

고려대학교 데이터 캠퍼스 4조

Gillajab-i

길라 잡이

외국인을 위한 한류 컨텐츠 기반 발음 교육 서비스

데이터 청년 캠퍼스

한국데이터산업진흥원

고려대학교

고려대학교 데이터 청년 캠퍼스

4조 이종현 박근형 이정훈 손소영 정세연

CONTENTS

목차

Design Background

고안 배경

발음 교육의 중요성

Service

서비스 소개

차별점

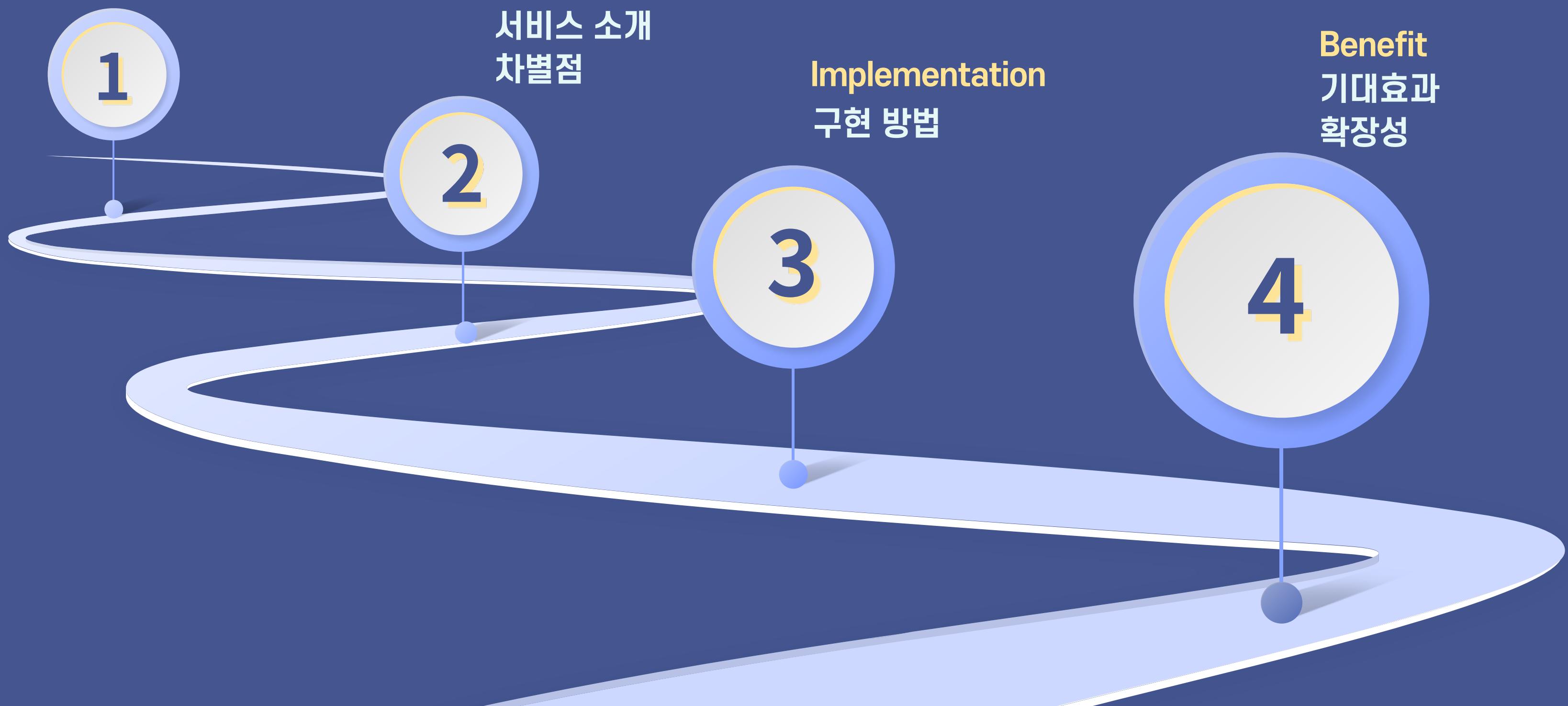
Implementation

구현 방법

Benefit

기대효과

확장성



01

Design Background

고안 배경
발음 교육의 중요성

고안 배경

실시간 소통의 어려움

01 실시간 번역 미제공



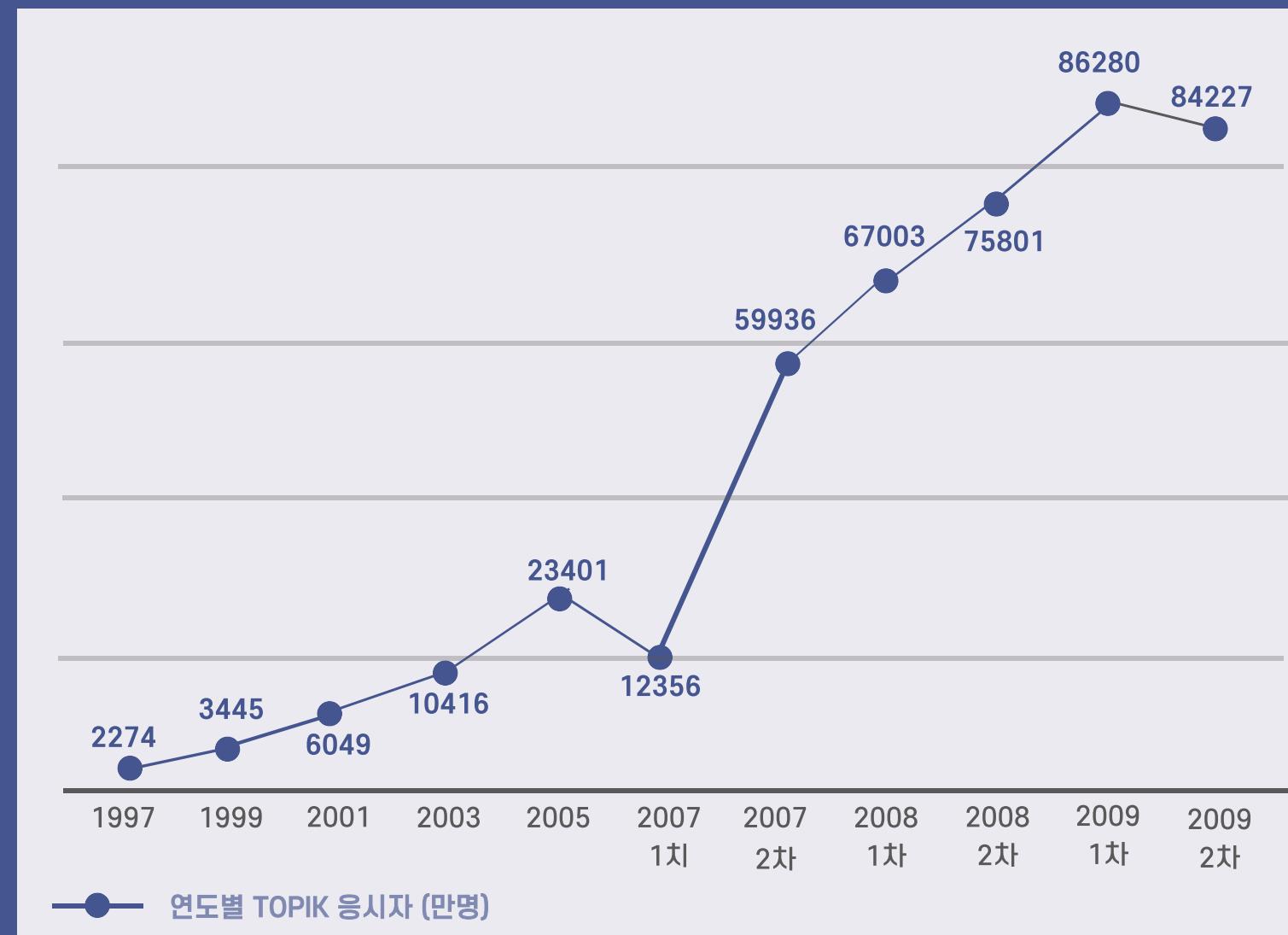
02 COVID - 19



고안 배경

한류 콘텐츠에 대한 세계적 인기 상승 => 한국어 교육 수요 상승

01 TOPIK 응시자 현황



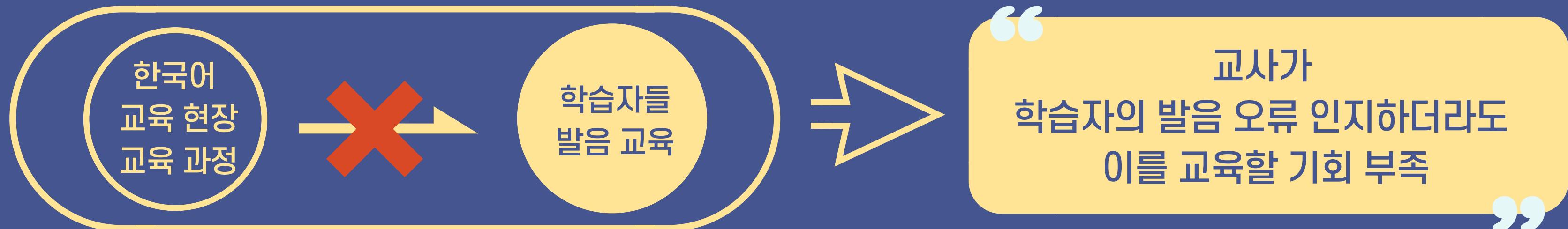
02 모어 사용자 수에 따른 전 세계 언어 순위

Rank	Language	Speakers (millions)	% of the World population (March 2019)	Language family branch
...
12	Wu Chinese	81.4	1.057	Sino-Tibetan Sinitic
13	Turkish	79.4	1.031	Turkic Oghuz
14	Korean	77.3	1.004	Koreanic Language isolate
15	French	77.2	1.003	Indo-European Romance

(출처 : Ethnologue 22nd edition)

한국어는 발음이 중요하지 않다?

발음 교육의 중요성



잘못된 발음 학습자의 문제

국립국어원의 「새국어생활 제25권 제1호」에 실린
한국어 학습자를 위한 발음 교육 방안

01

말하기 영역과 더불어 쓰기 영역에도 생기는 문제

02

의사소통 기능 전반에 걸친 문제

02

Service

서비스 소개
차별점

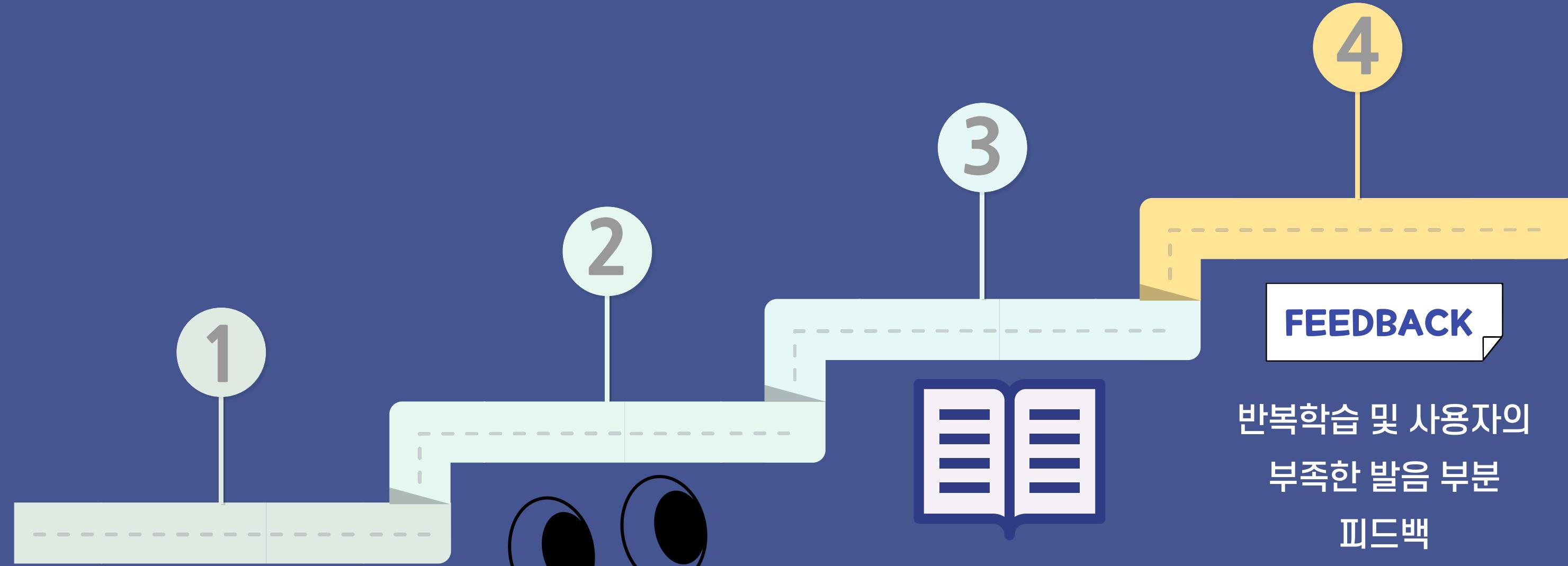
서비스 소개

로고
자리

길라잡이

외국인의 한국어 학습에 있어서 올바른 인도자가 되어주겠다

서비스 소개



사용자 5-10초 정도의
영상 클립 신청

한국어 문장, 번역 문장,
그리고 로마자 발음 변환
문장 보기

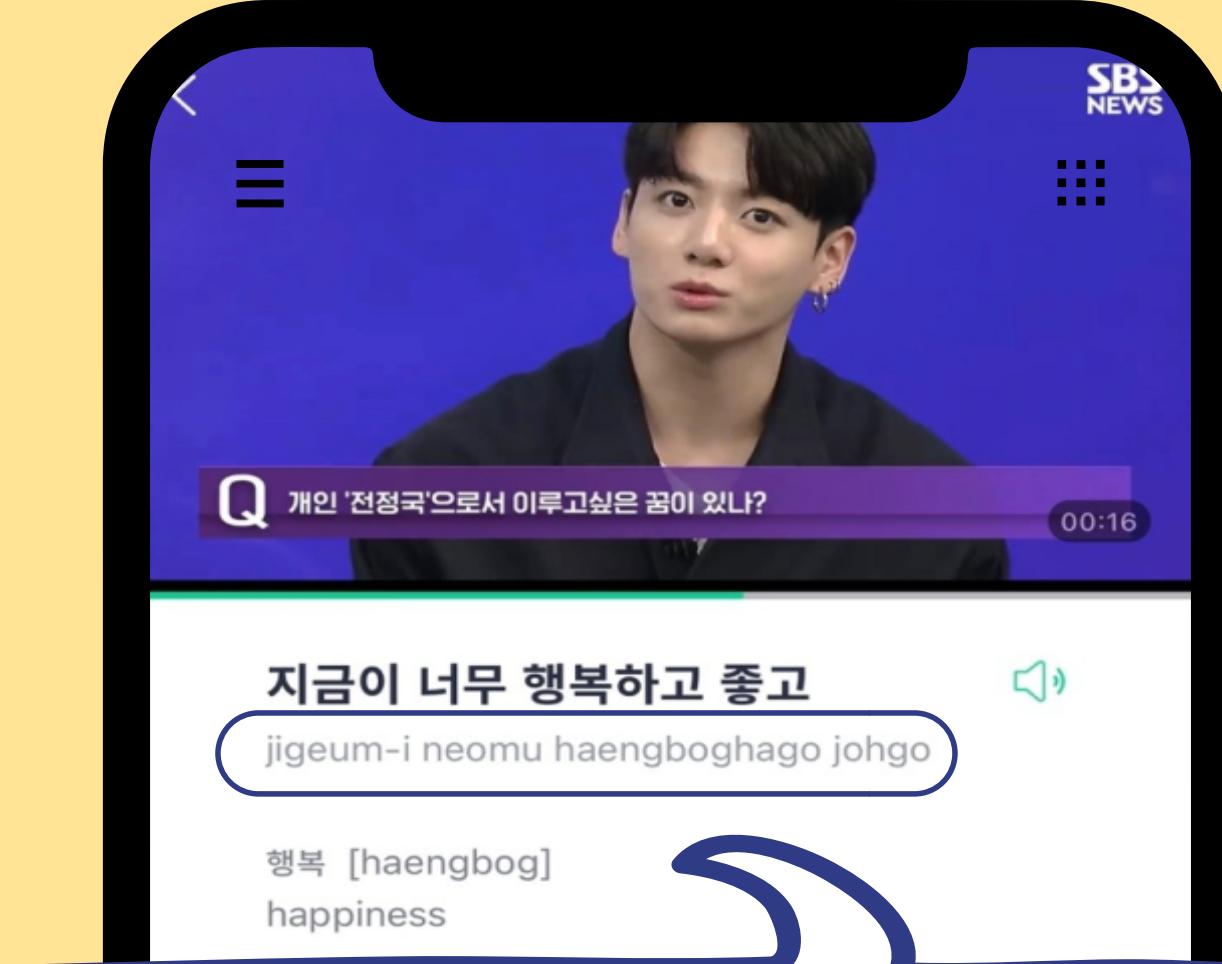
사용자 직접 문장 읽기

반복학습 및 사용자의
부족한 발음 부분
피드백

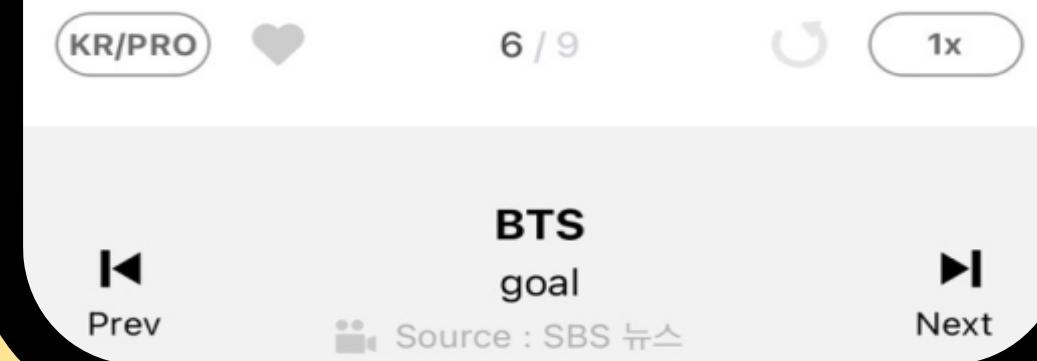
차별점

기존 헤이스타즈의 문제점

- 01 로마자 변환이 제대로 되지 않음



 jigeum-i neomu haengboghago johgo
0 **Jigeumi neomu haengbokago joko**

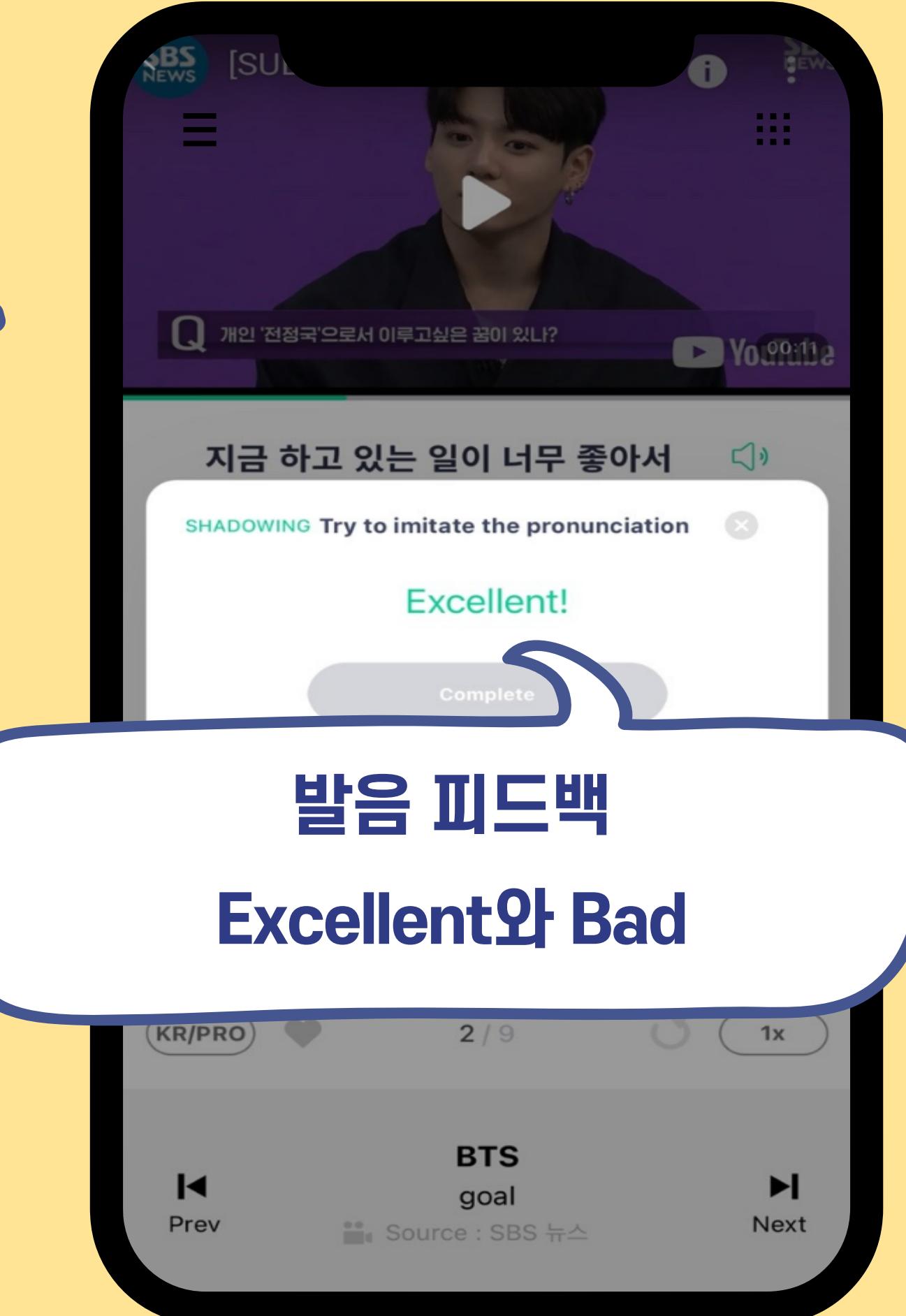


차별점

기존 헤이스타즈의 문제점

02

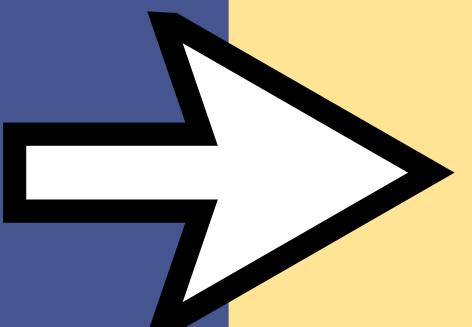
발음 피드백 부실



차별점

기존 헤이스타즈의 문제점

- 01 로마자 변환이 제대로 되지 않음
- 02 발음 피드백 부실



길라잡이

올바른 발음을
가르쳐준다

03

Implementation

구현 방법

구현 방법에 대한 전체 흐름도



데이터 수집

AI hub의 한국인 대화 음성 데이터

jupyter 데이터 수집



In [1] : AI hub의 '한국인 대화 음성 (2020)' 1000시간

The screenshot shows the AI Hub website interface. On the left, there's a sidebar with categories: 개방 데이터, 비전, 음성 / 자연어, 교육, 국토환경, and 농축수산. The '음성 / 자연어' category is currently selected. The main content area has a breadcrumb navigation: 홈 < 개방 데이터 < 음성/자연어. The page title is '음성/자연어'. Below the title, there are several data card boxes. One box is highlighted with a red border and features a large cursor icon pointing at it. This highlighted box contains the following information:

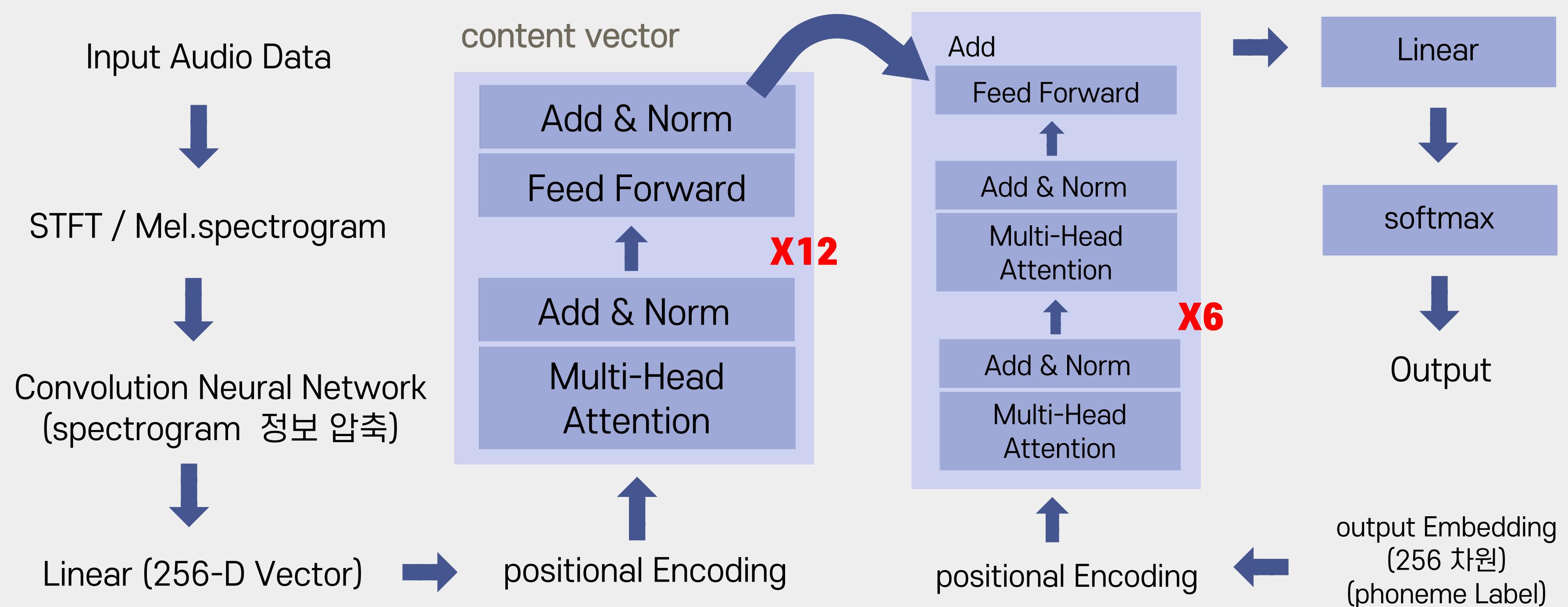
음성/자연어	한국어-일본어 번역 말뭉치	텍스트	2020
음성/자연어	한국어-중국어 번역 말뭉치(기술과학)	텍스트	2020
음성/자연어	한국어-중국어 번역 말뭉치(사회과학)	텍스트	2020
음성/자연어	한국인 대화 음성	텍스트 오디오	2020
음성/자연어	한국인 외래어 발화	텍스트 오디오	2020
음성/자연어	회의 음성	텍스트 오디오	2020

The '한국인 대화 음성' dataset is the one currently selected, as indicated by the highlighted box and the cursor icon.

모델 구조

발화자가 음성 녹음을 하면 모델에 적용시켜 발음의 정확도 점수 및 틀린 부분을 출력하도록 함

jupyter 모델 구조



데이터 전처리

ESPnet에서는 Kaldi 형식 요구
이에 맞게 파일 포맷 변환 필요

jupyter 데이터 전처리



In [1] : text : 파일 이름과 스크립트 (음소 단위)

: 음소 단위의 결과 출력이 목적이기 때문에
음소 단위로 변경

In [2] : spk2gender : 화자 ID와 성별

In [3] : spk2utt : 화자 ID와 파일 이름

In [4] : utt2spk : 파일 이름과 화자 ID

In [5] : wav.scp : 파일 이름과 절대 경로

데이터 전처리 순서

1. 발음대로 변환

script 정제 후
g2pk 사용하여
발음대로 변환

2. 음소단위로 변환

jamotools 사용하여
음소단위로 변환

6. 저장

AI hub 오디오 데이터
압축 해제하여
Kaldi-style 경로로
변환 후 저장

3. Kaldi format

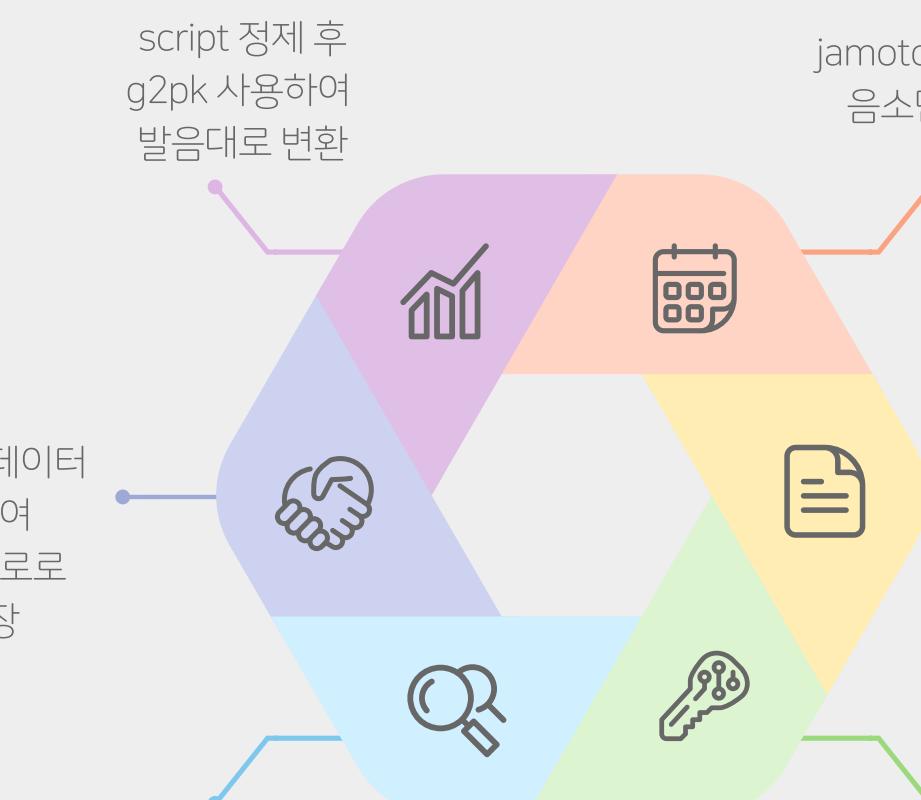
각 카테고리 별로
ESPnet에서 요구하는
파일 형태로 변환

5. 추가 작업

validation 데이터
test/valid로 분할 후
wav.scp 경로 변경 등

4. 파일 병합

변환한 파일
하나로 합치기



모델 학습

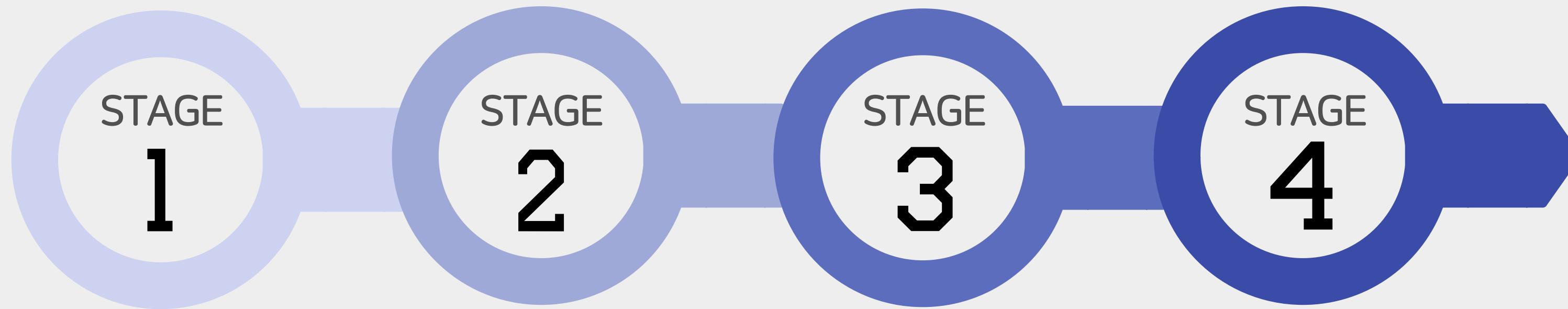
*ESPnet의 zeroth_korean의
asr.sh 수정하여 학습 진행*

jupyter 모델 학습



In [1] : # 변경한 변수

In [2] : stage (1 -> 2), stop_stage (10000 -> 12), audio_format(flac -> wav), token_type(bpe -> char)



zeroth_korean
자체 데이터를 다운로드하여
가공(Kaldi 형식에 맞게)하는
과정이므로 생략

speed_perturb_factors
변수에 대한 입력 값을 지
정하지 않았으므로 생략

format wav.scp
wav.scp 파일을 Kaldi 형식에
맞는지 보고 수정,
없으면 feat.scp 생성

remove long/short data
발화가 너무 길거나 짧은 데
이터 삭제

모델 학습

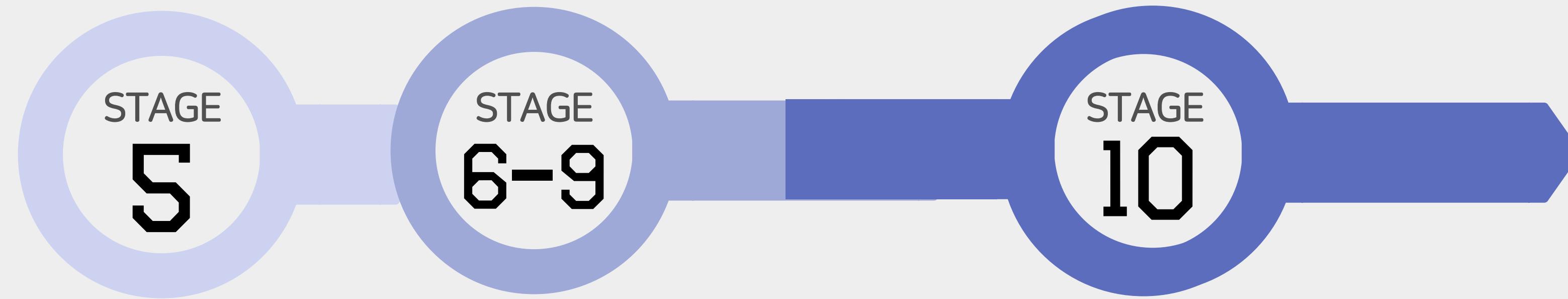
*ESPnet의 zeroth_korean의
asr.sh 수정하여 학습 진행*

jupyter 모델 학습



In [1] : # 변경한 변수

In [2] : stage (1 -> 2), stop_stage (10000 -> 12), audio_format(flac -> wav), token_type(bpe -> char)



generate token_list
character 레벨의 token list가 생성됨

language model(lm) 및
ngram을 사용하지 않으므로
생략

ASR collect stats
train/valid 데이터를 32개로 나누고
batch_keys(speech, text), stats_keys(feats, feats_length)에 맞게
speech_shape와 text_shape, feats_stats.npz와
feats_lengths_stats.npz 생성

모델 학습

*ESPnet의 zeroth_korean의
asr.sh 수정하여 학습 진행*

jupyter 모델 학습



```
In [1] : # 변경한 변수
```

```
In [2] : stage (1 -> 2), stop_stage (10000 -> 12), audio_format(flac -> wav), token_type(bpe -> char)
```



인코더

jupyter 인코더



디코더

jupyter 디코더



04

Benefit

기대 효과
확장성

기대효과

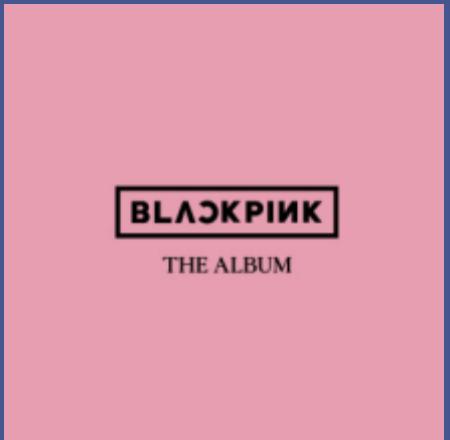
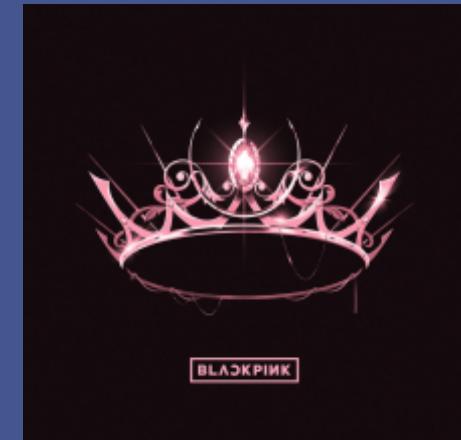
01

코어 팬 증가로 인한 한류 시장 수익 증가

* 코어팬

비용을 지출하면서 가수를 좋아하는 팬

여기서 말하는 비용은 앨범, 콘서트 티켓,
각종 굿즈, 공식 팬 클럽 등에 드는 비용을 말한다.
코어 팬은 가능한 모든 콘텐츠에
비용을 지불하는 경향이 있다.



기대효과



02

**한류 콘텐츠로
학습 흥미를 부여함에 따라
학습 지속성 증가**

요인별	수업 흥미도	수업 기억도	수업 집중도	수업내용 이해도	자료제시 방법 선호도
수업 흥미도(통제집단)	1				
"(실험집단 1)"	1				
"(실험집단 2)"	1				
수업 기억도(통제집단)	.268	1			
"(실험집단 1)"	.717**	1			
"(실험집단 2)"	.554*	1			
수업 기억도 평균	0.51				
수업 집중도(통제집단)	.346	.690**	1		
"(실험집단 1)"	.498**	.590**	1		
"(실험집단 2)"	.260	.385**	1		
수업 집중도 평균	0.37	0.56			
수업내용 이해도(통제집단)	.538**	.729**	.667**	1	
"(실험집단 1)"	.586**	.694**	.584**	1	
"(실험집단 2)"	.564**	.405*	.293	1	
수업내용 이해도 평균	0.56	0.61	0.51		
자료 제시방법 선호도(통제집단)	.221	.654**	.695**	.691**	1
"(실험집단 1)"	.585**	.566**	.627**	.793**	1
"(실험집단 2)"	.387*	.452*	.442*	.466**	1
자료 제시방법 선호도 평균	0.40	0.56	0.59	0.65	

** p < 0.01, * p < 0.05

<표4> 강의식 수업(통제집단), 멀티미디어 활용 교사 수업(실험집단1)과 내레이션 동영상 교실수업(실험집단2)의 상관관계

* 출처

기술가정 교과 '수송기술' 단원에서

수업자료의 제시 방법에 따른

학업 성취도에 미치는 영향

기대효과



03

한국 및 한국어의 위상을
높이는 데 큰 역할

* 사진 설명

태극기 한옥

한글 남산

남대문

확장성



고려대학교 데이터 청년 캠퍼스 4조

이상으로 발표를 마치겠습니다

들어주셔서 감사합니다 ——

Q & A Feedback

궁금한 점이 있다면
부담없이 물어봐주세요!

4조 이종현 박근형 이정훈 손소영 정세연