국립국어원 2025-01-08

발간등록번호
11-1371944-100002-01

2025년 묵자-점자 병렬 말뭉치 구축

연구 책임자 | 이연주

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 용역사업 계약에 따라 ‘2025년 묵자-점자 병렬 말뭉치 구축’ 사업 최종보고서를 작성하여 제출합니다.

■ 사업기간: 2025년 3월 28일 ~ 2025년 10월 29일

2025년 10월

사업수행기관	(사)한국시각장애인연합회
사 업 총 괄	이연주
사업수행인력	김은주, 이민정, 오혜은

목 차 …

제1장 서론

제2장 묵자-점자 병렬 말뭉치 데이터 구축 체계

제3장 묵자-점자 병렬 말뭉치 검수 및 통계 현황

제4장 묵자-점자 병렬 말뭉치 적합/오류/보류 사례

제5장 결론

표 목 차 …

그 림 목 차 …

제 1 장

서 론

1. 사업 배경
2. 사업 목적 및 내용 범위
3. 사업 방법

1 사업 배경

시각장애인의 정보 접근성과 학습권 보장을 위한 점자 사용 환경의 중요성이 지속적으로 강조되고 있다. 특히 디지털 전환 가속화에 따라 점자는 독립적인 독서 매체를 넘어 디지털 정보의 접근 권리를 실현하는 핵심 수단으로 자리매김하고 있다.

일부에서는 디지털 기술의 발전으로 인해 점자 활용도가 감소할 것이라는 우려도 있었지만 실제로는 다양한 정보 통신 기술과 점자의 융합이 가속화되며 점자의 활용 범위가 오히려 확대되고 있다. 국내외에서 점역·역점역 소프트웨어, 점자 정보 단말기, 전자 점자 파일, 웹 접근성을 보장하는 점자 변환 시스템 등 다양한 형태의 점자 연계 디지털 서비스가 개발되고 있으며, 스마트 기기 및 온라인 기반의 점자 애플리케이션 또한 꾸준히 상용화되고 있다.

점역·역점역 소프트웨어의 도입은 점자 자료 제작의 생산성과 접근성을 획기적으로 개선하는 데 기여하고 있다. 기존 점자 입력 방식에 비해 문서 입력 속도가 빠른 일반 묵자 입력을 기반으로 점역·역점역 소프트웨어를 활용하면 비전문가도 일정 교육을 거쳐 점자 자료 제작에 쉽게 참여할 수 있으며, 편집 및 수정도 쉽게 할 수 있다. 이는 점자 자료 생산의 효율성을 높이고 다양한 사용자의 요구에 따른 맞춤형 자료 제공을 가능하게 한다.

더 나아가, 21세기 지식 정보화 시대의 점자는 단순한 정보 접근 수단을 넘어 시각장애인의 학습권·정보권·문화권·노동권을 포괄하는 권리 실현의 기반이다. 「점자법」 제정 이후 제도적 기반이 지속적으로 확충되고 있음에도 불구하고 시각장애인의 학습권과 정보 접근권의 불균형은 여전히 구조적으로 존재한다. 특히 최신 정보나 다양한 콘텐츠 접근에 있어 점자 자료의 양적·질적 불균형이 주요한 문제로 지적되고 있다.

이에 따라 본 사업은 시각장애인의 생활 전반에서 점자 접근성을 제고하고 다방면으로 연계할 수 있는 기계 학습용 점자 데이터를 체계적으로 구축하는 것을 목표로 한다. 인공 지능 학습용으로 활용할 수 있는 묵자-점자 병렬 말뭉치 구축을 통해 점역·역점역 기술 고도화 기반을 마련하고, 이를 통해 시각장애인의 일상과 학습, 사회 참여 전반에서 점자 활용도를 실질적으로 향상시키는 토대를 마련하고자 한다.

특히 본 사업은 시각장애인의 정보 주권을 실현하고 장애-비장애 통합 사회를 향한 실질적 평등 구현에 기여하는 핵심 사회적 투자임을 지향한다. 궁극적으로는 인공 지능 기반 점역 기술의 발전과 함께 시각장애인의 점자 생활 환경 개선 및 다양한 디지털 점자 서비스 확산을 견인하고자 한다.

2 사업 목적 및 내용 범위

본 사업은 시각장애인의 일상생활에서 점자 사용 편의성과 학습 접근성을 실질적으로 높이기 위해 고품질 묵자-점자 병렬 말뭉치 데이터를 구축하고 체계적인 검수 체계를 마련하는 것을 목표로 한다. 이를 통해 향후 점자 친화적 사회 환경 조성과 인공 지능 학습용 점자 데이터 기반 마련 등 지속 가능한 발전 방향을 모색하는 데 목적이 있다.

구체적인 목표는 다음과 같다.

첫째, 한국어·영어 혼용 자료를 포함한 묵자-점자 병렬 말뭉치 구축이다. 국립국어원 신문 말뭉치를 활용하여 총 8만 문장(100만 어절 이상)의 묵자-점자 대응 문장을 신규 구축하고, 2021년 구축된 93,510문장(3,339,239어절)의 말뭉치 데이터를 2024년 개정 '한국 점자 규정'에 맞춰 수정한다. 중복 문장 정제 결과 발견된 결손 수량은 동일 출처에서 신규 문장으로 대체 구축하여 품질을 보완한다.

둘째, 검수·통계·관리 체계 고도화이다. 말뭉치 구축 전 과정에 말뭉치 웹 기반 관리 시스템을 적용하고 실시간 검수와 통계 모니터링 기능을 활용하여 검수 품질을 관리한다. 데이터 정제 및 검수 과정에서 발생하는 오류 유형을 체계적으로 기록·분석하고 데이터 집계 특성에 따른 품질 점검과 통계 분석 방안을 병행한다.

셋째, 오류 및 보류 사례 유형화와 기초 자료 확보이다. 검수 과정에서 수집된 오류와 보류 사례를 점자 규정 위반 유형과 연계 분석하여 향후 점자 규정 정비 및 실무 적용 시 참고 자료로 활용하고 사업 종료 후에는 결과 데이터를 점자 규정 개선 및 관련 연구 자료로 활용하도록 체계화한다.

3 사업 방법

본 사업의 목표 달성을 위해 다음과 같은 추진 방법을 적용한다.

첫째, 말뭉치 원천 데이터 확보 및 데이터 구축 방식이다. 데이터 저작권 및 민감 정보 보호 문제를 사전에 차단하기 위해 국립국어원 〈모두의 말뭉치〉 내 신문 분야 데이터를 활용하였으며 개인정보 및 기업 정보 비식별화를 전제 조건으로 최소 8만 문장 이상을 확보한다. 시나리오 5종을 기준으로 세분화하여 추출하였고 영어 단어 포함 여부와 통일영어점자 규정 적용 가능성을 고려하여 학습 효율성이 높은 데이터를 선별한다.

둘째, 데이터 구축 및 품질 관리 절차 구축이다. 구축된 데이터는 말뭉치 웹 기반 관리 시스템을 통해 정량적 검수 및 보정이 이루어지며 오류 및 중복 검출 기능, 시나리오별 문장 분배, 실시간 통계 추출 기능 등을 통해 데이터 품질을 체계적으로 관리한다. 구축된 데이터는 향후 점역·역점역 기능 고도화 및 인공 지능 학습용 데이터로 확장할 수 있도록 가공 및 정제한다.

셋째, 구축 전 과정에 걸친 품질 보증 체계 확립이다. 데이터 수집, 정제, 검수, 보정, 품질 점검까지 전 단계별 표준 프로세스를 수립하고 주기적 품질 모니터링과 오류 유형 피드백 과정을 병행한다. 이를 통해 인공 지능 기반 점자 기술 개발뿐 아니라 실무 점역 품질 향상에도 활용할 수 있는 고품질 말뭉치를 구축한다.

제 2 장

묵자-점자 병렬 말뭉치 데이터 구축 체계

1 말뭉치 데이터 구축 단계

본 사업에서 구축하는 묵자-점자 병렬 말뭉치는 고품질 데이터 확보를 위해 총 4단계의 구축 절차를 적용하였다. 각 단계별 주요 목적과 세부 절차는 다음과 같으며 데이터 품질 향상을 위해 단계별 관리 기준과 검수 절차를 명확히 설정하였다.

표 2-1 묵자-점자 병렬 말뭉치 데이터 구축 단계

구축 단계	설 명
1단계 데이터 수집	- 원천 데이터 선정: 저작권이 확보되고 개인 정보·기업 정보 등이 비식별화된 안전한 원천 데이터를 우선 선정 - 데이터 수집 기준: 한국어와 영어, 기호·숫자 혼용 문장 포함 여부 확인 후 수집 - 수집 방식: 국립국어원 신문 말뭉치 등 공개 말뭉치 활용, 시나리오 5종 기준 문장 분류
2단계 데이터 정제	- 부적합 데이터 삭제: 수집 과정에서 형식 오류가 생긴 문장 삭제 - 정제 절차: 기계 필터링 후 수동 검토를 병행하여 최종 적합 데이터 추출 - 중복 문장 제거: 웹 기반 말뭉치 관리 시스템 내 중복 자동 탐지 기능 활용
3단계 점자 가공	- 초벌 점역(MTPE): 말뭉치 웹 기반 관리 시스템 내 점역 엔진을 활용해 초벌 점역 진행 - 점자 규칙 적용: 점자 규정(개정 한국 점자 규정, 통일영어점자 규정) 자동 점검
4단계 품질 검사	- 1차 검수: 점역·교정사 68명과 말뭉치 사업 수행 인력 3명이 참여해 문장별 적합/오류/보류 판정 - 2차 검수: 동일 문장에 대해 교차 검수 실시 - 오류·보류 교정: 1차·2차 검수에서 단 1회라도 오류/보류 판정 시 사업 수행 인력이 직접 교정 - 통계·품질 점검: 위반 유형별 통계 추출, 주별 품질 모니터링

※ MTPE(Machine Translation Post-Editing): 기계 번역 후편집, 기계가 자동으로 번역한 점역문을 사람이 검수하는 과정

2 데이터 수집

본 사업의 데이터 수집 과정은 원천 데이터 선정과 원천 데이터 수집의 두 가지 주요 단계로 구분된다. 이를 통해 고품질의 묵자-점자 병렬 말뭉치 구축에 최적화된 자료를 확보하였다.

표 2-2 데이터 수집 단계

세부 절차	작 업
1. 원천 데이터 선정	- 저작권 확인 - 문장 적절성: 분야, 길이, 분량 - 기술 문제 검토: 수집 작업 적절성
2. 원천 데이터 수집	- 수집 방법 및 기준 설정: 문장 분절, 길이, 시나리오 조건 선정, 기호 처리 등 - 메타데이터 결정 - 1차 변환: 제이슨(JSON)에서 묵자 문장 추출 과정 - 2차 변환: 정의된 시나리오 5종에 해당하는 묵자 문장 추출 과정

※ JSON(JavaScript Object Notation): 데이터를 저장하거나 전송할 때 많이 사용되는 경량의 데이터 교환 형식으로 자바스크립트(JavaScript)에서 객체를 만들 때 사용하는 표현식을 의미

1) 원천 데이터 선정

원천 데이터를 선정할 때 가장 먼저 저작권 문제와 개인 정보, 상품 정보, 기업 정보 등의 포함 여부를 고려하였다. 국립국어원 〈모두의 말뭉치〉에 공개된 제이슨(JSON, JavaScript Object Notation) 형식의 말뭉치를 원천 데이터로 활용하기로 하였으며, 데이터 내 개인 정보, 상품 정보, 기업 정보 등 비식별화 대상을 확인하였다.

이에 따라 국립국어원 〈모두의 말뭉치〉 누리집에 등재된 데이터 중 활용 가능한 자료를 검토하였으며, '신문 말뭉치 2024'와 '국회 회의록 요약 말뭉치 2023'을 분석 대상으로 삼았다. 두 자료에 대하여 문장 길이, 시나리오 적합성, 영어 혼용 여부, 점자 변환 가능성 등을 종합적으로 평가한 결과, 최종적으로 '신문 말뭉치 2024'를 묵자-점자 병렬 말뭉치 구축의 주된 원천 데이터로 선정하였다.

특히 구축 과정에서 발생할 수 있는 데이터 부족 문제를 예방하기 위해 당초 목표량(8만 문장)의 105%인 84,000문장 규모를 확보하였다. 이는 향후 가공 및 검수 과정에서 불합격 판정이 발생할 경우에도 안정적으로 목표치를 달성할 수 있도록 대비한 것이다.

또한 원천 데이터의 기술적 처리 가능성도 함께 검토하였다. 제이슨(JSON) 형식으로 제공되는 원천 데이터 파일을 말뭉치 웹 기반 관리 시스템에 올려 문장 자동 추출, 점역 결과 연동, 검수자별

작업 분배, 검수 및 교정 이력을 관리할 수 있도록 하였다.

‘2021년 묵자-점자 병렬 말뭉치’의 경우 비학습 분야 구축 결과물을 그대로 선정하였고, 마찬가지로 수정·보완하여 동일한 시스템상에서 병렬 관리할 수 있도록 연동 구축하였다. 말뭉치 웹 기반 관리 시스템의 주요 기능은 아래와 같다.

(1) 검수 진행

말뭉치 구축의 체계적인 진행을 위해 말뭉치 웹 기반 관리 시스템을 활용하고 있다. 해당 시스템은 사용자 권한에 따라 관리자와 검수자로 구분된다.

관리자는 국립국어원 말뭉치 원천 데이터 파일(제이슨(JSON) 형식)을 내려받아 관리 시스템에 등록하며, 등록 과정에서 서버 기반 점역 엔진을 통해 문장과 점역 결과가 함께 자동 저장된다. 등록된 말뭉치는 메타 정보가 함께 기록되며 전체 문장 수 통계도 자동 관리된다.

관리자는 검수자 계정의 생성·삭제·수정 권한을 가지며, 각 검수자에게 균등한 분량의 문장이 할당되도록 작업량을 지정할 수 있다. 할당 방식은 사전 배분이 아닌 실시간 자동 배정으로 검수자 로그인 시 미검수 문장 중 남은 문장이 자동 할당되는 구조이다.

그림 2-1 말뭉치 웹 기반 관리 시스템 첫 화면

검수자는 해당 시스템에 접속해 문제 풀이 형식으로 검수 작업을 수행한다. 묵자(일반 문자)-점자 변환 결과를 기준으로 적합/오류/보류로 판정하며 오류 및 보류 시에는 필수로 사유를 입력하도록 설정되어 있다.

시스템에서는 검수 진행률과 남은 문장 수를 실시간으로 확인할 수 있어 검수자가 본인의 작업량과 진도를 스스로 조율할 수 있다. 또한 시각장애인 검수자의 접근성을 위해 키보드 접근과 화면 낭독기(Screen Reader) 사용이 가능하다.

수정 구축 대상인 '2021년 묵자-점자 병렬 말뭉치' 및 오류/보류 판정 문장의 교정 과정 역시 동일 시스템을 통해 동일한 절차로 운영하였다.

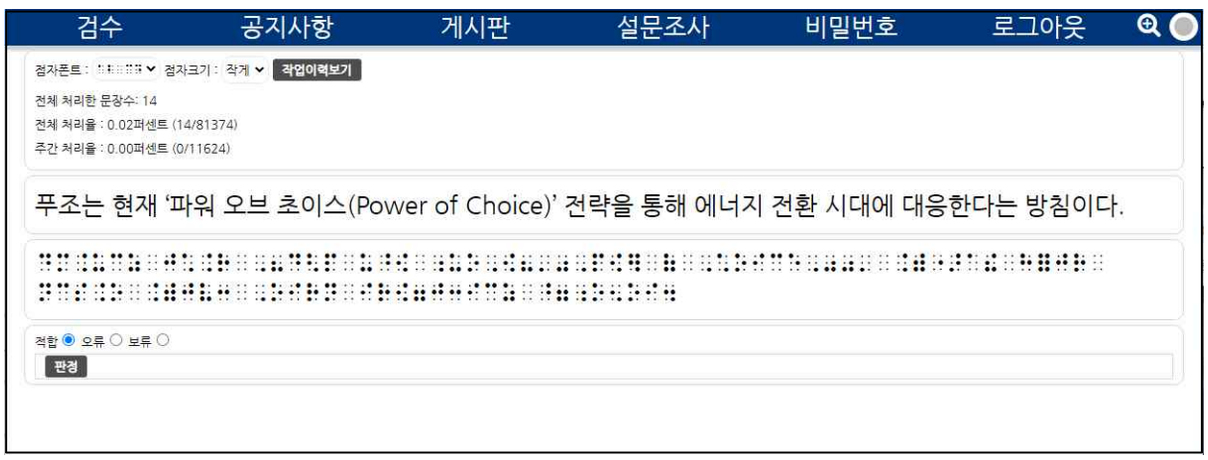


그림 2-2 말뭉치 검수자 검수 화면

그림 2-3 말뭉치 오류/보류 판정 문장 교정 화면

검수 결과는 주 단위 집계로 관리되며 교차 검수 기능을 통해 검수자가 상호 점검할 수 있도록 하였다. 검수자는 자신의 과거 검수 이력을 조회하거나 수정하면서 개별적으로 점역 결과를 학습하면서 오류를 줄일 수 있다.

그림 2-4 말뭉치 검수자 작업 내역 보기 화면

(2) 검수 관리

검수자는 주 단위 검수 진행 상황을 시스템에서 실시간으로 확인할 수 있으며 검수자의 검수 결과는 시스템에 자동으로 누적된다. 검수 정확도와 진행률에 따라 관리자 화면에서 검수자별 적합률, 오류율, 보류율, 작업 속도 등을 수치화하여 모니터링할 수 있다.

검수자의 검수 이력 및 통계는 주기적으로 집계되며 동일 검수 집단 내 교차 검수를 하여 특정 검수자 개인의 오류 발생 가능성을 상시 점검한다. 또한 검수자 교체가 필요한 상황에서는 검수자 계정을 실시간으로 생성하거나 비활성화하는 기능을 통해 검수자 관리의 유연성을 확보하였다.

검수자는 본인의 전체 검수 이력과 판정 결과를 확인할 수 있으며 필요시 개별 문장에 대해 판정 결과를 수정할 수도 있다. 이 기능은 검수 품질을 개선하고 자기 검토를 하도록 돕는다.

말뭉치 내역 현황 [조회]

검수작업	2025년 신문말뭉치 / 1차 검수자(A)	주차	전체

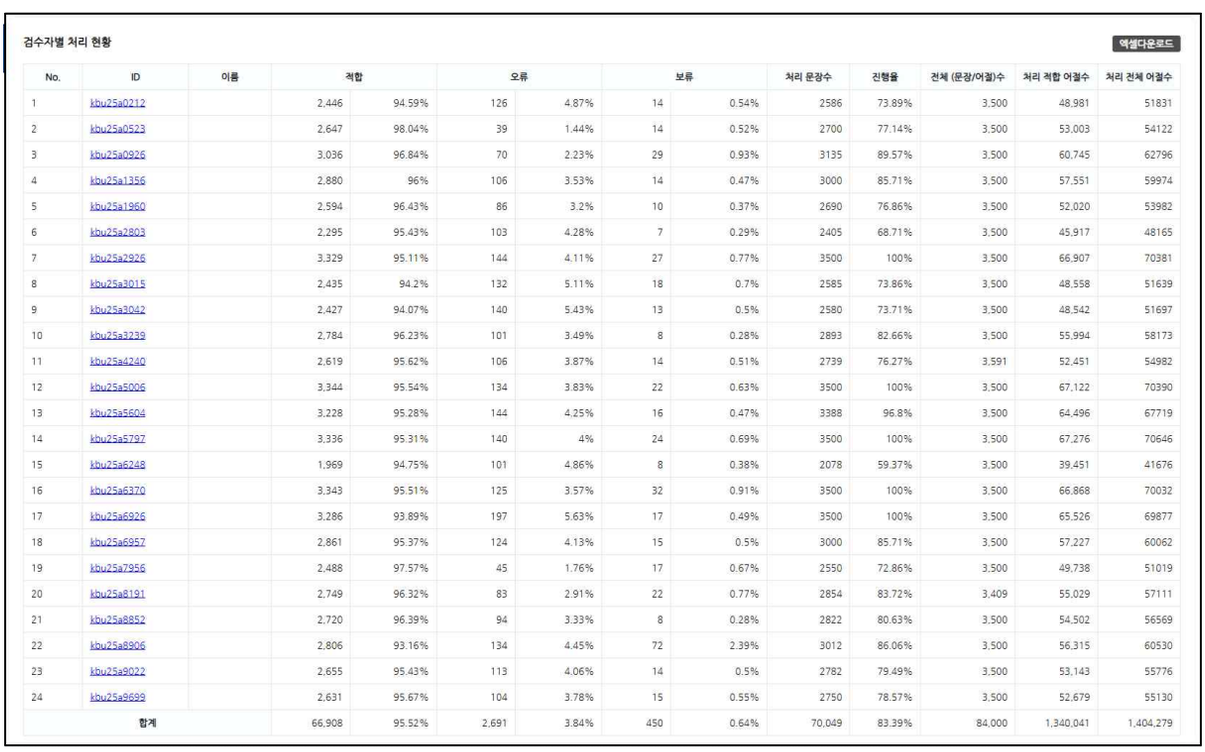
전체 처리 현황 [엑셀다운로드]

주차	기간	총 검수수 (84,000문장)	누적 검수율(%)	검수결과									총 어절수	총 문장수
				적합				오류	보류	계	오류율(%)	보류율(%)		
				적합	적합율(%)	어절수	문장수							
1 주차	20250609 ~ 20250615	13035	15.52%	12,336	94.64%	248,377	12,336	584	115	13035	4.48%	0.88%	262,682	13035
2 주차	20250616 ~ 20250622	13267	15.8%	12,722	95.89%	253,726	12,722	451	94	13267	3.4%	0.71%	264,885	13267
3 주차	20250623 ~ 20250629	13141	15.66%	12,497	95.1%	248,892	12,497	516	128	13141	3.93%	0.97%	262,125	13141
4 주차	20250630 ~ 20250706	13385	15.94%	12,840	95.93%	258,445	12,840	510	35	13385	3.81%	0.26%	269,465	13385
5 주차	20250707 ~ 20250713	12486	14.86%	11,971	95.88%	240,113	11,971	485	30	12486	3.88%	0.24%	250,682	12486
6 주차	20250714 ~ 20250720	4735	5.64%	4,542	95.92%	90,488	4,542	145	48	4735	3.06%	1.01%	94,440	4735
합계		70,049	83.39%	66,908	95.52%	1,340,041	66,908	2,691	450	70,049	3.84%	0.64%	1,404,279	70,049

시나리오 처리 현황 [엑셀다운로드]

No.	시나리오	적합		오류		보류		적합처리 어절 수	처리 문장수	처리 어절수	전체 문장수	전체 어절수
1	한글(영어)	50,280	95.01%	2,259	4.27%	381	0.72%	1,011,108	52,920	1,065,138	52,920	1,065,138
2	영어(한글)	5,962	93.62%	351	5.51%	55	0.86%	120,166	6,368	128,387	6,368	128,387
3	한글(숫자)	8,967	99.39%	47	0.52%	8	0.09%	175,061	9,022	176,214	9,022	176,214
4	숫자(한글)	303	91.54%	23	6.95%	5	1.51%	6,039	331	6,619	331	6,619
5	영어(숫자)	1,396	99.15%	11	0.78%	1	0.07%	27,667	1,408	27,921	1,408	27,921
6	기타	0	0%	0	0%	0	0%	0	0	0	0	0
전체		66,908	95.52%	2,691	3.84%	450	0.64%	1,340,041	70,049	1,404,279	70,049	1,404,279

그림 2-5 말뭉치 웹 기반 관리 시스템 검수 관리 화면



검수자별 처리 현황 [엑셀다운로드]

No.	ID	이름	적합		오류		보류		처리 문장수	진행율	전체 (문장/어절)수	처리 적합 어절수	처리 전체 어절수
1	kbu25a0212		2,446	94.59%	126	4.87%	14	0.54%	2586	73.89%	3,500	48,981	51831
2	kbu25a0523		2,647	98.04%	39	1.44%	14	0.52%	2700	77.14%	3,500	53,003	54122
3	kbu25a0926		3,036	96.84%	70	2.23%	29	0.93%	3135	89.57%	3,500	60,745	62796
4	kbu25a1356		2,880	96%	106	3.53%	14	0.47%	3000	85.71%	3,500	57,551	59974
5	kbu25a1960		2,594	96.43%	86	3.2%	10	0.37%	2690	76.86%	3,500	52,020	53982
6	kbu25a2803		2,295	95.43%	103	4.28%	7	0.29%	2405	68.71%	3,500	45,917	48165
7	kbu25a2926		3,329	95.11%	144	4.11%	27	0.77%	3500	100%	3,500	66,907	70381
8	kbu25a3015		2,435	94.2%	132	5.11%	18	0.7%	2585	73.86%	3,500	48,558	51639
9	kbu25a3042		2,427	94.07%	140	5.43%	13	0.5%	2580	73.71%	3,500	48,542	51697
10	kbu25a3239		2,784	96.23%	101	3.49%	8	0.28%	2893	82.66%	3,500	55,994	58173
11	kbu25a4240		2,619	95.62%	106	3.87%	14	0.51%	2739	76.27%	3,591	52,451	54982
12	kbu25a5006		3,344	95.54%	134	3.83%	22	0.63%	3500	100%	3,500	67,122	70390
13	kbu25a5604		3,228	95.28%	144	4.25%	16	0.47%	3388	96.8%	3,500	64,496	67719
14	kbu25a5797		3,336	95.31%	140	4%	24	0.69%	3500	100%	3,500	67,276	70646
15	kbu25a6248		1,969	94.75%	101	4.86%	8	0.38%	2078	59.37%	3,500	39,451	41676
16	kbu25a6370		3,343	95.51%	125	3.57%	32	0.91%	3500	100%	3,500	66,868	70032
17	kbu25a6926		3,286	93.89%	197	5.63%	17	0.49%	3500	100%	3,500	65,526	69877
18	kbu25a6957		2,861	95.37%	124	4.13%	15	0.5%	3000	85.71%	3,500	57,227	60062
19	kbu25a7956		2,488	97.57%	45	1.76%	17	0.67%	2550	72.86%	3,500	49,738	51019
20	kbu25a8191		2,749	96.32%	83	2.91%	22	0.77%	2854	83.72%	3,409	55,029	57111
21	kbu25a8852		2,720	96.39%	94	3.33%	8	0.28%	2822	80.63%	3,500	54,502	56569
22	kbu25a8906		2,806	93.16%	134	4.45%	72	2.39%	3012	86.06%	3,500	56,315	60530
23	kbu25a9022		2,655	95.43%	113	4.06%	14	0.5%	2782	79.49%	3,500	53,143	55776
24	kbu25a9689		2,631	95.67%	104	3.78%	15	0.55%	2750	78.57%	3,500	52,679	55130
합계			66,908	95.52%	2,691	3.84%	450	0.64%	70,049	83.39%	84,000	1,340,041	1,404,279

그림 2-6 말뭉치 웹 기반 관리 시스템 검수자별 처리 현황 관리 화면

(3) 문장 관리

누적 검수 완료 문장의 전체 목록도 시스템에서 조회할 수 있다. 관리자는 검수 완료 문장 데이터를 제이슨(JSON), 엑셀(Excel) 등 다양한 포맷으로 일괄 추출할 수 있으며 검수 적합률, 오류 사례, 보류 건 등 주요 지표에 기반한 집계 결과도 주기적으로 확인할 수 있다.

시나리오별, 어절 수별, 오류 유형별 필터링 기능이 제공되어 특정 조건에 따라 문장을 추출할 수 있고, 점자 규정 개정 등 정책 변화에 따라 특정 유형의 문장을 선별·추적하는 데에도 유용하게 활용할 수 있다.

그림 2-7 말뭉치 웹 기반 관리 시스템 문장 관리 화면

(4) 검수 문장 비교

검수자 간 동일 문장에 대한 판정이 상이한 경우 교차 검수 기능을 통해 판정 결과 비교가 가능하다. 이 기능은 동일 문장에 대한 판정 이력을 모두 조회할 수 있으며 각 검수자의 선택 사유 및 판정 결과를 확인할 수 있다.

시스템 내 비교 화면에서는 원문 묵자와 기존 점역 결과, 그리고 수정된 점역 결과를 1:1로 나란히 비교하는 방식으로 제공되며 특정 시점에 수정이 이루어진 변경 이력도 확인할 수 있다.

검수자의 판정 오차를 최소화하고 교정자 검수 과정의 정확성을 확보하기 위해 위 기능을 전 과정에 적용하였다.

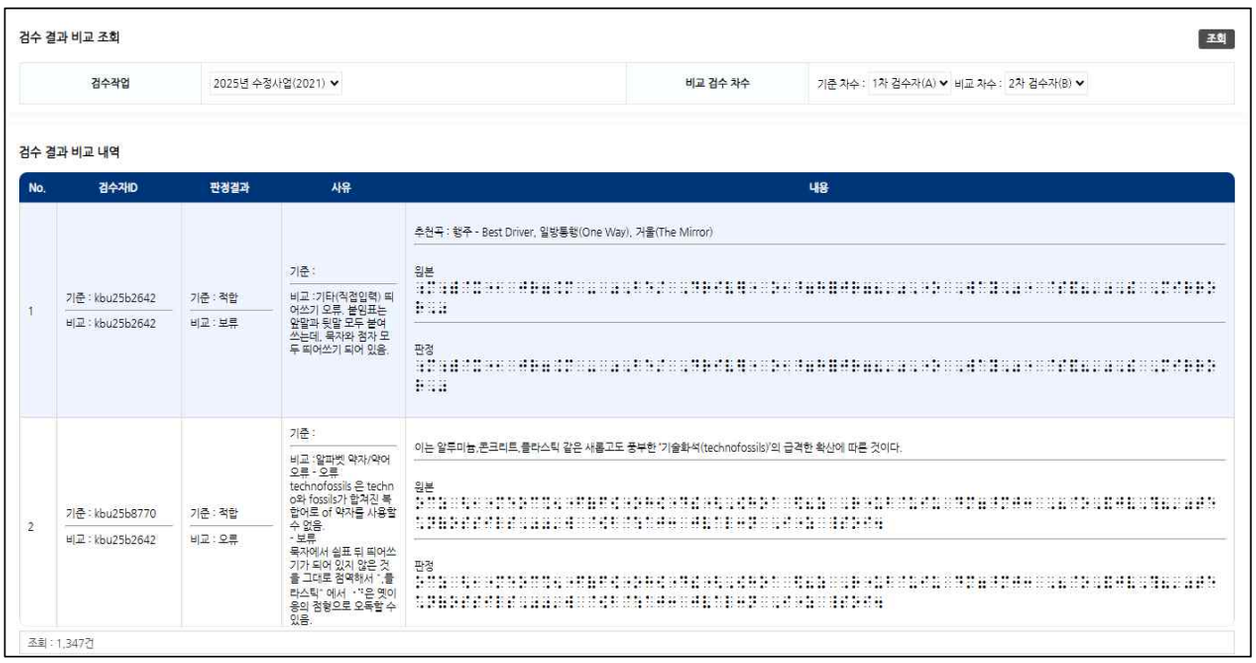


그림 2-8 말뭉치 웹 기반 관리 시스템 검수 문장 비교 화면

2) 원천 데이터 수집

원천 데이터 선정 이후 실제 데이터 수집 절차가 진행된다. 수집 단계에서는 원천 데이터의 형식과 활용 목적에 따라 수집 방식과 처리 기준을 세분화하여 정의한다.

수집 과정은 주로 자바(Java) 기반의 배치 프로그램(Batch Processing)을 활용하여 자동화하였다. 배치 프로그램은 순차적 처리 방식으로 제이슨(JSON) 파일에서 묵자 문장을 추출하는 데 사용된다. 이때 문장 길이, 불필요한 기호 처리, 분절 규칙 등 다양한 조건이 사전 정의되어 반영되었다.

데이터 구축 목적에 따라 분야별 메타데이터 기준이 수립되며 추출된 문장은 해당 기준에 따라 메타데이터와 함께 데이터베이스(DB)에 저장된다. 추출 단계에서는 가공 및 검수 이전의 원천 데이터 상태를 유지하되 품질 관리를 위한 최소 기준을 적용한다. 단 '2021년 묵자-점자 병렬 말뭉치'는 이미 구축 완료된 자료이므로 별도의 수집 과정을 생략하고 기존 데이터 그대로 활용하였다.

(1) 말뭉치 분석

데이터 수집 대상 말뭉치는 정확성, 활용도, 정보성, 사용 빈도 등을 기준으로 분석 후 선정하였다. 본 사업에서는 국립국어원 '신문 말뭉치 2024'를 주 원천 데이터로 활용하였다.

'신문 말뭉치 2024'는 총 1,119,753건의 신문 기사로 구성되어 있으며, 기사 텍스트가 제이슨(JSON) 형식으로 저장되어 있다.

또한 2021년 구축된 '2021년 묵자-점자 병렬 말뭉치'는 총 93,777개의 묵자-점자 대응 문장으로 구성되어 있으며, 이번 사업에서는 중복 문장을 제외한 93,510문장을 2024년 개정한 한국 점자 규정을 반영하여 수정 구축하였다.

'신문 말뭉치 2024'의 기본 제이슨(JSON) 구조는 아래와 같다.

표 2-3 신문 말뭉치 제이슨(JSON) 파일 예시


```
{
  "id": "NPRW2400000002.4889",
  "metadata": {
    "title": "매일경제 2023년 기사",
    "author": "이선희 기자",
    "publisher": "매일경제",
    "date": "20230120",
    "topic": "경제",
    "original_topic": "부동산"
  },
  "paragraph": [
    {
      "id": "NPRW2400000002.4889.1",
      "form": "“제발 우리 아파트 좀 사주세요”…금융위기 때보다 집 팔기 힘들다 [매부리레터]"
    },
    {
      "id": "NPRW2400000002.4889.2",
      "form": "은퇴를 앞둔 회사원 이모씨(57)는 서울 잠실 아파트를 처분하고 시골로 내려갈 계획이었다. 그러나 지난해 1년 내내 집을 내놨지만 가끔 문의만 올뿐 매도가 성사되지 않았다. “가격도 많이 내렸는데 이렇게 집이 안나가네요. 집이 팔려야 고향으로 내려갈텐데….” 이씨는 “집이 팔려야 서울 생활을 정리하고 내려갈텐데 집이 안팔려서 너무 답답하다. 올해는 정부가 규제를 푼다고 하니 제발 집이 팔렸으면 좋겠다”고 했다."
    },
    {
      "id": "NPRW2400000002.4889.3",
      "form": "전국 부동산 시장에 '거래 한파'가 계속되고 있다. 지난해 서울 아파트 거래량은 역대 최저를 기록한 것으로 나타났다. 2008년 금융위기 때보다 더 적은 거래량이다. 공인중개사들은 “이렇게 거래량이 없기는 처음”이라고 입을 모은다. 각종 규제가 중첩된 데다 고금리까지 덮치면서 거래가 완전 '실종'된 것이다. 매수자가 실종되면서 집을 팔고 이사를 가려던 사람들은 발이 묶였다. 집이 안팔리니 급한대로 전세로 돌리는데 전세는 더 가파른 하락세다. 역대급 거래량 실종에 지자체는도 비상이다. 취득세가 마르다보니 세수 확보에 '비상'이 걸렸다."
    },
```

이러한 구조를 고려하여 데이터 수집 단계에서 필요한 필터링과 문장 단위 전처리 작업이 용이하도록 하였다.

(2) 1차 변환 묵자 텍스트 파일

말뭉치 분석이 완료된 후 추출된 문장들은 대량의 기계 점역 작업이 가능하도록 묵자 문장 텍스트 파일 형태로 1차 변환된다. 해당 과정은 자바 기반 배치 프로그램을 활용하여 자동으로 처리하였다.

1차 변환 과정에서는 묵자 문장에 필요한 최소 정보만 추출하며 유형이나 출처 등 부가 정보는 원본인 국립국어원 말뭉치 데이터에서 필요시 역추적이 가능하도록 하였다. 주요 추출 요소는 문장 식별자(예: NIRW2400000001.189.4)와 문장(예: 지원 조건은 2년 거치 일시상환으로 업체당 3억원 한도이며, 프리(PRE)·명품 강소기업, 일자리우수기업, 광주형일자리기업, 우수중소기업인, 산업안전보건 우수기업 등 우수기업은 5억원 이내에서 지원한다.)이다.

특히 가운뎃점(·), 마침표(.), 쉼표(,) 등 시각적으로 유사하지만 유니코드상 서로 다른 문자인 경우 전처리 작업을 하였다. 이를 통해 기계 점역 과정에서 발생할 수 있는 오류 가능성을 원천 차단하였다.

표 2-4 1차 변환 단계 예시

NIRW2400000001.189.4:	지원 조건은 2년 거치 일시상환으로 업체당 3억원 한도이며, 프리(PRE)·명품 강소기업, 일자리우수기업, 광주형일자리기업, 우수중소기업인, 산업안전보건 우수기업 등 우수기업은 5억원 이내에서 지원한다.
NIRW2400000001.189.5:	이와 함께, 광주시는 중소기업의 어려움 해소를 위해 시비로 2%의 이자차액을 보전하며, 프리(PRE)·명품 강소기업, 일자리우수기업, 광주형일자리기업, 우수중소기업인, 산업안전보건 우수기업은 1%를 추가 지원한다.

변환된 묵자 문장 데이터는 단순 텍스트(txt) 파일로 저장되며, [문장 식별자]:[문장] 구조를 따른다. 해당 파일은 이후 2차 변환 및 점자 데이터 구축을 위한 중간 데이터로 활용된다.

(3) 묵자 추출 시나리오 조건

2차 변환 과정에서는 데이터 구축 목적에 따라 다양한 조건이 추가로 적용된다. 특히 시나리오별 조건(총 21종)을 통해 대상 문장을 정교하게 필터링하는데, 주요 시나리오 조건은 다음과 같다.

① 한글 + (영어 + … +)

② 영어 + (한글 + … +)

③ 한글 + (숫자 + … +) → 열고 닫는 괄호가 반드시 쌍으로 존재할 것

④ 숫자 + (한글 + … +)

⑤ 영어 + (숫자 + ... +)

⑥ 한글 + (숫자 + ... +) + 영어 → ③의 조건에 포함되어 추후 삭제

⑦ 한글 + 숫자

⑧ 영어 + 숫자 → ⑦과 ⑧은 경우의 수가 많으므로 추출 조건을 더 상세히 제시

⑨ 한글 + 영어 + 숫자

⑩ 한글 + 숫자 + 영어

⑪ 영어 + 한글 + 숫자

⑫ 영어 + 숫자 + 한글

⑬ 숫자 + 영어 + 한글

⑭ 숫자 + 한글 + 영어

⑮ 한글 + 문장부호 + 영어

⑯ 영어 + 문장부호 + 한글

* ⑮와 ⑯은 플러스(+) 부분을 붙여 쓴 경우임. 즉, 괄호나 빈칸이 들어가는 경우는 제외

⑰ 한글 + 영어 철자 나열 + 한글

⑱ 한글 문장 내에 로마자가 있는 것

⑲ 한글 + 가운뎃점 의미로 사용되는 아래아 + 한글

⑳ 화폐 단위가 포함된 문장

㉑ 문장부호 이외의 점자 규정에 포함된 부호

표 2-5 시나리오 조건

구 분	대상 조건(21종)	선정 조건(5종)
시나리오 조건	한글(영어), 영어(한글), 한글(숫자), 숫자(한글), 영어(숫자)+한글, 한글(숫자)+영어, 한글+숫자, 영어+숫자, 한글+영어+숫자, 한글+숫자+영어, 영어+한글+숫자, 영어+숫자+한글, 숫자+영어+한글, 숫자+한글+영어, 한글+문장부호+영어, 영어+문장부호+한글, 한글+ 영어 철자 나열 +한글, 한글 문장 내에 로마자가 있는 것, 한글+가운뎃점 의미로 사용되는 아래아+한글, 화폐 단위, 문장부호 이외의 점자 규정에 포함된 부호 등	한글 + (영어), 영어 + (한글), 한글 + (숫자), 숫자 + (한글), 영어 + (숫자)

(4) 2차 변환 묵자 파일

2차 변환 단계에서는 실제 검수 대상 문장 선별을 목적으로 사전 정의된 시나리오 5종 기준에 따라 묵자 문장 84,000개를 최종 추출하였다. 본 과정에서는 단순 문장 수 기준이 아닌 어절 수 필터링(15어절 이상 25어절 이하) 조건을 추가로 적용하여 지나치게 짧거나 길어 점역 검수에 부적합한 문장을 사전에 제외하였다. 추출된 문장 데이터는 '2021년 묵자-점자 병렬 말뭉치'(93,510문장)과 별도로 검수를 진행한다.

표 2-6 말뭉치 종류별 추출 문장 수

말뭉치 종류	말뭉치 파일명	유형	문장 수
신문 말뭉치 2024	NIRW2400000003~ NIRW2400000005	신문 〉 인터넷 기반 신문	20,550
	NLRW2400000001~ NLRW2400000008	신문 〉 지역 종합지	12.219
	NPRW2400000001~ NPRW2400000004	신문 〉 전문지	41,066
	NWRW2400000001~ NWRW2400000002	신문 〉 전국 종합지	10,165
소 계			84,000
2021년 묵자-점자 병렬 말뭉치 (신문 말뭉치)	NIRW1900000001~ NIRW1900000008	신문 〉 인터넷 기반 신문	62,722
	NLRW1900000001~ NLRW1900000083	신문 〉 지역 종합지	23,970
	NPRW1900000035~ NPRW1900000059	신문 〉 전문지	3,406
	NWRW1900000014~ NWRW1900000035	신문 〉 전국 종합지	3,412
소 계			93,510
합 계			177,510

표 2-7 시나리오별 추출 문장 수

말뭉치 종류	시나리오	비율(%)	추출 문장 수(개)
신문 말뭉치 2024	한글 + (영어) 형태 포함	76.46	64,222
	영어 + (한글) 형태 포함	8.47	7,115
	한글 + (숫자) 형태 포함	12.84	10,784
	숫자 + (한글) 형태 포함	0.47	398
	영어 + (숫자) 형태 포함	1.76	1,481
소 계		100	84,000
2021년 묵자-점자 병렬 말뭉치 (신문 말뭉치)	한글 + (영어) 형태 포함	26.98	25,230
	영어 + (한글) 형태 포함	28.71	26,850
	한글 + (숫자) 형태 포함	33.32	31,151
	숫자 + (한글) 형태 포함	7.20	6,732
	영어 + (숫자) 형태 포함	3.79	3,547
소 계		100	93,510
합 계			177,510

3) 유형 구성 및 파일 명명 규칙

본 사업에서 구축한 묵자-점자 병렬 말뭉치 데이터는 대분류 및 중분류 체계를 기준으로 유형화하여 체계적으로 관리된다. 이는 데이터의 다양성을 확보하고 향후 인공 지능 학습 및 점역 활용의 효율성을 높이기 위함이다.

표 2-8 데이터 유형

대분류	중분류	메타데이터
신문	가. 전국 종합지	title(도서명), author(저자), publisher(출판사), date(작성 연도), original topic(주제)
	나. 지역 종합지	
	다. 전문지	
	라. 인터넷 기반 신문	
	마. 기타	

구축된 묵자-점자 병렬 말뭉치 데이터는 검수 및 교정 과정을 모두 완료한 이후 최종 납품 시 표준화된 파일 명명 규칙을 적용하여 관리된다. 이는 파일 관리의 일관성과 추후 데이터 활용 시 식별 용이성을 확보하기 위한 조치이다.

최종 파일은 아래와 같은 규칙에 따라 명명된다.

표 2-9 파일 명명 규칙(총 17자리)

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
속성	유형	매체 및 장르	분석 층위		출발 자료		도착 자료		사업 연도		일련번호(7자리)						
정의값	N: 신문	W: 전국 종합지 L: 지역 종합지 P: 전문지 I: 인터넷 기반 신문 Z: 기타	PA: 병렬		KO: 한국어 (묵자)		KB: 점자		25		0000001~9999999 (일곱 자리 일련번호)						

4) 최종 적합 파일 유형

검수 및 교정 완료 후 최종 구축된 데이터는 제이슨(JSON) 형식뿐만 아니라 엑셀(Excel) 파일로도 별도로 가공하였다. 엑셀 파일은 후속 통계 분석, 사용자 교육 등 다양한 목적에 활용될 수 있도록 문장별 주요 메타 정보와 유형 분류를 포함하여 작성된다.

표 2-10 최종 데이터 적합 엑셀 파일 유형

dataset	메타데이터	
	language_info	parallel
국립국어원 묵자-점자 병렬 말뭉치	source_language(원자료 언어, kl), target_language(번역 대상 언어, kb)	id(묵자-점자 병렬 말뭉치 문장 고유 ID), original_id(원문 말뭉치 문장 고유 ID), source(묵자 문장), target(점자 문장), revision1(검수된 점자 문장)

3 데이터 정제

수집 완료된 묵자 데이터는 점역 및 검수 과정에 앞서 불필요한 오류를 사전에 제거하고 가공에 최적화된 형태로 교정·정제된다. 정제 과정은 크게 부적합 데이터 삭제와 데이터 정제의 두 단계로 구분하였다.

표 2-11 데이터 정제 수행 과정

세부 절차	작업
1. 부적합 데이터 삭제	- 분절 오류 문장 삭제 - 고유 명사 처리 기준 미달 문장 삭제(개인 정보, 미등록어 과다 등) - 심각한 비문 삭제
2. 데이터 정제	- 점형에 부적합한 기호 처리 - 문장 종료 처리(분리, 병합, 마침표 등) - 문장 오타 주석 오류 확인

1) 부적합 묵자 데이터 삭제

수집된 묵자 문장 중 점자 변환 및 검수에 적합하지 않은 문장을 선별해 삭제하였다. 주요 삭제 기준은 다음과 같다.

- 문장 분절 오류: 데이터 수집 과정에서 비정상적으로 끊어진 문장
- 개인 정보 포함: 개인 정보, 상호명, 상품명, 기업명 등 식별 가능 정보 포함 문장
- 미등록어 과다 포함 문장: 신조어, 비표준어가 과도하게 포함된 문장

정제 과정에서 이러한 문장을 선별적으로 삭제하여 검수 효율성을 확보하고 불필요한 오류 교정 부담을 최소화하였다.

2) 데이터 정제

부적합 데이터 삭제 후 남은 데이터는 추가 정제를 통해 가공하였다. 주요 정제 작업은 다음과 같다.

- 중복 문장 제거: 말뭉치 전체에서 중복된 문장 제거
- 점형 불가 기호 처리: 점역 시 오류가 예상되는 특수문자, 이모티콘 등 전처리
- 문장 종료 일괄 처리: 마침표 누락, 문장 병합·분리, 종결 부호 보정

정제 완료 후 최종 묵자 문장은 웹 기반 관리 시스템에 탑재된 후 기계 점역 및 검수 절차를 거치게 된다.

4 데이터 가공

데이터 가공 단계에서는 수집·정제된 묵자 문장을 기계 번역 후편집(MTPE, Machine Translation Post-Editing) 방식으로 점자 변환을 수행한다. 기계 번역 후편집(MTPE) 방식은 반드시 후행 검수 절차를 전제로 하며 기계 번역 오류를 보완하는 작업이 필수적으로 이루어진다.

표 2-12 가공 세부 절차

세부 절차	작업
기계 번역 후편집 (MTPE)	- 분야별 최적 기계 번역 엔진 선택 - 기계 번역 결과 후편집 - 오류 심각 시 수작업 초벌 점역 적용 - 최종 묵자 문장 기계 점역 처리

1) 기계 번역 후편집(MTPE)

본 사업에서는 점자 변환의 효율성과 품질 향상을 위해 기계 번역 후편집(MTPE) 방식을 기본 가공 절차로 채택하였다. 주요 작업 흐름은 다음과 같다.

- 분야별 최적 기계 번역 선택: 신문 기사 데이터 특성을 반영하여 점역 정확도가 높은 기계 번역 엔진 선택
- 기계 번역 후편집: 기계 점역 결과를 원문과 비교 검토 후 오류 수정
- 심각 오류 문장 대응: 정제 문장 여유량을 확보한 상태에서 품질 미달 문장은 제외 처리
- 최종 점역 처리: 최종 형태의 묵자 문장은 기계 점역 후 검수 전 단계 데이터로 변환

최종 가공 결과는 묵자-점자 문장 한 쌍으로 구성되며 대응 관계가 명확하도록 설계되었다. 문자 세트는 유니코드 기반의 한글 낱자를 점자 규칙에 맞춰 대응한 변환 테이블을 활용한다. 최종 데이터는 제이슨(JSON) 등 기계가 읽을 수 있는(Machine Readable) 형식으로 저장하여 후속 인공지능 학습 및 응용이 가능하도록 처리한다.

표 2-13 가공 작성 예시


```
{
  "source": {
    "text": "바이어 04 레버쿠젠의 페르난도 카로 CEO가 팀을 성공적으로 이끌고 있는 사비 알론소
감독(42)의 미래에 대해 언급했다.",
    "text_language": "korean",
    "category": "normal",
  },
  "braille_translation_result": {
    "unicode": "⠘⠣⠕⠎ ⠼⠚⠙ ⠐⠝⠘⠎⠋⠍⠨⠝⠒⠺ ⠙⠝⠐⠪⠉⠒⠊⠥ ⠋⠐⠥ ⠴⠠⠠⠉
⠑⠕⠲⠫ ⠓⠕⠢⠮ ⠠⠻⠈⠿⠨⠹⠪⠐⠥ ⠕⠠⠈⠮⠈⠥ ⠕⠌⠉⠵ ⠇⠘⠕ ⠣⠂⠐⠷⠠⠥ ⠫⠢
⠊⠭⠦⠄⠼⠙⠃⠠⠴⠺ ⠑⠕⠐⠗⠝ ⠊⠗⠚⠗ ⠾⠈⠪⠃⠚⠗⠌⠊⠲",
    "ttb_option": "none",
    "characters_per_line": "16"
  }
}
```


2) 3차 변환 묵자-점자 말뭉치 파일

최종 구축용 말뭉치 데이터 생성을 위해 3차 변환 절차를 진행하였다. 이 단계에서는 정제 및 가공이 완료된 묵자 문장 데이터를 대상으로 기계 점역 과정을 수행하여 묵자-점자 병렬 말뭉치 최종 파일을 구축한다.

기계 점역 도구로는 ㈜ACNS의 Docu Braille V1.0 솔루션을 사용하였는데, 이는 웹 기반 관리 시스템과 에이피아이(API) 방식으로 연계되어 실시간 점역 결과를 자동 반영하는 구조이다. 기계 점역 결과는 점자 유니코드 값으로 변환되어 데이터베이스에 저장된다.

(1) 자바(JAVA) 객체 변환

최종 변환 과정에서 데이터 처리 효율성을 높이기 위해 자바(Java) 환경에서 다음과 같은 단계로 객체 변환을 수행하였다.

① 1차 변환: 묵자 문장 84,000개를 자바에서 처리 가능한 제이슨 객체(JSON Object) 형태로 변환

② 2차 변환: 가공된 제이슨 객체(JSON Object)에서 점역 대상인 묵자 문장 값을 추출하여 제이슨 배열(JSON Array)로 변환

③ 3차 변환: 최종 제이슨 배열(JSON Array)에서 메타데이터는 제이슨 객체(JSON Object)로, 본문(Paragraph)은 제이슨 배열(JSON Array)로 분리하여 저장

이를 통해 최종 구축된 묵자-점자 말뭉치는 기계 학습용 학습 데이터로 활용할 수 있도록 구조화하였으며, 점자 규정 반영 및 통계 분석 기반 검수가 용이한 형태로 가공하였다.

#	이름	데이터 유형	길이/설정	부호 없음	NULL 허용	0으로 채움	기본값	코멘트
1	year	VARCHAR	4	☐	☐	☐	기본값 없음	년도
2	brl_id	VARCHAR	20	☐	☐	☐	기본값 없음	검수ID (코드+일련번호)
3	brl_ser	INT	11	☐	☐	☐	기본값 없음	검수일련번호
4	process_ser	TINYINT	4	☐	☑	☐	'1'	검수작업차수
5	json_id	VARCHAR	14	☐	☑	☐	NULL	검수작업파일 및 메타정보 테이블의 JSON ID 값
6	subcategory	VARCHAR	50	☐	☑	☐	NULL	중분류
7	title	VARCHAR	200	☐	☑	☐	NULL	도서명
8	author	VARCHAR	100	☐	☑	☐	NULL	저자
9	publisher	VARCHAR	50	☐	☑	☐	NULL	출판사
10	write_date	VARCHAR	8	☐	☑	☐	NULL	작성년도
11	original_topic	VARCHAR	200	☐	☑	☐	NULL	주제
12	original_id	VARCHAR	30	☐	☑	☐	NULL	원본 paragraph id
13	scenario	TINYINT	4	☐	☑	☐	NULL	시나리오(정규식에 의한 분류값) CD010
14	form	TEXT		☐	☑	☐	기본값 없음	문장
15	orignal_braille	TEXT		☐	☑	☐	기본값 없음	기계점역점자
16	word_count	SMALLINT	6	☐	☑	☐	NULL	어절 수
17	form_length	SMALLINT	6	☐	☑	☐	NULL	문장 길이
18	target_yn	VARCHAR	1	☐	☑	☐	NULL	작업대상여부
19	job_userid	VARCHAR	20	☐	☑	☐	NULL	작업자ID (검수 중 인 작업자ID)
20	reg_date	DATETIME		☐	☑	☐	NULL	등록일시
21	reg_userid	VARCHAR	20	☐	☑	☐	NULL	등록사용자ID

그림 2-9 자바(JAVA) 객체 변환 화면

(2) 데이터베이스 등록

최종 변환 과정을 거쳐 생성된 묵자-점자 병렬 말뭉치 데이터는 데이터베이스에 항목별로 체계적으로 등록하였다. 등록 대상 파일은 아래와 같이 2개 유형의 제이슨(JSON) 객체로 구분된다.

- Metadata 객체: title(문서명), author(저자), publisher(출판사), data(작성일), original_topic (주제) 등 항목으로 데이터베이스에 저장
- Paragraph 객체: ID(원본 paragraph ID), form(묵자) 항목으로 데이터베이스에 저장

두 유형의 데이터는 최종 데이터베이스에서 메타 정보와 본문 데이터가 분리된 형태로 관리되며 각각 고유 ID를 기반으로 묵자-점자 병렬 말뭉치와의 연계 조회가 가능하도록 하였다.

최종 데이터베이스 등록 화면은 다음 그림 〈2-10〉과 같이 구성된다.

brl_id	json_id	subcategory	title	author	publisher	write_date	original_id	original_topic	scenario	word_count	form_length	form	original_braille
KBUCORPUS0000000	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 곽경식 ...	노컷뉴스	20210101	NRW2200000001.1.2	지역,전북|지역,경기|지역,전남	1	16	80	전북 고창과 부안, 경기 김포, 경북 ...	
KBUCORPUS0000001	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.2.6	국제,의료_복지|국제,유럽_EU|	2	15	74	발표 전망치는 더 어둡다. CDC(질병...	
KBUCORPUS0000002	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.2-2.12	국제,의료_복지|국제,유럽_EU|	1	11	53	시온 스티븐스 영국 국민보건서비스(...	
KBUCORPUS0000003	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.3.3	국제,유럽_EU|사회,의료_간호|	2	20	109	31일(현지시간) 회선보도를 종합하...	
KBUCORPUS0000004	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.4.2	IT_과학,IT_과학일반||	2	22	97	WHO(세계보건기구)가 파이자와 바...	
KBUCORPUS0000005	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.4.7	IT_과학,IT_과학일반||	1	23	101	WHO는 코백스(COVAX)를 통해 자개...	
KBUCORPUS0000006	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	홍단CBS 이상홍 기자	노컷뉴스	20210101	NRW2200000001.5.9	지역,경남|지역,대구|지역,부산	3	29	132	북구에 거주하는 최윤자(68)씨는 "한...	
KBUCORPUS0000007	NRW2200000001	신문 > 인터넷 기반 신문	홍단CBS 이상홍 기자	홍단CBS 이상홍 기자	노컷뉴스	20210101	NRW2200000001.5.13	지역,경남|지역,대구|지역,부산	3	23	92	남구에 사는 이효준(48)씨는 "아내와 ...	
KBUCORPUS0000008	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	홍단CBS 이상홍 기자	노컷뉴스	20210101	NRW2200000001.5.14	지역,경남|지역,대구|지역,부산	3	23	114	북구에 거주하는 이해준(35)씨는 "처...	
KBUCORPUS0000009	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	경남CBS 최효정 기자	노컷뉴스	20210101	NRW2200000001.6.3	사회,사건_사고|국제,아시아국제...	5	18	82	경남 마산동부경찰서는 음주단속(32)...	
KBUCORPUS0000010	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.9.3	국제,러시아|국제,중국|국제,유럽...	2	23	115	31일(현지시간) 로이터통신에 따르...	
KBUCORPUS0000011	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.9.8	국제,러시아|국제,중국|국제,유럽...	2	19	88	알서 파태프 시어트로 팀가리 외교장...	
KBUCORPUS0000012	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	전북CBS 송승민 기자	노컷뉴스	20210101	NRW2200000001.11.3	사회,사건_사고|지역,대구|사회,...	5	16	71	1일 전주지법 등에 따르면 검찰은이...	
KBUCORPUS0000013	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	전북CBS 송승민 기자	노컷뉴스	20210101	NRW2200000001.11.4	사회,사건_사고|지역,대구|사회,...	5	25	99	A씨는 지난 4월 24일 밤 전북 전주시...	
KBUCORPUS0000014	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	전북CBS 송승민 기자	노컷뉴스	20210101	NRW2200000001.11.12	사회,사건_사고|지역,대구|사회,...	2	20	109	B씨는 충남 논산의 호남고속도로 별...	
KBUCORPUS0000015	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 정준플 ...	노컷뉴스	20210101	NRW2200000001.12.2	스포츠,야구|스포츠,축구|스포츠,...	3	10	55	한국 야구 국가대표팀 주전 포수이...	
KBUCORPUS0000016	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 장성주 ...	노컷뉴스	20210101	NRW2200000001.14.8	IT_과학,IT_과학일반|국제,일본|...	1	30	126	국방기고 중 우기 과학의 최고 연구기...	
KBUCORPUS0000017	NRW2200000001	신문 > 인터넷 기반 신문	제주CBS 김대희기자	노컷뉴스	노컷뉴스	20210101	NRW2200000001.22.6	지역,제주|지역,경남|지역,충남	3	13	82	추가된 확진자 5명은 현라사루나 감...	
KBUCORPUS0000018	NRW2200000001	신문 > 인터넷 기반 신문	노컷뉴스 2021년 기사	CBS노컷뉴스 곽경식 ...	노컷뉴스	20210101	NRW2200000001.33.2	지역,경북|지역,전북|지역,경남	1	11	55	전북 고창의 유흥오리 농장과 경북 ...	

그림 2-10 데이터베이스 예시

5 데이터 검사

데이터 검사 단계에서는 최종 가공된 묵자-점자 병렬 말뭉치에 대한 정밀 검수와 교정 작업을 수행한다. 검수자는 사업에서 설정한 검수 지침에 따라 문장별로 적합/오류/보류로 판정하며 오류와 보류 사례에 대해서는 구체적인 사유를 필수로 기재한다.

검수 품질 확보를 위해 2단계 검수 체계를 도입하였다. 한 번이라도 오류/보류 판정을 받은 문장은 말뭉치 사업 수행 인력 3명이 최종 점형 교정(수정)을 통해 최종 적합 문장으로 재확인하는 절차를 마련한 것이다. 이러한 다중 검수와 교정 과정을 통해 최종 말뭉치 데이터의 완전성, 유일성, 유효성이 유지되도록 관리하였다.

표 2-14 데이터 검사 세부 절차

세부 절차	작 업	해당 작업자
1차 검수	- 검수 단계 '1단계' 설정 - 지침에 따른 점형 문장 적합/오류/보류 판정	말뭉치 구축 검수자(68명) + 말뭉치 사업 수행 인력(3명)
2차 검수	- 검수 단계 '2단계' 설정 - 1차 검수된 적합 문장을 재검수 후 최종 판정	말뭉치 구축 검수자(66명) + 말뭉치 사업 수행 인력(3명)
오류/보류 문장 교정	- 오류 및 보류 문장을 올바른 점형으로 최종 교정 - 교정 후 재검수 없이 적합 반영	말뭉치 사업 수행 인력(3명)

1) 검수 지침

사람이 직접 검수할 때에는 일관성이 일정 수준 확보된 형태로 작업을 하는 것이 중요하다. 이를 위한 검수 지침은 아래와 같다.

(1) 사이트 접속 및 로그인

- 말뭉치 웹 기반 관리 시스템 접속
- 컴퓨터(PC)/모바일/태블릿 모두 접속할 수 있으나, 모바일은 휴대전화 기종에 따라 점자 화면이 잘려서 보이는 경우가 많으므로 컴퓨터(PC)/태블릿에서 작업 권고

그림 2-11 말뭉치 관리 시스템 로그인 화면

(2) 내비게이션 구성

- 검수: 묵자-점자 병렬 말뭉치 검수 화면
- 공지사항: 공지사항 확인 페이지
- 게시판: 검수자가 직접 글을 작성하고 댓글도 작성할 수 있는 다용도 게시판 페이지
- 설문조사: 설문조사에 참여할 수 있는 페이지
- 비밀번호: 비밀번호 변경(새 창 열림)
- 로그아웃: 검수 사이트에서 로그아웃
- 돋보기 아이콘: 글자 돋보기 기능(이스케이프키로 종료), 돋보기 상태에서 클릭은 불가능
- 회색 원형 아이콘: 다크 모드 켜기/끄기 기능
- 알트키+M키를 통해 내비게이션 메뉴 비활성화 가능

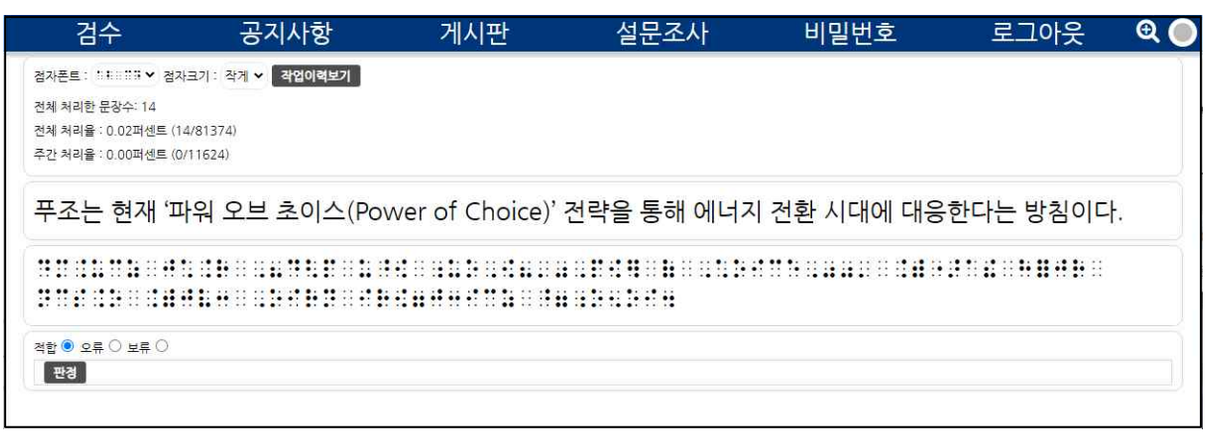


그림 2-12 말뭉치 관리 시스템 내비게이션 화면

(3) 검수 화면

- 화면에는 묵자-점자 한 쌍만 나오며 윗줄에는 묵자, 아랫줄에는 점자가 표기됨
- 화면 낭독기(센스리더) 사용자는 점자 영역에 초점 접근 후 홈(Home)키 입력 후 방향키로 이동하면 아스키코드(ex. 135점, 24점) 형태로 음성 출력됨
- 가독성 및 편의를 위해 점자 글꼴은 3종, 점자 크기/묵자 크기는 4종으로 변경 가능
- 인터넷 창 가로 길이를 조절하여 묵자-점자 배치를 양옆 또는 위아래로 변경 가능

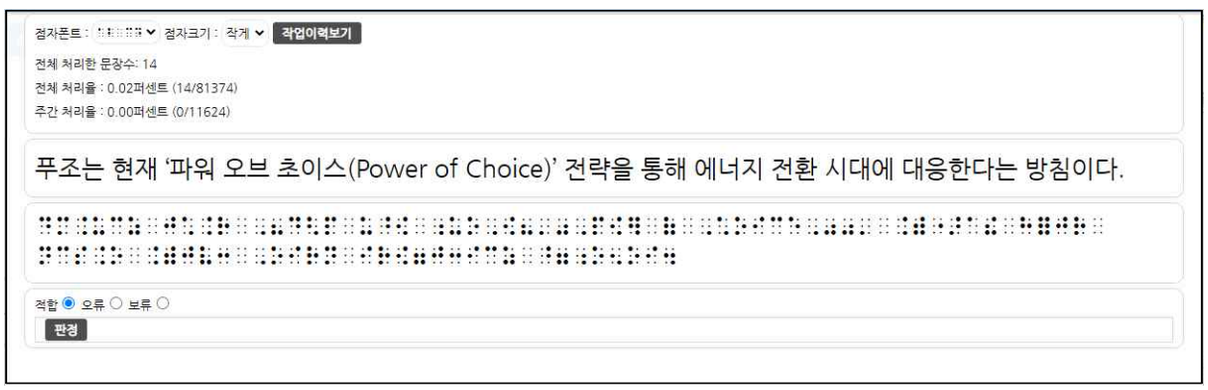


그림 2-13 말뭉치 관리 시스템 묵자-점자 검수 화면

(4) 적합/오류/보류 버튼 체크

- 묵자-점자 일치 여부에 따라 말뭉치 하단에 있는 적합/오류/보류 라디오 버튼 체크
- 적합: 묵자와 점자 점역이 규정과 일치하며 모두 올바르게 점역된 경우
- 오류: 묵자와 점자 점역이 규정과 다르거나 묵자 내용이 바르게 점역되지 않은 경우
- 보류: 한국 점자 규정에 정의되지 않은 내용이나 기호가 있거나, 묵자 내용이 제2외국어, 수식, 악보 등과 같이 한글 점자 또는 통일영어점자 범위를 벗어나는 경우

(5) 오류/보류 사유 기재

- 적합의 경우 바로 적합 라디오 버튼 체크 후 판정 버튼 선택
- 오류/보류의 경우 콤보 박스에서 사유 선택 후 텍스트 박스에 틀린 단어/사유 기재
- 사유 유형: 1급 점자표 오류, 로마자 시작표/종료표 오류, 알파벳 약자/약어 오류, 문장부호 오류, 단위/연산기호 오류, 영어 대문자 표기 오류, 제2외국어/한자 오류, 기타
- 오류/보류 시 틀린 단어를 텍스트 박스에 기재해야 판정 버튼이 활성화되며 다음 문장으로 넘어갈 수 있음

- 한 문장에 오류와 보류가 섞여 있는 경우, 오류 버튼을 누른 후 오류 사유와 보류 사유 모두 작성
- 틀린 사항이 여러 개인 경우, 기타(직접 입력) 사유 콤보 박스 선택 후 틀린 단어 모두 작성
- 묵자 원문의 띄어쓰기가 잘못되어 있으나 점자 확인 시 그대로 점역되어 있다면 '적합'이지만, 띄어쓰기가 한국 점자 규정에 어긋난다면 '오류'로 판정

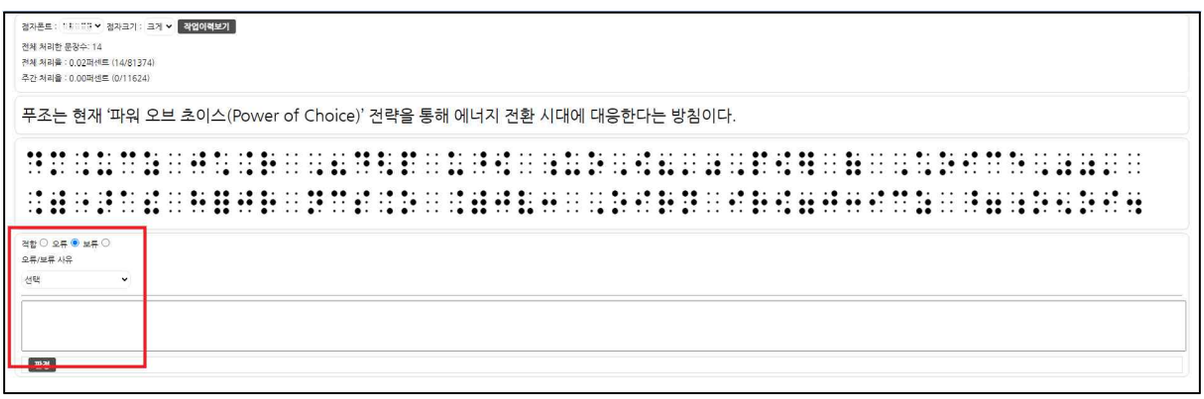


그림 2-14 오류/보류 버튼 선택 시 하단에 나오는 콤보 박스 및 텍스트 박스 화면

(6) 검수 작업 이력 보기

- 본인이 검수한 문장은 작업 이력 보기 버튼 클릭 후 목록에서 확인 및 수정 가능
- 화면에 최대 10개의 문장이 보이며, 화살표 버튼 선택 시 다른 문장이 10개씩 추가로 보임
- 화면에 보이는 10개의 문장은 숫자키(1~0)를 통해 빠른 이동이 가능함

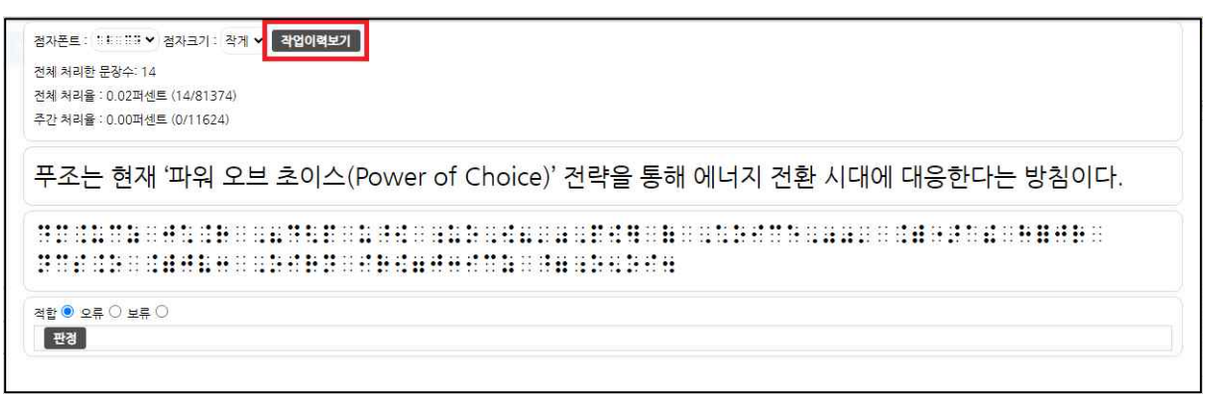


그림 2-15 작업 이력 보기 화면

그림 2-16 작업 이력 보기 클릭 후 화면

(7) 작업 이력 보기 상세

- 주차별/판정별 선택 확인
- 특정 주차별 내역 또는 특정 판정 내역만 확인이 가능하며 해당 버튼 클릭
- 주차별: 전체 보기, 각 주차별 버튼
- 판정별: 전체 보기, 적합, 오류, 보류 버튼
- 판정 내역 옆에 소괄호로 해당 판정 숫자가 기재되어 있음(예: 적합(513))
- 판정별 버튼 우측 콤보 박스에서 사유를 선택하면 해당 사유 문장만 확인 가능

그림 2-17 주차별/판정별 검수 문장 분류 기능

- 문자 입력 후 엔터키를 누르면 해당 문자가 포함된 문장만 확인 가능
- 묵자: 묵자 검색 → 텍스트 입력창에 묵자 기재 후 엔터
- 점자: 점자 선택 → 텍스트 입력창에 점자 붙여넣기 후 엔터키 또는 알트키+슬래시를 통해 점자 직접 입력 후 엔터

그림 2-18 단어 검색 기능

- 작업 이력 화면에서 수정 희망 시 검수 방법과 동일하게 판정 후 수정 가능
- 적합 → 오류/보류: 콤보 박스에서 해당 사유 선택 후 텍스트 박스에 틀린 단어 기재 뒤 수정하기 눌러야 변경 적용
- 오류/보류 → 적합: 적합 라디오 버튼 선택 후 수정하기 눌러야 변경 적용
- '수정하기' 클릭 시 '수정하시겠습니까?' 팝업이 뜨며 확인/취소 버튼 활성화

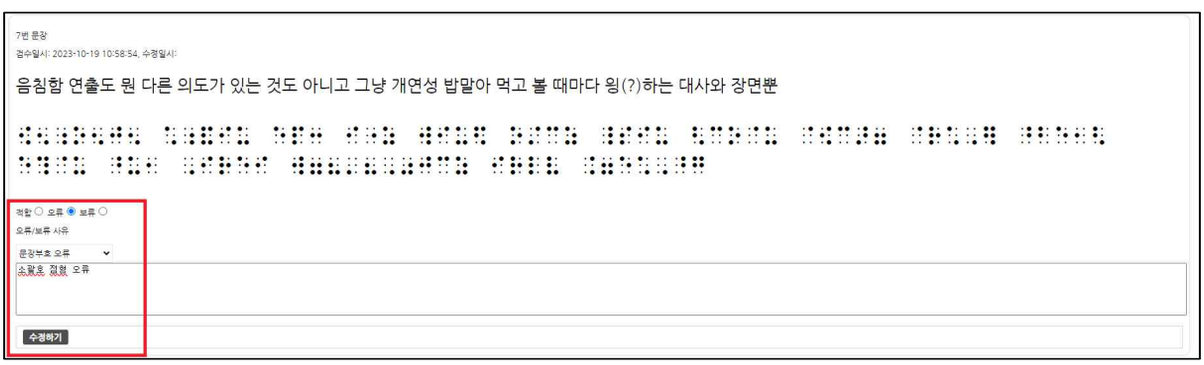


그림 2-19 작업 이력 화면 내 수정 시 화면

(8) 오류/보류 점형 교정

- 사업 수행 인력이 작업하는 화면은 검수자 검수 화면과 동일하게 윗줄에는 묵자, 아랫줄에는 점자가 표기됨
- 묵자 상단에는 본인이 받은 적합/오류/보류 문장 수가 표기되어 있음
- 상단에 주차별 내역이 있으며, 메뉴 건너뛰기 버튼 선택 시 바로 문장으로 이동
- 비시각장애인 검수자의 경우, 점자 글꼴 가독성 및 편의를 위해 3종의 점자 글꼴로 변경 가능

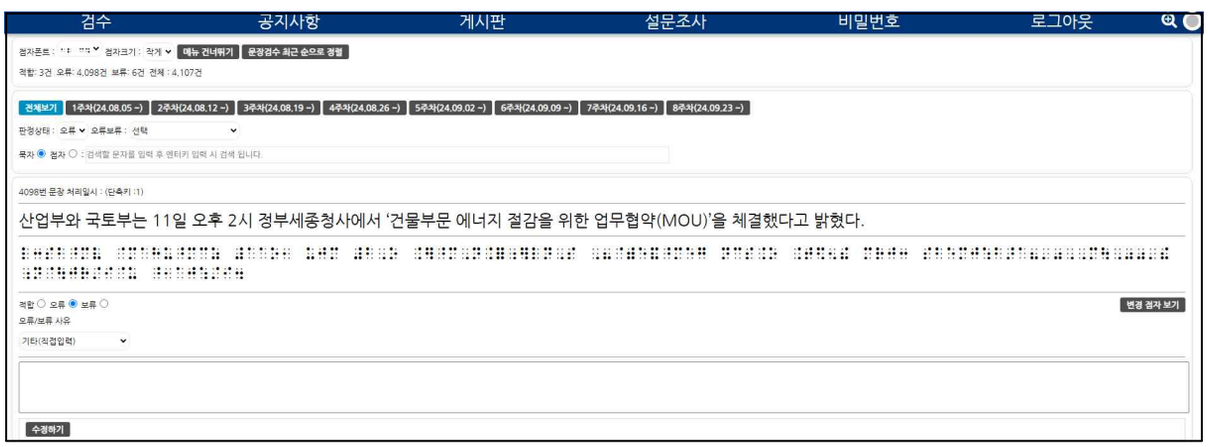


그림 2-20 말뭉치 관리 시스템 오류/보류 점형 교정 화면

- 점자 점형을 변경하고자 하는 경우 점자가 표시된 영역을 더블클릭 혹은 엔터키를 입력하면 점형 수정 가능한 새 창으로 이동
- 화면 낭독기(센스리더) 사용 시 엔터키가 작동되지 않을 경우 알트키+엔터키로 이동 가능
- 점자 편집 창에서 오류/보류 부분에 해당하는 점형을 삭제한 뒤 키보드 6점 키를 이용해 올바른 점형 입력
- 점형 입력 후 엔터키를 누르면 이전 검수 화면에 수정 적용된 점자가 나타남
- 새 창이 열릴 때 초점이 상단 묵자에 위치하므로 화면 낭독기(센스리더) 사용자는 탭키를 이용해 점자 편집 창 접근 가능
- 점자 점형 수정 후 판정에서 적합으로 변경 후 수정하기 버튼 선택

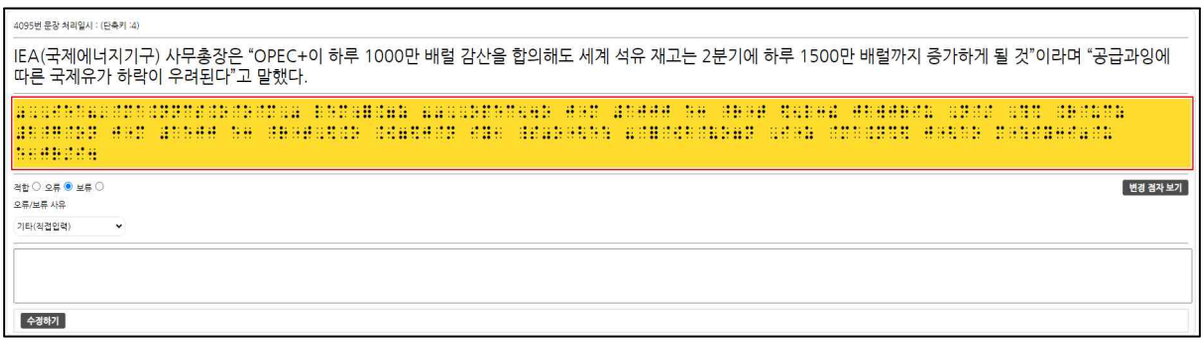


그림 2-21 오류/보류 검수 화면에서 수정 전 선택 화면

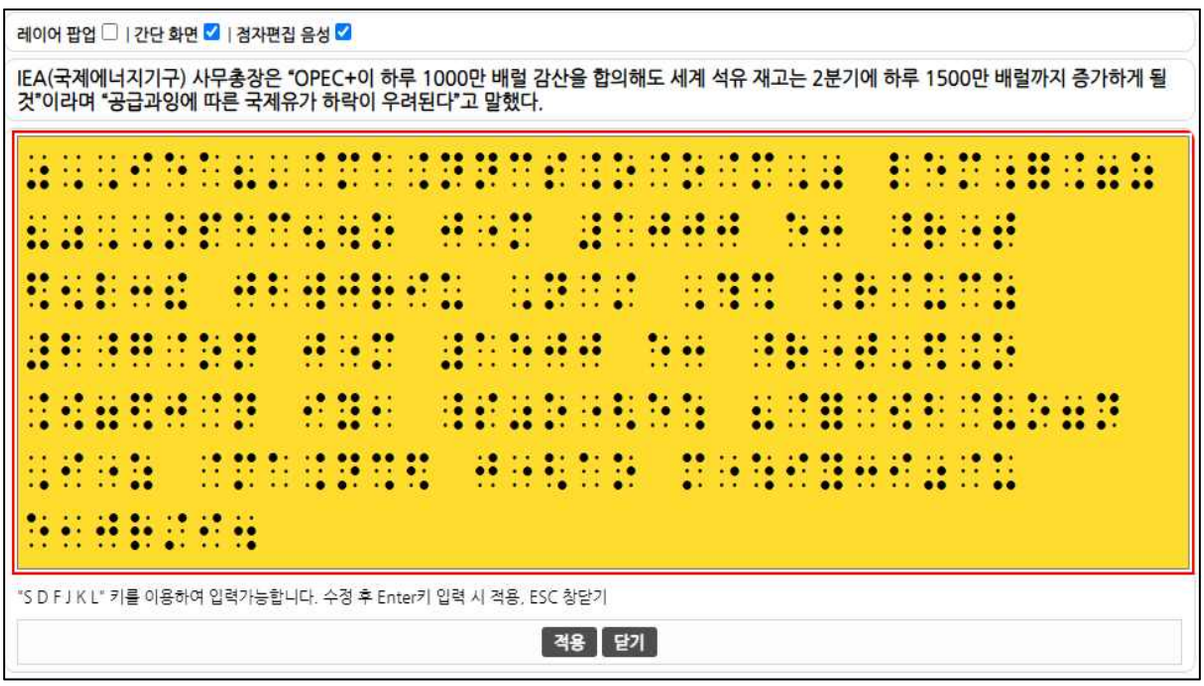


그림 2-22 점자 점형 교정 새 창 화면

(9) 교정 작업 이력 보기

- 작업 이력 보기 버튼 대신 상단에 주차별 내역이 있으며, 메뉴 건너뛰기 버튼 선택 시 바로 문장으로 이동
- 주차별/판정별/수정 이력별 선택 확인은 없으나 메뉴 건너뛰기 버튼 우측에 판정별 선택 콤보 박스로 선택 확인 가능

그림 2-23 오류/보류 점형 교정 작업 이력 보기 화면

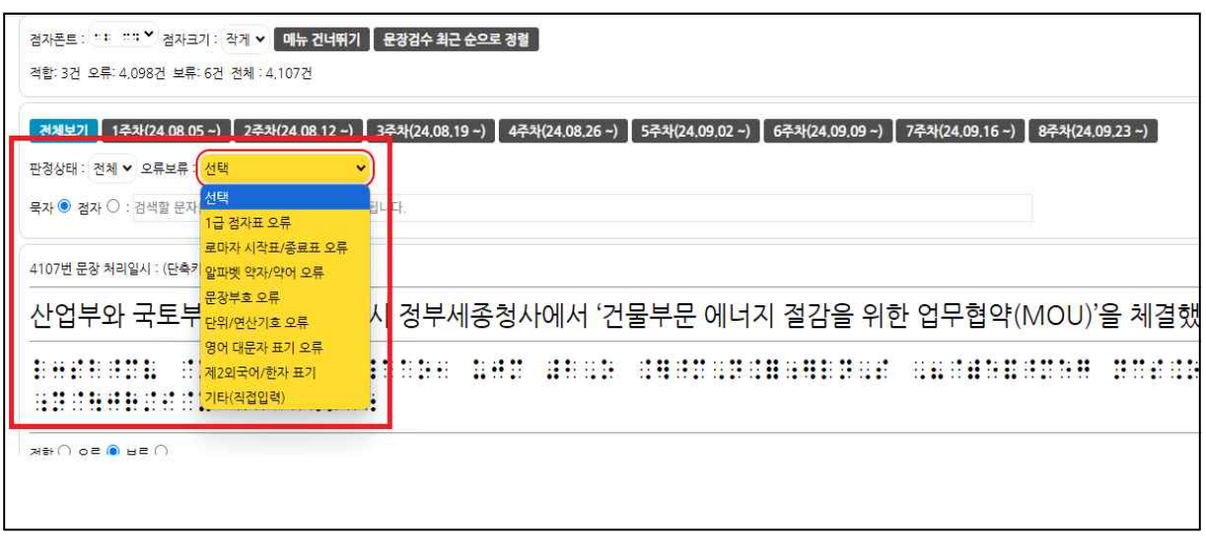


그림 2-24 판정 상태 및 오류/보류 유형별 콤보 박스

제 3 장

묵자-점자 병렬 말뭉치 검수 및 통계 현황

1. 검수 방법
2. 조사 결과

1 검수 방법

1) 검수 참여자

본 사업의 검수 참여자는 목적 표집(Purposeful Sampling) 방식으로 선정하였다. 검수자는 주로 시각장애인 관련 기관 및 점자 편집 실무 현장에서 활동하고 있는 점역·교정사를 대상으로 선발하였으며 검수 참여자의 전문성과 다양성을 확보하는 데 중점을 두었다.

검수자 선발 기준으로는 점역·교정사 자격증 보유 여부, 자격 취득 과목, 시각장애인 당사자 여부 등을 고려하였다. 최종적으로 총 68명의 검수자가 선발되었는데, 이 중 시각장애인 점역·교정사는 17명, 비시각장애인 점역·교정사는 51명이었다.

검수자의 주요 업무는 묵자(일반 문자)와 점자 변환 결과를 대조하여 변환의 정확성을 확인하고 적합/오류/보류 판정을 수행하는 것이다. 수정 구축의 경우, 검수자 작업 품질 관리 및 편의성을 위해 87어절 이상인 2,487문장(259,239어절)을 말뭉치 사업 수행 인력 3명이 직접 검수하였다. 검수 과정에서 오류 또는 보류로 판정된 문장의 최종 교정 작업도 말뭉치 사업 수행 인력 3명이 수행하였다.

구체적인 검수자 정보는 〈표 3-1〉과 같다.

표 3-1 검수 참여자 정보

연번	이 름	담당 분야	1차 검수 할당 문장(어절) 수	2차 검수 할당 문장(어절) 수	장애 구분	자격증 급수	취득 과목
1	강*혜	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
2	강*원	수정	70,000어절	70,000어절	-	1급	국어/영어/음악
3	공*	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학/음악/일본어
4	곽*선	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
5	김*리	신규	3,500문장	-	-	1급	국어/영어/수학·과학/음악
6	김*진	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
7	김*정	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/음악
8	김*경	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
9	김*연	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
10	김*연	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
11	김*기	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/음악

연번	이 름	담당 분야	1차 검수 할당 문장(어절) 수	2차 검수 할당 문장(어절) 수	장애 구분	자격증 급수	취득 과목
12	김 * 주	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
13	김 * 정	신규	3,500문장	3,818문장	-	1급	국어/영어/수학·과학
14	김 * 준	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
15	김 * 혜	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/음악
16	김 * 나	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
17	김 * 영	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
18	김 * 란	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
19	김 * 리	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/일본어
20	류 * 정	신규	3,500문장	3,818문장	-	1급	국어/영어/수학·과학
21	민 * 영	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/음악/일본어
22	박 * 화	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
23	박 * 규	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
24	박 * 재	신규	3,500문장	3,818문장	중증	1급	국어/영어/음악
25	박 * 미	수정	70,000어절	70,000어절	-	2급	국어/영어
26	박 * 연	수정	70,000어절	70,000어절	-	1급	국어/영어/음악
27	박 * 영	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
28	박 * 영	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
29	박 * 림	수정	70,000어절	70,000어절	-	1급	국어/영어/음악
30	봉 * 용	신규	3,500문장	3,818문장	중증	1급	국어/영어/음악
31	서 * 옥	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
32	서 * 희	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/일본어
33	송 * 혜	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
34	신 * 비	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
35	신 * 규	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
36	안 * 진	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
37	안 * 령	신규	3,500문장	3,818문장	-	1급	국어/영어/수학·과학
38	양 * 훈	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학/음악/일본어
39	양 * 경	신규	3,500문장	3,818문장	-	1급	국어/영어/수학·과학
40	양 * 린	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
41	오 * 리	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/일본어

연번	이 름	담당 분야	1차 검수 할당 문장(어절) 수	2차 검수 할당 문장(어절) 수	장애 구분	자격증 급수	취득 과목
42	유 * 연	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
43	윤 * 정	신규	3,500문장	3,818문장	-	1급	국어/영어/음악
44	윤 * 정	수정	70,000어절	-	중증	1급	국어/영어/수학·과학
45	이 * 진	신규	3,500문장	3,818문장	-	1급	국어/영어/수학·과학
46	이 * 연	수정	70,000어절	70,000어절	-	1급	국어/영어/일본어
47	이 * 내	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
48	이 * 진	수정	70,000어절	70,000어절	중증	1급	국어/영어/수학·과학
49	이 * 정	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
50	이 * 희	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
51	이 * 수	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
52	이 * 연	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
53	이 * 선	수정	70,000어절	70,000어절	-	1급	국어/영어/음악
54	이 * 민	신규	3,500문장	3,818문장	-	1급	국어/영어/수학·과학
55	전 * 희	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
56	전 * 민	수정	70,000어절	70,000어절	-	1급	국어/영어/음악
57	정 * 현	신규	3,591문장	3,818문장	중증	1급	국어/영어/수학·과학/음악/일본어
58	조 * 영	신규	3,409문장	70,000어절	-	1급	국어/영어/음악
59	조 * 자	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
60	진 * 라	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/음악
61	하 * 섭	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학/음악/일본어
62	하 * 리	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학/음악/일본어
63	하 * 주	수정	70,000어절	70,000어절	-	1급	국어/영어/음악
64	하 * 총	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학
65	하 * 영	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
66	한 * 수	신규	3,500문장	3,818문장	중증	1급	국어/영어/수학·과학
67	한 * 은	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학/음악
68	한 * 진	수정	70,000어절	70,000어절	-	1급	국어/영어/수학·과학

2) 검수 단계

구축된 묵자-점자 병렬 말뭉치의 품질을 확보하기 위해 2단계 검수 체계를 운영하였다. 1차 검수에서는 점자 변환 결과의 기본 적합성을 검토하고, 2차 검수에서는 타 검수자의 1차 판정 결과를 재검토하는 방식으로 품질 신뢰도를 향상한다. 검수 결과는 주 단위 통계로 관리되며 일정 기준 이하의 검수량이나 오류율이 높은 경우 추가 조치를 통해 품질을 보완하였다.

(1) 1차 검수 단계

1차 검수는 묵자와 점역된 점자 문장 간의 일치 여부를 확인하는 과정으로 두 가지 기록 체계로 운영되었다.

첫째, 검수자는 적합/오류/보류 중 하나의 판정을 선택하도록 되어 있다.

- 적합: 점역이 원문 내용과 정확히 일치하는 경우
- 오류: 점역 과정에서 내용 왜곡, 점자 규정 위반 등 명백한 오류가 확인된 경우
- 보류: 한국 점자 규정의 한글 점자에서 정의되지 않은 표현(제2외국어, 수식, 화학식, 악보 등)이 있는 경우

둘째, 오류 및 보류 판정 시 반드시 사유를 기재하도록 구성하였다. 검수자는 단순히 오류를 지적하는 것에 그치지 않고 점역 오류의 원인과 유형을 상세히 기술하도록 하였다. 예를 들어 영어 단어 다음에 로마자 종료표가 빠짐, 영어 단어 앞에 1급 점자표 오기, 한국 점자 규정에 나와 있지 않은 글머리 기호, 한국 점자 규정에서 정의되지 않은 점형 등 어떤 내용에서 점역이 잘못됐는지 해당 내용을 표기하도록 하였다.

1차 검수 참여자는 〈표 3-1〉에 명시된 검수자 68명과 말뭉치 사업 수행 인력 3명으로 총 71명이며, 전체 문장 분량에 따라 검수 대상이 배정되었다.

- 검수 기간: 2025년 6월 9일~7월 27일(총 7주)
- 검수 분량: 신규 구축 84,000문장, 수정 구축 93,510문장(3,339,239어절)
- 예비 기간: 2025년 7월 28일~8월 3일, 미달자 추가 작업 및 휴식 기간 운영

(2) 2차 검수 단계

2차 검수는 1차 검수에서 판정된 문장을 대상으로 교차 검토를 하는 단계이다. 검수 방법은 1차와 동일하게 적합/오류/보류 판정 방식을 적용하며, 검수의 일관성과 정확도를 추가로 검증하는 목적을 가진다.

- 검수 기간: 2025년 8월 4일~9월 21일(총 7주)
- 검수 대상: 1차 검수 완료 후 판정 문장 전량

2차 검수는 검수자 66명과 말뭉치 사업 수행 인력 3명이 참여하였다.

표 3-2 1차 검수 및 2차 검수 일정

구분	날짜	내용	비고
1차 검수	6월 9일~7월 27일 (총 7주)	- 검수자 68명 1차 검수 진행 - 검수량: 매주 500문장 또는 10,000어절	1인당 3,000문장 또는 70,000어절
-	7월 28일~8월 3일 (총 1주)	- 휴식 - 1차 검수 검수량 미달자 추가 작업	미달자 불성실 적용
2차 검수	8월 4일~9월 21일 (총 7주)	- 검수자 66명 2차 검수 진행 - 검수량: 매주 500문장 또는 10,000어절	1인당 3,818문장 또는 70,000어절
-	9월 22일~9월 28일	- 2차 검수 검수량 미달자 추가 작업	미달자 불성실 적용

(3) 오류/보류 점형 교정

1차 및 2차 검수 과정에서 한 번이라도 오류/보류 판정을 받은 문장은 별도의 교정 작업을 거친다. 이는 최종 구축 말뭉치의 품질을 극대화하고 오류율을 최소화하기 위한 조치이다.

오류/보류 판정 문장은 2차 검수 기간과 맞물려 말뭉치 사업 수행 인력 3명이 점형 오류를 직접 교정하며, 정확한 점자 변환 예시 결과물을 구축하는 것을 목표로 하였다. 최종 교정 작업 완료 시 해당 교정 문장은 최종 말뭉치 데이터에 포함되었다.

표 3-3 오류/보류 점형 교정 검수 일정

구분	날짜	내용	비고
교정	8월 11일~9월 28일 (총 7주)	- 오류/보류 판정 문장 점형 교정 - 검수량: 매주 약 423문장(오류/보류율 5% 기준)	1인당 2,958문장

3) 검수 교육 절차

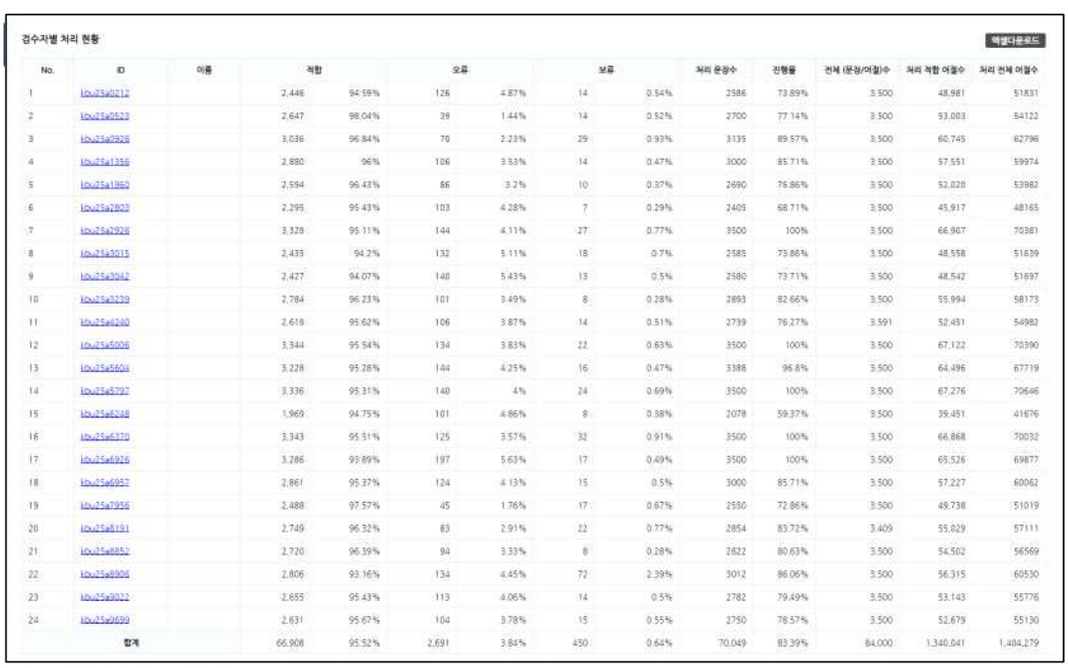
검수 품질을 확보하기 위해 사전 검수자 교육을 필수적으로 실시하였다. 검수 참여자는 묵자와 점역된 점자 문장을 비교하여 점역의 정확성을 검증하는 역할을 수행하므로 이를 위해 이론 교육과 실습 교육으로 사전 교육을 구성하였다. 검수 참여자 교육은 온라인 비대면 사전교육으로 진행하였다. 이론 교육은 2025년 5월 31일 오후 2시부터 오후 5시까지 3시간 실시하였다. 주요 교육 내용은 다음과 같다.

- 사업 설명: 국립국어원 '2025년 묵자-점자 병렬 말뭉치 구축' 사업의 목적 및 추진 방향 설명
- 검수 기준 안내: 묵자-점자 병렬 말뭉치 검수 대상 문장 수 및 수정 구축 대상 설명
- 검수 시스템 사용법 실습: 웹 기반 관리 시스템 접속 방법, 검수 판정법(적합/오류/보류), 오류 및 보류 사유 기재 요령 실습
- 점자 규정 교육: 개정 한국 점자 규정 주요 변경 사항 및 검수 시 유의 사항 집중 교육
- 자주 하는 질문 및 질의응답: 교육 종료 후 질의응답 및 자주 발생하는 오류 사례 공유

4) 자료 처리

검수자는 웹 기반 관리 시스템에 접속하여 실시간으로 검수 작업을 진행한다. 시스템은 접속 시점에 검수되지 않은 문장을 자동으로 배정하는 방식으로 운영되며 검수자의 개별 작업량은 시스템을 통해 실시간으로 확인하고 관리할 수 있다. 검수 기준은 적합/오류/보류 세 가지로 구분되며 오류/보류 판정 시에는 반드시 사유를 입력하도록 설정되어 있다. 검수 진행 상황은 실시간으로 시각화되어 제공되며 검수자는 본인의 총 검수 문장 수, 누적 검수율, 오류 및 보류 건수, 어절 수 등을 수시로 확인할 수 있다.

관리자 화면에서는 전체 검수 진행률과 주별 검수량 현황이 시각화되어 제공된다. 검수 결과는 엑셀 파일 형태로 변환이 가능하며, 데이터는 개수 기준으로 표기된다. 집계 항목은 검수 결과별 적합, 오류, 보류 수량뿐 아니라 어절 수, 문장 수 기준의 세부 검수 현황까지 포함되며, 오류/보류 판정 시 기록한 사유별 세부 통계도 함께 제공된다.



검수자별 처리 현황

엑셀다운로드

No.	ID	이름	적합		오류		보류		처리 문장수	진행률	전체 (문장/어절)수	처리 적합 어절수	처리 전체 어절수
1	kbu25a0212		2,446	94.59%	126	4.87%	14	0.54%	2586	73.89%	3,500	48,981	51831
2	kbu25a0523		2,647	98.04%	39	1.44%	14	0.52%	2700	77.14%	3,500	53,003	54122
3	kbu25a0926		3,036	96.84%	70	2.23%	29	0.93%	3135	89.57%	3,500	60,745	62796
4	kbu25a1356		2,880	96%	106	3.53%	14	0.47%	3000	85.71%	3,500	57,551	59974
5	kbu25a1960		2,594	96.43%	86	3.2%	10	0.37%	2690	76.86%	3,500	52,020	53982
6	kbu25a2803		2,295	95.43%	103	4.28%	7	0.29%	2405	68.71%	3,500	45,917	48165
7	kbu25a2928		3,328	95.11%	144	4.11%	27	0.77%	3500	100%	3,500	66,967	70381
8	kbu25a3015		2,435	94.2%	132	5.11%	18	0.7%	2585	73.86%	3,500	48,558	51639
9	kbu25a3042		2,427	94.07%	140	5.43%	13	0.5%	2580	73.71%	3,500	48,542	51697
10	kbu25a3239		2,784	96.23%	101	3.49%	8	0.28%	2893	82.66%	3,500	55,994	58173
11	kbu25a4240		2,618	95.62%	106	3.87%	14	0.51%	2739	76.27%	3,591	52,451	54982
12	kbu25a5006		3,344	95.54%	134	3.83%	22	0.63%	3500	100%	3,500	67,122	70390
13	kbu25a5608		3,228	95.28%	144	4.25%	16	0.47%	3388	96.8%	3,500	64,496	67719
14	kbu25a5797		3,336	95.31%	140	4%	24	0.69%	3500	100%	3,500	67,276	70646
15	kbu25a6248		1,969	94.75%	101	4.86%	8	0.38%	2078	59.37%	3,500	39,451	41676
16	kbu25a6370		3,343	95.51%	125	3.57%	32	0.91%	3500	100%	3,500	66,868	70032
17	kbu25a6926		3,286	93.89%	197	5.63%	17	0.49%	3500	100%	3,500	65,526	69877
18	kbu25a6957		2,861	95.37%	124	4.13%	15	0.5%	3000	85.71%	3,500	57,227	60062
19	kbu25a7956		2,488	97.57%	45	1.76%	17	0.67%	2550	72.86%	3,500	49,738	51019
20	kbu25a8191		2,749	96.32%	83	2.91%	22	0.77%	2854	83.72%	3,409	55,029	57111
21	kbu25a8852		2,720	96.39%	94	3.33%	8	0.28%	2822	80.63%	3,500	54,502	56569
22	kbu25a8906		2,806	93.16%	134	4.45%	72	2.39%	3012	86.06%	3,500	56,315	60530
23	kbu25a9022		2,655	95.43%	113	4.06%	14	0.5%	2782	79.49%	3,500	53,143	55776
24	kbu25a9699		2,631	95.67%	104	3.78%	15	0.55%	2750	78.57%	3,500	52,679	55130
합계			66,908	95.52%	2,691	3.84%	450	0.64%	70,049	83.39%	84,000	1,340,041	1,404,279

그림 3-1 검수자별 말뭉치 검수 처리 현황 화면

2 조사 결과

1) 1차 검수 데이터 통계 현황

1차 검수 데이터 통계 현황을 조사하였다. 신규 구축과 수정 구축의 1차 검수 데이터 통계 현황은 〈표 3-4〉, 〈표 3-5〉와 같다.

신규 구축 데이터의 검수 결과(1주~8주: 2025년 6월 9일~8월 3일 기준), 총 검수 수는 전체 84,000문장으로 누적 검수율은 100%로 조사되었다.

적합 수는 전체 84,000문장 중 80,347문장으로, 적합률은 95.65%이고 적합 어절 수는 1,610,647개로 조사되었다. 오류 수는 전체 84,000문장 중 3,155문장으로 오류율은 3.76%, 보류 수는 498문장으로 보류율은 0.59%로 조사되었다.

표 3-4 신규 구축 1차 검수 데이터 통계 현황

주차	기간	총 검수 수 (문장)	검수율 (%)	검수 결과								총 어절 수 (개)
				적합			오류		보류		계 (문장)	
				적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1주	6월 9일~ 6월 15일	13,035	15.52	12,338	248,418	94.65	581	4.46	116	0.89	13,035	262,682
2주	6월 16일~ 6월 22일	13,267	15.8	12,723	253,753	95.90	453	3.41	91	0.69	13,267	264,885
3주	6월 23일~ 6월 29일	13,141	15.66	12,499	248,930	95.11	516	3.93	126	0.96	13,141	262,125
4주	6월 30일~ 7월 6일	13,385	15.94	12,836	258,356	95.90	514	3.84	35	0.26	13,385	269,465
5주	7월 7일~ 7월 13일	12,486	14.86	11,976	240,190	95.92	487	3.90	23	0.18	12,486	250,682
6주	7월 14일~ 7월 20일	9,511	11.33	9,181	181,974	96.53	308	3.24	22	0.23	9,511	188,717
7주	7월 21일~ 7월 27일	8,604	10.25	8,282	168,708	96.26	263	3.06	59	0.69	8,604	175,352
8주	7월 28일~ 8월 3일	571	0.68	512	10,318	89.67	33	5.78	26	4.55	571	11,526
합 계		**84,000**	**100**	**80,347**	**1,610,647**	**95.65**	**3,155**	**3.76**	**498**	**0.59**	**84,000**	**1,685,434**

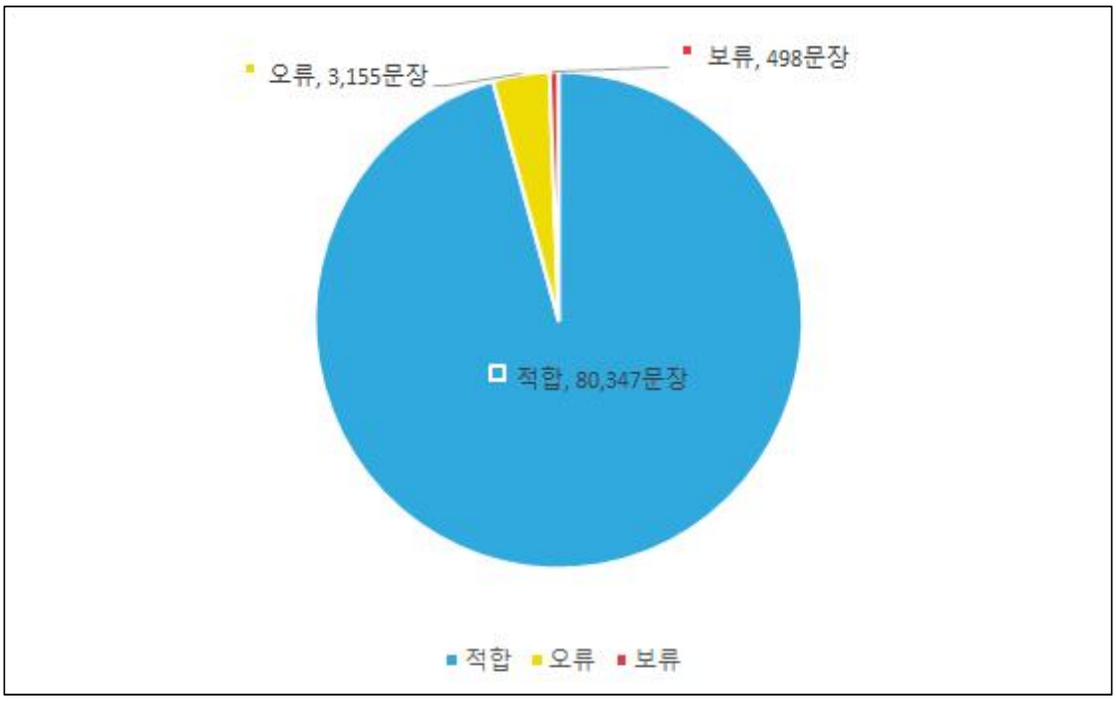


그림 3-2 신규 구축 1차 검수 적합/오류/보류 문장 수

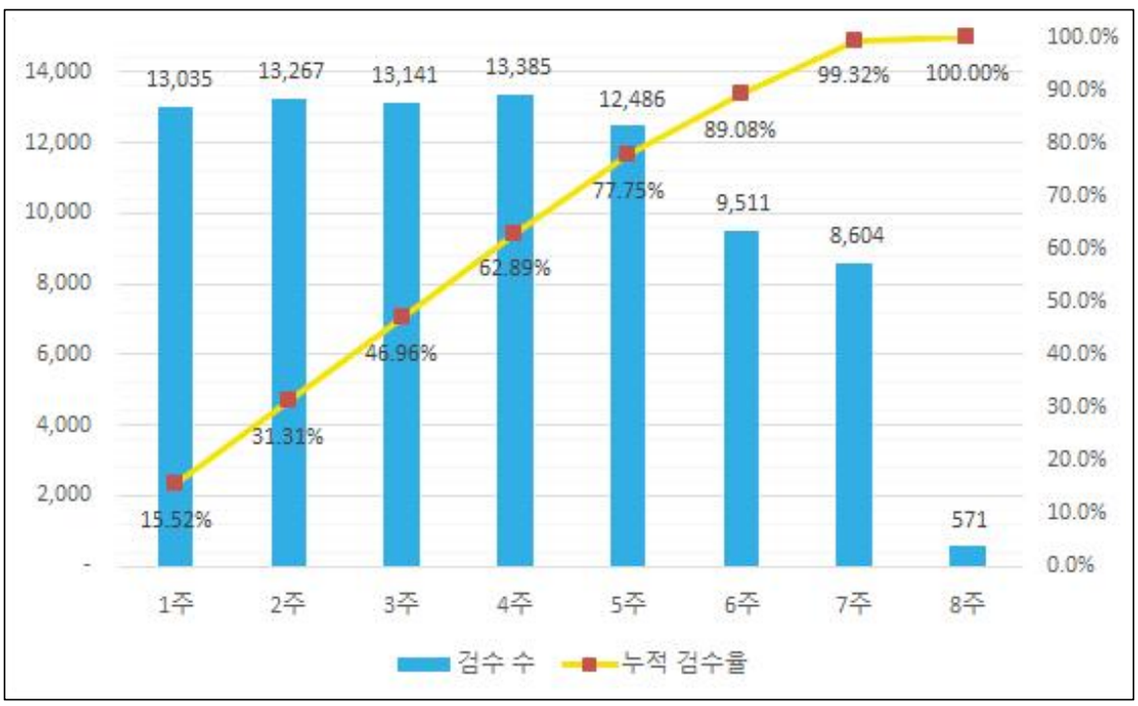


그림 3-3 신규 구축 1차 검수 주차별 검수 문장 수 및 누적 검수율

그림 3-4 신규 구축 1차 검수 주차별 적합/오류/보류 문장 수

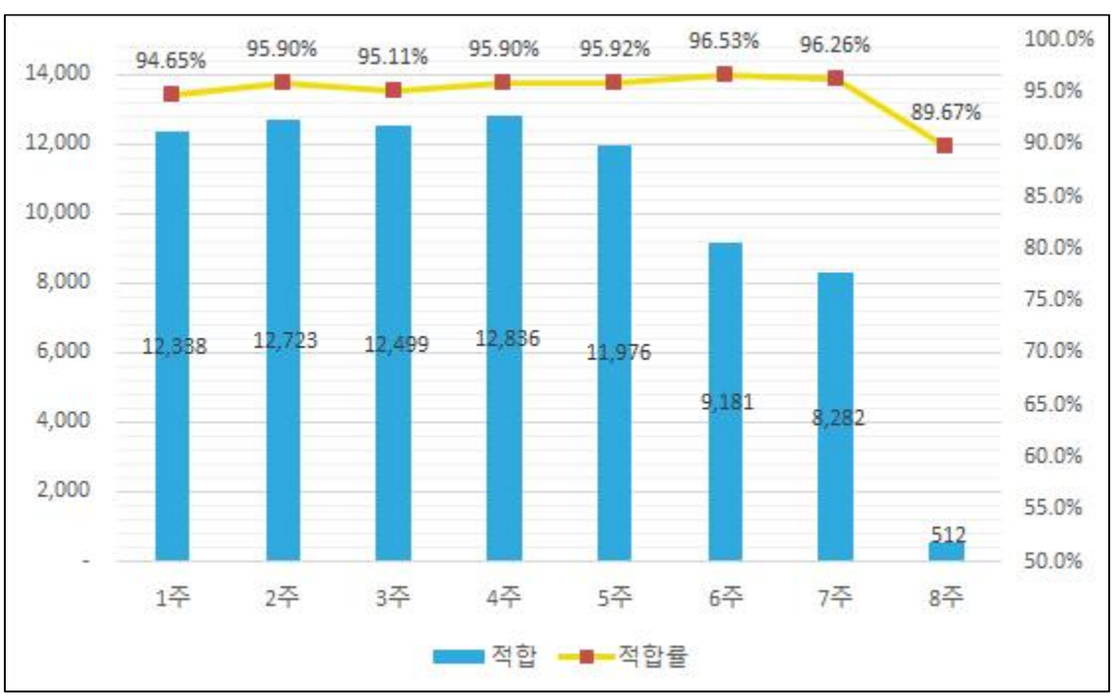


그림 3-5 신규 구축 1차 검수 주차별 적합 문장 수 및 적합률

수정 구축 데이터의 검수 결과(1주~7주: 2025년 6월 9일~7월 27일 기준, 8주 작업 없음), 총 검수 수는 전체 93,510문장으로 누적 검수율은 100%로 조사되었다.

적합 수는 전체 93,510문장 중 91,342문장으로, 적합률은 97.68%이고 적합 어절 수는 3,249,681개로 조사되었다. 오류 수는 전체 93,510문장 중 1,712문장으로 오류율은 1.83%, 보류 수는 456문장으로 보류율은 0.49%로 조사되었다.

표 3-5 수정 구축 1차 검수 데이터 통계 현황

주차	기간	총 검수 수 (문장)	검수율 (%)	검수 결과								총 어절 수 (개)
				적합			오류		보류		계	
				적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1주	6월 9일~ 6월 15일	25,407	27.17	24,880	491,763	97.93	426	1.68	101	0.40	25,407	504,242
2주	6월 16일~ 6월 22일	16,952	18.13	16,671	474,832	98.34	230	1.36	51	0.30	16,952	482,961
3주	6월 23일~ 6월 29일	14,255	15.24	13,966	482,995	97.97	234	1.64	55	0.39	14,255	494,032
4주	6월 30일~ 7월 6일	12,281	13.13	11,983	490,727	97.57	242	1.97	56	0.46	12,281	504,075
5주	7월 7일~ 7월 13일	10,331	11.05	10,066	475,554	97.43	201	1.95	64	0.62	10,331	488,533
6주	7월 14일~ 7월 20일	8,620	9.22	8,316	465,639	96.47	228	2.65	76	0.88	8,620	483,213
7주	7월 21일~ 7월 27일	5,664	6.06	5,460	368,171	96.40	151	2.67	53	0.94	5,664	382,183
합 계		**93,510**	**100**	**91,342**	**3,249,681**	**97.68**	**1,712**	**1.83**	**456**	**0.49**	**93,510**	**3,339,239**

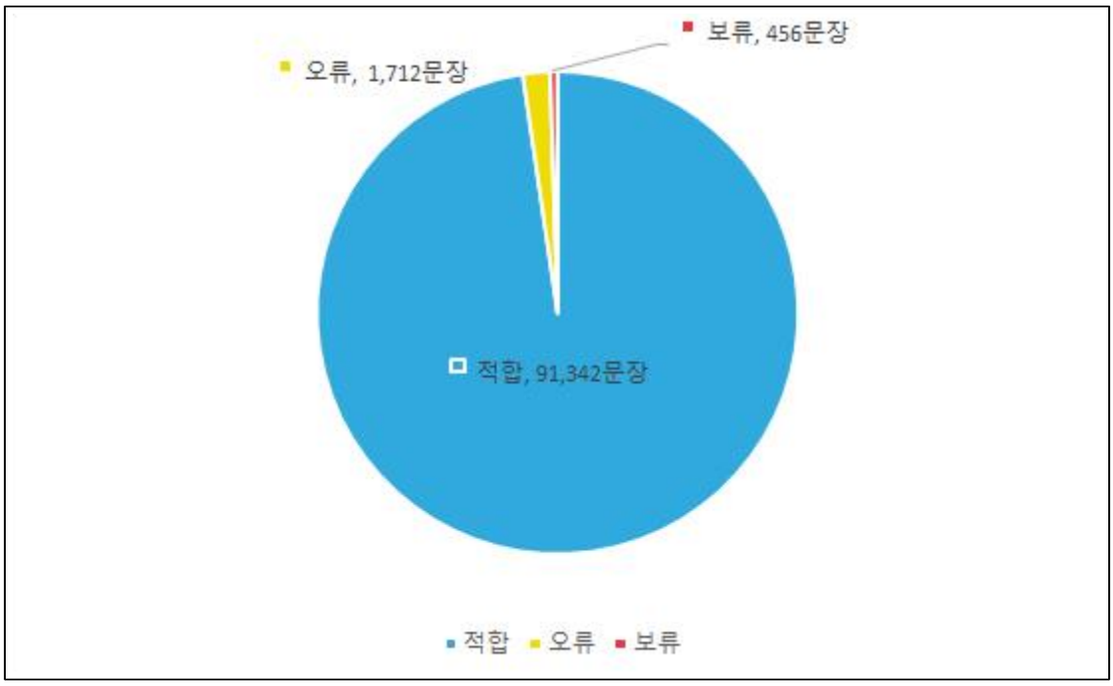


그림 3-6 수정 구축 1차 검수 적합/오류/보류 문장 수

그림 3-7 수정 구축 1차 검수 주차별 검수 문장 수 및 누적 검수율

그림 3-8 수정 구축 1차 검수 주차별 적합/오류/보류 문장 수

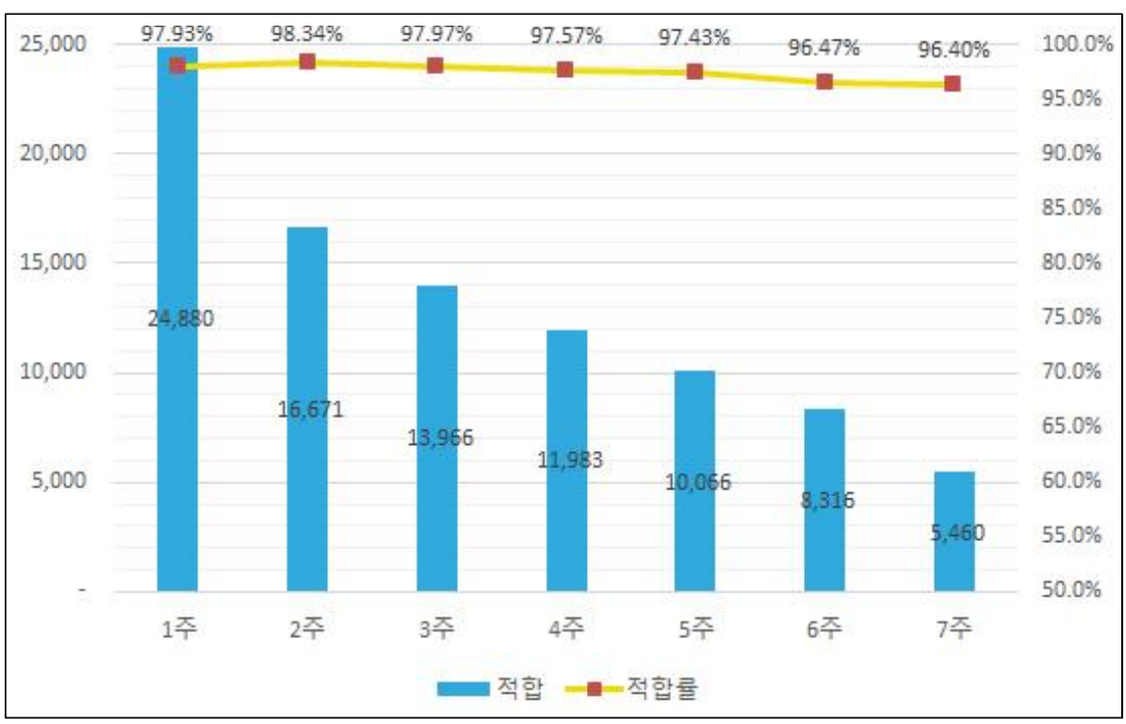


그림 3-9 수정 구축 1차 검수 주차별 적합 문장 수 및 적합률

3) 2차 검수 데이터 통계 현황

2차 검수 데이터 통계 현황을 조사하였다. 신규 구축과 수정 구축의 2차 검수 데이터 통계 현황은 〈표 3-6〉, 〈표 3-7〉과 같다.

신규 구축 데이터의 2차 검수 결과(1주~8주: 2025년 8월 4일~9월 21일 기준), 총 검수 수는 전체 84,000문장으로 누적 검수율은 100%로 조사되었다.

적합 수는 전체 84,000문장 중 80,224문장으로, 적합률은 95.50%이고 적합 어절 수는 1,608,091개로 조사되었다. 오류 수는 전체 84,000문장 중 3,356문장으로 오류율은 4.00%, 보류 수는 420문장으로 보류율은 0.50%로 조사되었다.

표 3-6 신규 구축 2차 검수 데이터 통계 현황

주차	기간	총 검수 수 (문장)	검수율 (%)	검수 결과							계	총 어절 수 (개)
				적합			오류		보류			
				적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1주	8월 4일~ 8월 10일	13,169	15.68	12,408	249,826	94.22	668	5.07	93	0.71	13,169	265,368
2주	8월 11일~ 8월 17일	13,515	16.10	12,951	257,931	95.83	484	3.58	80	0.59	13,515	269,673
3주	8월 18일~ 8월 24일	14,077	16.75	13,400	266,994	95.19	564	4.01	113	0.80	14,077	280,885
4주	8월 25일~ 8월 31일	13,527	16.11	12,945	260,942	95.70	545	4.03	37	0.27	13,527	272,698
5주	9월 1일~ 9월 7일	13,006	15.49	12,492	250,195	96.05	489	3.76	25	0.19	13,006	260,686
6주	9월 8일~ 9월 14일	10,288	12.25	9,911	197,813	96.34	360	3.50	17	0.17	10,288	205,487
7주	9월 15일~ 9월 21일	6,418	7.65	6,117	124,390	95.31	246	3.83	55	0.86	6,418	130,637
합 계		**84,000**	**100**	**80,224**	**1,608,091**	**95.50**	**3,356**	**4.00**	**420**	**0.50**	**84,000**	**1,685,434**

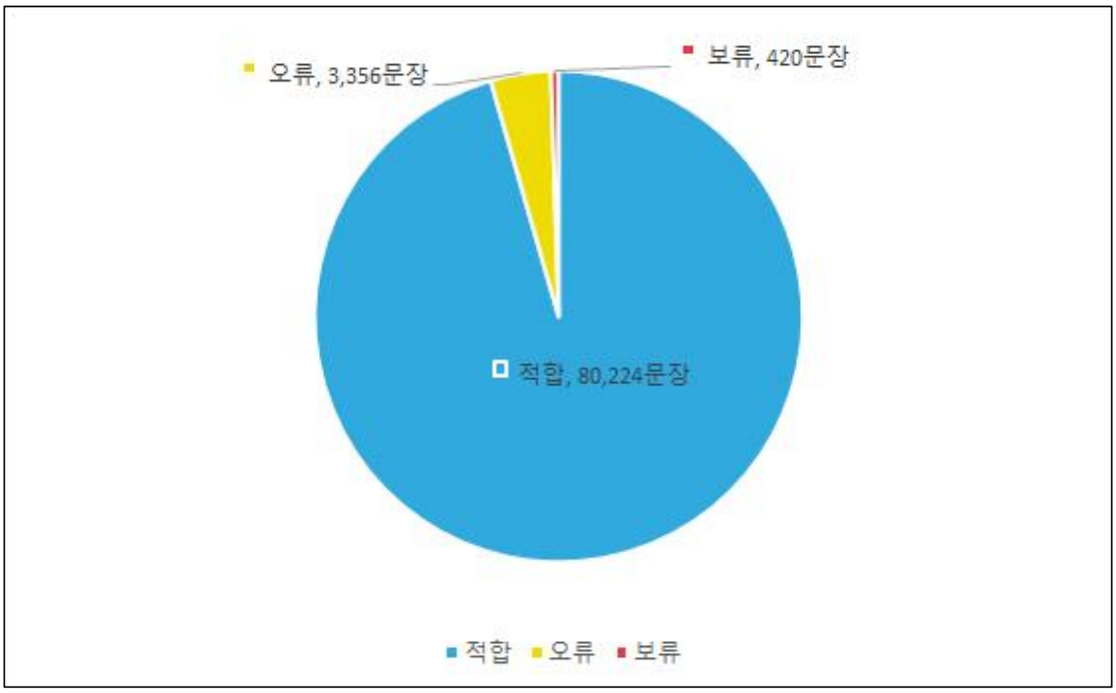


그림 3-10 신규 구축 2차 검수 적합/오류/보류 문장 수

그림 3-11 신규 구축 2차 검수 주차별 검수 문장 수 및 누적 검수율

그림 3-12 신규 구축 2차 검수 주차별 적합/오류/보류 문장 수

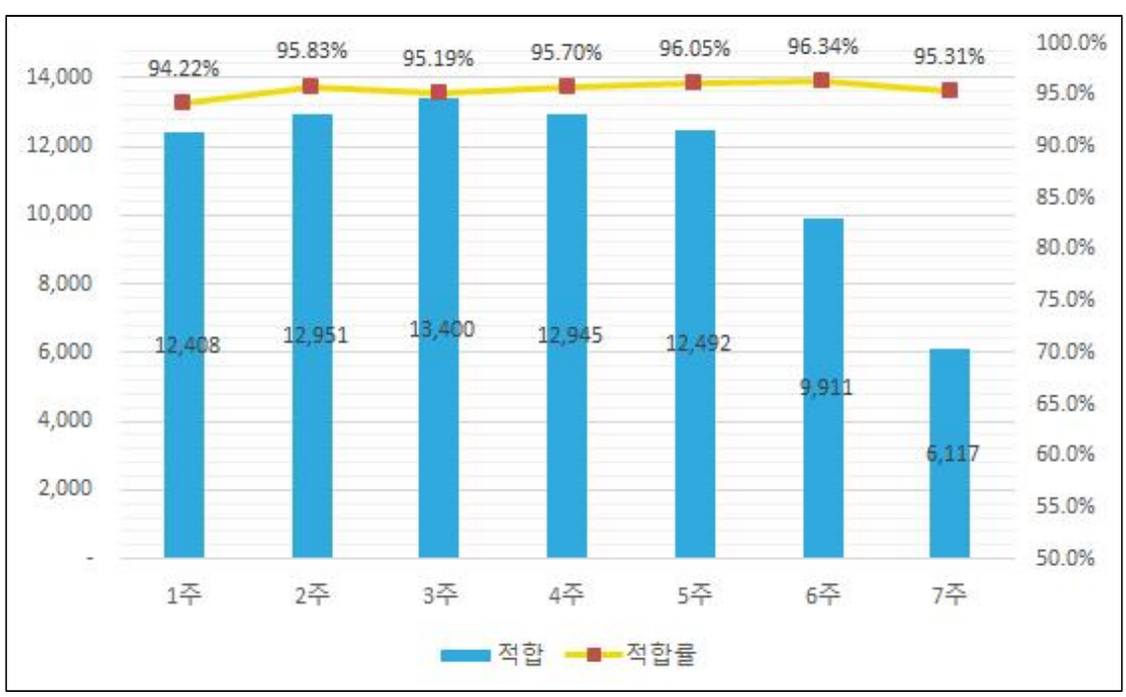


그림 3-13 신규 구축 2차 검수 주차별 적합 문장 수 및 적합률

수정 구축 데이터의 2차 검수 결과(1주~7주: 2025년 8월 4일~9월 21일 기준, 8주 작업 없음), 총 검수 수는 전체 93,510문장으로 누적 검수율은 100%로 조사되었다.

적합 수는 전체 93,510문장 중 91,418문장으로 적합률은 97.76%, 적합 어절 수는 3,251,659개로 조사되었다. 오류 수는 전체 93,510문장 중 1,691문장으로 오류율은 1.81%, 보류 수는 401문장으로 보류율은 0.43%로 조사되었다.

표 3-7 수정 구축 2차 검수 데이터 통계 현황

주차	기간	총 검수 수	검수율 (%)	검수 결과								총 어절 수
				적합			오류		보류		계	
				적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1주	8월 4일~ 8월 10일	25,126	26.87	24,625	479,778	98.01	421	1.68	80	0.32	25,126	491,708
2주	8월 11일~ 8월 17일	16,946	18.12	16,646	479,120	98.23	236	1.39	64	0.25	16,946	488,472
3주	8월 18일~ 8월 24일	14,629	15.64	14,341	499,361	98.03	242	1.65	46	0.18	14,629	510,701
4주	8월 25일~ 8월 31일	12,352	13.21	12,076	491,705	97.77	212	1.72	64	0.25	12,352	503,540
5주	9월 1일~ 9월 7일	10,897	11.65	10,620	512,917	97.46	215	1.97	62	0.25	10,897	526,966
6주	9월 8일~ 9월 14일	8,582	9.18	8,297	471,814	96.68	232	2.70	53	0.21	8,582	490,122
7주	9월 15일~ 9월 21일	4,978	5.32	4,813	316,964	96.69	133	2.67	32	0.13	4,978	327,730
합 계		**93,510**	**100**	**91,418**	**3,251,659**	**97.76**	**1,691**	**1.81**	**401**	**0.43**	**93,510**	**3,339,239**

그림 3-14 수정 구축 2차 검수 적합/오류/보류 문장 수

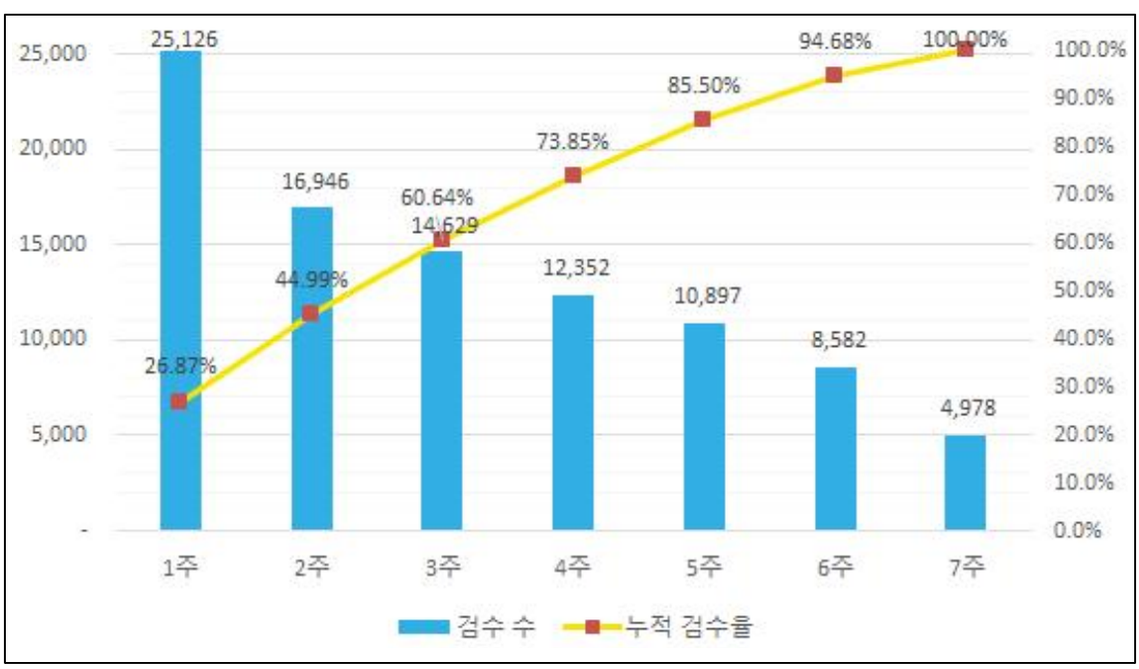


그림 3-15 수정 구축 2차 검수 주차별 검수 문장 수 및 누적 검수율

그림 3-16 수정 구축 2차 검수 주차별 적합/오류/보류 문장 수

그림 3-17 수정 구축 2차 검수 주차별 적합 문장 수 및 적합률

3) 오류/보류 점형 교정 데이터 통계 현황

신규 구축 1차 검수와 2차 검수에서 둘 다 적합 판정을 받은 문장 수를 조사한 결과, 총 79,381문장으로 적합률은 94.50%로 조사되었다. 1차 검수 및 2차 검수에서 한 번이라도 오류/보류 판정을 받은 문장은 총 4,619문장으로, 사업 수행 인력 3명이 교정 작업을 진행하였다. 또한, 1차 검수와 2차 검수에서 둘 다 적합 판정을 받은 문장 중에서도 교정이 필요한 문장 2,776문장을 선별하여 추가 교정하여 적합의 정확도를 높였다.

수정 구축 중에서 1차 검수와 2차 검수에서 둘 다 적합 판정을 받은 문장 수를 조사한 결과, 총 90,734문장으로 적합률은 97.03%로 조사되었다. 1차 검수 및 2차 검수에서 한 번이라도 오류/보류 판정을 받은 문장은 총 2,776문장으로, 사업 수행 인력 3명이 점형을 바르게 교정하여 적합 처리하였다. 1차 검수와 2차 검수에서 둘 다 적합 판정을 받은 문장 중에서도 교정이 필요한 문장 451문장을 선별하여 추가 교정하였다.

표 3-8 신규 구축과 수정 구축 교정 문장 수

내용	신규 구축	수정 구축 (검수자)	수정 구축 (사업 수행 인력)
① 전체 문장 수	84,000문장	91,023문장	2,487문장
② 1차 검수와 2차 검수에서 모두 적합 판정	79,381문장	88,447문장	2,287문장
③ 1차와 2차 검수에서 한 번이라도 오류/보류로 판정(①-②)	4,619문장	2,576문장	200문장
④ 적합 문장 표본 추출하여 정확성 재검토 * 추출: 특정 단어 검색(ex. SDS, ADPi 등)	2,776문장	388문장	63문장
교정 총 문장 수 (③+④)	7,395문장	2,964문장	263문장

신규 구축 데이터의 교정 데이터 구축 결과는 〈표 3-9〉와 같다. 교정 기간(1주~7주: 8월 11일~9월 28일 기준) 동안 작업한 총 교정 데이터는 전체 7,395문장이었다.

신규 구축 교정 결과, 총 7,395문장 중 6,923문장이 적합으로 판정되었고, 적합 어절 수는 141,060개였다. 오류 수는 전체 7,395문장 중 0문장으로 오류율은 0%이고, 보류 수는 472문장인 6.38%로 조사되었다.

표 3-9 신규 구축 교정 통계 현황

주차	기간	총 교정 수	검수 결과								총 어절 수
			적합			오류		보류		계	
			적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1주	8월 11일~ 8월 17일	0	0	0	0	0	0	0	0	0	0
2주	8월 18일~ 8월 24일	564	502	10,251	89.01	0	0	62	10.99	564	11,552
3주	8월 25일~ 8월 31일	911	847	17,362	92.97	0	0	64	7.03	911	18,694
4주	9월 1일~ 9월 7일	745	675	13,850	90.60	0	0	70	9.40	745	15,313
5주	9월 8일~ 9월 14일	847	691	13,904	81.58	0	0	156	18.42	847	17,162
6주	9월 15일~ 9월 21일	892	855	17,495	95.85	0	0	37	4.15	892	18,289
7주	9월 22일~ 9월 28일	3,436	3,353	68,198	97.58	0	0	83	2.42	3,436	69,904
합 계		**7,395**	**6,923**	**141,060**	**93.62**	**0**	**0**	**472**	**6.38**	**7,395**	**150,914**

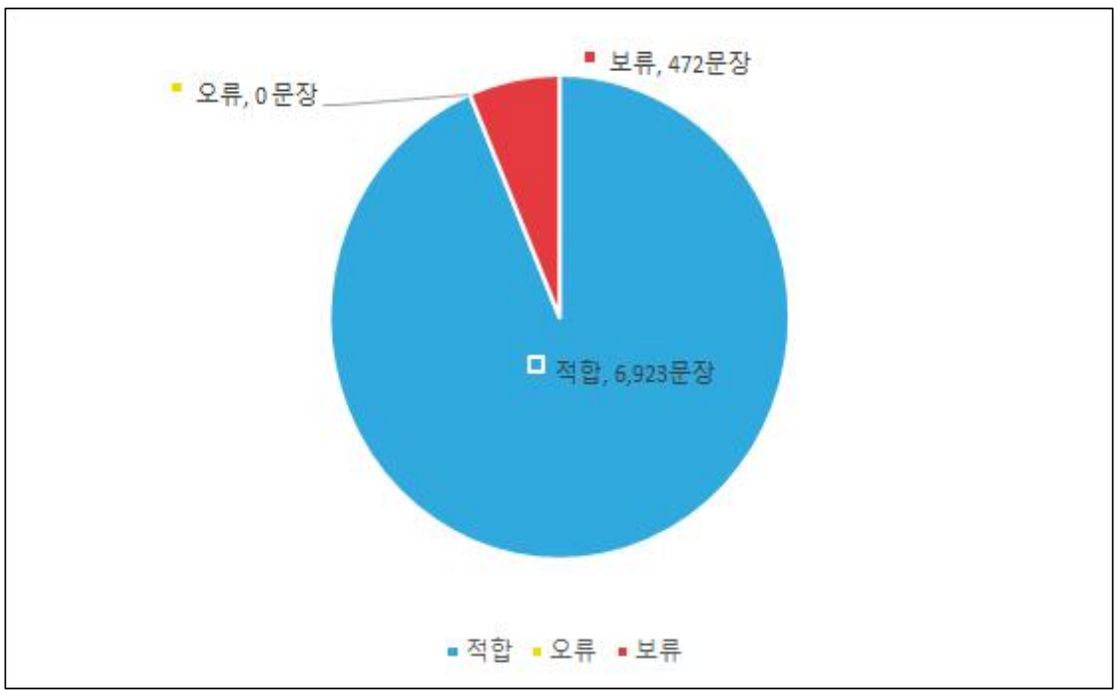


그림 3-18 신규 구축 교정 적합/오류/보류 문장 수

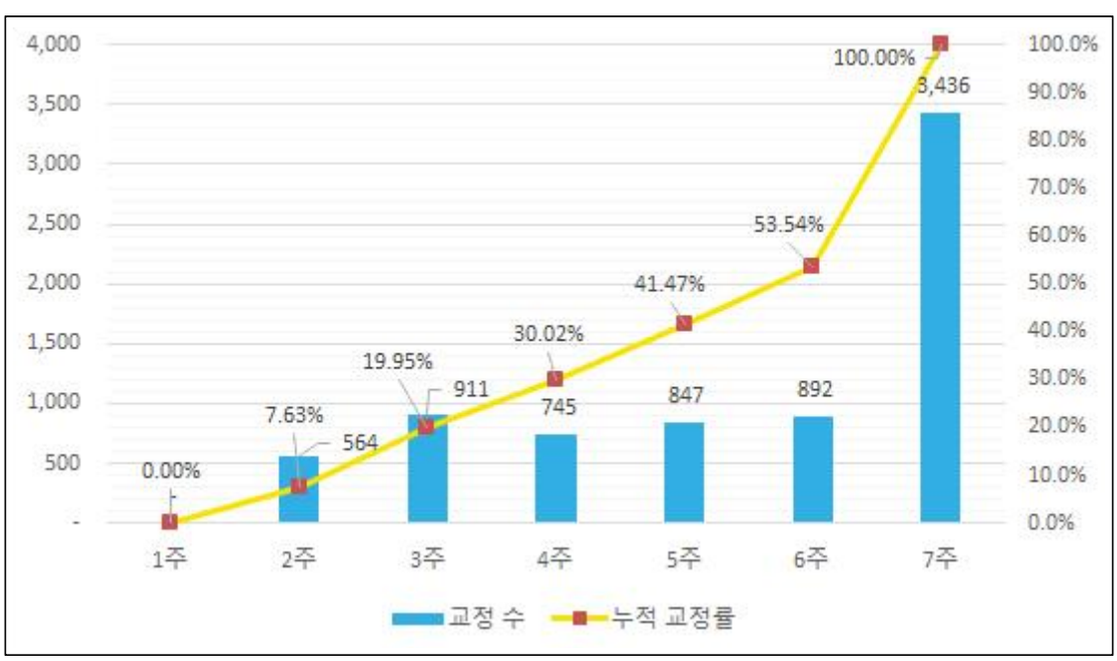


그림 3-19 신규 구축 교정 주차별 교정 문장 수 및 누적 교정률

그림 3-20 신규 구축 교정 주차별 적합/오류/보류 문장 수

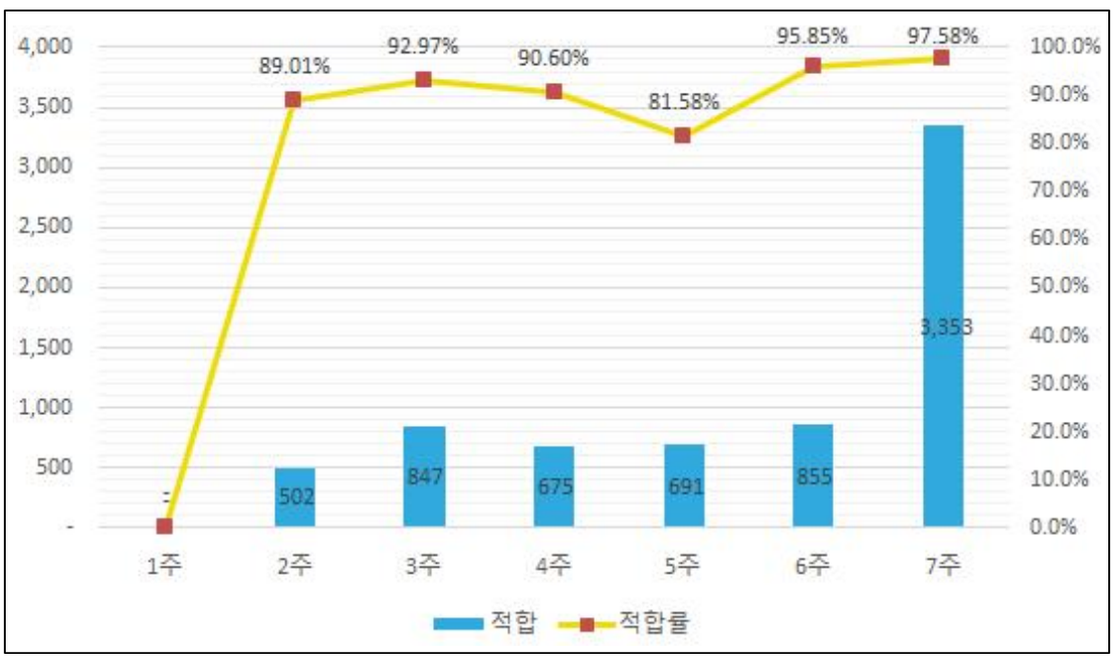


그림 3-21 신규 구축 교정 주차별 적합 문장 수 및 적합률

수정 구축 데이터의 교정 데이터 구축 결과는 〈표 3-10〉과 같다. 교정 결과(1주~7주: 8월 11일~9월 28일 기준), 총 교정 수는 전체 3,227문장이었다.

적합 수는 전체 3,227문장 중 2,832문장으로 적합률은 87.76%로 조사되었다. 적합 어절 수는 115,023개이다. 오류 수는 전체 3,227문장 중 0문장으로 오류율은 0%이고, 보류 수는 395문장인 12.24%로 조사되었다.

표 3-10 수정 구축 교정 통계 현황

주차	기간	총 검수 수	검수 결과								총 어절 수
			적합			오류		보류		계	
			적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1주	8월 11일~ 8월 17일	0	0	0	0	0	0	0	0	0	0
2주	8월 18일~ 8월 24일	544	512	11,796	94.12	0	0	32	5.88	544	7,902
3주	8월 25일~ 8월 31일	403	376	10,952	93.30	0	0	27	6.70	403	9,321
4주	9월 1일~ 9월 7일	341	330	13,041	96.77	0	0	11	3.23	341	9,785
5주	9월 8일~ 9월 14일	348	311	12,700	89.37	0	0	37	10.63	348	12,154
6주	9월 15일~ 9월 21일	667	439	22,628	65.82	0	0	228	34.18	667	27,205
7주	9월 22일~ 9월 28일	924	864	43,906	93.51	0	0	60	6.49	924	38,458
합 계		**3,227**	**2,832**	**115,023**	**87.76**	**0**	**0**	**395**	**12.24**	**3,227**	**104,825**

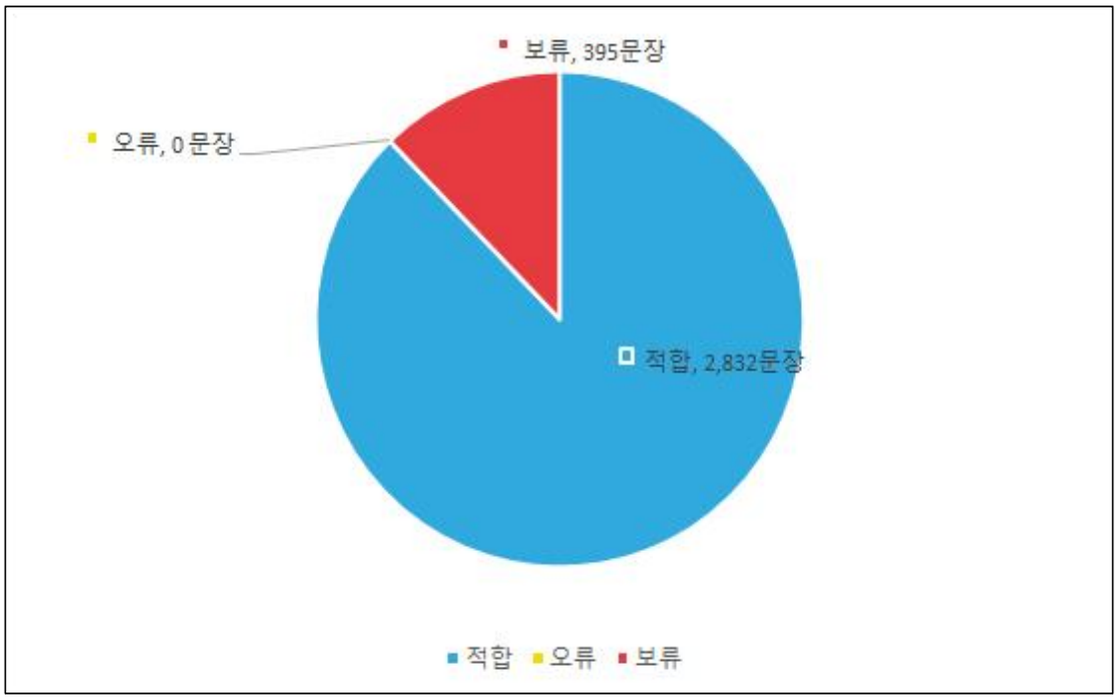


그림 3-22 수정 구축 교정 적합/오류/보류 문장 수

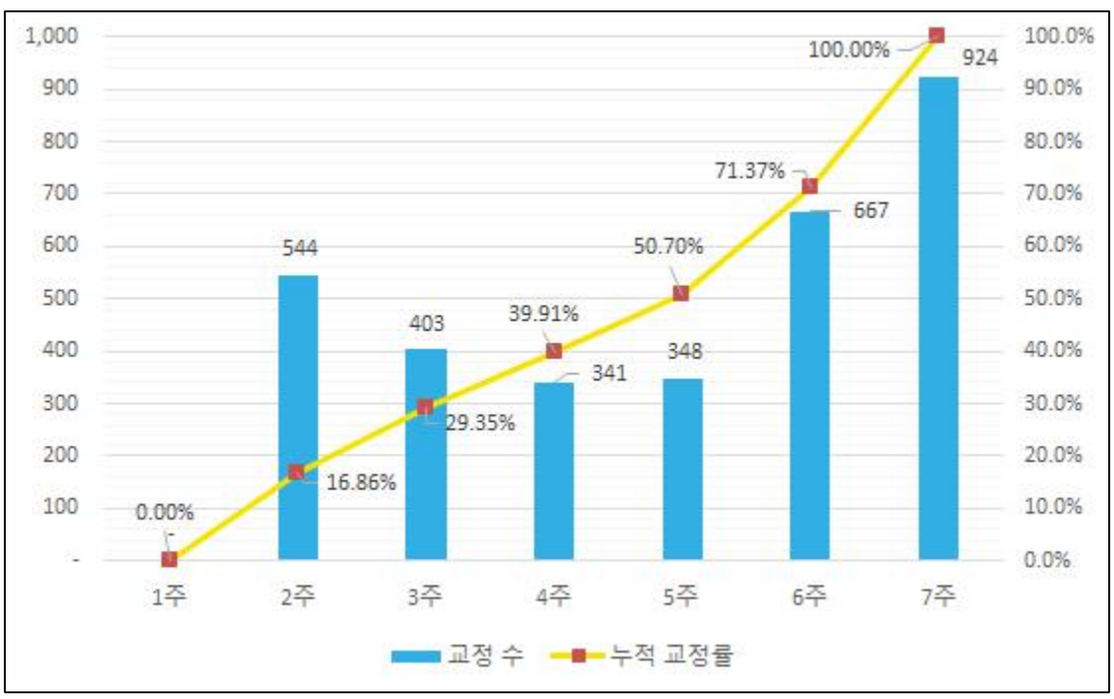


그림 3-23 수정 구축 교정 주차별 교정 문장 수 및 누적 교정률

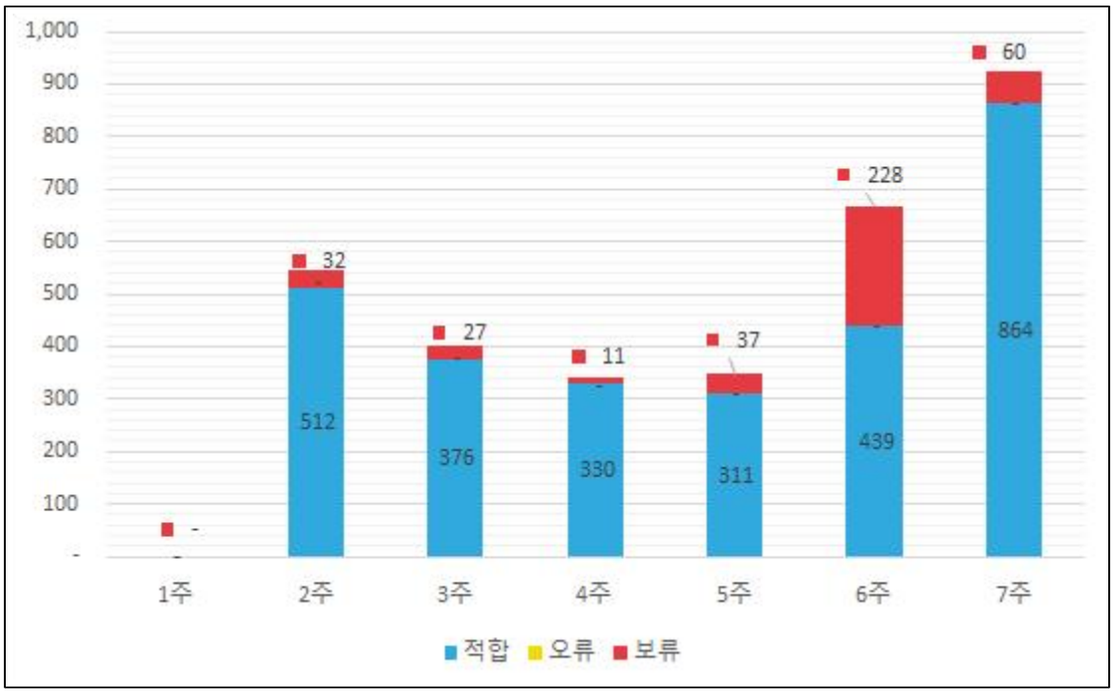


그림 3-24 수정 구축 교정 주차별 적합/오류/보류 문장 수

그림 3-25 수정 구축 교정 주차별 적합 문장 수 및 적합률

4) 최종 데이터 통계 현황

검수와 교정 과정을 거친 최종 데이터 통계 현황을 정리하였다. 최종 데이터 통계 현황을 제시하면 〈표 3-11〉과 같다.

신규 구축 데이터의 1차 검수 결과(총 8주), 총 검수 수는 전체 84,000문장이고, 이 중 적합 수는 80,347문장으로 적합률은 95.65%로 조사되었다. 적합 어절 수는 1,610,647개이다. 오류 수는 3,155문장으로 오류율은 3.76%로 조사되었고, 보류 수는 498문장인 0.59%로 조사되었다.

신규 구축 데이터의 2차 검수 결과(총 7주), 총 검수 수는 전체 84,000문장이고, 이 중 적합 수는 80,224문장으로 적합률은 95.50%로 조사되었다. 적합 어절 수는 1,608,091개이다. 오류 수는 3,356문장으로 오류율은 4.00%로 조사되었고, 보류 수는 420문장인 0.50%로 조사되었다.

신규 구축 1차 검수 및 2차 검수에서 한 번이라도 오류/보류 판정을 받은 데이터를 사업 수행 인력이 총 7주에 걸쳐서 교정한 결과, 총 교정 수는 전체 7,395문장으로, 이 중 적합 수는 6,923문장으로 적합률은 93.62%가 되었고 적합 어절 수는 141,060개였다. 오류 수는 0문장, 보류 수는 472문장으로 조사되었다.

오류 수가 0문장, 보류 수가 472문장으로 교정되어, 신규 구축 84,000문장 중에서 적합 문장 수는 83,528문장으로 확정되었다. 따라서 실제 말뭉치에 포함된 최종 적합 문장 수는 83,528문장(1,675,580어절)이 되었다.

표 3-11 신규 구축 최종 데이터 통계 현황

구분	총 검수 수	검수 결과								총 어절 수
		적합			오류		보류		계	
		적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1차 검수	84,000	80,347	1,610,647	95.65	3,155	3.76	498	0.59	84,000	1,685,434
2차 검수	84,000	80,224	1,608,091	95.50	3,356	4.00	420	0.50	84,000	1,685,434
교 정	7,395	6,923	141,060	93.62	0	0	472	6.38	7,395	150,914
최 종	**84,000**	**83,528**	**1,675,580**	**99.44**	**0**	**0**	**472**	**0.56**	**84,000**	**1,685,434**

수정 구축 데이터의 1차 검수 결과(총 7주), 총 검수 수는 전체 93,510문장이고, 적합 수는 91,342문장으로 적합률은 97.68%로 조사되었다. 적합 어절 수는 3,249,681개이다. 오류 수는 1,712문장으로 오류율은 1.83%로 조사되었고, 보류 수는 456문장인 0.49%로 조사되었다.

수정 구축 데이터의 2차 검수 결과(총 7주), 총 검수 수는 전체 93,510문장이고, 적합 수는 91,418문장으로 적합률은 97.76%로 조사되었다. 적합 어절 수는 3,251,659개이다. 오류 수는 1,691문장으로 오류율은 1.81%로 조사되었고, 보류 수는 401문장인 0.43%로 조사되었다.

수정 구축 1차 검수 및 2차 검수에서 한 번이라도 오류/보류 판정을 받은 데이터를 사업 수행 인력이 총 7주에 걸쳐서 교정한 결과, 총 교정 수는 전체 3,227문장으로, 이 중 적합 수는 2,832문장으로 적합률은 87.76%가 되었고 적합 어절 수는 115,023개였다. 오류 수는 0문장으로 오류율은 0%이고, 보류 수는 395문장인 12.24%로 조사되었다.

오류 수가 0문장, 보류 수가 395문장으로 교정되어, 수정 구축 93,510문장 중에서 적합 문장 수는 93,115문장이 확정되었다. 따라서 실제 말뭉치에 포함된 최종 적합 문장 수는 93,115문장(3,322,242어절)이 되었다.

표 3-12 수정 구축 최종 데이터 통계 현황

구분	총 검수 수	검수 결과								총 어절 수
		적합			오류		보류		계	
		적합	어절 수	적합률(%)	오류	오류율(%)	보류	보류율(%)		
1차 검수	93,510	91,342	3,249,681	97.68	1,712	1.83	456	0.49	93,510	3,339,239
2차 검수	93,510	91,418	3,251,659	97.76	1,691	1.81	401	0.43	93,510	3,339,239
교 정	3,227	2,832	115,023	87.76	0	0	395	12.24	3,227	131,430
최 종	**93,510**	**93,115**	**3,322,242**	**99.58**	**0**	**0**	**395**	**0.01**	**93,510**	**3,339,239**

제 4 장

묵자-점자 병렬 말뭉치 적합/오류/보류 사례

1. 오류/보류 주요 유형
2. 적합 사례(예시)
3. 데이터 주요 오류/보류 사례
4. 데이터 주요 오류/보류 처리

1 오류/보류 주요 유형

신규 구축 데이터와 수정 구축 데이터에서 나타난 오류/보류 주요 유형을 조사하였다. 1차 검수와 2차 검수 데이터 중 오류/보류 주요 유형은 〈표 4-1〉과 같다.

신규 구축 데이터의 경우, 1차 검수(총 8주간 진행) 데이터의 오류 수는 3,155문장, 보류 수는 498문장으로 총 오류/보류 수는 3,653문장이었다. 오류/보류 주요 유형으로는 문장부호 오류가 1,240건으로 가장 많으며 전체 오류/보류의 33.94%를 차지하였다. 그다음으로 기타 오류가 767건으로 21%, 알파벳 약자/약어 오류가 560건으로 15.33%를 차지하였다.

2차 검수(총 7주간 진행) 데이터의 오류 수는 3,356문장, 보류 수는 420문장으로 총 오류/보류 수는 3,776문장이었다. 오류/보류 주요 유형으로는 문장부호 오류가 1,395건으로 가장 많으며 전체 오류/보류의 36.94%를 차지하였다. 그다음으로 기타 오류가 741건으로 19.62%, 알파벳 약자/약어 오류가 553건으로 14.65%를 차지하였다.

수정 구축 데이터의 경우, 1차 검수(총 7주간 진행) 데이터의 오류 수는 1,712문장, 보류 수는 456문장으로 총 오류/보류 수는 2,168문장이다. 오류/보류 주요 유형으로는 문장부호 오류가 568건으로 가장 많으며 전체 오류/보류의 26.20%를 차지하였다. 그다음으로 기타 오류가 552건으로 25.46%, 알파벳 약자/약어 오류가 334건으로 15.40%를 차지하였다.

2차 검수(총 7주간 진행) 데이터의 오류 수는 1,691문장, 보류 수는 401문장으로 총 오류/보류 수는 2,092문장이다. 오류/보류 주요 유형으로는 문장부호 오류가 592건으로 가장 많으며 전체 오류/보류의 28.30%를 차지하였다. 그다음으로 기타 오류가 452건으로 21.61%, 알파벳 약자/약어 오류가 334건으로 15.97%를 차지하였다.

표 4-1 데이터 오류/보류 주요 유형 현황

오류/보류 사유	신규 구축				수정 구축			
	1차 검수		2차 검수		1차 검수		2차 검수	
	건수 (개)	비율 (%)	건수 (개)	비율 (%)	건수 (개)	비율 (%)	건수 (개)	비율 (%)
1급 점자표 오류	47	1.29	38	1.01	43	1.98	39	1.86
알파벳 약자/약어 오류	560	15.33	553	14.65	334	15.40	334	15.97
문장부호 오류	1,240	33.94	1,395	36.94	568	26.20	592	28.30
단위/연산기호 오류	379	10.38	401	10.62	193	8.90	189	9.03
영어 대문자 표기 오류	454	12.43	431	11.41	214	9.87	225	10.76
제2외국어/한자 표기	20	0.55	28	0.74	117	5.40	115	5.50
로마자 시작표/종료표 오류	186	5.09	189	5.01	147	6.78	146	6.98
기 타	767	21.00	741	19.62	552	25.46	452	21.61
합 계	**3,653**	**100**	**3,776**	**100**	**2,168**	**100**	**2,092**	**100**

표 4-2 데이터 오류/보류 주요 유형 순위

순위	신규 구축		수정 구축	
	오류/보류 사유	비율(%)	오류/보류 사유	비율(%)
1	문장부호 오류	35.47	문장부호 오류	27.23
2	기타	20.30	기타	23.57
3	알파벳 약자/약어 오류	14.98	알파벳 약자/약어 오류	15.68
4	영어 대문자 표기 오류	11.91	영어 대문자 표기 오류	10.30
5	단위/연산기호 오류	10.50	단위/연산기호 오류	8.97
6	로마자 시작표/종료표 오류	5.05	로마자 시작표/종료표 오류	6.88
7	1급 점자표 오류	1.14	제2외국어/한자 표기	5.45
8	제2외국어/한자 표기	0.65	1급 점자표 오류	1.92

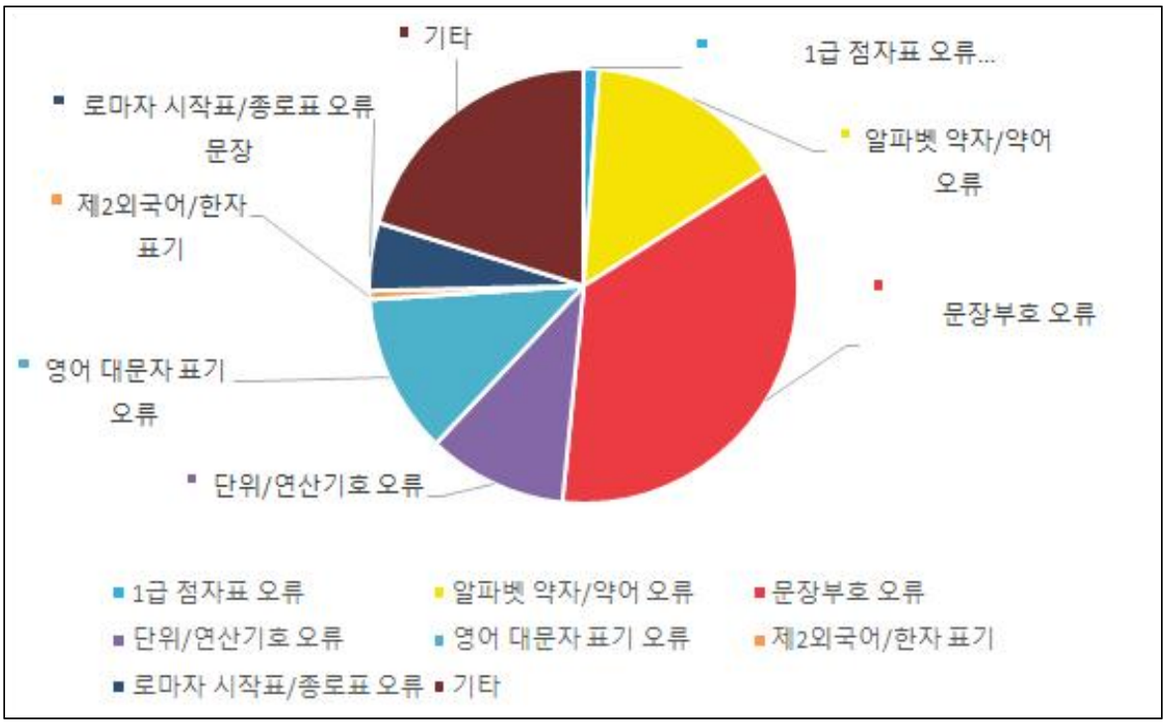


그림 4-1 신규 구축 검수 데이터 오류/보류 주요 유형

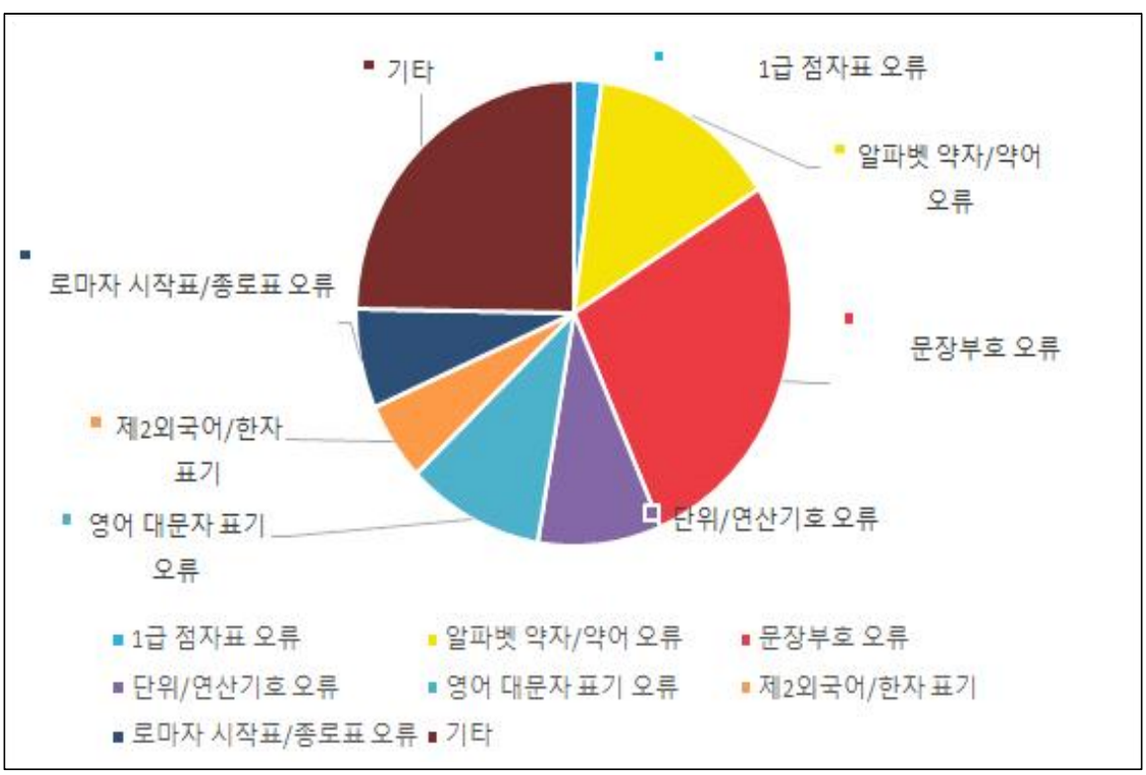


그림 4-2 수정 구축 데이터 오류/보류 주요 유형

2 적합 사례(예시)

1) 시나리오 1: 한글 + (영어 + ... +)

- 한글 뒤에 괄호가 있고, 그 속에 영어 단어가 포함되어 있는 문장
- 점자에서 괄호는 한국 점자 규정에 따라야 하며, 영어 단어 앞에 로마자 시작표가 있는지 확인해야 함.

표 4-3 시나리오 1(한글+(영어)) 적합 사례 ①

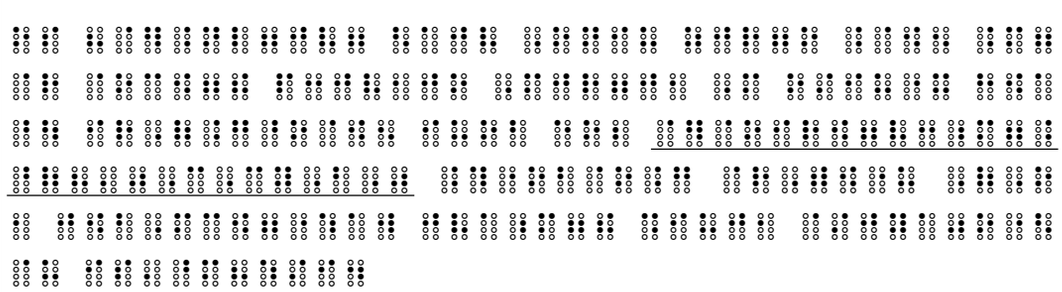
시나리오 1(한글+(영어))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전국 종합지
Metadata	* title: 경향신문 2023년 기사 * author(저자): 박채영 기자 * publisher(출판사): 경향신문 * date(작성 연도): 20230317 * original_topic(주제): 경제|금융·재테크
Form	이어 “경제상황 악화 시에도 은행이 자금 중개 기능을 원활히 수행할 수 있도록 특별대손준비금 도입 및 경기대응완충자본(CCyB) 적립기준 개선으로 손실 흡수능력의 확충을 유도할 계획”이라고 덧붙였다.
Braille	

표 4-4 시나리오 1(한글+(영어)) 적합 사례 ②

시나리오 1(한글+(영어))

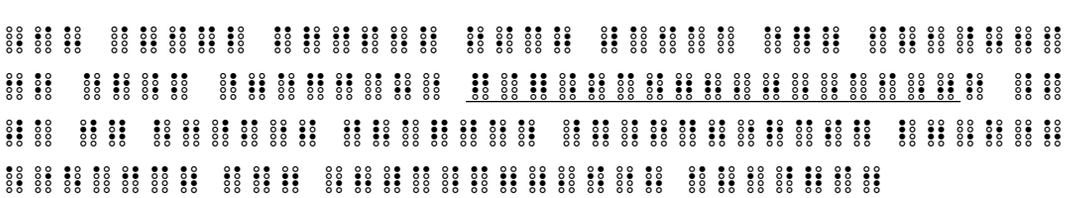
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전국 종합지
Metadata	* title: 경향신문 2023년 기사 * author(저자): 김보미 기자 * publisher(출판사): 경향신문 * date(작성 연도): 20230317 * original_topic(주제): 지역|서울
Form	또 고독사 위험이 있는 1911명은 돌봄단이 매주 방문하고, 인공지능(AI)이 주 1회 안부를 확인하며 복지플래너가 상시로 연락하는 등 '3중'으로 돌본다.
Braille	

표 4-5 시나리오 1(한글+(영어)) 적합 사례 ③

시나리오 1(한글+(영어))

Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전국 종합지
Metadata	* title: 경향신문 2023년 기사 * author(저자): 박은경 기자 * publisher(출판사): 경향신문 * date(작성 연도): 20230314 * original_topic(주제): 정치|국방·외교
Form	북한이 14일 동해상으로 단거리탄도미사일(SRBM) 2발을 발사했다.
Braille	

2) 시나리오 2: 영어 + (한글 + ... +)

- 영어 단어 뒤에 괄호가 있고, 그 속에 한글이 포함되어 있는 문장

표 4-6 시나리오 2(영어+(한글)) 적합 사례 ①

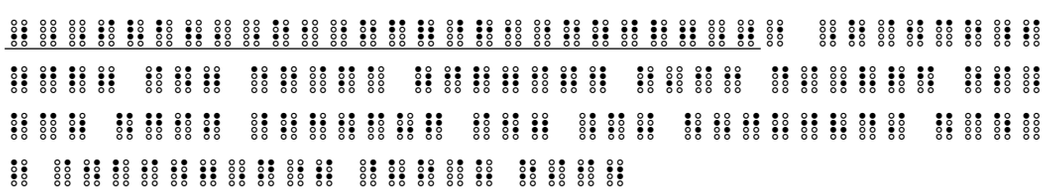
시나리오 2(영어+(한글))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전국 종합지
Metadata	* title: 경향신문 2023년 기사 * author(저자): 이재덕 기자 * publisher(출판사): 경향신문 * date(작성 연도): 20230314 * original_topic(주제): 경제|경제 일반
Form	SVB(실리콘밸리은행), 시그니처은행 등 미국 은행들의 잇단 폐쇄가 이어지는 가운데 비트코인 등 주요 암호화폐 가격이 급등세를 보이고 있다.
Braille	

표 4-7 시나리오 2(영어+(한글)) 적합 사례 ②

시나리오 2(영어+(한글))

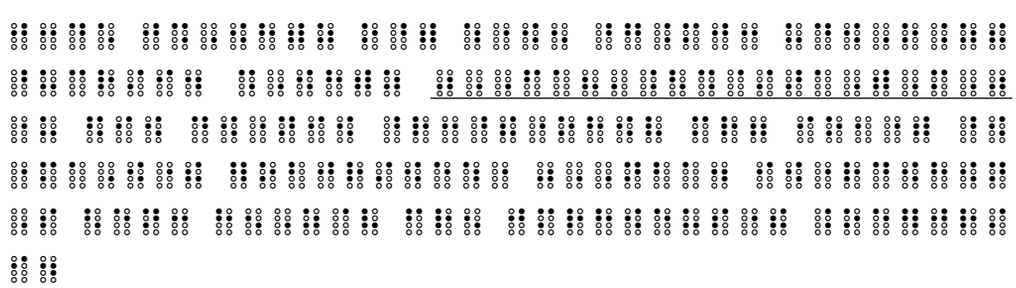
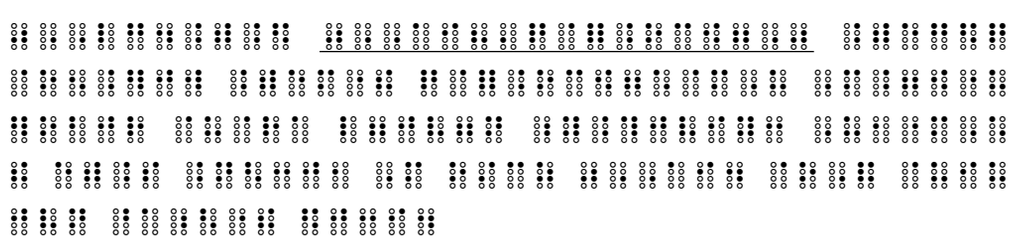
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전문지
Metadata	* title: 스포츠경향 2023년 기사 * author(저자): 김하영 기자 * publisher(출판사): 스포츠경향 * date(작성 연도): 20231018 * original_topic(주제): 스포츠종합|농구
Form	한편 휴스턴은 이번 여름 분주한 오프시즌을 보냈다. 휴스턴은 FA(자유계약선수)로 가드 프레드 밴블리트와 윙 딜런 브룩스를 영입했으며, NBA 드래프트에서 아멘 탐슨과 캠 휘트모어를 지명했다.
Braille	

표 4-8 시나리오 2(영어+(한글)) 적합 사례 ③

시나리오 2(영어+(한글))

Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 인터넷 기반 신문
Metadata	* title: 아이뉴스24 2023년 기사 * author(저자): 장유미 기자 * publisher(출판사): 아이뉴스24 * date(작성 연도): 20230101 * original_topic(주제): 경제|산업/재계
Form	LG전자가 AI(인공지능) 전문가인 김정희 전무를 인공지능연구소 수장으로 영입해 고객 상황에 최적화된 솔루션을 먼저 제안할 수 있는 AI 기술 고도화에 박차를 가한다.
Braille	

3) 시나리오 3: 한글 + (숫자 + ... +)

- 한글 뒤에 괄호가 있고, 그 속에 숫자가 포함되어 있는 문장

표 4-9 시나리오 3(한글+(숫자)) 적합 사례 ①

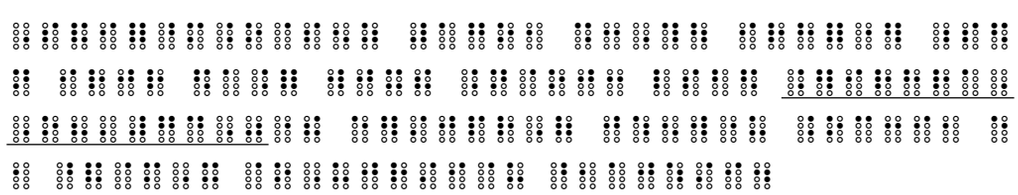
시나리오 3(한글+(숫자))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전국 종합지
Metadata	* title: 경향신문 2023년 기사 * author(저자): 김송이 기자 * publisher(출판사): 경향신문 * date(작성 연도): 20230317 * original_topic(주제): 사회|사회 일반
Form	서울동부지검은 17일 윤석열 대통령 처가에 대해 각종 의혹을 제기한 사업가 정대택씨(74)를 명예훼손 혐의로 지난달 말 불구속 기소했다고 밝혔다.
Braille	

표 4-10 시나리오 3(한글+(숫자)) 적합 사례 ②

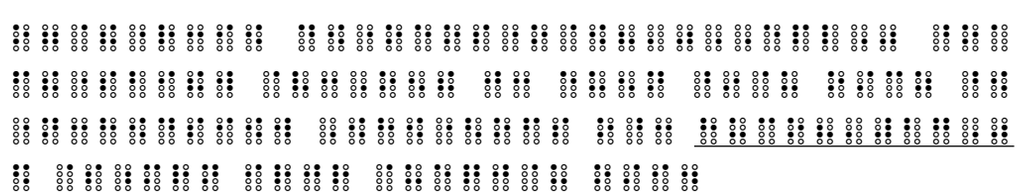
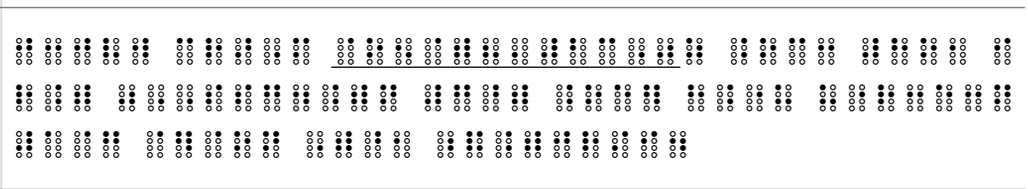
시나리오 3(한글+(숫자))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전문지
Metadata	* title: 스포츠경향 2023년 기사 * author(저자): 박찬기 온라인기자 * publisher(출판사): 스포츠경향 * date(작성 연도): 20231120 * original_topic(주제): 축구|해외축구
Form	잉글랜드 프리미어리그(EPL) 빅클럽들의 관심을 한 몸에 받고 있는 브렌트퍼드의 스트라이커 이반 토니(27)가 거취에 대해 고민하고 있다.
Braille	

표 4-11 시나리오 3(한글+(숫자)) 적합 사례 ③

시나리오 3(한글+(숫자))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전문지
Metadata	* title: 스포츠경향 2023년 기사 * author(저자): 사직=배재흥 기자 * publisher(출판사): 스포츠경향 * date(작성 연도): 20230714 * original_topic(주제): 야구|국내야구
Form	한화의 내야수 김건(23)은 지난 8일 대전 SSG전에 7번 지명 타자로 오랜만에 1군 경기에 선발 출전했다.
Braille	

4) 시나리오 4: 숫자 + (한글 + ... +)

- 숫자 뒤에 괄호가 있고, 그 속에 한글이 포함되어 있는 문장

표 4-12 시나리오 4(숫자+(한글)) 적합 사례 ①

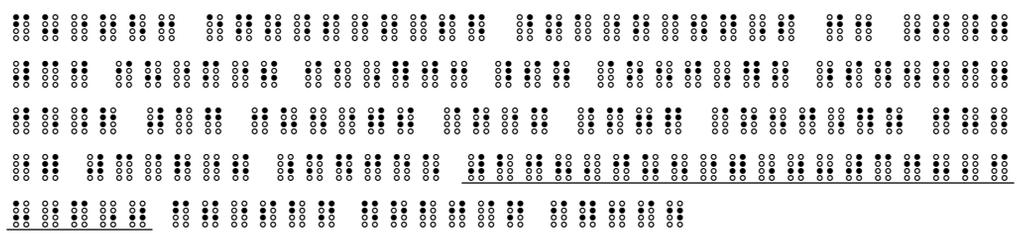
시나리오 4(숫자+(한글))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전문지
Metadata	* title: 스포츠투데이 2023년 기사 * author(저자): 이상필 기자 * publisher(출판사): 스포츠투데이 * date(작성 연도): 20230904 * original_topic(주제): 스포츠|해외야구
Form	코리안 메이저리거가 빅리그에서 한 시즌 30도루를 달성한 것은 김하성이 처음이다. 현재 17홈런을 기록 중인 김하성은 홈런 3개를 추가하면 20(홈런)-30(도루) 클럽에 가입하게 된다.
Braille	

표 4-13 시나리오 4(숫자+(한글)) 적합 사례 ②

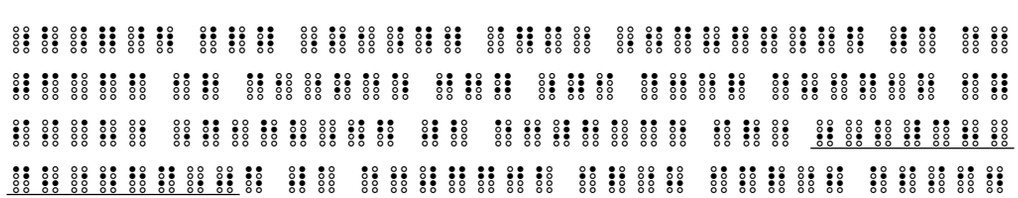
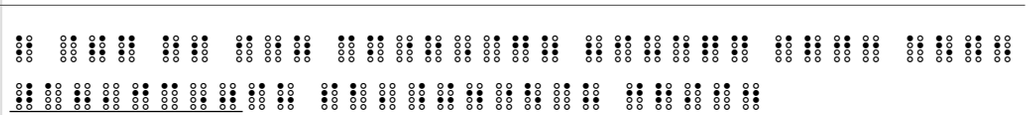
시나리오 4(숫자+(한글))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 인터넷 기반 신문
Metadata	* title: 아이뉴스24 2023년 기사 * author(저자): 김종성 기자 * publisher(출판사): 아이뉴스24 * date(작성 연도): 20230901 * original_topic(주제): 경제|산업/재계
Form	차종별 해외 실적의 경우, 스포티지가 3만6166대 팔리며 해외 최다 판매 모델이 되었고, 셀토스가 2만8135대, K3(포르테)가 1만7702대로 뒤를 이었다.
Braille	

표 4-14 시나리오 4(숫자+(한글)) 적합 사례 ③

시나리오 4(숫자+(한글))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 지역 종합지
Metadata	* title: 충청매일 2023년 기사 * author(저자): 최영덕 기자 * publisher(출판사): 충청매일 * date(작성 연도): 20230719 * original_topic(주제): 정치
Form	이 글에 이어 다른 누리꾼은 “도민에 대한 마음 1(하나)도 없죠.”라고 했다.
Braille	

5) 시나리오 5: 영어 + (숫자 + ... +)

- 영어 뒤에 괄호가 있고, 그 속에 숫자가 포함되어 있는 문장

표 4-15 시나리오 5(영어+(숫자)) 적합 사례 ①

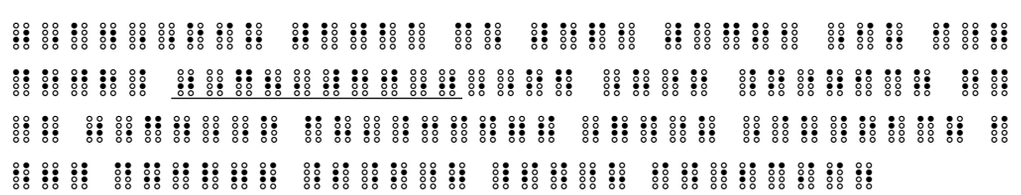
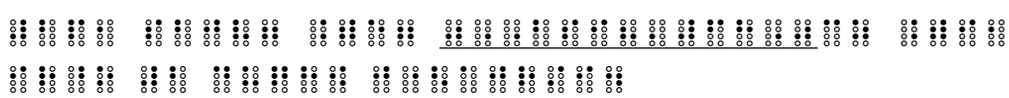
시나리오 5(영어+(숫자))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전문지
Metadata	* title: 매일경제 2023년 기사 * author(저자): 백지연 매경닷컴 기자 * publisher(출판사): 매일경제 * date(작성 연도): 20231230 * original_topic(주제): 사회
Form	B씨도 2021년 5월 17일 또 다른 피해자 G(80)씨가 소리 지른다는 이유로 G씨 콧잔등을 손으로 꼬집는 등의 폭행을 저지른 것으로 드러났다.
Braille	

표 4-16 시나리오 5(영어+(숫자)) 적합 사례 ②

시나리오 5(영어+(숫자))	
Category	국립국어원 묵자-점자 병렬 말뭉치 / 신문 / 전국 종합지
Metadata	* title: 경향신문 2023년 기사 * author(저자): 최희진 기자 반기웅 기자 * publisher(출판사): 경향신문 * date(작성 연도): 20230127 * original_topic(주제): 경제|경제 일반
Form	2월 업황 전망 BSI(68)는 전달 대비 2포인트 하락했다.
Braille	

3 데이터 주요 오류/보류 사례

1) 1급 점자표 오류

표 4-17 축어와 혼동될 수 있는 단어의 1급 점자표 오류 사례

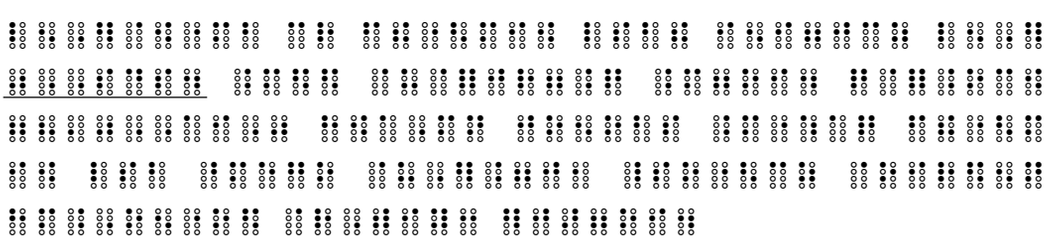
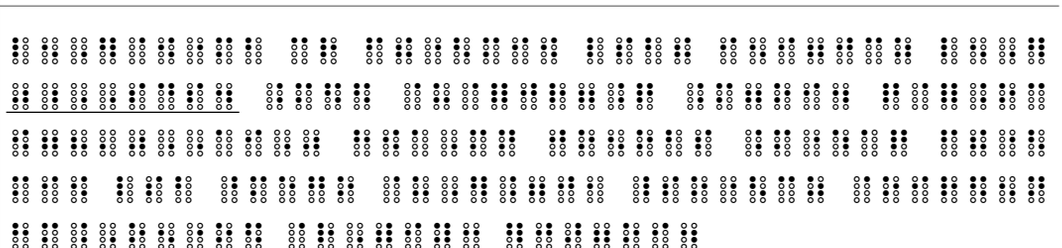
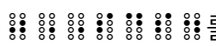
시나리오 1(한글+(영어))	
Form	삼성그룹 내 클라우드 사업을 담당하는 삼성SDS 주가가 고공행진 중이다. 인공지능(AI) 특수에 힘입어 주력인 클라우드 사업 부문이 고성장할 것이라는 기대감에 투자심리가 개선된 영향이다.
Braille	
오점역 설명	축어 said를 사용한 saids와 혼동이 될 수 있는 SDS에 1급 점자표를 사용하지 않음. 
Braille 수정	
수정 설명	 를 로 수정함.

표 4-18 알파벳 ‘GTX-D’ 1급 점자표 오류 사례

시나리오 1(한글+(영어))

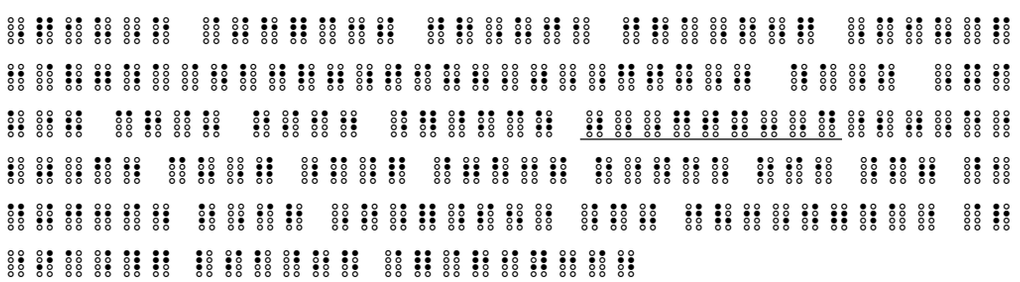
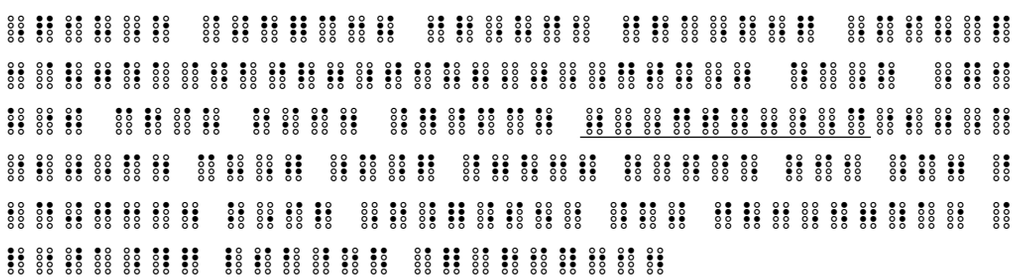
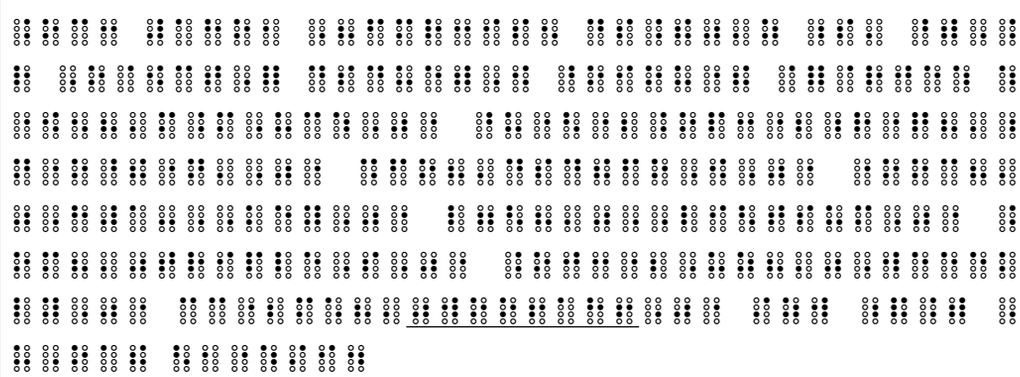
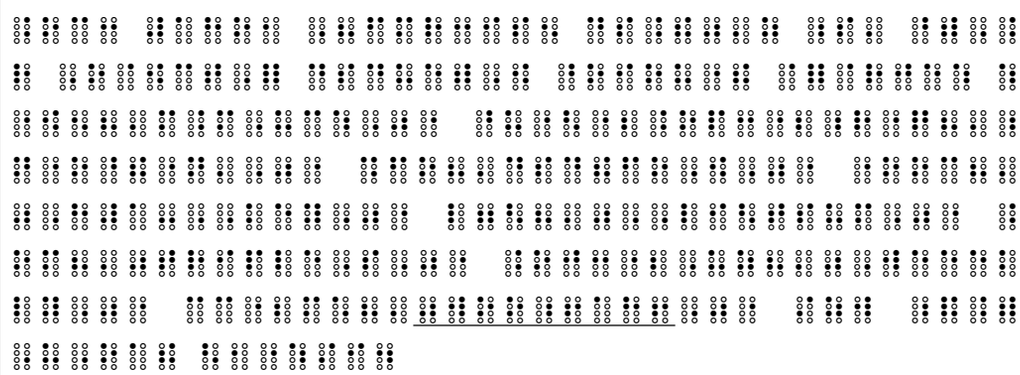
Form	신도시 교통난을 해소할 핵심인 수도권광역급행철도(GTX) 역시 속도를 내고 있다. 정부는 GTX-D·E·F 노선 추진 방안을 이르면 이달 중 발표한다. 이때 시종점, 주요 환승역, 개략적인 사업비가 공개된다.
Braille	
오점역 설명	단독으로 쓰인 로마자 D 앞에 1급 점자표 ⠰가 없음.
Braille 수정	
수정 설명	단독으로 쓰인 로마자 D 앞에 1급 점자표 ⠰을 함께 적음.

표 4-19 약자와 혼동될 수 있는 느낌표의 1급 점자표 오류 사례

시나리오 1(한글+(영어))

Form	지난 18일 '퀸덤 퍼즐'은 첫 번째 시그널송 퍼포먼스 비디오를 공개하며 여름(우주소녀), 보라·지원·채린(체리블렛), 유키(퍼플키스), 리이나(H1-KEY), 상아(LIGHTSUM), 지우(tripleS), 지한·소은(Weeekly), 나나·우연(woo!ah!), 예은의 출전 소식을 알렸다.
Braille	
오점역 설명	woo!ah!에서 약자 ff와 혼동될 수 있는 첫 번째 느낌표(!) 앞에 1급 점자표 ⣸을 사용하지 않음.
Braille 수정	
수정 설명	woo!ah!에서 첫 번째 느낌표(!) 앞에 1급 점자표 ⣸을 사용하여 약자 ff와 혼동을 방지함.

2) 알파벳 약자/약어 오류

표 4-20 ow 약자 오류 사례

시나리오 1(한글+(영어))

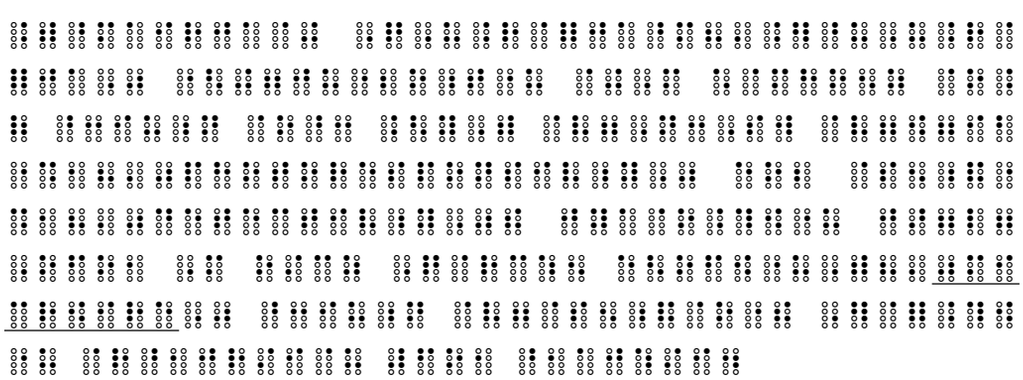
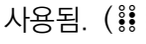
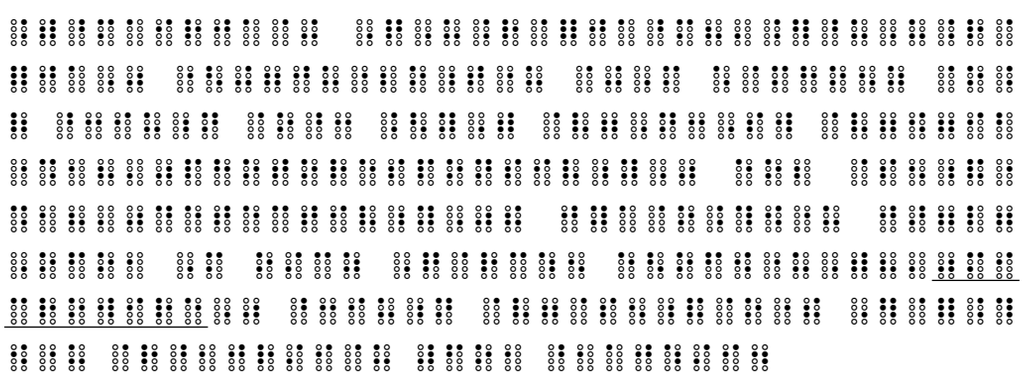
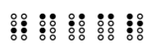
Form	전북대학교 신소재공학부(정보소재공학) 라용호·이철로 교수 연구팀은 기존 반도체 기반 자외선 광센서의 광응답률(photoresponsivity) 및 검출률(detectivity)을 획기적으로 향상시킬 수 있는 신개념 마이크로선(microwire) 반도체 광검출기를 성공적으로 개발했다고 6일 밝혔다.
Braille	
오점역 설명	microwire는 micro와 wire의 복합어이므로 ow 약자를 사용할 수 없으나 ow 약자 사용됨. ()
Braille 수정	
수정 설명	 에서 ow 약자를 풀어 로 변경함.

표 4-21 one 어두 약자 오류 사례

시나리오 1(한글+(영어))

Form	또 알츠하이머 치매 치료제인 도네페질(Donepezil) 패치 제품도 2021년 말 국내 허가를 획득하고 셀트리온제약을 통해 공급하고 있다.
Braille	[illegible]
오점역 설명	Donepezil은 one이 한 음절로 발음되지 않으므로 어두 약자 one을 사용할 수 없으나 사용됨. [illegible]
Braille 수정	[illegible]
수정 설명	one 어두 약자를 풀어 쓰는 것으로 수정함. [illegible]

표 4-22 LED 묶음 약자 오류 사례

시나리오 1(한글+(영어))

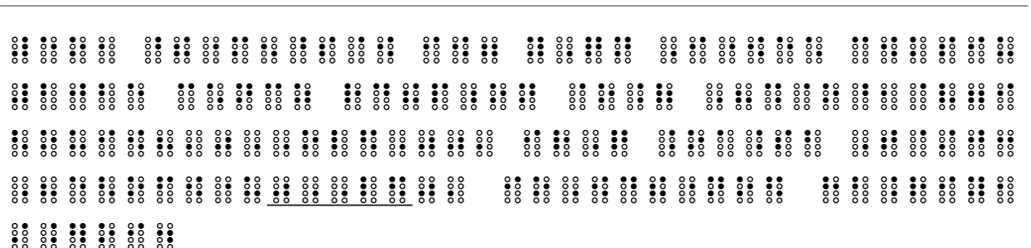
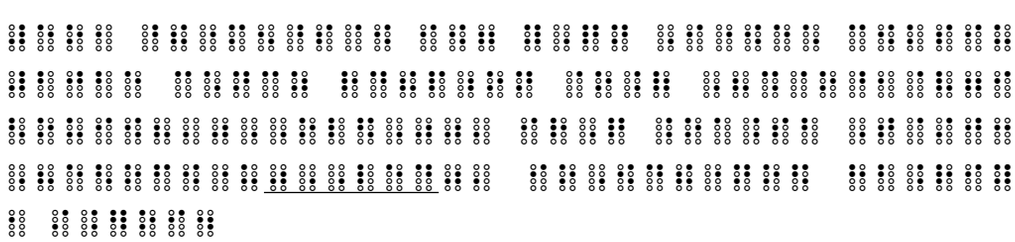
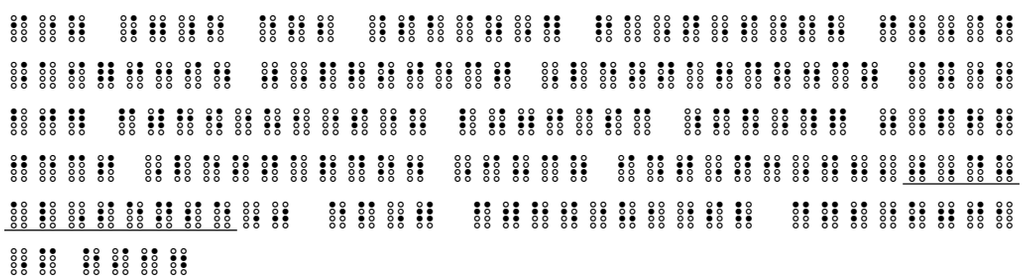
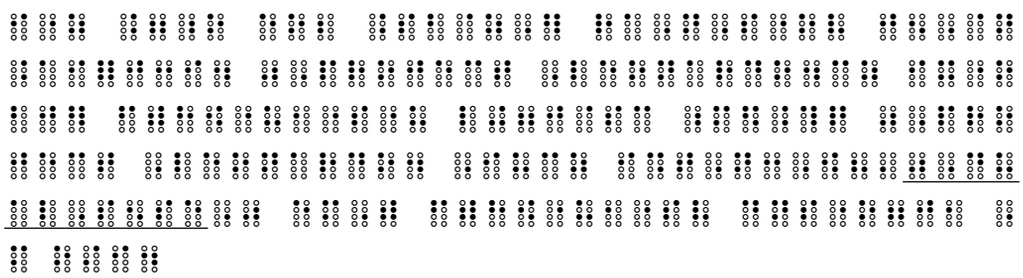
Form	5일 블룸버그 등 외신에 따르면 늦어도 2025년에는 애플워치에 기존 '유기발광다이오드(OLED)' 대신 직접 설계한 '마이크로LED' 디스플레이가 탑재될 예정이다.
Braille	
오점역 설명	마이크로LED에서 ed가 약자로 되어 있음.
Braille 수정	
수정 설명	'엘이디'로 읽히는 경우 약자를 사용하지 않으므로 'ed'로 수정함. ⠑⠙

표 4-23 en 약자 미사용 오류 사례

시나리오 1(한글+(영어))

Form	다른 장치 및 접근성 액세서리와 함께 작동한다. Project Leonardo는 독립형 컨트롤러로 사용하거나 추가적인 Project Leonardo 또는 듀얼센스(DualSense) 무선 컨트롤러와 페어링할 수 있다.
Braille	
오점역 설명	DualSense에서 en 약자가 사용되지 않음.
Braille 수정	
수정 설명	

3) 문장부호 오류

표 4-24 쌍점 띄어쓰기 오류 사례

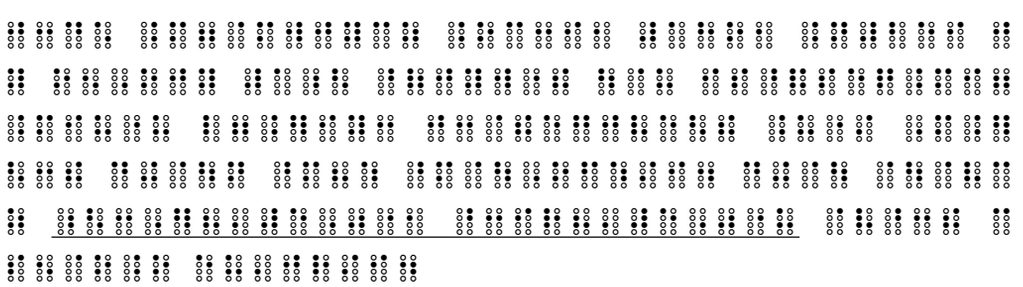
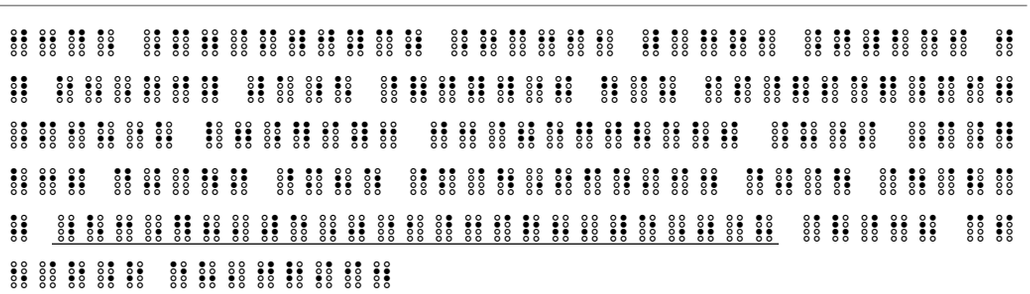
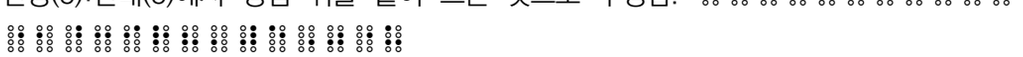
시나리오 3(한글+(영어))	
Form	한편 중구의회는 지난달 18일 제259회 임시회 1차 본회의를 열고 더불어민주당 주도로 상정된 한글문화마을 조례 수정안을 표결에 붙여 부결시켰다. 표결 결과는 찬성(5):반대(5)로 과반을 넘기지 못했다.
Braille	
오점역 설명	대비를 나타내는 쌍점에 띄어쓰기 있음.
Form 수정	
수정 설명	찬성(5):반대(5)에서 쌍점 뒤를 붙여 쓰는 것으로 수정함. 

표 4-25 아포스트로피 점역 오류 사례

시나리오 1(영어+(한글))

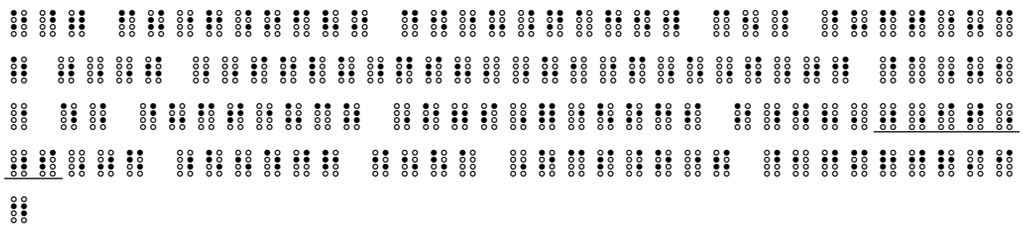
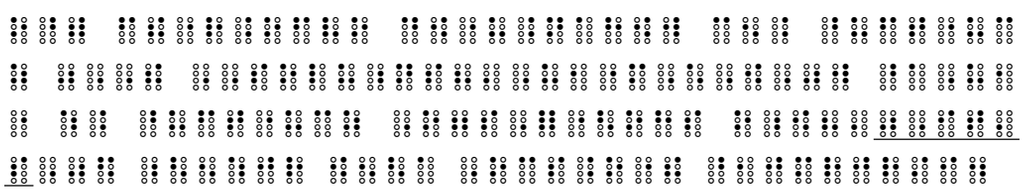
Form	이번 클래지콰이 프로젝트의 남자 보컬로는 THE SOLUTIONS(솔루션스)의 박솔, 여자 보컬로는 싱어송라이터 이츠(It's)가 참여해 음악 시너지를 발휘했다.
Braille	
오점역 설명	아포스트로피가 영어 닫는 작은따옴표로 점역됨.
Braille 수정	
수정 설명	영어 닫는 작은따옴표 ⠴⠄ 점형을 아포스트로피 점형 ⠄로 수정함.

표 4-26 마침표 뒤에 온 숫자와 혼동되는 첫소리 글자의 띄어쓰기 오류 사례

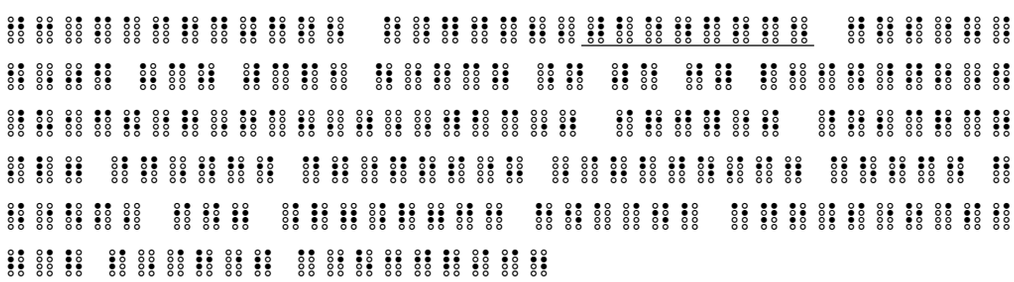
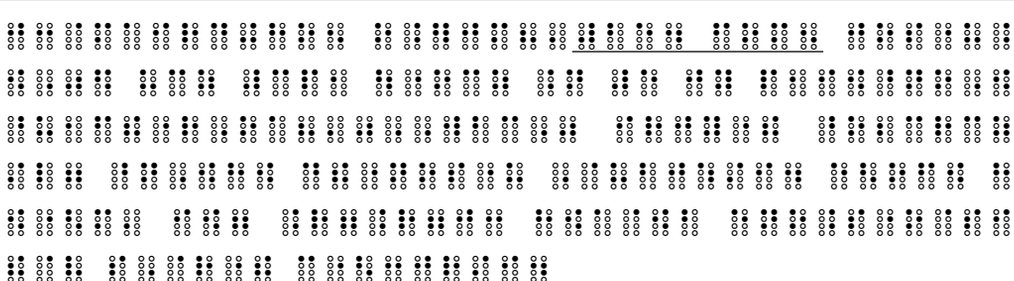
시나리오 1(한글+(영어))	
Form	한국대표팀 이정후(25.키움 히어로즈)가 오는 3월 열리는 제 5회 월드베이스볼클래식(WBC) 대회를 빛내는 예상 베스트 플레이어로 꼽혔다. 마이크 트라웃 등 쟁쟁한 특급 메이저리거들과 어깨를 나란했다.
Braille	
오점역 설명	25.키움에서 마침표 뒤에 띄어쓰기가 없어 25.6이움으로 오독할 수 있기 때문에 한국 점자 규정 제44항 [다만]에 의거하여 마침표 뒤를 띄어 써야 함.
Braille 수정	
수정 설명	'25.'와 '키움' 사이에 빈칸을 삽입함.

표 4-27 'A+'에서 덧셈표 점역 오류 사례

시나리오 1(한글+(영어))	
Form	평가는 17개 분야에서 40개 지표를 기준으로 이뤄졌다. 한국은 연구개발(R&D) 투자(A+), 원격 의료(A+), 디지털 자산(A), 드론(A), 기업 활동(A+) 인적 자원(A−) 등에서 높은 평가를 받았다.
Braille	⠙⠻⠫⠉⠵ ⠼⠁⠛⠈⠗ ⠘⠛⠜⠝⠠⠎ ⠼⠙⠚⠈⠗ ⠨⠕⠙⠬⠐⠮ ⠈⠕⠨⠛⠪ ⠐⠥ ⠕⠐⠏⠨⠱⠌⠊⠲ ⠚⠒⠈⠍⠁⠵ ⠡⠈⠍⠈⠗⠘⠂⠦⠄⠴⠠⠗⠈⠯⠠⠙⠠ ⠴ ⠓⠍⠨⠦⠄⠴⠠⠁⠢⠠⠴⠐ ⠏⠒⠈⠱⠁ ⠺⠐⠬⠦⠄⠴⠠⠁⠢⠠⠴⠐ ⠊⠕ ⠨⠕⠓⠾ ⠨⠇⠒⠦⠄⠴⠠⠁⠠⠴⠐ ⠊⠪⠐⠷⠦⠄⠴⠠⠁⠠⠴⠐ ⠈⠕⠎⠃ ⠚ ⠧⠂⠊⠿⠦⠄⠴⠠⠁⠢⠠⠴ ⠟⠨⠹ ⠨⠣⠏⠒⠦⠄⠴⠠⠁⠤⠠⠴ ⠊⠪⠶⠝⠠⠎ ⠉⠥⠲⠵ ⠙⠻⠫⠐⠮ ⠘⠔⠣⠌⠊⠲
오점역 설명	문맥상 A+와 A−의 덧셈표와 뺄셈표가 앞의 글자와 결합하여 하나의 의미 단위를 구성하고 있기 때문에 영어 속성이 이어지나 각각 한국 점자 규정의 덧셈표와 붙임표로 점역됨.
Braille 수정	⠙⠻⠫⠉⠵ ⠼⠁⠛⠈⠗ ⠘⠛⠜⠝⠠⠎ ⠼⠙⠚⠈⠗ ⠨⠕⠙⠬⠐⠮ ⠈⠕⠨⠛⠪ ⠐⠥ ⠕⠐⠏⠨⠱⠌⠊⠲ ⠚⠒⠈⠍⠁⠵ ⠡⠈⠍⠈⠗⠘⠂⠦⠄⠴⠠⠗⠈⠯⠠⠙⠠ ⠴ ⠓⠍⠨⠦⠄⠴⠠⠁⠐⠖⠠⠴⠐ ⠏⠒⠈⠱⠁ ⠺⠐⠬⠦⠄⠴⠠⠁⠐⠖⠠⠴⠐ ⠊⠕⠨⠕⠓⠾ ⠨⠇⠒⠦⠄⠴⠠⠁⠠⠴⠐ ⠊⠪⠐⠷⠦⠄⠴⠠⠁⠠⠴⠐ ⠈⠕⠎⠃ ⠚⠧⠂⠊⠿⠦⠄⠴⠠⠁⠐⠖⠠⠴ ⠟⠨⠹ ⠨⠣⠏⠒⠦⠄⠴⠠⠁⠐⠤⠠⠴ ⠊⠪⠶ ⠝⠠⠎ ⠉⠥⠲⠵ ⠙⠻⠫⠐⠮ ⠘⠔⠣⠌⠊⠲
수정 설명	⠴⠠⠁⠢를 ⠴⠠⠁⠐⠖로, ⠴⠠⠁⠤를 ⠴⠠⠁⠐⠤로 변경함.

표 4-28 ‘25bp(1bp=0.01%p)’에서 등호와 %p 점역 오류 사례

시나리오 1(한글+(영어))

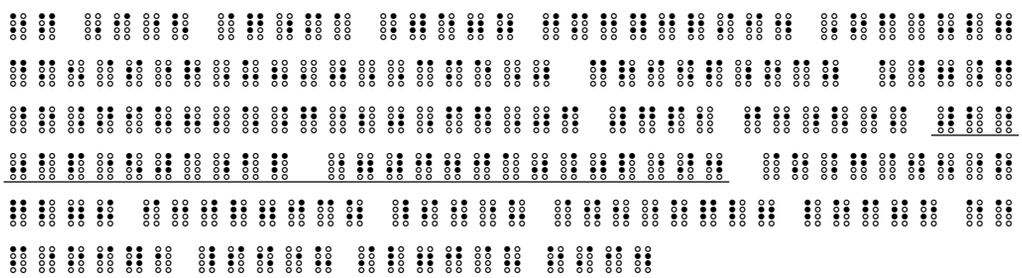
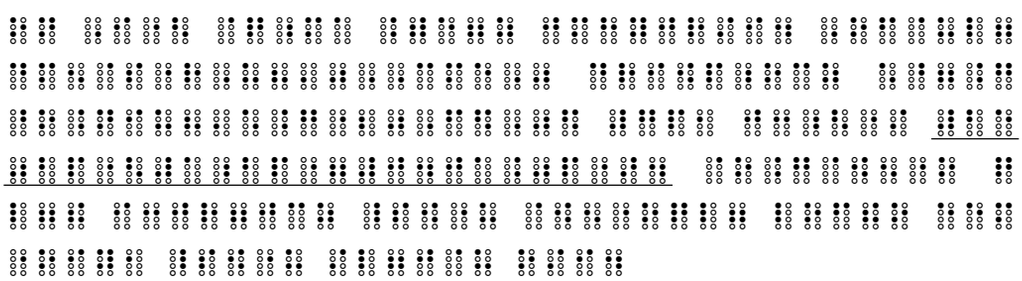
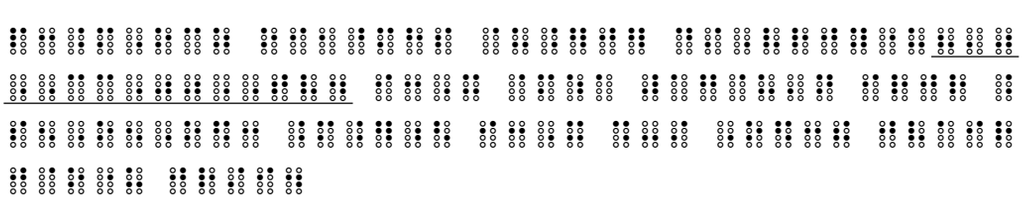
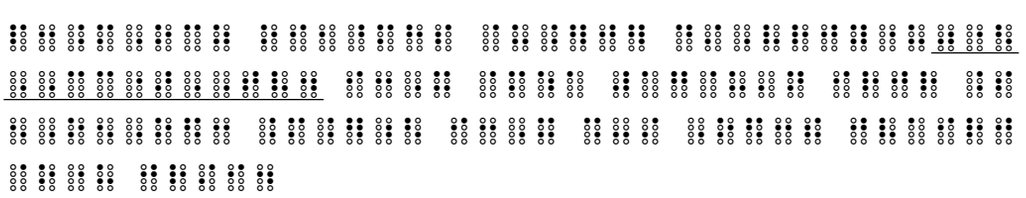
Form	이에 따라 긴축 전망은 후퇴했다. 시카고상품거래소(CME) 패드워치는 연방준비제도(연준·Fed)가 7월 한차례 25bp(1bp=0.01%p) 기준금리 인상을 단행하는 것으로 금리인상 사이클이 마무리될 것으로 예상하고 있다.
Braille	
오점역 설명	통일영어점자 11.2.1에 의하여 영어 연산기호의 띄어쓰기는 묵자를 따르지만 등호 앞을 한 칸 띄어 쓰고 %p 앞에 불필요한 로마자 시작표가 사용됨.
Braille 수정	
수정 설명	등호 앞 빈칸과 %p 앞 불필요한 로마자 시작표를 삭제함.

표 4-29 '(영어)영어' 형식의 소괄호 사용 오류 사례

시나리오 1(한글+(영어))

Form	원주시는 이달부터 고정형 폐쇄회로(CC)TV 단속 구역 17곳에 대해 점심시간 주정차 단속 유예 시간을 확대하기로 했다.
Braille	
오점역 설명	(CC)TV에서 영어 속성이 이어지기 때문에 한국 점자 규정 제29항과 제32항에 의거하여 영어 소괄호로 점역해야 하나 한글 소괄호로 점역됨.
Braille 수정	
수정 설명	 를 로 영어 소괄호로 수정함.

4) 단위/연산기호 오류

표 4-30 소괄호 안 연산기호 점역 오류 사례

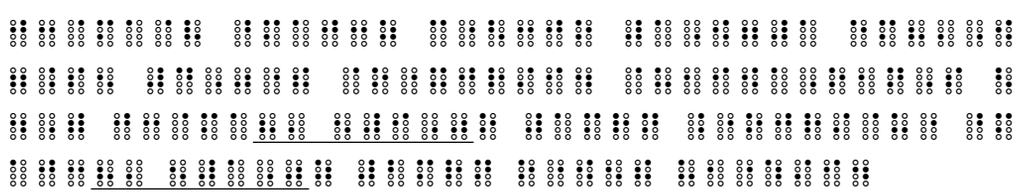
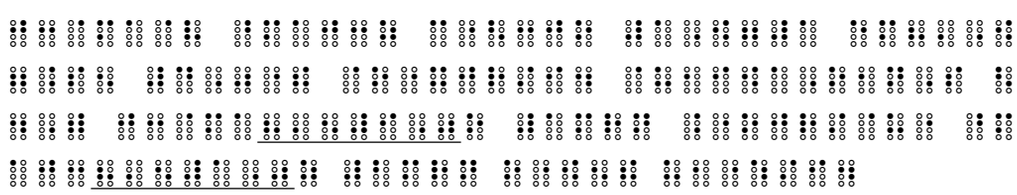
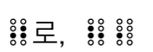
시나리오 1(한글+(영어))	
Form	한국과 북한은 나란히 1승1무(승점 4)를 기록했다. 골득실에서 앞선 한국(+9)이 1위에 자리했고, 북한(+1)은 2위에 이름을 올렸다.
Braille	
오점역 설명	(+9)와 (+1)에서 여는 소괄호와 연산기호 띄어쓰기하여 점역함.
Braille 수정	
수정 설명	여는 소괄호와 연산기호 띄어쓰기 없게 수정함.  를 로, 를 로 변경함.

표 4-31 단위기호 ㎡의 점역 오류 사례

시나리오 5(영어+(숫자))

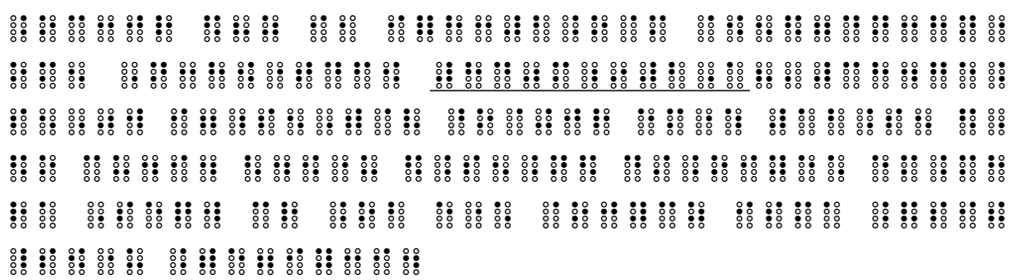
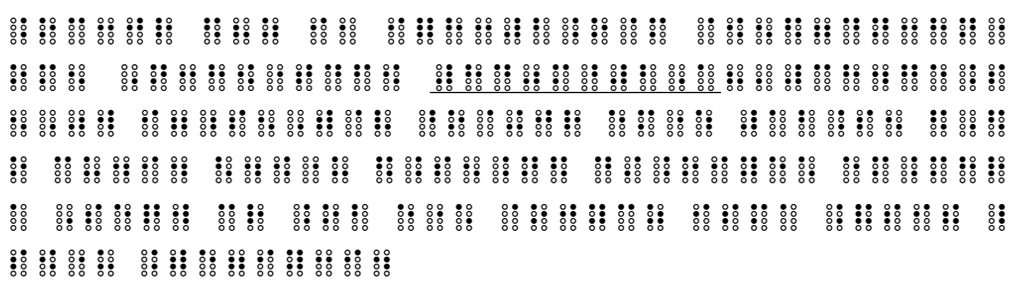
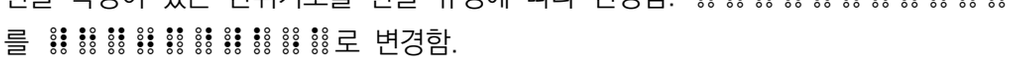
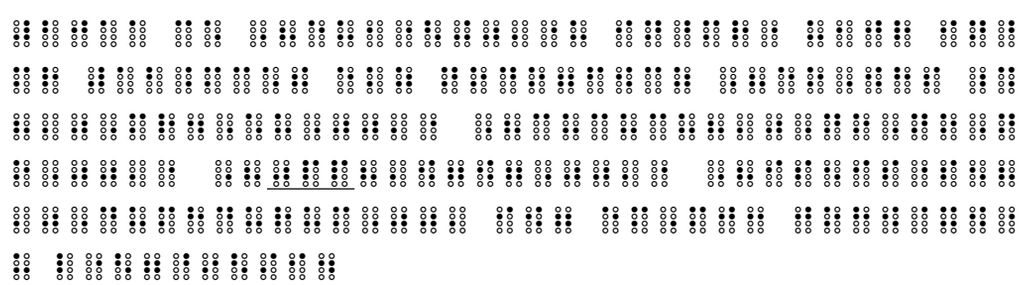
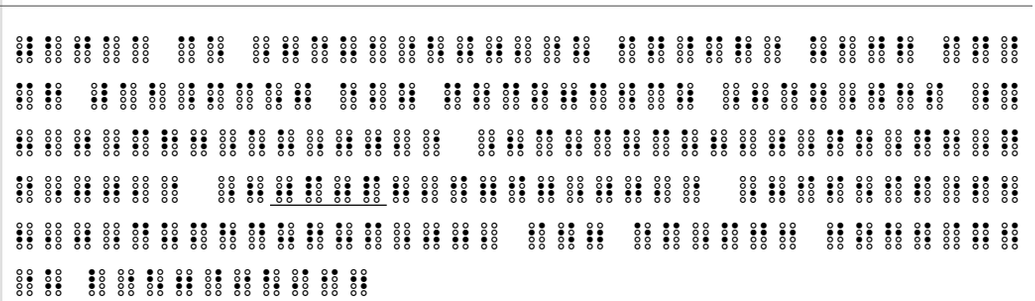
Form	지난해 같은 달 동탄2지구 금강펜테리움 센트럴파크 84㎡A(38.75점)의 당첨선과 비교해 무려 12점 가까이 높다. 앞으로 가점제가 폐지되면 무주택 서민의 내 집 마련 기회는 더욱 줄어들 것으로 전망된다.
Braille	
오점역 설명	단위 기호인 ㎡를 영어 규정에 따라 점역함.
Braille 수정	
수정 설명	한글 속성이 있는 단위기호를 한글 규정에 따라 변경함.  를  로 변경함.

표 4-32 연속으로 사용된 비로마자 단위기호 오류 사례

시나리오 1(한글+(영어))

Form	2011년 ‘몰라요’로 데뷔, 올해 데뷔 12주년을 맞은 에이핑크는 ‘미스터 츄(Mr.Chu)’, ‘노노노(NoNoNo)’, ‘%%(응응)’, ‘덤더럼(Dumhdurum)’ 등 무수한 히트곡으로 사랑받았다.
Braille	
오점역 설명	%%가 %p로 점역됨.
Braille 수정	
수정 설명	두 번째 % 기호 앞에도 단위표를 삽입하여 점역함.

5) 영어 대문자 표기 오류

표 4-33 대문자 종료표 표기 오류 사례 ①

시나리오 2(영어+(한글))	
Form	트라이비는 8일 오후 방송된 MBC M '쇼! 챔피언'에서 두 번째 미니앨범 'W.A.Y(웨이)'의 타이틀곡 'WE ARE YOUNG(위 아 영)' 무대를 선보였다.
Braille	⠓⠪⠐⠣⠕⠘⠕⠉⠵ ⠼⠓⠕⠂ ⠥⠚⠍ ⠘⠶⠠⠿⠊⠽⠒ ⠴⠠⠠⠍⠃⠉ ⠰⠠⠍ ⠲ ⠠⠦⠠⠬⠖ ⠰⠗⠢⠙⠕⠾⠴⠄⠝⠠⠎ ⠊⠍ ⠘⠾⠠⠨⠗ ⠑⠕⠉⠕⠗⠂⠘⠎ ⠢ ⠠⠦⠴⠠⠺⠲⠠⠁⠲⠠⠽⠦⠄⠏⠗⠕⠠⠴⠴⠄⠺ ⠓⠣⠕⠓⠮⠈⠭ ⠠⠦⠴⠠ ⠠⠠⠺⠑ ⠜⠑ ⠐⠽⠦⠄⠍⠗ ⠣ ⠻⠠⠴⠴⠄ ⠑⠍⠊⠗⠐⠮ ⠠⠾⠘⠥⠱⠌⠊ ⠲
오점역 설명	대문자 구절표 사용 시 대문자 종료표를 사용해야 하지만 미입력함. ⠴⠠⠠⠠⠺⠑ ⠜⠑ ⠐⠽
Braille 수정	⠓⠪⠐⠣⠕⠘⠕⠉⠵ ⠼⠓⠕⠂ ⠥⠚⠍ ⠘⠶⠠⠿⠊⠽⠒ ⠴⠠⠠⠍⠃⠉ ⠰⠠⠍ ⠲ ⠠⠦⠠⠬⠖ ⠰⠗⠢⠙⠕⠾⠴⠄⠝⠠⠎ ⠊⠍ ⠘⠾⠠⠨⠗ ⠑⠕⠉⠕⠗⠂⠘⠎ ⠢ ⠠⠦⠴⠠⠺⠲⠠⠁⠲⠠⠽⠦⠄⠏⠗⠕⠠⠴⠴⠄⠺ ⠓⠣⠕⠓⠮⠈⠭ ⠠⠦⠴⠠ ⠠⠠⠺⠑ ⠜⠑ ⠐⠽⠠⠄⠦⠄⠍⠗ ⠣ ⠻⠠⠴⠴⠄ ⠑⠍⠊⠗⠐⠮ ⠠⠾⠘⠥⠱ ⠌⠊⠲
수정 설명	대문자 종료표를 입력하는 것으로 수정함. ⠴⠠⠠⠠⠺⠑ ⠜⠑ ⠐⠽⠠⠄

표 4-34 대문자 종료표 표기 오류 사례 ②

시나리오 1(한글+(영어))

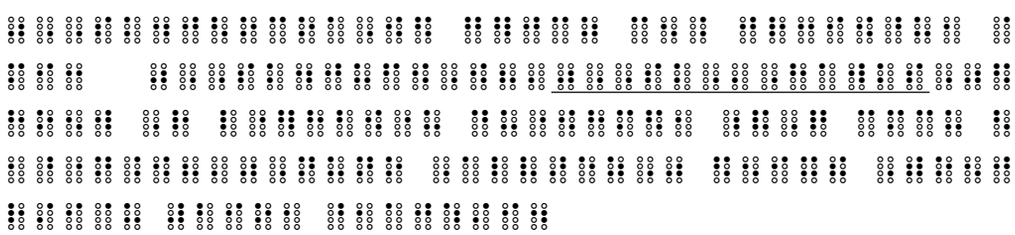
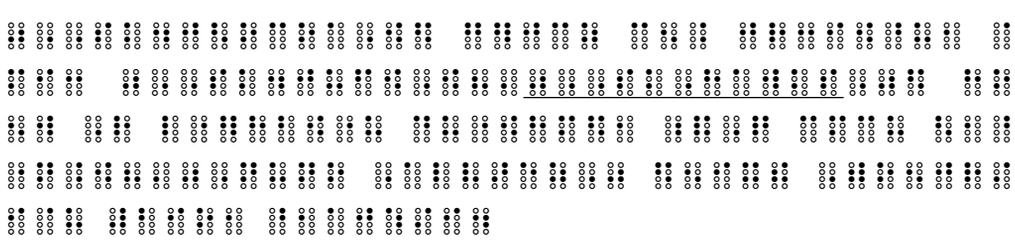
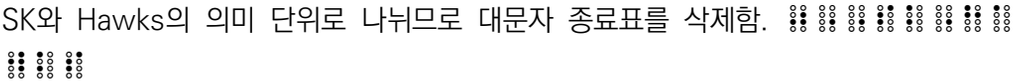
Form	SK하이닉스가 운영하는 남자 핸드볼 구단 SK호크스(SKHawks)가 팀의 새 사령탑으로 포르투갈 출신 누노 알바레즈(Nuno Alvarez) 감독을 선임했다고 29일 밝혔다.
Braille	
오점역 설명	대문자 단어표의 효력은 단일 대문자에 의해 종료되나 대문자 종료표도 삽입됨.
Braille 수정	
수정 설명	SK와 Hawks의 의미 단위로 나뉘므로 대문자 종료표를 삭제함. 

표 4-35 대문자 종료표 표기 오류 사례 ③

시나리오 1(한글+(영어))

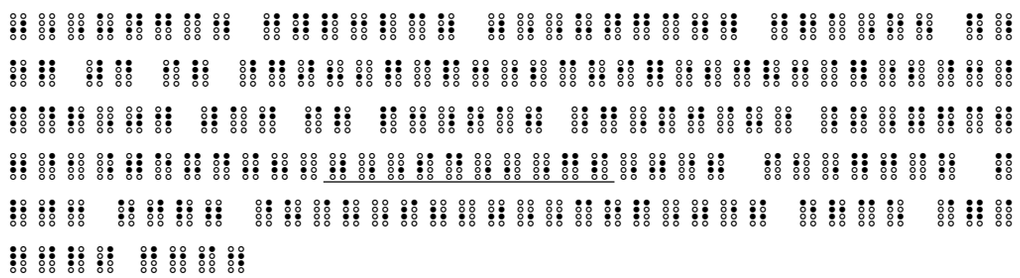
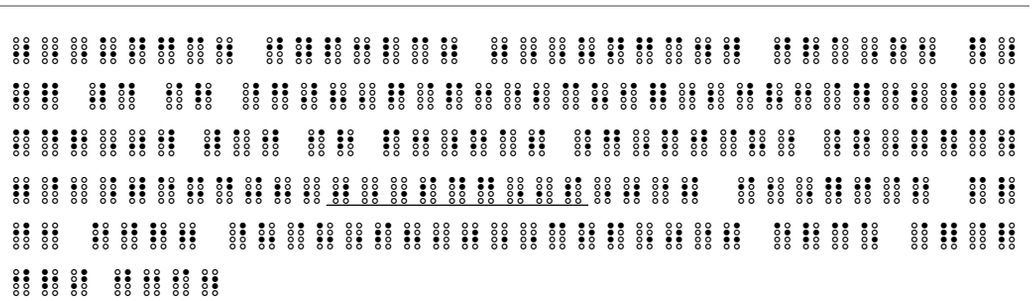
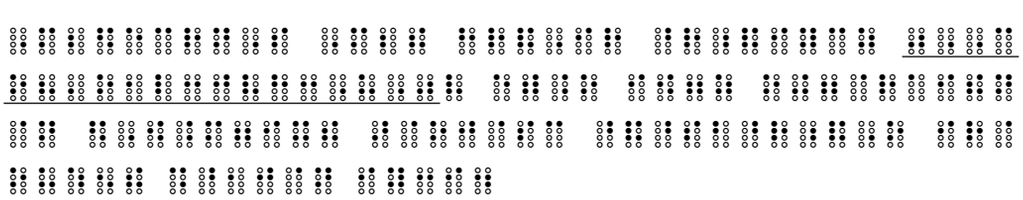
Form	UNGC 회원사는 UNGC의 핵심 가치인 4대 분야(인권·노동·환경·반부패)의 10대 원칙을 준수하고, 지속가능발전목표(SDGs)를 달성하기 위한 이행 보고서(COP)를 매년 공개해야 한다.
Braille	
오점역 설명	SDGs에서 대문자 종료표가 D 뒤에 위치해 있고 G 앞에 대문자 글자표가 사용되어 있음.
Braille 수정	
수정 설명	SDGs에서 대문자 종료표를 G 뒤에 위치하도록 수정하고 G 앞의 대문자 글자표를 삭제함. 

표 4-36 화학 원소 기호에서의 대문자표 표기 보류 사례

시나리오 2(영어+(한글))

Form	숯가마내에서 숯을 태울때 배출되는 CO(일산화탄소)와 먼지 등은 이용객들에게 가스 중독을 야기하거나 불법배출시 대기오염을 유발하게 된다.
Braille	
보류 이유	화학식 CO의 대문자 표기에 있어 한국 점자 규정 제28항 [붙임]과 과학 점자 규정 제7항 2조의 내용이 충돌함.

6) 로마자 시작표/종료표 오류

- 로마자의 시작표/종료표의 위치를 어디에 두어야 하는지와 관련된 오류 사례

표 4-37 로마자 시작표/종료표 오류 사례 ①

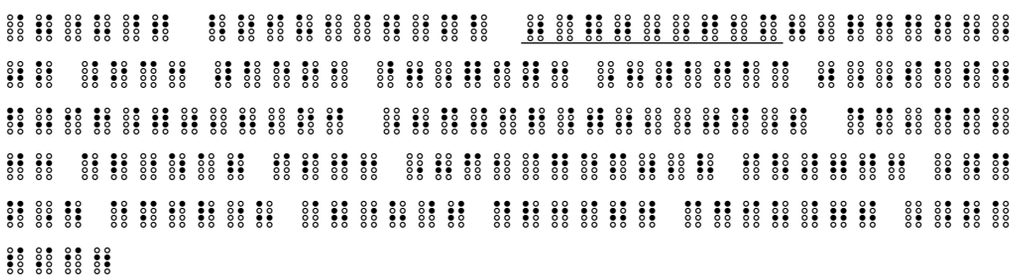
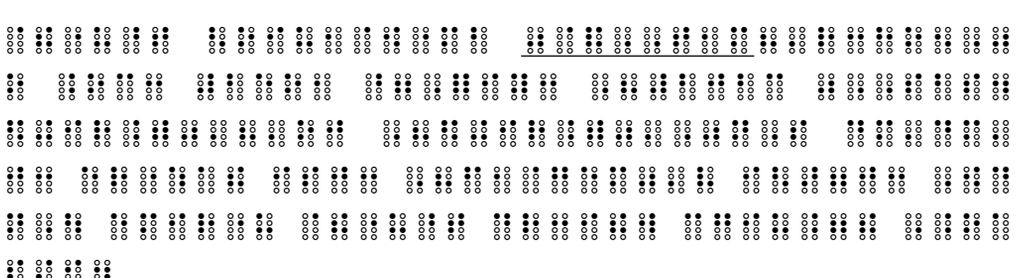
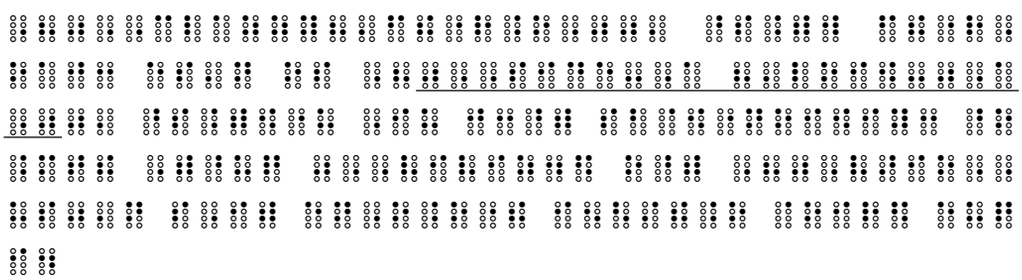
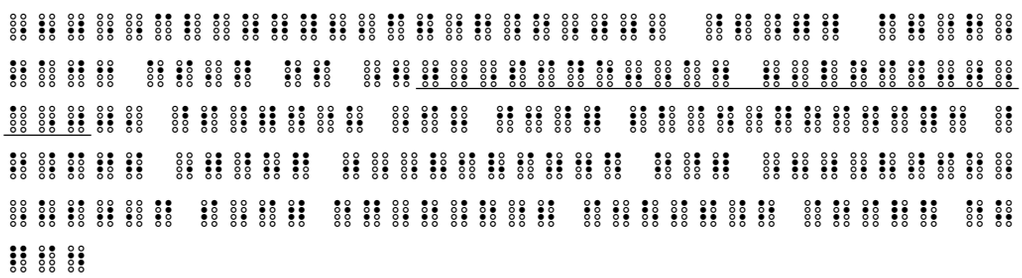
시나리오 2(영어+(한글))	
Form	글로벌 아이돌그룹 &TEAM(앤팀)이 지난 25일 방송된 '2023 SBS 가요대전'(이하 '가요대전')에서 풋풋한 매력을 더한 '칼군무'와 다양한 스페셜 무대로 글로벌 팬들의 눈도장을 찍었다.
Braille	
오점역 설명	&TEAM에서 T 앞에 로마자 시작표가 다시 사용됨.
Braille 수정	
수정 설명	&에서 이미 로마자 시작표를 사용했기 때문에 T 앞의 로마자 시작표를 삭제함.

표 4-38 로마자 시작표/종료표 오류 사례 ②

시나리오 2(영어+(한글))

Form	‘CLAZZY(클래지)’ 버전의 클래식한 멋에 이어 ‘SIDE-A (사이드-A)’ 버전으로 또 한번 업그레이드된 비주얼을 선보인 VIVIZ가 이번 ‘VarioUS’에 어떤 메시지를 담았을지 기대가 모인다.
Braille	
오점역 설명	SIDE-A 뒤에 띄어쓰기가 있으므로 로마자 종료표를 삽입해야 함.
Braille 수정	
수정 설명	

7) 제2외국어/한자 표기

- 프랑스어, 러시아어 등 외국어 점역 시 통일영어점자 내 변음 부호(제2외국어)를 적용하는데 검수자의 한계가 있어 제2외국어는 심화 범위로 판단 후 보류
- 한자의 경우 점역으로 표기할 때 소리를 나타내는 음만 표기할지, 동일한 음으로 오역을 방지하기 위해 음과 훈을 같이 사용할지 논의가 필요하여 보류

표 4-39 영어 외 외국어 사례 ①

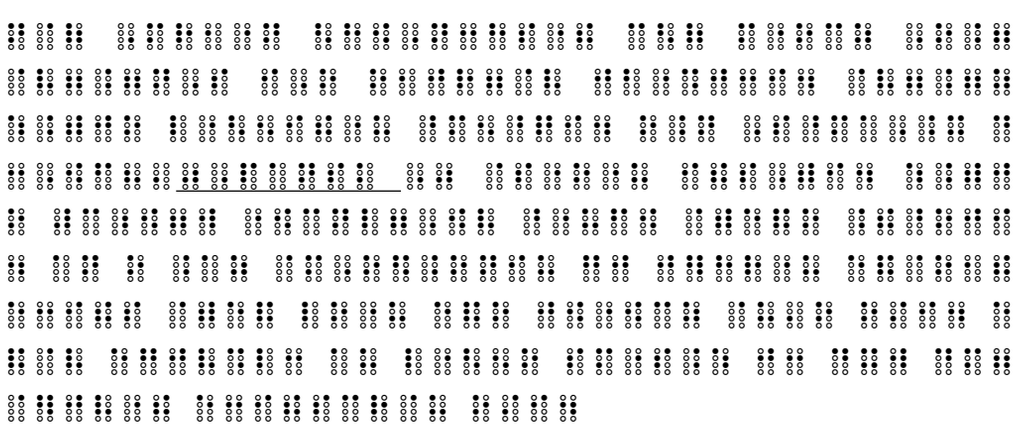
시나리오 1(한글+(영어))	
Form	에곤 쉴레 아트센터를 나온 우리는 시장 광장에서 다시 일행과 합류한다. 광장은 여전히 사람들로 붐빈다. 이제 서북쪽 판스카(Panská) 거리로 들어선다. 여전히 3층의 르네상스와 바로크 건물이 즐비하다. 그런데 이 작은 구시가지에도 운하 형태로 물길을 만들어 건물 사이로 물이 흐르는 곳이 있다. 물과 문화유산 그리고 사람이 어우러져 한 폭의 풍경화를 만들어내고 있다.
Braille	
보류 이유	Panská에 표기된 변음부호는 제2외국어이므로 어떻게 점역해야 할지 어려움이 있음.

표 4-40 영어 외 외국어 사례 ②

시나리오 2(영어+(한글))

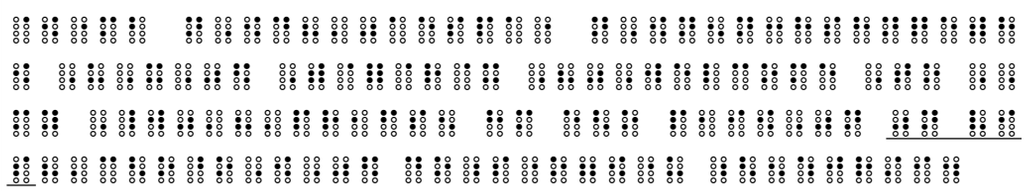
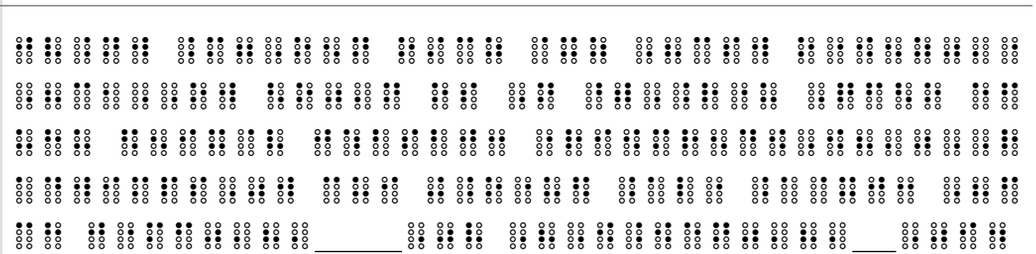
Form	그룹 에스파(aespa, 에스엠엔터테인먼트 소속)가 선공개곡 'Welcome To MY World'(웰컴 투 마이 월드)에 nævis(나이비스)가 피처링으로 참여했다.
Braille	
보류 이유	nævis에 표기된 합자 æ는 제2외국어이므로 어떻게 점역해야 할지 어려움이 있음.

표 4-41 영어 외 외국어 사례 ③

시나리오 1(한글+(영어))

Form	화제의 중심에 있는 곡은 '너의 이름은', '날씨의 아이'에 이어 세 번째로 신카이 마코토 감독과 협업한 래드윔프스(RADWIMPS)의 노다 요지로가 작사, 작곡한 '카나타 하루카'(カナタハルカ)와 '스즈메'(すずめ)다.
Braille	
보류 이유	カナタハルカ와 すずめ는 제2외국어이므로 어떻게 점역해야 할지 어려움이 있음.

8) 기타

- 삼각형, 사각형, 동그라미 등이 빠짐표나 숨김표가 아니라 기타 기호가 사용된 경우 점자 규정에서 점형이 정해지지 않았기에 오류가 아닌 보류로 판단
- 한국 점자 규정 외 기호(└ , ☎, ◇, ◆, ■, ▲, ▽, ▼, ▷, ▶ 등)

표 4-42 한국 점자 규정 외 기타 기호 보류 사례 ①

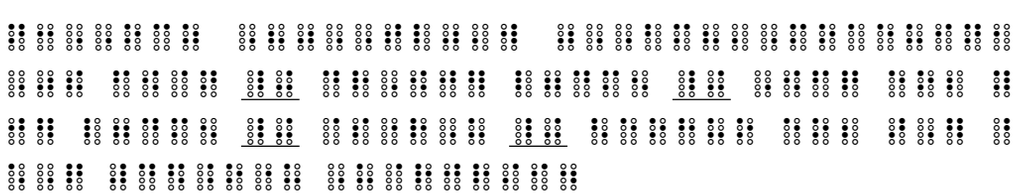
시나리오 2(영어+(한글))	
Form	엔씨는 'TL'의 BM(수익모델)을 크게 ▲ 패스형 상품 ▲ 스킨 및 외형 상품 ▲ 거래소 ▲ 아미토이 및 야성 변신 4가지로 소개했다.
Braille	
보류 이유	검은 위쪽 삼각형이 글머리 기호로 사용되었으나 한국 점자 규정 제72항에서 정의된 세모 글머리 기호가 예시의 흰 위쪽 삼각형만을 의미하는 것인지 삼각형 형태의 모든 기호를 포함하는 것인지 알 수 없음.

표 4-43 한국 점자 규정 외 기타 기호 보류 사례 ②

시나리오 1(한글+(영어))

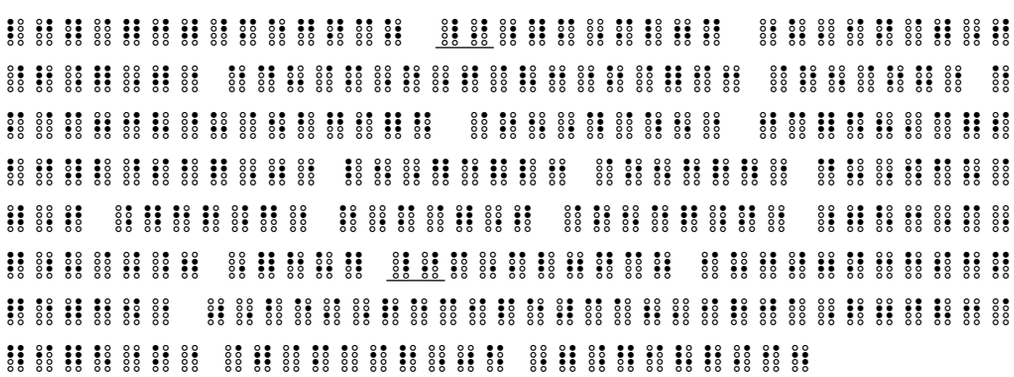
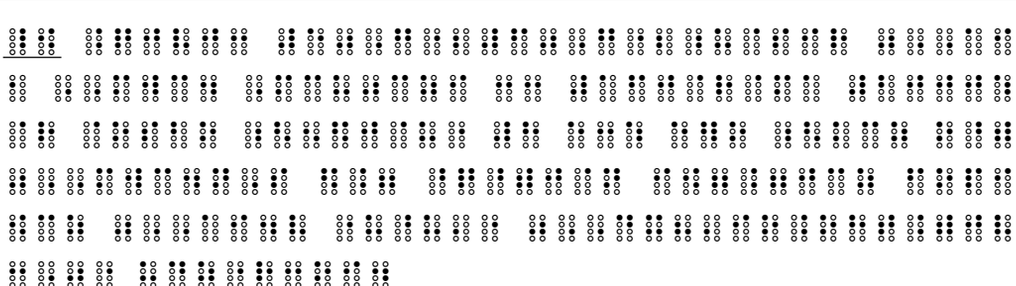
Form	사회공헌부문에는 ▷최우수상에 롯데건설 배종선, 마포구시설관리공단 김기옥, 무궁화보급운동가 김석겸, 하농가(농업회사법인), 삼성물산 김태열, 포스코건설 정태준, 이수건설 김민제, 천안수신초교장 송토영 ▷우수상에는 도향엔터테인먼트, BestRecipe31(대학생환경동아리, 건국대)가 선정됐다.
Braille	
보류 이유	오른쪽을 향한 흰색 삼각형이 글머리 기호로 사용되었으나 점자 규정에서 점형이 정해지지 않았기에 어떻게 점역해야 할지 어려움이 있음.

표 4-44 한국 점자 규정 외 기타 기호 보류 사례 ③

시나리오 2(영어+(한글))

Form	◇ 진화한 5G·6G·초거대 AI, MWC 수놓는다 = 160개국 2000여개 기업이 참가하고, 8만여 명이 찾는 이번 MWC에서 가장 빈번하게 등장하는 키워드는 AI와 로봇, DX(디지털전환), 오픈랜이다.
Braille	
보류 이유	마름모가 글머리 기호로 사용되었으나 점자 규정에서 점형이 정해지지 않았기에 어떻게 점역해야 할지 어려움이 있음.

- 쉼표 뒤에 빈칸 없이 모음이나 수표가 이어서 오는 경우 등 한국 점자 규정 제49항과 같이 묵자 기준으로 점역하면 오독할 요소가 있기에 보류로 판단

표 4-45 표기 오독으로 인한 보류 사례

시나리오 2(영어+(한글))

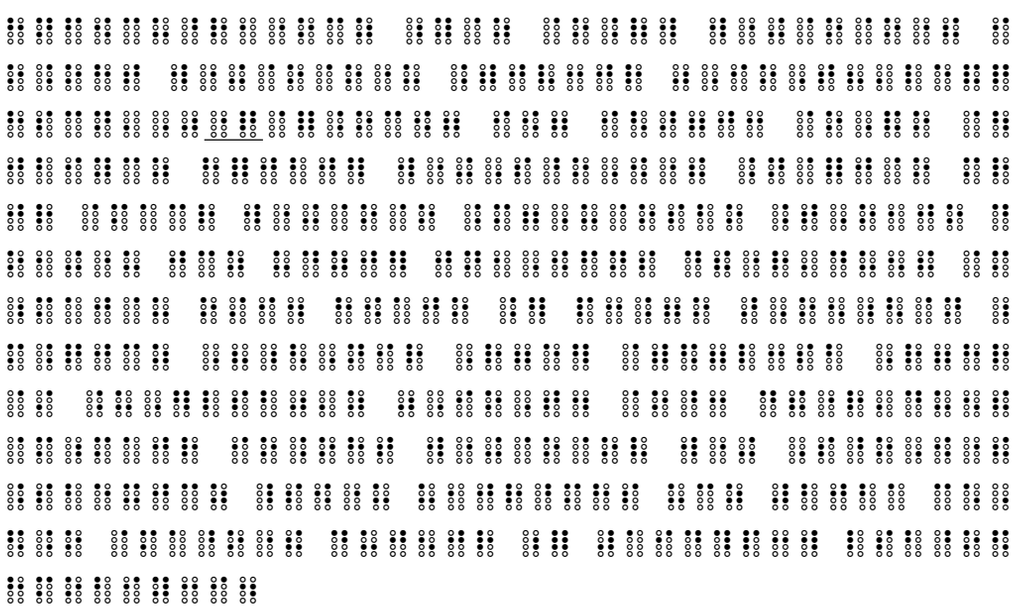
Form	테크노밸리는 최근 기존의 의료기기를 디지털 의료기기로 전환해 IoT(사물인터넷),인공지능 등 다양한 기술이 결합되는 통합형 의료서비스를 제공하기 위해 국내 의료기기 중소기업이 절실히 필요로 하는 오픈형 헬스케어 플랫폼을 구축하고 있다. 특히 정 원장이 야심차게 추진하는 '차세대 생명 건강산업 생태계 조성사업'은 IoT 기반 플랫폼을 구축해 디지털 의료기기와 의료 서비스를 접목하는 것으로 올해부터 오는 2021년까지 국비를 포함해 총 194억원의 사업비가 투입된다.
Braille	
보류 이유	빈칸 없이 묵자 기준으로 점역할 경우 린으로 오독할 요소가 있어 어떻게 점역해야 할지 어려움이 있음.

표 4-46 잘못 표기된 문장부호 보류 사례

시나리오 3(한글+(숫자))

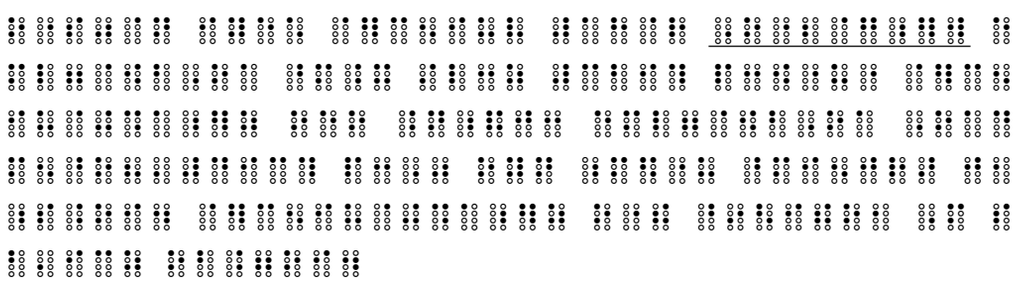
Form	이렇게 되면 경남도와 18개 시·군청의 무상급식 부족 예산은 329억 원으로, 경남도교육청은 이미 제출한 무상급식 식품비(493억 원) 이외에 추가로 부담해야 할 처지다. 경남도교육청은 이를 받아들일 수 없다는 입장이다.
Braille	
보류 이유	1점 리더가 가운뎃점으로 점역되었으나 점의 형태와 위치가 마침표와 유사하여 논의가 필요함. ※ 시.군청

표 4-47 연속으로 사용된 쉼표 보류 사례

시나리오 3(한글+(숫자))

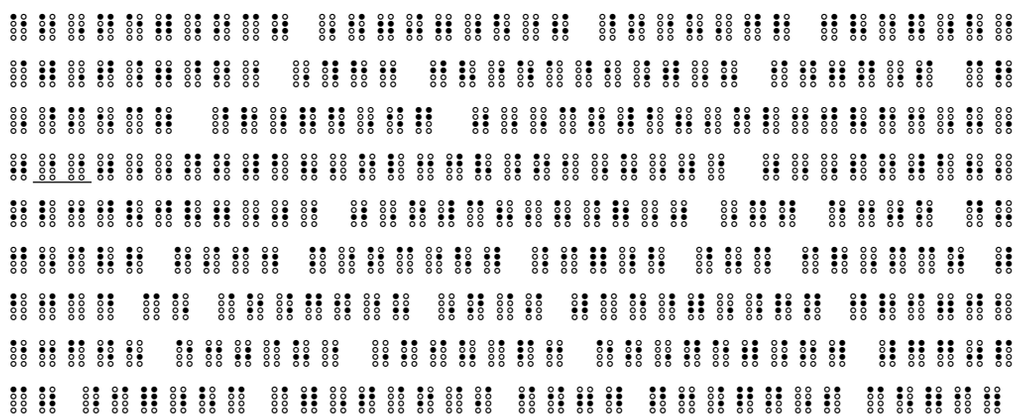
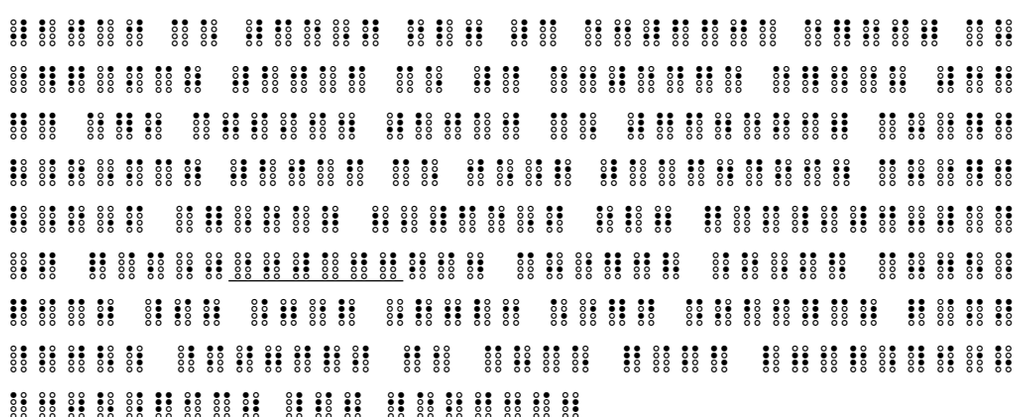
Form	미세먼지는 승용차를 비롯해 화물차, 건설장비, 석탄 화력발전소 등에서 내뿜는 배출 가스인 CO2(이산화탄소),,,NO2(이산화질소), SO2(이산화황), O3(오존) 속에 많이 포함돼 있다. 우리나라의 자동차 보유 대수는 2014년 기준으로 세계 15번째에 해당할 만큼 많고, 수도권 미세먼지의 77%는 자동차나 건설기계 등의 엔진에서 나온다.
Braille	
보류 이유	쉼표가 연속으로 두 번 사용되었으며, 사이에 띄어쓰기가 있으나 쉼표는 앞말에 붙인다는 규정으로 인하여 띄어쓰기가 사라짐.

표 4-48 곱셈표로 사용된 별표 보류 사례

시나리오 3(한글+(숫자))

Form	2010년 65세 이상 3만6601명이던 노령인구는 2016년 4만5475명으로 8874명이 늘었다. 2010년 76.5이던 노령화지수는 2016년 현재 116.4이다. 노령화지수 공식은 (65세 이상 인구/0-14세 인구)*100이다. 노령화 지수가 높아진다는 것은 장래 생산 연령에 유입되는 인구에 비하여 부양해야 할 노년 인구가 상대적으로 많아진다는 것을 의미한다.
Braille	
보류 이유	별표가 곱셈표의 의미로 사용되었으며, 수학 점자 규정의 별표와 용례가 다르기에 논의가 필요함.

4 데이터 주요 오류/보류 처리

검수 과정에서 발견된 오류 및 보류 데이터에 대하여 주요 유형을 조사하고 유형별로 차등하여 처리하였다.

1급 점자표 오류, 알파벳 약자/약어 오류, 문장부호 오류, 단위/연산기호 오류, 영어 대문자 표기 오류, 로마자 시작표/종료표 오류 등 한국 점자 규정에 점형 규정이 명확히 제시된 항목에 대해서는 1차 검수 또는 2차 검수 과정에서 단 한 번이라도 오류 또는 보류 판정을 받은 경우, 말뭉치 전담 인력이 직접 점형을 바르게 교정하여 데이터 변환을 진행한다.

반면 기타 부호 오류, 제2외국어 및 한자 표기 오류와 같이 개정 한국 점자 규정 내 명확한 규정이 부재한 유형에 대해서는 교정 처리가 불가하므로 해당 문장은 최종 데이터에서 보류 상태로 남겨 관리한다.

검수 과정에서 자주 발생하는 검색 엔진 점형 오류는 웹 기반 관리 시스템의 점역 엔진 업데이트를 통해 지속적으로 개선하였다. 검수 중 발견되는 유형의 오류 및 사이트 접속 오류 등의 오작동 사항이 있을 때마다 신속히 개선하고 이후 추가 업데이트를 통해 시스템의 적합률을 높였다.

이와 함께 웹 기반 관리 시스템 공지 사항 기능을 적극 활용하여 업데이트 패치 내역, 오류 유형별 대응 방안, 자주 하는 질문(FAQ) 등을 주기적으로 게시하였다. 검수자 커뮤니티 활성화를 위해 자유 게시판 기능을 신설하여 검수자들이 자유롭게 질의응답하고 사례를 공유하도록 지원하였다. 이를 통해 검수자 간 정보 교류를 촉진하고, 실시간 피드백과 노하우 축적을 통해 전반적인 검수 정확도와 적합률을 높였다.

그림 4-3 검수자 소통 자유 게시판

제 5 장

결 론

1. 결 론

1 결 론

본 사업의 목적은 묵자-점자 병렬 말뭉치 구축을 통해 궁극적으로 시각장애인의 일상생활 속 점자 사용 및 학습 편의를 증진하고 나아가 점자 친화적 생활환경 조성을 위한 개선 방안을 마련하는 데 있다. 이러한 목적 달성을 위해 다음과 같은 구체적인 목표를 설정하고 추진하였다.

첫째, 묵자-점자 병렬 말뭉치 데이터 구축이다. 한국어·영어 혼용 자료(묵자-점자 일대다 대응 자료 포함)를 확보하고, 신규로 묵자-점자 대응 문장 총 8만 쌍(100만 어절 이상)을 구축하였다. 또한 2021년 구축된 묵자-점자 병렬 말뭉치 93,510문장(3,339,239어절, 중복 제거)에 대해 2024년 개정된 한국 점자 규정을 반영하여 수정 구축하였다.

둘째, 검수 및 통계 분석 체계 마련이다. 구축된 병렬 말뭉치의 품질을 관리하기 위해 웹 기반 관리 시스템을 활용한 체계적인 검수 절차를 도입하였으며, 오류 방지를 위해 데이터 집계 특성에 따른 통계 분석 체계를 구축하고 이를 지속적으로 운영하였다.

셋째, 오류 및 보류 사례 유형화이다. 검수 과정에서 발견된 오류와 보류 사례를 유형화하여 점자 규정 및 지침과의 연계성을 분석하고 추후 규정 정비와 연계될 수 있는 기초 데이터를 확보하였다.

본 사업에서는 위의 목표를 달성하기 다음과 같은 절차를 거쳤다.

첫째, 묵자-점자 병렬 말뭉치 구축 및 검수 시스템을 고도화하였다. 웹 기반 관리 시스템을 통해 수집, 정제, 가공, 검수까지의 말뭉치 구축 전 단계에 걸쳐 효율적으로 관리하였다. 이렇게 구축된 병렬 말뭉치는 향후 점역 엔진 학습용 데이터로 활용할 수 있다.

둘째, 점자를 실제 사용하며 업무를 수행하고 있는 점역·교정 전문 인력을 동원하여 구축하였다. 총 68명의 점역·교정사 검수자와 3명의 말뭉치 사업 전담 인력이 개정 한국 점자 규정을 기반으로 검수 지침 교육, 예시 공유와 일대일 피드백 관리 등을 통해 해당 기간 동안 신규 말뭉치 84,000문장과 2021년 수정 구축 말뭉치 93,510문장을 개별 할당받아 검수하였다. 신규 구축 말뭉치에서 적합 판정을 받은 문장 수는 총 83,528문장으로 적합 검수율은 99.44%, 적합 어절 수는 1,675,580개에 달한다. 수정 구축 말뭉치에서 적합 판정을 받은 문장 수는 93,115문장으로 적합 검수율은 99.58%, 적합 어절 수는 3,322,242개에 달한다.

셋째, 데이터 오류 및 보류 유형을 분석하였다. 검수 과정에서 오류 및 보류 사유는 다음 8개 유형으로 분류하였다. ① 1급 점자표 오류, ② 알파벳 약자/약어 오류, ③ 문장부호 오류, ④ 단위/연산기호 오류, ⑤ 영어 대문자 표기 오류, ⑥ 로마자 시작표/종료표 오류, ⑦ 제2외국어 및 한자 표기 오류, ⑧ 기타 기호 표기 오류로 각 유형에 따라 데이터가 재교정되어 적합 처리되거나 불가피한 경우 보류 데이터로 분석하였다.

넷째, 본 사업에서는 최종 구축 결과물을 제이슨(JSON) 형식으로 가공하였으며, 작성 형식은 아래 〈표 5-1〉의 예시와 같이 구성하였다.

표 5-1 제이슨(JSON) 파일 작성 예시


```
{
   "language_info" : {
      "source_language" : "ko",
      "target_language" : "kb"
   },
   "parallel" : [ {
      "id" : "NIPAKOKB2500000003.3.11",
      "original_id" : "NIRW2400000003.3.11",
      "source" : "LG전자는 ‘업(UP)가전’을 무기로 내세우고 있다. 이번 CES에서도 LG 씽큐 앱에서 터치만으
로 제품 색상을 바꿀 수 있는 무드업 냉장고를 비롯한 다양한 업가전을 선보일 예정이다.",
      "target" : "⠴⠠⠠⠇⠛⠲⠨⠾⠨⠉⠵ ⠠⠦⠎⠃⠦⠄⠴⠠⠠⠥⠏⠠⠴⠫⠨⠾⠴⠄⠮ ⠑⠍⠈⠕⠐⠥ ⠉⠗
⠠⠝⠍⠈⠥ ⠕⠌⠊⠲ ⠕⠘⠾ ⠴⠠⠠⠉⠑⠎⠲⠝⠠⠎⠊⠥ ⠴⠠⠠⠇⠛⠲ ⠠⠠⠕⠶⠋⠩ ⠗⠃⠝⠠⠎ ⠓
⠎⠰⠕⠑⠒⠪⠐⠥ ⠨⠝⠙⠍⠢ ⠠⠗⠁⠇⠶⠮ ⠘⠠⠈⠯ ⠠⠍ ⠕⠌⠉⠵ ⠑⠍⠊⠪⠎⠃ ⠉⠗⠶⠨⠶⠈⠥
⠐⠮ ⠘⠕⠐⠥⠄⠚⠒ ⠊⠣⠜⠶⠚⠒ ⠎⠃⠫⠨⠾⠮ ⠠⠾⠘⠥⠕⠂ ⠌⠨⠻⠕⠊⠲",
      "revision" : {
         "revision1" : "⠴⠠⠠⠇⠛⠲⠨⠾⠨⠉⠵ ⠠⠦⠎⠃⠦⠄⠴⠠⠠⠥⠏⠠⠴⠫⠨⠾⠴⠄⠮ ⠑⠍⠈⠕⠐⠥
⠉⠗⠠⠝⠍⠈⠥ ⠕⠌⠊⠲ ⠕⠘⠾ ⠴⠠⠠⠉⠑⠎⠲⠝⠠⠎⠊⠥ ⠴⠠⠠⠇⠛⠲ ⠠⠠⠕⠶⠋⠩ ⠗⠃⠝⠠⠎
⠓⠎⠰⠕⠑⠒⠪⠐⠥ ⠨⠝⠙⠍⠢ ⠠⠗⠁⠇⠶⠮ ⠘⠠⠈⠯ ⠠⠍ ⠕⠌⠉⠵ ⠑⠍⠊⠪⠎⠃ ⠉⠗⠶⠨⠶⠈
⠥⠐⠮ ⠘⠕⠐⠥⠄⠚⠒ ⠊⠣⠜⠶⠚⠒ ⠎⠃⠫⠨⠾⠮ ⠠⠾⠘⠥⠕⠂ ⠌⠨⠻⠕⠊⠲"
          }
         ……
       ]
     }
     …….
   ]
}
```


이와 같은 방식으로 말뭉치 구축 사업의 체계적인 진행과 품질 관리를 수행하였으며, 최종 검수 및 데이터 검증을 통해 보다 정교한 병렬 말뭉치 데이터 세트를 구축하였다.

사업 총괄(PM)
이연주(한국시각장애인연합회 사무총장)

사업 수행 인력(PL)
김은주(한국시각장애인연합회 한국점자교육문화원 주임)

사업 수행 인력
이민정(한국시각장애인연합회 한국점자교육문화원 담당)
오혜은(한국시각장애인연합회 한국점자교육문화원 담당)

사업 담당자
홍혜진(국립국어원 수어점자진흥과 학예연구관)
김민정(국립국어원 수어점자진흥과 연구원)

2025년 묵자-점자 병렬 말뭉치 구축

2025년 10월 29일 인쇄
2025년 10월 29일 발행

| **발행인** | 국립국어원장
| **발행처** | 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9737

※ 이 책은 국립국어원의 용역비로 수행한 '2025년 묵자-점자 병렬 말뭉치 구축' 사업 결과를 발간한 것입니다.