서의 된소리 표기

예) 가스→까스, 달러→딸라, 다운→따운, 케이크→케익 플러스→쁘라스

7. 유행어.신조어 등의 표기

○ 기존 사전에 표제어로 등재되어 있지 않은 신조어, 유행어, 비어, 속어 등의 경우는 표준 표기형을 정하고 목록화 작업을 거친 후 이후의 전사에 일관되게 반영한다.

예) 궁시렁거리다, 뽕가다, 꼬장부리다, 뽕세다, 짬밥, 꾀살이끼다, 뽕본, 무대뽕,
뽕사시하다, 뽕때리다

8. 질문 사항 정리

*전사자들의 질문 사항을 수집한다.

○ ‘-마는’과 ‘-면은’ : ‘-마는’은 하나의 어미이다. 이 둘은 자주 나오는 형태이므로 주의한다.

예) 뭐~ 확이야 그대로 그을 수는 있겠지마는(있겠지만은x)
군사 물자를 가지고 오면은(오며는x)

○ 띄어쓰기 정리 : ~부작, ~년생, ~원대

- 이십사 부 작
- 구십칠 년 작
- 구십 년생
- 만 원대
- 구십구 점 오 퍼센트

<Abstract>

Spoken data collection and raw corpus establishment project

This research is the spoken data collection and raw corpus establishment project and the purpose of this research is to establish a new corpus of 15,000 hours of spoken words, including lectures and debates, and 15.4 million words, including drama scripts, according to the data collection and corpus establishment guidelines. The main tasks and research results in accordance with such purposes are as follows.

Spoken/semi-spoken data collection: Spoken data from media such as TV, radio and Internet broadcasting, and semi-spoken data from drama scripts, are collected and an agreements for permission to use copyrights for improving data utilization by the private sector is concluded. It is verified whether or not collected data are valid data for industries and the academic world, and data which are not valid are excluded from the target for collection objective.

Raw corpus establishment: Collected data with verified validity are transcribed in compliance with the corpus establishment guidelines. A raw corpus of 15,000 hours of spoken words, including meta information, and over 15.4 million semi-spoken words is established targeting the transcribed result.

Raw corpus utilization: The cases and direction of established raw corpus utilization are presented by utilizing a raw corpus in the voice recognition engine and the language recognition engine in the form of learning data.

A raw corpus of spoken data is a state-led corpus and it will be utilized in research on AI technology and Korean language education, contributing to the reinforcement of global competitiveness of the AI industry.

Keywords: Raw corpus, spoken corpus, semi-spoken corpus, spoken corpus collection, semi-spoken corpus collection, raw corpus utilization.

Project Director: Kyung-il Lee(Saltflux)

사업 책임자	이경일((주)솔트룩스)
사업 참여자	박지혜((주)솔트룩스)
	주재현((주)솔트룩스)
	박선희((주)솔트룩스)
	김소정((주)솔트룩스)
	조성현((주)솔트룩스)
	배소영((주)솔트룩스)
	강수빈((주)솔트룩스)
	오지희((주)솔트룩스) 외 10명
담당 연구원	이승재(국립국어원 언어정보과장)
	홍혜진(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장
 발행처: 국립국어원
 서울시 강서구 금남화로 154
 전화 02-2669-9775, 전송 02-2669-9727
 인쇄일: 2019년 12월 27일
 발행일: 2019년 12월 27일
 인 쇄: H&J 인쇄

※ 이 책은 국립국어원의 용역비로 수행한 ‘구어 자료 수집 및 원시 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.