

외국어 표기법 연구안

훈민바름 전산본

정본을 잃지 않고 저장하고 나르고 다시 읽는 규격

영어확장판 제3본 · Unicode 직렬화 · 글꼴과 셰이핑 · 국가표준 정합성

2026년 10월 12일 개정판 · 연구·검토용

Park Il Hwan

연구·검토용 비구속 문서

우리 글, 바른 소리

훈민바름 영어확장판 4본의 구성

훈민바름 영어확장판은 하나의 체계를 네 역할로 나눈다. 선언본은 뜻과 공공적 필요를 밝히고, 해례본은 손으로 기록할 수 있는 정본의 구조와 발음 원리를 풀며, 전산본은 정본의 디지털 보존·처리 규격을 세우고, 구현본은 웹 참조구현과 시험으로 실제 작동을 검증한다.

문서	핵심 내용	답하는 질문
제1본 선언본	역사·문화·사회·정책·교육·접근성·AI·표준화와 세계화 비전	왜 필요한가
제2본 해례본	영어 RP·GA 정본, 음가-자모 대응과 손글씨 시범 조자법	어떤 소리를 어떻게 적고 읽는가
제3본 전산본	구조 데이터·Unicode·글꼴·셰이핑·검색·복사·역변환	정본을 어떻게 보존하고 처리하는가
제4본 구현본	웹 참조구현·실제 시험·감사·공개 배포와 오류 제보	규격이 실제로 어떻게 작동하는가

이 문서의 지위와 명칭

‘훈민바름’은 옛한글·확장 한글 자모를 활용한 「외국어 표기법」의 가능성을 연구·개발하는 동안 사용하는 내부 코드명이다. 공적 규범의 명칭을 선점하지 않는다. 현재 훈민바름은 하나의 가표본(working specimen) 연구이자 참조 실험(reference experiment)이며, 이 체계를 그대로 채택하도록 요구하는 규범이 아니다.

첫 참조범위는 영어 RP·GA다. 다른 영어 방언은 독립 프로파일로, 다른 언어는 각자의 원문·source IPA·음운체계와 화자 검토를 갖춘 독립 모듈로 연구한다. 영어의 자모 대응을 다른 언어에 자동으로 상속하지 않는다.

현재 개발·구현 상태

제1-4본은 연구·검토용 판본이다. 이 제3본은 재현 가능한 사전 전수조사와 현재 참조구현 상태를 반영한 교정본이다. 제4본과 digital conversion engine, web app은 계속 개발·시험하며, 규격 제안은 최종 표준 채택이나 전체 구현 완료를 뜻하지 않는다.

문서 근거: [I1] 제1본 2쪽, [I2] 제2본 2·6쪽, [I3] 커버레터 2-3쪽.

이 문서의 읽는 순서와 근거

먼저 제1부에서 무엇을 보존하는지 정하고, 제2부에서 그 구조를 어떤 부호열로 운반하는지 살핀다. 제3·4부는 표시와 정보 회수를 다루며, 제5부는 이 기술 규격을 국가표준 정합성과 개정 제안으로 연결한다. 제6부와 부록은 구현자가 확인할 항목을 제공한다.

표지	이 문서에서의 뜻
잠정 고정	현 가표본의 기준값. 공개 검토와 시험에 따라 개정 가능
시범	비교·사용자 시험을 위한 잠정 형식·수치·절차
미결	최종 판정을 유보한 공동 연구 과제

근거를 구분하는 방식

제1·2본의 정책적 지위와 정본 규칙은 [I1]·[I2]로 인용한다. 이전 전산본과 대화 기록은 변경 경위를 확인하는 내부 자료 [I4]·[I5]로 구분한다. Unicode·OpenType·국가표준 관련 설명에는 외부 문헌 [E1]–[E6]를 붙인다. mappingUnits, 오류 상태와 교환 봉투의 세부 필드는 이번 개정에서 제안하는 전산 설계다.

셰이핑 폭은 2026-09-08 작성 환경의 계측값을 유지한다. 사전 전수조사는 저장소의 data/dict-rp.json·data/dict-ga.json과 tools/cluster_census.mjs로 다시 실행해 §21의 집계가 재현됨을 확인했다. 사전판·원출처·라이선스의 별도 확인이 필요하다는 한계는 그대로 남긴다.

열람 한계

KS X 1026-1:2007은 한국 측이 공개한 영문 번역본의 관련 조항을 확인했다. 번역본은 국문 정식 표준이 아니며 국문판이 최종 권위임을 스스로 밝힌다. 현행 국문판의 조문·개정 이력은 이번 집필에서 확정하지 않았다. 개정 제안은 이 한계를 명시한 연구안이다.

이번 개정의 범위

2026-08-14의 41쪽 전산본이 가진 6부·28절·부록 A–D의 조직을 유지한다. 데이터 모델, 원문 보존, 직렬화와 역변환의 한계, 동치 판정, 계측 근거와 국가표준 논증을 현재 제2본에 맞춰 다시 썼다. 웹앱의 현재 출력과 규격 사이의 차이는 제4본의 감사 대상이다.

차례

1. 4본 안에서 전산본의 지위	5
2. 다섯 층위 모델	5
3. 정본 데이터 모델	6
4. 자모값 등록부 · 자음	9
5. 직렬화의 기본 원칙	12
6. 자리별 직렬화 규칙	13
7. 직렬화 갤러리	14
8. 정규화와 동치	16
9. 금지되는 연쇄와 오류 검출	17
10. 표시층의 지위	18
11. 글꼴 요구사항	18
12. 세이핑 요구사항	19
13. 현재 환경에서 무엇이 합성되는가 · 재측정	20
14. 대체 표시	22
15. 검색	23
16. 복사와 붙여넣기	24
17. 역변환 · 무엇을 되찾는가	25
18. 교환 형식	27
19. 이미 표준 안에 있는 것	28
20. KS X 1026-1에서 확인한 조항	29
21. 훈민바름 연쇄가 막히는 지점	30
22. 두 갈래의 해법과 그 비용	31
23. 개정 제안	32
24. 국제표준과의 관계와 이행 경로	34
25. 적합성 등급	35
26. 시험 벡터와 27. 버전·변경 관리	36
28. 잠정 고정·시범·미결 목록	37
맷음말 · 부호는 정본을 섬긴다	38
부록 A. 복합 음절핵과 코드포인트 식별	39
부록 B. 시험 벡터와 기대 판정	41
부록 C. 측정 방법과 재현 절차	42
부록 D. 용어 대조	44
참고문헌 및 작성 근거 · 내부 자료	45

제1부 · 정보를 계산 가능하게 만드는 층

1. 4본 안에서 전산본의 지위

제2본이 세운 정보는 사람이 손으로 적을 수 있는 자모값과 배열 원리다. Unicode 부호열과 글꼴은 이를 저장하고 나르는 수단이다. 디지털 장치가 어떤 자형을 잘 그린다는 이유로 음가를 바꾸거나, 한 음절의 자음군을 여러 음절로 나누지 않는다.

동시에 제2본이 시범·미결로 남긴 조자법을 전산본에서 확정 규칙으로 승격하지 않는다. 현재 저장형의 무초성 중성 연쇄를 부호열로 시험할 수는 있지만, 그것이 복합 음절핵의 최종 손글씨 조자법을 결정하지는 않는다.

2. 다섯 층위 모델

층	이름	보존·처리하는 정보
1	canonicalModel	음운 토큰·mappingUnits·자모 요소·음절 경계·운율·판정 상태
2	unicodeSerialization	지원되는 정보 구조의 결정적 조합형 자모 연쇄
3	displayRuns	블록별 표시 명령, 상대 크기, 대체 표시 상태
4	displayText	글꼴·조판 환경에 따라 나타나는 파생 문자열
5	fallbackText	합성 불가·미배당 시 제공하는 선형 표기·IPA·설명

다섯 층 앞에는 원문과 source IPA를 보관하는 source 레코드가 있다. source에서 normalized IPA를 만들고 음운 토큰을 분석한 뒤 층 1을 구성한다. 이 변환에는 언어모듈·정규화 규칙·음절분석의 버전과 근거를 남긴다.

같은 입력만으로는 재현 조건이 충분하지 않다. 같은 source, 언어모듈, registry, 정규화·음절분석 규칙과 직렬화 프로파일 버전이 함께 고정되어야 같은 결과를 재현할 수 있다.

층 1과 source가 보존의 권위다. 층 2는 구조의 일부를 옮긴 파생 형식이며, 층 3–5는 이를 읽게 하는 출력이다. 표시 실패는 표시 상태를 바꿀 뿐 source나 음운 분석을 바꾸지 않는다.

[I1] 2·29–30쪽; [I2] 6·8–10·26쪽. 세부 데이터 필드: 본 개정 설계안.

3. 정보 데이터 모델

3.1 원문과 정규화 결과를 함께 보존한다

source IPA는 출처가 제공한 문자열 그대로 보존한다. normalized IPA는 선택된 언어모듈이 매핑 직전에 사용하는 분석형이다. 정규화는 철자를 고치는 작업과 음운 분석을 구분하며, 변경 규칙과 원문상의 위치를 기록한다. Unicode 정규화와 IPA 분석 정규화도 서로 다른 절차다.

예를 들어 GA source /'wɒtə/는 해당 규칙에 따라 normalized /'wɒtə/로 분석한다. source의 /ə/를 삭제하지 않고 변환 이력으로 연결한다. /ɜ/의 기저 모음은 제2본에서 미결이므로 임의로 /ə/나 /ʌ/로 바꾸지 않는다.

3.2 음운 토큰과 자모 요소를 분리한다

phonemeTokens는 선택된 모듈이 인정한 음운 단위의 순서다. mappingUnits는 이 토큰들과 정보 자모값의 대응 관계다. 하나의 음운 단위가 여러 자모 요소를 가질 수 있고, 여러 음소가 하나의 등록 복합 자모값을 가질 수 있다. 따라서 음운 토큰 수와 자모 요소 수를 같다고 가정하지 않는다.

음운 입력	자모값 배열	관계
/p/	[JV-P]	1:1
/u:/	[JV-U, JV-U]	1:N
/eɪ/	[JV-E, JV-KIT]	1:N
/t/ + /s/	[JV-TS]	N:1

각 mapping unit은 입력 token_ids의 순서, 출력 jamo_values의 순서, 적용 rule_id와 판정 상태를 가진다. 장모음은 기존 모음 요소를 반복하고 length=long을 보존한다. 복합값 ʌ은 /t/와 /s/ 두 토큰을 남긴다. 역변환은 이 관계와 등록된 역규칙을 함께 따른다.

3.3 /l/과 무음 초성

play /pleɪ/의 /l/은 같은 초성군에서 /p/ 뒤에 한 번만 기록한다. 초성군 자모값은 [JV-P, JV-L]이다. 모음으로 시작하는 음절의 ɔ은 음소가 아닌 자리표지이며, JV-ZERO와 kind=orthographic_zero로 기록한다. Unicode 채움 문자와 구별한다.

[I2] 7·25-27·30·37·58-59쪽; [I5] Master 3, 137-140쪽. 식별자 철자와 필드 구성은 시범 스키마.

3. 정보 데이터 모델 · 음절과 상태

3.4 블록은 자모가 아니라 선택된 음절분석을 따른다

한 영어 음절은 하나의 훈민바름 음절블록에 대응한다. 각 블록은 onset[], nucleus[], coda[]를 가진다. 배열 원소는 mapping unit과 그 출력 요소의 번호를 가리킨다. 장모음 /u:/의 두 중성 요소는 동일한 음운 토큰에 연결되지만, 각각 구별되는 출력 요소다.

영어의 음절분석은 철자열이나 자모 종류만으로 새로 추정하지 않는다. 출처 사전의 경계, 언어מוד의 분석 규칙 또는 검토자의 판정을 syllabification에 기록한다. 둘 이상의 합리적 분석이 있으면 선택된 분석과 후보를 구분한다. 부호열에서 읽는 경계는 이미 결정된 분석을 회수하는 경계다.

3.5 완성·시범·미결을 데이터로 표현한다

필드	예시 값	용도
status	complete / partial / unresolved	이 레코드의 구조 완성 상태
rule_status	provisional / trial / unresolved	개별 규칙의 잠정 고정·시범·미결
serialization.status	complete / unavailable	평문 전체를 만들 수 있는지
render_status	composed / linear / missing	화면 표시의 상태
issues[]	issue_id, token_ids, reason	실패 위치와 원인을 보존

구조가 유효하더라도 중성 위치형 부호가 미배당이면 평문은 unavailable일 수 있다. 반대로 평문을 생성할 수 있어도 최종 압축 자형은 아직 만들지 못할 수 있다. 미결 항목이 포함된 레코드를 단순히 성공 또는 실패 한 값으로 압축하지 않는다.

3.6 Narrow와 운율

Broad 구조를 먼저 보존하고 Narrow는 underlying token, surface phone, feature_id, 환경 규칙과 함께 별도 층에 추가한다. GA [ɹ]와 dark L [ɫ]은 Broad의 /t,d/와 /l/을 덮어쓰지 않는다. [ʔ] 자형과 음절자음의 최종 처리처럼 미결인 항목에는 완성된 자모열을 만들어 넣지 않는다.

stress는 primary·secondary·none·unknown을 구분한다. 길이는 핵의 음운 속성이다. 제1강세·제2강세 핵은 100%, 무강세 핵은 75%이며 같은 단어에 제2강세가 있을 때만 제1강세를 125%로 표시한다. 강세기호 없는 단음절을 자동으로 primary로 판정하지 않는다.

[I2] 31–37쪽. 상태 필드와 unknown 처리: 본 개정 설계안.

3. 정보 데이터 모델 · 교환 예시

다음은 boost의 핵심 구조를 보이는 유효한 JSON 예시다. 문서 예시이므로 source의 사전 출처는 비워 두며, 실제 사전 레코드에는 부록 C의 출처 정보를 채워야 한다. 버전 문자열은 이 초안의 시범 식별자다.

```
{
  "schema_version": "draft-2026-09-08",
  "module_id": "en-RP", "registry_version": "draft-2026-09-08",
  "source": {"spelling": "boost", "ipa": "/bu:st/",
    "provenance": {"kind": "document-example", "ref": "I2:p9"}},
  "normalized_ipa": "/bu:st/",
  "normalization_rules": [],
  "canonicalModel": {
    "status": "complete",
    "phonemeTokens": [
      {"id": "t1", "ipa": "b"}, {"id": "t2", "ipa": "u:"},
      {"id": "t3", "ipa": "s"}, {"id": "t4", "ipa": "t"}],
    "mappingUnits": [
      {"id": "m1", "token_ids": ["t1"], "jamo_values": ["JV-B"],
        "rule_id": "consonant-b"},
      {"id": "m2", "token_ids": ["t2"],
        "jamo_values": ["JV-U", "JV-U"], "length": "long",
        "rule_id": "long-u", "rule_status": "trial"},
      {"id": "m3", "token_ids": ["t3"], "jamo_values": ["JV-S"],
        "rule_id": "coda-s"},
      {"id": "m4", "token_ids": ["t4"], "jamo_values": ["JV-T"],
        "rule_id": "consonant-t"}],
    "blocks": [{"id": "b1", "stress": "unknown",
      "onset": [["m1", 0]], "nucleus": [["m2", 0], ["m2", 1]],
      "coda": [["m3", 0], ["m4", 0]]}]
  },
  "serialization": {"profile": "hmbr-jamo-trial-2026-09-08",
    "status": "complete", "normalization": "NFD",
    "codepoints": [4359, 4462, 4462, 4538, 4544]}
}
```

자리 참조 ["m2", 0]은 m2의 첫 자모 요소를 가리킨다. 두 중성 요소를 모두 소비한 뒤 같은 출력 요소를 다른 자리에 다시 사용하지 않는다.

위 예시의 source와 canonicalModel, serialization을 합친 전체 레코드가 보존 대상이다. serialization만 따로 복사하면 source와 강세·출처가 함께 전달되는 것은 아니다.

4. 자모값 등록부 · 자음

등록부는 문서·구현·시험이 공유하는 자모값 ID와 위치형의 연결표다. 음가-자모 대응은 제2본을 따르고, 전산본은 위치별 Unicode 부호와 처리 상태를 명시한다. 자형만 닮았다는 이유로 부호를 대체하지 않는다.

IPA	자모값 ID	초성형	종성형
/p/	JV-P	U+1111	U+11C1
/b/	JV-B	U+1107	U+11B8
/t/	JV-T	U+1110	U+11C0
/d/	JV-D	U+1103	U+11AE
/k/	JV-K	U+110F	U+11BF
/g/	JV-G	U+1100	U+11A8
/f/	JV-F	U+114B	U+11F4
/v/	JV-V	U+1144	U+11E6
/θ/	JV-TH	U+1145	미배당
/ð/	JV-DH	U+1142	미배당
/s/	JV-S	U+1109 / U+110A	U+11BA
/z/	JV-Z	U+1136	U+D7EF
/ʃ/	JV-SH	U+1140	U+11EB
/ʒ/	JV-ZH	U+1146	미배당

/s/의 초성 위치형은 /u:, ʊ/ 앞 ㅅ, 그 밖의 모음 앞 ㅆ, 자음군 안 ㅅ이라는 현재 시범 경계를 사용한다. 경계의 최종 범위는 사용자·지각 시험 대상이다. 종성형은 U+11BA이며, 표시층에서 한국어식 발음 중화를 적용하지 않는다.

등록부의 상태는 두 축으로 관리한다. 음가-자모값의 판정과 해당 위치형의 Unicode 배당 상태는 서로 다르다. 종성 부호가 미배당이라는 말은 해당 음소나 손글씨 자모값 자체가 없다는 뜻이 아니다.

[I2] 11-12·25·29·36·59쪽; [I4] 9-11쪽. 코드포인트 이름은 부록 A의 처리 원칙에 따라 확인한다.

4. 자모값 등록부 · 자음과 특수값

IPA·기능	자모값 ID	초성형	종성형
tʃ	JV-CH	U+1149	미배당
dʒ	JV-JH	U+1148	미배당
t+s	JV-TS	해당 없음	U+11BE
h	JV-H	U+1112	해당 없음
m	JV-M	U+1106	U+11B7
n	JV-N	U+1102	U+11AB
ŋ	JV-NG	해당 없음	U+11BC
l	JV-L	U+1105	U+11AF
ɹ	JV-R	U+A976	U+D7DB
j	JV-Y	U+1159	해당 없음
w	JV-W	U+1147	해당 없음
[ɾ]	JV-FLAP	U+111B	해당 없음
[ɬ]	JV-DARKL	해당 없음	U+11D0
무음 ɔ	JV-ZERO	U+110B	해당 없음

JV-TS는 두 음소 /t+s/에 대응하는 종성 복합값이다. 이 프로파일에서는 초성 자리에 적용하지 않는다. /tʃ/ 한 음소의 JV-CH와 구분한다. 어두 /j/, /w/의 기저값은 보존하되 모음과의 결합·축약형은 미결이므로 명시형 예시를 최종 결합규칙으로 선언하지 않는다.

4.1 미배당 위치형의 처리

/θ, ð, ʒ, tʃ, dʒ/의 종성은 이 초안에 계승한 등록부에서 전용 부호를 확정하지 않았다. 구조형은 해당 음소·자모값·coda 위치를 그대로 저장한다. 평문 전체는 unavailable로 두고, 누락 음소와 이유를 issues[]에 기록한다.

초성 부호를 종성 자리에 끼워 넣거나, filler를 붙여 임시 평문을 정상 직렬화처럼 내보내지 않는다. 그렇게 하면 L/V/T 순서에 의한 경계 판별과 위치별 역변환의 전제가 무너진다. 부호 후보를 시험하려면 별도 실험 프로파일과 버전을 사용한다.

[I2] 30–37쪽; [I4] 10–11쪽. 미배당 시 반환 상태: 본 개정 설계안.

4. 자모값 등록부 · 모음

음가·요소	ID	부호	적용
i	JV-I	U+1175	단모음·장모음의 구성 요소
ɪ	JV-KIT	U+119D	KIT 대립 보존
ɛ / e	JV-E	U+1166	DRESS와 FACE 구성 요소
æ	JV-AE	U+1162	TRAP
ɑ / a	JV-A	U+1161	GA LOT·PALM 및 복모음 요소
ɒ / o	JV-O	U+1169	RP LOT, GOAT·CHOICE 요소
ɔ	JV-AW	U+1182	GA THOUGHT 시범값
u	JV-U	U+116E	장모음의 구성 요소
ʊ	JV-FOOT	U+118D	FOOT 및 복모음 끝 요소
ʌ	JV-EO	U+1165	STRUT·RP NURSE 구성 요소
ə	JV-SCHWA	U+119E	약모음·RP GOAT 구성 요소

JV-KIT·JV-FOOT 등은 이번 시범 등록부의 식별자다. 식별자 철자 자체가 새 음가 판정은 아니다. 여러 음운 환경이 같은 기본 자모 요소를 쓰는 경우, 그 관계는 개별 mapping unit의 rule_id로 구분한다.

한 자모 부호를 거꾸로 읽는 것만으로 원래 음운 단위를 항상 알 수는 없다. U+1169는 RP /ɒ/의 단모음, GA /oʊ/의 첫 요소, 공통 /ɔɪ/의 첫 요소로 쓰인다. U+1166 역시 /ɛ/와 /eɪ/의 첫 요소에 쓰인다. 역변환은 모듈과 전체 음절핵 규칙을 필요로 한다.

4.2 장모음과 복모음

장모음과 복모음은 단일 부호 이름에서 분해하지 않고 제2본의 구성표에서 명시적으로 가져온다. /u:/는 [JV-U, JV-U], /ʊ/는 [JV-FOOT]이며 둘은 다르다. RP /ɔ:/는 [JV-O, JV-O], GA /ɔ/는 [JV-AW]다. 전체 대응표는 부록 A에 둔다.

등록부에는 codepoint, Unicode 이름, Hangul_Syllable_Type, 허용 위치, module_id, 규칙 상태와 근거를 함께 둔다. 등록되지 않은 음가를 가장 비슷한 자모로 추정하여 채우지 않는다.

[I2] 13-14·35-37·50-53쪽. 자모 ID와 registry 구조: 본 개정 설계안.

제2부 · Unicode 직렬화 규격

5. 직렬화의 기본 원칙

한 음절블록 = 한 자모 연쇄

지원되는 Broad 블록을 L+ V+ T* 형태로 직렬화한다. L은 하나 이상, V는 하나 이상, T는 없거나 하나 이상이다. 이것은 현 시범 조자법을 운반하는 전산 프로파일이며 최종 압축 자형의 완성을 뜻하지 않는다.

음가를 가진 초성군이 비어 있으면 JV-ZERO의 U+110B를 앞에 둔다. 뒤 중성의 저장 부호열에는 ◦을 넣지 않는다. 2026-09-17 저자 결정으로 읽기용 표시에만 무음 ◦을 붙인다. 무음 초성은 U+115F와 다르다.

층 2에는 이 프로파일의 등록부가 허용한 조합형 자모만 사용한다. 현대 완성자, 호환 자모, filler와 임의 제어문자는 허용하지 않는다. 이는 훈민바름의 저장 기준형을 정한 것이며, 해당 문자들이 Unicode에서 잘못된 문자라는 뜻은 아니다.

5.1 직렬화 단위와 지원 범위

직렬화 단위는 하나의 어휘 레코드다. 단어 안의 블록 사이에는 구분 부호를 넣지 않는다. 여러 단어·문장·구두점을 교환할 때는 레코드 배열이나 별도 문서 구조를 사용한다. [], 세로줄과 공백은 설명용 표에서 경계를 보일 수 있으나 정본 자모열에는 포함하지 않는다.

nucleus[]가 없는 음절자음 구조는 이 모음핵 기반 평문 프로파일의 지원 범위 밖이다. source와 syllabic 속성은 구조형에 남기고, 음절핵에 임의의 모음을 추가하지 않는다. 미배당 coda가 있으면 어휘 전체의 직렬화 상태를 unavailable로 돌려준다.

5.2 경계 판별

L을 읽는 동안 같은 초성군에 더하고, V가 나오면 음절핵을 연다. V 다음의 V는 같은 핵, T는 종성군이다. V 또는 T 다음의 L은 다음 블록의 시작이다. T 다음의 V, 첫 위치의 T, V 없이 끝나는 L 연쇄는 오류다. 이 판별은 부호열에 보존된 경계를 읽는 절차이며 영어 음절을 새로 분석하는 알고리즘은 아니다.

[I2] 8-10·26·36쪽; [E1] §3.12 D134. 프로파일 제한은 본 연구의 설계다.

6. 자리별 직렬화 규칙

6.1 초성군

초성군의 각 요소를 등록부의 L 위치형으로 순서대로 출력한다. friend의 초성군은 U+114B U+A976, play는 U+1111 U+1105다. /l/을 한 번만 출력하며 자음 사이에 모음을 넣지 않는다. /s/의 위치형 선택은 해당 모듈의 명시된 시범 규칙을 따른다.

6.2 음절핵

단모음은 하나의 V를, 장모음·복모음은 구성표에 등록된 V 배열을 출력한다. /eɪ/는 U+1166 U+119D, /aʊ/는 U+1161 U+118D다. 코드포인트 이름에 U-U나 O-O가 들어 있다고 하여 같은 문자의 두 번 반복과 동치로 해석하지 않는다.

6.3 종성군과 /s/

종성군은 T 위치형을 음소 순서대로 출력한다. bank는 U+11BC U+11BF, text는 U+11BF U+11BA U+11C0이다. /s/는 같은 음절블록의 독립된 종성 요소로 보존한다. 디지털 표시에서는 제2본의 선형 시범 원칙에 따라 패스처럼 /s/를 드러내며, 한국어 완성자 팻을 정상 HMBR 표시로 대체하지 않는다.

평문 pass(GA)의 U+1111 U+1162 U+11BA는 Unicode NFC에서 팻으로 합성될 수 있다. 그 부호상의 정준 동등성과 훈민바름이 요구하는 학습용 표시 규칙은 서로 다른 문제다. 저장에서는 기준형을 지키고 표시에서는 자모값과 /s/의 가시성을 지킨다.

6.4 /t+s/ 복합값

같은 종성군 안에서 인접한 /t/와 /s/가 등록된 복합값 규칙에 해당하면 JV-TS를 먼저 적용하고 U+11BE를 출력한다. cats는 /k æ t s/ 네 음소와 세 자모 요소를 가진다. 뒤따르는 자음이 있으면 그 요소를 이어 기록한다. 음절·어휘 경계를 가로질러 결합하지 않는다.

```
{ "id": "m3", "token_ids": ["t3", "t4"],
  "jamo_values": ["JV-TS"], "rule_id": "coda-ts",
  "reverse_ipa": ["t", "s"] }
```

/tʃ/를 /t/와 /ʃ/로 쪼개 이 규칙에 넣지 않는다. tokenizer가 인정하는 파찰음 표기 변이는 원문과 정규화 이력을 남기고 하나의 음소로 분석한다.

[I2] 25-30·58-59쪽; [E2] 정준 동등성. /s/ 표시와 직렬화의 역할을 분리한다.

7. 직렬화 갤러리

아래는 제2본의 예시를 현 시범 프로파일의 코드포인트 열로 옮긴 것이다. 코드포인트 사이의 공백은 읽기용이며, 실제 문자열에는 없다. 모든 행은 한 영어 음절이며 한 훈민바름 음절블록이다.

낱말	IPA	코드포인트 (각 항 U+)
pick	/pɪk/	1111 119D 11BF
pat	/pæt/	1111 1162 11C0
sing	/sɪŋ/	110A 119D 11BC
bank	/bæŋk/	1107 1162 11BC 11BF
milk	/mɪlk/	1106 119D 11AF 11BF
text	/tɛkst/	1110 1166 11BF 11BA 11C0
spin	/spɪn/	1109 1111 119D 11AB
friend	/frɛnd/	114B A976 1166 11AB 11AE
three	/θri:/	1145 A976 1175 1175
play	/pleɪ/	1111 1105 1166 119D
boy	/bɔɪ/	1107 1169 119D
boost	/bu:st/	1107 116E 116E 11BA 11C0
food	/fu:d/	114B 116E 116E 11AE

장모음·복모음의 두 V는 같은 핵의 저장 요소다. 읽기용 표시에는 뒤 V 앞에 무음 ㅇ을 붙인다. 이 표시용 자리는 음소·음절 수나 저장 부호열을 바꾸지 않는다. /s/의 위치형 경계는 시범이다.

[I2] 7-17·26-30쪽. 실제 표시 자형과 부호열은 별개로 검증한다.

7. 직렬화 갤러리 · 복합값과 모듈 대조

날말·모듈	normalized IPA	코드포인트 (각 항 U+)
cats RP·GA	/kæts/	110F 1162 11BE
boots RP·GA	/bu:ts/	1107 116E 116E 11BE
blue RP·GA	/blu:/	1107 1105 116E 116E
pass RP	/pɑ:s/	1111 1161 1161 11BA
pass GA	/pæs/	1111 1162 11BA
car RP	/kɑ:/	110F 1161 1161
car GA	/kɑɪ/	110F 1161 D7DB
go RP	/gəʊ/	1100 119E 118D
go GA	/goʊ/	1100 1169 118D
ask RP	/ɑ:sk/	110B 1161 1161 11BA 11BF
about RP·GA	/ə'baʊt/	110B 119E 1107 1161 118D 11C0
very RP·GA	/'vɛɹi/	1144 1166 A976 1175

세로줄은 이 표에서만 쓰는 편집 기호다. about과 very는 각각 두 블록이며 실제 부호열에는 세로줄이 없다. very의 출처형 /'vɛri/를 쓰는 자료는 source에 그대로 두고, 본 표에서는 분석에 사용하는 /ɹ/를 명시했다.

7.1 미결 레코드도 버리지 않는다

catch /kætʃ/는 source와 /tʃ/ 토큰, JV-CH와 coda 위치를 구조형에 보존할 수 있다. 다만 이 등록부에서 해당 종성 부호가 미배당이므로 완성 평문을 제공하지 않는다. nurse(GA)의 /ɜ:/는 기저 모음이 미결이므로 normalized IPA와 mapping 결과를 확정하지 않는다.

예시의 위계

제2본의 § 9.3·부록 A와 현 교정본 예시에 따라 ‘초성군의 ㄹ 한 번’을 적용한다. 이전 예시의 외관을 역으로 분석해 폐기된 중복 규칙을 되살리지 않는다.

[I2] 20–24·27·30·37·58쪽; [I5] Master 3, 138쪽.

8. 정규화와 동치

8.1 기준형은 NFD 조합형이다

이 프로파일은 등록된 조합형 자모 연쇄를 저장 기준형으로 삼는다. 기준형의 코드포인트 일치와 Unicode의 정준 동등성을 구분한다. 전자는 규칙이 고른 직렬화가 같은지, 후자는 Unicode 정규화 기준에서 같은 문자열인지를 검사한다.

입력·처리	본 프로파일의 판정
규격에 맞는 NFD 자모열	기준형 후보, registry·배열 규칙 추가 검사
NFC로 합성된 완성자 포함 입력	원입력을 보존한 뒤 NFD로 분해하고 검증하여 수용 가능
호환 자모·NFKC/NFKD 적용 입력	출처와 변환 이력을 확인. 원래 위치를 추정하여 자동 확정하지 않음
U+118D와 U+116E U+116E	서로 다른 값. 동치 치환 금지
U+1182와 U+1169 U+1169	서로 다른 값. 동치 치환 금지

NFC는 지원되는 일부 자모를 현대 완성자로 합칠 수 있으나 그 사실만으로 음운 정보가 소실됐다고 단정하지 않는다. 이 문서의 시험 자모열은 NFC를 거친 뒤 NFD로 되돌리면 원래 코드포인트 열이 복원됐다. 그러므로 수신 어댑터는 NFD 후 검증할 수 있다. 다만 source·강세·규칙 이력은 처음부터 평문에 없었으므로 정규화로 복원할 수 없다.

8.2 프로파일 의미 동치

등록 겹낱자와 낱자 연쇄의 의미 동치는 Unicode 이름이나 모양으로 자동 결정하지 않는다. 해당 모듈·버전이 동일한 값과 역규칙을 명시한 경우만 별도 별칭 관계로 둘 수 있다. 이 관계를 Unicode 정규화라고 부르지 않으며 UAX #15의 알고리즘을 바꾸지 않는다.

현재 기준형에는 겹낱자 별칭의 포괄적 치환을 도입하지 않는다. 따라서 등록부의 허용 목록에 없는 복합 부호를 받아 임의로 분해하지 않는다. 특히 /ʊ/와 /u:/의 대립을 지키기 위해 U-U라는 이름을 분해 지시로 사용해서는 안 된다.

[E2] UAX #15 §§1-2-6. 정준 정규화와 연구 프로파일의 의미 규칙은 구분한다.

9. 금지되는 연쇄와 오류 검출

검사기는 입력 오류를 발견하면 원입력과 위치를 보존한다. 미배당이나 미결을 다른 음가로 바꾸어 성공 처리하지 않는다. 수신 어댑터의 정규화와 기준형 검사를 구분하여, 고쳐 받은 입력이 원래부터 기준형이었던 것처럼 기록되지 않게 한다.

오류 ID	검출 대상	처리
E-CHAR	기준형의 완성자·호환 자모·filler·제어문자	원입력 보존, 수신 어댑터 여부 확인
E-ORDER	첫 T, T 다음 V, V 없는 블록	해당 위치의 직렬화 거부
E-REGISTRY	미등록 자모값·미배당 위치형	구조 보존, 평문 unavailable
E-REFERENCE	없는 token_id·mapping id·요소 번호	구조형 검증 실패
E-COVERAGE	입력 음소 누락·중복 소비, 출력 요소 누락	mapping 관계 재검토
E-BOUNDARY	구조형 경계와 자모열 경계 불일치	전체 레코드 일관성 실패
E-AMBIGUOUS	둘 이상의 역규칙 후보	후보 반환, 자동 단일화 금지
E-UNRESOLVED	/ʒ/·음절자음 등 미결 단계	원문·미결 사유 보존

9.1 구문만으로는 잡지 못하는 오류

음절핵 도중에 잘못 삽입된 L은 자모열만 보면 새 음절의 시작처럼 보일 수 있다. 따라서 정규식 검사만으로 구조가 맞다고 선언하지 않는다. 원 canonicalModel의 블록 수·자리 순서·토큰 대응과 다시 대조해야 한다.

한 mapping unit이 여러 토큰을 소비할 수 있지만, 같은 입력 토큰을 관계없는 두 mapping unit에서 중복 소비하지 않는다. 출력 요소 참조도 정확히 한 자리에 속한다. 무음 조성처럼 음소를 소비하지 않는 요소는 예외의 종류를 명시해야 한다.

9.2 성공 조건

complete 평문을 내보낼 때는 모든 블록과 모든 위치형이 직렬화 가능해야 한다. 부분 결과는 완성 평문 필드에 쓰지 않는다. 표시용 설명에서는 이미 처리한 부분과 미결 부분을 함께 보여 줄 수 있으나, 그 전체를 유효한 정보 문자열로 표시하지 않는다.

오류 ID와 검증 절차: 본 개정의 시범 규칙. 제2본의 미결사항을 자동 보정하지 않는다.

제3부 · 글꼴과 세이핑

10. 표시층의 지위

표시층은 정본을 읽게 만드는 층이다. 부호열의 글리프가 화면에 나타난다는 것, 음절블록이 하나의 표시 영역으로 묶인다는 것, 자모값이 식별된다는 것은 서로 다른 조건이다. 한 조건의 성공으로 다른 조건까지 충족했다고 판단하지 않는다.

displayRuns는 block_id, 자모 요소 참조, 표시 프로파일, 강세값과 render_status를 가진다. displayText는 그 명령을 문자열로 나타낼 수 있을 때의 파생 결과다. SVG나 canvas로 그리는 경우 문자열만으로는 실제 배치를 모두 설명할 수 없으므로, 자모 참조와 접근성 텍스트를 함께 유지한다.

11. 글꼴 요구사항

검사	확인할 것	그것만으로 알 수 없는 것
cmap	요구 코드포인트가 글리프에 연결되는가	결합 후 올바른 모양과 위치
글리프	/v/·/f/·KIT·FOOT 등 자형이 구별되는가	임의 복합 연쇄의 조판 성공
GSUB·GPOS	필요한 치환·배치 기능과 lookup	해당 연쇄에 실제 적용되는지
글꼴 로드	실제 선택된 파일과 버전	다른 환경의 동일 결과
라이선스	사용·수정·배포 조건	정본 규칙의 타당성

Microsoft의 Hangul 설명은 등록 조합을 처리하는 전처리와 ljmo·vjmo·tjmo 등의 적용을 설명한다. 기능 태그가 존재한다는 사실만으로 모든 L+ V+ T* 연쇄의 배치가 보장되는 것은 아니다. 실제 엔진·글꼴·적용 기능을 함께 시험한다.

11.1 이번 시험 글꼴

이번 재측정에는 NanumMyeongjoOTF-YetHangul.otf, Version 1.0.0.9 Build 20140917을 사용했다. 글리프 수는 26,030개이며 GSUB에는 aalt·ccmp·liga·ljmo·tjmo·vjmo가 있다. 이전 보고의 28,844개와 다른 파일이므로, 이번 결과를 이전 글꼴 파일의 동일 재현이라고 부르지 않는다.

실제 서비스에서는 글꼴 다운로드 성공과 폰트 선택, 코드포인트 범위, 필요한 위치형의 표시를 확인한다. 다른 글꼴의 자동 대체로 우연히 보인 결과와 목표 글꼴 자체의 지원을 구분한다.

[E3] Microsoft Hangul 설명. 파일 정보와 수치: 2026-09-08 직접 측정, 부록 C.

12. 셰이핑 요구사항

12.1 보존해야 할 것

셰이핑은 블록 안 자모의 순서와 음가 식별을 지키면서 표시 영역을 구성해야 한다. 블록 폭이나 모양을 맞추기 위해 /ʊ/를 /u:/의 부호로 바꾸거나 /t+s/를 /tʃ/로 읽게 하지 않는다. 여러 글리프로 한 블록을 그릴 수도 있으므로 글리프 수 하나를 합적 기준으로 삼지 않는다.

최종 압축 손글씨 배치형은 제2본의 미결 과제다. 그 상태에서 전산본이 임의의 압축 모양을 최종 정본으로 고정할 수는 없다. 현 단계에서는 블록 경계와 내부 순서가 보이는 선형 시범 표시를 지원하고, 압축 배치 후보는 별도 프로파일에서 비교한다.

12.2 세 종류의 경계

경계	결정 기준	시험
음운 음절 경계	선택된 언어모듈·사전·검토	canonicalModel과 source 분석 대조
텍스트 처리 경계	Unicode 분절 규칙과 앱의 명시적 조정	선택·삭제·커서 이동·줄바꿈
실제 자형 영역	글꼴·셰이핑·배치 프로파일	페이지·화면의 자모 식별과 배치

UAX #29의 GB6–GB8은 한글 연쇄 내부의 그래핌 경계를 다룬다. 이 경계는 원어의 음운 분석을 대신하지 않으며, 여러 자모를 한 네모로 배치하는 기능 그 자체도 아니다. 앱은 실제 커서와 삭제 동작을 따로 시험한다.

12.3 폭 측정의 역할

한 음절의 기준 폭으로 나눈 상대 폭은 분산 표시를 탐지하는 보조 지표다. 약 1자리 폭이라고 해도 자모가 겹쳐 안 보이거나 다른 음가처럼 보이면 실패다. 최종 표시 판정에는 자모 누락·겹침·읽는 순서·/s/ 가시성·모음 대립을 시각적으로 확인하는 절차가 필요하다.

표시 후보가 성공하더라도 canonicalModel과 기준 자모열은 유지한다. 프로파일 이름과 버전을 바꾸지 않은 채 새로운 축약 모양을 배포하지 않는다.

[I2] 26·36쪽; [E4] UAX #29 §3.1.1, GB6–GB8. 표시 요구: 본 연구의 제안.

13. 현재 환경에서 무엇이 합성되는가 · 재측정

2026-09-08, HarfBuzz 8.3.0과 §11.1의 OTF 파일을 사용해 문자열의 가로 advance 합계를 측정했다. 글꼴 좌표계는 1,000 units/em이며 기준 문자열 가(U+AC00)의 advance는 950이다. 상대 폭은 시험 문자열의 advance를 950으로 나눈 값이다.

입력	부호 수	글리프 수	advance	상대 폭
pick	3	3	973	1.0242
bank	4	2	1923	2.0242
boost	5	4	3869	4.0726
spin	4	4	1946	2.0484
milk	4	4	1946	2.0484
friend	5	5	2919	3.0726
play	4	3	2896	3.0484
boy	3	2	1923	2.0242
cats	3	1	950	1.0000

13.1 확인된 결과

pick은 세 글리프를 가지면서 약 1.02자리, cats는 한 글리프로 1자리 폭을 차지했다. 반면 bank·milk·spin·boy는 약 2자리, friend·play는 약 3자리, boost는 약 4자리다. 이 환경에서는 여러 자모를 포함한 연쇄가 하나의 압축된 표시 영역으로 모이지 않았다.

이는 이번 입력과 글꼴·엔진 조합에서의 폭 측정 결과다. 모든 조합형 자모, 다른 글꼴·운영체제·브라우저를 대표하는 수치가 아니다. 또한 상대 폭은 가독성·음성학적 타당성·교육 효과의 점수가 아니다.

13.2 이전 보고와의 관계

이전 41쪽 전산본의 그림과 보고도 비슷한 분산 표시를 보여 주지만, 글꼴 파일과 엔진 버전의 동일성이 입증되지 않았다. 따라서 새 측정값과 이전 보고를 혼합해 하나의 실험처럼 제시하지 않는다. §21의 사전 전수조사는 셰이핑 실험과 별개로 현재 저장소의 코드와 자료에서 재현했다.

이번 직접 측정, 글꼴 해시·실행 조건·재현 절차는 부록 C. 이전 보고: [I4] 17–20쪽.

13. 현재 환경에서 무엇이 합성되는가 · 대조

같은 환경에서 일부 자모 연쇄를 등록된 복합 부호로 바꾼 대조 입력도 측정했다. 이 입력은 글리프 처리의 차이를 살피는 실험용이며, 훈민바름 정본의 동치 표기로 승인한 것이 아니다.

대조 이름	실험용 부호열 (각 항 U+)	상대 폭
bank	1107 1162 11EF	1.0242
boost	1107 118D D7F1	1.0242
spin	113A 119D 11AB	1.0242
milk	1106 119D 11D8	1.0242
friend	114B A976 1166 11C6	2.0484

bank 대조에서는 종성 두 부호 대신 U+11EF, spin에서는 초성 두 부호 대신 U+113A, milk에서는 종성 두 부호 대신 U+11D8을 넣었다. friend는 겹종성 U+11C6만 적용했으므로 초성군 부분의 분산이 남았다.

boost 대조의 한계

U+118D는 이 연구에서 /ʊ/의 부호다. U+116E U+116E인 /u:/를 U+118D로 바꾼 대조 입력은 boost의 정상 표기가 아니다. 약 1자리 폭이 나왔다는 사실은 /u:/를 손실 없이 표현했다는 증거가 아니다.

13.3 이 실험으로 말할 수 있는 것

이 환경은 일부 등록 복합 부호가 포함된 입력을 약 1자리 폭으로 처리했다. 이는 동일한 엔진에서도 입력 부호의 구성이 표시 결과에 영향을 준다는 근거다. 그러나 임의의 미등록 조합을 현재 글꼴이 이미 그릴 수 있다거나, 셰이핑의 유일한 원인이 부호 배당이라는 결론은 나오지 않는다.

글꼴의 lookup, 엔진 전처리, 클러스터 인식과 최종 배치는 함께 검토해야 한다. 새 문자 부호를 배당하는 일과 기존 글꼴에 그 글리프를 설제하는 일도 별개다. 부호를 등록하면 즉시 모든 글꼴이 지원한다는 주장은 성립하지 않는다.

13.4 다음 시험

최종 배치 후보마다 원 자모의 식별과 읽기 순서를 확인하고, 원본 자모열 유지 여부와 복사 결과를 검사한다. 다른 글꼴·브라우저·운영체제의 시험은 제4본에서 환경별로 축적한다.

대조 입력과 폭: 이번 직접 측정. 의미 판정은 [I2]의 모음 대립과 §8의 동치 원칙을 따른다.

14. 대체 표시

대체 표시는 표시 실패 때도 source와 정본 구조를 읽을 수 있게 하는 출력이다. 블록을 합성하지 못했다고 영어 음절을 새로 나누지 않는다. 여러 자모가 선형으로 보이더라도 한 음절의 범위를 [] 로 묶고 IPA를 함께 제시한다.

상태	출력 원칙	기록
복합 블록 압축 불가	블록별 선형 자모·음소 순서 표시	render_status=linear
특정 글리프 없음	코드포인트·자모값·IPA 병기	missing_glyph와 해당 요소
종성 부호 미배당	구조와 IPA 설명, 완성 평문 없음	serialization.status=unavailable
원문 분석 미결	source IPA와 미결 사유 표시	status=unresolved

boost는 한 블록 [부우스트] 으로 표시한다. 2026-09-17 저자 결정에 따라 뒤 모음에 무음 ㅇ을 붙여 읽기 쉽게 한다. 우는 같은 /u:/의 뒤 요소이며 새 음절이 아니다. 저장형 L V V T T와 /b u: s t/는 그대로 보존한다.

14.1 정상 표시와 시범 표시

단순 블록이 잘 합성되면 composed로 기록할 수 있다. 복합 블록의 선형 시범형은 linear로 알린다. 미결인 압축 모양은 experimental 표시 프로파일로 구분한다. ‘옛한글 지원’이라는 한 문구로 이 상태들을 합쳐 알리지 않는다.

14.2 접근성

자모의 모양만으로 정보를 전달하지 않는다. source spelling·선택한 발음 모듈·IPA·블록 수·강세를 텍스트로 제공하고, 장모음·복모음은 핵의 구성과 길이를 설명할 수 있어야 한다. 색상이나 글자 크기만으로 상태나 강세를 전달하지 않는다.

화면낭독기가 훈민바름 문자열을 정확히 발음한다고 가정하지 않는다. 원문·IPA 읽기, 음성자료 연결과 구조 설명을 별도로 시험한다. 단순 선형 표시를 복사하려는 사용자의 선택은 허용하되, 정본 교환용 복사와 구별하여 어떤 형식이 전달되는지 알린다.

[I1] 접근성 원칙; [I2] 9-10·26쪽. display/fallback 세부 요구는 본 개정 설계안.

제4부 · 검색·복사·역변환

15. 검색

검색은 목적에 따라 키를 나눈다. 부호열 완전일치, 같은 음소 찾기, 같은 블록 구조 찾기는 서로 다른 질문이다. 어느 키든 모듈·등록부·규칙 버전을 포함하여 다른 분석을 우연히 합치지 않는다.

검색 종류	키	보존하는 구별
원문 검색	spelling, source language/script	원철자·원문 문자
출처 IPA 검색	source.ipa + source id	사전 표기 관행
음운 검색	normalized tokens + module/version	선택된 음운 대립
자모값 검색	mapping 결과의 자모 요소열	자모값·출력 순서
정본 부호열 일치	profile/version + NFD codepoints	기준 직렬화
구조 검색	onset/nucleus/coda와 prosody	음절·자음군·핵·강세
표시 별칭 검색	승인된 alias + profile/version	검색 편의, 별칭 출처

soon /su:n/의 자모값 키는 JV-S, JV-U, JV-U, JV-N이다. 여기에는 장모음 구성요소가 둘 있지만 음운 키에는 /u:/ 하나가 있다. 이 차이를 유지해야 ‘자모가 네 개인 형태’와 ‘음소가 세 개인 형태’를 각각 검색할 수 있다.

15.1 동음어와 동형 자모열

같은 발음을 가진 원철자는 여러 개일 수 있다. 따라서 자모열에서 단어 철자를 유일하게 복원한다고 선언하지 않는다. 또한 서로 다른 언어모듈이 같은 자모열을 다른 분석으로 사용할 수 있으므로 module_id 없는 검색 결과에는 분석 범위를 표시한다.

15.2 정규화와 검색 인덱스

입력 원문은 보존하고 별도의 검색 기준키를 만든다. NFC 입력을 NFD로 변환해 검색할 수 있으나, 원 source IPA를 그 과정에서 덮어쓰지 않는다. 표시 별칭은 별칭 검색에만 쓰고 정본의 의미 동치로 승격하지 않는다. 새로운 registry 버전은 기존 키를 재생성할 수 있어야 한다.

[I1] 원문·발음·데이터 연결 원칙; [E2] 정규화. 검색 키 분리는 본 개정의 시범 규칙.

16. 복사와 붙여넣기

복사는 사용자가 선택한 정보 수준을 전달해야 한다. 화면 문자열만 보내는 복사, 정본 자모열 복사, source와 구조 전체를 보내는 복사는 보존 범위가 다르다. 구현은 이를 명확히 구별한다.

형식	답는 것	한계
text/plain	지원되는 기준 자모열	source·강세·버전이 스스로 따라가지 않음
JSON 파일·지원되는 구조 형식	source·canonicalModel·규칙 버전·직렬화	수신 측의 스키마 지원 필요
text/html	표시 마크업·IPA·기준 자모열·버전 표시	수신 앱이 속성을 제거할 수 있음
읽기용 표시 복사	블록 경계·선형 표기·설명	정본 교환 형식으로 재해석하지 않음

16.1 기본 복사

평문을 생성할 수 있는 레코드의 정본 복사는 기준 자모열을 제공한다. 가능한 환경에서는 구조 형식을 함께 제공하되, 임의 MIME 형식을 모든 클립보드와 앱이 지원한다고 전제하지 않는다. 구조형 파일 다운로드를 위한 별도 경로다.

평문이 unavailable인 레코드에는 완성 자모열을 제공하는 것처럼 보이는 단추를 두지 않는다. 대신 구조형 내보내기 또는 IPA·미결 설명 복사를 제공한다. 표시 문자열에서 빠진 음소가 복사 과정에서도 조용히 사라지지 않게 한다.

16.2 부분 선택과 붙여넣기

정본 블록 복사는 선택 범위를 블록 경계까지 넓힐 수 있으며 그 동작을 사용자에게 알려야 한다. 자모 요소만 필요한 고급 복사는 fragment임을 명시한다. 조각을 완성 음절로 읽지 않는다.

붙여넣은 평문은 원입력을 보존한 뒤 NFD 변환과 프로파일 검사를 거친다. 형식·버전을 알 수 없거나 역규칙이 여러 개면 후보로 남긴다. HTML의 보조 속성이 사라졌다면 구조 전체를 복구했다고 선언하지 않는다. 복사 왕복은 실제 수신 앱을 포함해 시험한다.

클립보드 형식은 교환 설계 요구다. 실제 브라우저·앱 지원과 시험 결과는 제4본에서 보고한다.

17. 역변환 · 무엇을 되찾는가

역변환은 정보 수준별로 정의한다. `source` 레코드에서 원문을 다시 읽는 것, `canonicalModel`에서 선택된 음운 분석을 회수하는 것, 평문에서 가능한 분석을 구하는 것은 서로 다른 작업이다.

경로	되찾을 수 있는 것	되찾지 못하는 것
전체 구조 레코드	보존한 <code>source</code> , 토큰, <code>mapping</code> , 강세와 버전	원래 기록하지 않은 정보
기준 자모열 + 모듈·역규칙	지원 범위의 자모 구조와 유일한 음운 후보	출처 표기·강세·분석 이력·원철자
기준 자모열만	부호와 L/V/T 구조	모듈을 모르면 음운 해석을 확정할 수 없음
표시·대체 문자열	보이는 정보의 참고적 해석	완전한 정보 복원 보장 없음

17.1 구조형 왕복

JSON을 쓰고 다시 읽었을 때 배열 순서, 문자열, 필드의 `null`과 생략, 상태값과 참조 관계가 보존되는지 검사한다. JSON 키의 직렬화 순서가 바뀐 것과 데이터 의미가 바뀐 것은 구분한다. `source IPA`의 원문 복원은 보관된 `source.ipa`를 읽는 절차다.

17.2 평문 왕복의 전제

부호열을 블록으로 나누고 위치형을 자모 요소로 되돌린 뒤, 모듈의 전체 음절핵·복합값 역규칙을 적용한다. 등록부가 단일 부호마다 하나의 자모값을 가리킨다는 조건만으로 충분하지 않다. 장모음·복모음은 여러 부호를 함께 읽으며 /t+s/는 한 부호에서 두 음소를 회수한다.

지원 범위에서 하나의 자모 구조가 둘 이상의 `normalized` 음운 분석을 허용하면, 반환값은 `ambiguous`와 후보 목록이다. 임의로 첫 후보를 고르지 않는다. 이 초안의 시험 예시가 모두 유일하더라도 영어 전체의 유일 역변환을 입증한 것은 아니다.

[I2] 13–14·30·37·50–53쪽의 대응에서 도출한 전산 요구. 완전 보존은 구조 레코드를 전제로 한다.

17. 역변환 · 복합 대응과 제한

17.3 boost와 cats

입력	구조 회수	음운 회수
boost의 L V V T T	B / U,U / S,T	/b/ + /u:/ + /s/ + /t/
cats의 L V T	K / AE / TS	/k/ + /æ/ + /t/ + /s/
play의 L L V V	P,L / E,KIT / Ø	/p/ + /l/ + /eɪ/

boost의 두 U는 long-u 역규칙으로 /u:/ 하나가 된다. 이를 /u/ 두 음소라고 읽지 않는다. cats의 TS는 coda-ts 규칙으로 /t/와 /s/ 두 음소가 된다. play의 초성 L은 한 번만 회수한다.

17.4 같은 기본 요소, 다른 음절핵

U+1169 U+119D는 공통 CHOICE 규칙에 따라 /ɔɪ/다. U+1169 U+118D는 GA GOAT 규칙에 따라 /oʊ/다. 단독 U+1169가 RP에서 /ɒ/를 나타낸다고 하여 모든 연쇄의 첫 요소를 /ɒ/로 읽지 않는다. 역규칙은 부록 A의 전체 핵 단위로 조회한다.

U+1165 U+1165는 현 RP NURSE 규칙에서 /ɜ:/다. U+1165 단독의 /ʌ/를 두 번 출력해 해결하지 않는다. 같은 원리로 모듈의 음운 단위와 자모 구성요소를 분리해야 한다.

17.5 복원하지 않는 정보

about의 평문에는 강세기호가 없으므로 /ə'baʊt/의 제2음절 강세를 평문만으로 되찾았다고 말할 수 없다. 규칙으로 음소 분절열 /əbaʊt/를 회수할 수 있더라도 source의 기호 배치와 출처는 구조 레코드가 보존한다.

Narrow [ɹ]가 보이더라도 기저 음소가 /t/인지 /d/인지는 surface phone만으로 정해지지 않을 수 있다. underlying token을 저장해야 한다. 연결발화 표면형 역시 사전형을 자동으로 유일하게 복원하는 자료가 아니다.

왕복 적합성의 보고 단위

원문 문자열, normalized 음운 토큰, 자모 구조, 코드포인트 열, 표시 결과를 각각 검사한다. 서로 다른 정보 수준을 ‘무손실’ 한 단어로 묶지 않는다.

18. 교환 형식

완전 보존의 기본 교환 단위는 `source`와 `canonicalModel`을 포함하는 구조 레코드다. 직렬화는 전체 정보가 아니라 함께 제공하는 파생 값이다. 받는 쪽은 해당 버전을 지원하는지 확인하고, 구조에서 다시 만든 자모열이 함께 온 직렬화와 일치하는지 검증한다.

필수 묶음	내용
식별·버전	<code>schema_version</code> , <code>module_id</code> , <code>registry_version</code> , 규칙·프로파일 버전
<code>source</code>	원문, <code>source IPA</code> , 언어·문자·출처와 이용조건
<code>analysis</code>	<code>normalized IPA</code> , 정규화 이력, 음절분석과 판정 근거
<code>canonicalModel</code>	음운 토큰, <code>mappingUnits</code> , 블록, 운율·Narrow와 상태
<code>serialization</code>	상태, <code>profile</code> , 기준 <code>codepoints</code> 또는 <code>unavailable</code> 사유
<code>validation</code>	검사 버전, 오류·미결 목록, 검사 범위

18.1 코드포인트 배열과 문자열

`codepoints[]`는 Unicode 스칼라값을 정수 배열로 담는다. 이 프로파일은 허용된 자모만 포함하므로 각 값이 `registry`와 L/V/T 자리 규칙을 충족해야 한다. `text` 필드를 함께 두면 `codepoints`로 만든 문자열과 같은지 검사한다.

정규화를 강제하는 전송 환경에서는 숫자 배열 또는 명시된 이스케이프 전송을 쓸 수 있다. JSON이라는 포장만으로 내부 문자열의 정규화가 자동 방지되는 것은 아니다. 수신 측에서 재구성 후 검증해야 한다.

18.2 불일치와 버전 호환

구조형과 함께 온 자모열이 다르면 수신자는 한쪽을 조용히 정본으로 채택하지 않는다. 원 레코드와 불일치를 보존하고, 선택한 `registry`로 재생성한 결과를 별도 후보로 기록한다. 알 수 없는 필드를 버리는 구현은 완전 왕복을 주장하지 않는다.

교환 과정의 `content hash`는 전송 손상을 확인하는 수단이다. 그것이 음운 분석의 타당성이나 출처 라이선스를 입증하지는 않는다. 동일 레코드의 바이트 해시가 필요하다면 JSON 직렬화 방식까지 고정해야 한다.

교환 봉투와 필드명은 이번 개정의 시범 설계다. 현재 웹앱이 이미 이 형식을 사용한다고 주장하지 않는다.

제5부 · 국가표준·국제표준 정합성

19. 이미 표준 안에 있는 것

정밀 표기층을 연구하려면 문자부호, 음절의 구조, 표시 기술과 적용 규범을 나누어 살펴야 한다. 문자가 Unicode에 있다는 사실은 표현의 출발점이다. 모든 언어의 음가 대응이나 모든 환경의 상호운용성이 이미 해결됐다는 뜻은 아니다.

19.1 Unicode 음절블록과의 구조적 정합성

Unicode 17.0 §3.12 D134는 표준 한국어 음절블록을 하나 이상의 L, 하나 이상의 V, 0개 이상의 T가 이어지는 연쇄 또는 그와 정준 동등한 연쇄로 정의한다. 본 프로파일의 L+ V+ T* 구조는 이 정의의 조합형 구조와 정합한다. 다만 Unicode의 정의가 훈민바름의 영어 음운 분석과 음가 배당을 승인하는 것은 아니다. [E1]

또한 이 정의는 실제 글꼴이 임의의 복합 연쇄를 읽기 좋은 한 블록으로 그린다는 보증이 아니다. 음절 경계, 자형 합성과 복사·검색의 상호운용성은 별도로 검증해야 한다.

19.2 사용 범위

Unicode 블록	범위	본 프로파일
Hangul Jamo	U+1100–U+11FF	등록부의 L/V/T 위치형만 허용
Hangul Jamo Extended-A	U+A960–U+A97F	등록된 초성 위치형
Hangul Jamo Extended-B	U+D7B0–U+D7FF	등록된 중성·종성 위치형
Hangul Syllables	U+AC00–U+D7A3	수신 어댑터에서 NFD 후 검사
Hangul Compatibility Jamo	U+3130–U+318F	설명·날자 표시용, 기준형에는 배제

범위 전체를 허용 문자 목록으로 쓰지 않는다. 미할당 코드포인트, 미등록 위치형과 filler는 별도 검사한다. Unicode 이름은 문자 식별을 위한 자료이며 HMBR 음가를 자동으로 계산하는 규칙이 아니다.

표준의 역할 구분

Unicode 적합성, KS 적용 프로파일과의 정합성, HMBR 규칙 적합성, 실제 표시 적합성은 각각 판정한다. 어느 한 층의 성공이나 제약을 나머지 모든 층의 결론으로 확대하지 않는다.

20. KS X 1026-1에서 확인한 조항

본 절의 직접 확인 자료는 한국 측 국제회의 제출문서 N3422에 실린 KS X 1026-1:2007 영문 번역본이다. 이 자료는 국문 정식 표준이 아니며, 국문판과 차이가 있으면 국문판이 최종 권위임을 밝힌다. 아래 절 번호와 요지는 이 번역본 기준이다. [E5]

번역본의 절	확인한 요지	검토할 관계
§5.1, §5.3(1)	복합 낱자를 단순 낱자 부호 여러 개의 연쇄로 나타내는 것을 제한	복수 L·V·T를 사용하는 확장 프로파일
§5.2, §5.3(3)	현대 한글 음절의 외부 교환은 완성형 사용	일관 조합형 기준 저장 방식
§5.3(2)	완성 음절과 조합형 낱자를 합쳐 다른 블록으로 표현하는 것을 제한	NFC 입력·NFD 기준형과의 관계
§6	정규화와 별도 권고 알고리즘	UAX #15와 적용 프로파일의 구분

20.1 연구 질문의 범위

이 조항들을 외국어 정밀 표기 프로파일에도 그대로 적용할 경우, 하나의 음절 안에서 여러 자모 요소를 보존하는 방식을 어떻게 수용할지 검토가 필요하다. 특히 등록된 복합 부호의 사용 관행과, 같은 기본 요소의 연쇄에 독립된 음가를 부여하는 연구 프로파일 사이에는 명시적 조정이 필요하다.

현행 정보교환 프로파일이 허용하는 범위만 적용하면, 이 연구가 요구하는 복수 자모 요소의 독립 음가와 순서를 완전하게 표현할 길이 없다. 그러나 그것이 곧 법적 금지를 뜻하지는 않는다. 실제 구속력·기관별 채택·현행 국문판 적용은 별도 확인 대상이며, 이 절은 정보교환 프로파일의 표현 범위와 정합성에 한정한다.

20.2 개정 전에 확인할 사항

현행 국문 정식본과 개정·확인 이력을 대조하고, 정밀 외국어 표기층을 기존 정의에 포함시킬지 부속 프로파일로 돌지 검토한다. 국제표준의 정규화 정의를 바꾸지 않으면서 허용 표현과 교환 조건을 명시할 수 있는지 확인한다. 그 결과에 따라 본문 개정, 부속서 또는 독립 응용 프로파일 가운데 적절한 형식을 선택한다.

[E5] PDF 5쪽의 번역본 지위, 12~14쪽의 관련 조항. 개정 방향과 적용 해석은 본 연구의 제안이다.

21. 훈민바름 연쇄가 막히는 지점

21.1 확실한 사례부터 분리한다

첫째는 자모값·위치형의 미배당이다. /θ/ 등의 종성형 문제는 현재 연구 등록부의 과제이므로, 이를 모두 국가표준의 제약으로 돌리지 않는다. 둘째는 구성 자모가 있어도 원하는 연쇄의 표시와 교환이 충분히 지원되지 않는 경우다. 셋째는 복합 부호와 구성 연쇄를 자동 동치로 취급할 때 음운 대립이 사라지는 경우다.

§13은 둘째 문제를 한 환경에서 직접 측정했다. /ʊ/와 /u:/, GA /ɔ/와 RP /ɔ:/의 부호 대비는 셋째 문제를 설명한다. 이 사례들은 대규모 사전의 비율을 확정하지 않아도 프로파일의 동치 규칙을 신중하게 설계해야 한다는 근거가 된다.

21.2 사전 전수조사의 재현 결과

자리·모듈	종류 총계	부호 있음	부호 없음	미배당	복합값 /t+s/
초성 RP	207	16	190	1	0
초성 GA	123	11	112	0	0
종성 RP	447	38	326	82	1
종성 GA	303	31	208	63	1

2026-09-16 저장소의 두 사전과 tools/cluster_census.mjs로 다시 계산했다. RP는 초성군 207종(부호 있음 16·없음 190·미배당 1), 종성군 447종(38·326·82·복합값 /t+s/ 1)이다. GA는 초성군 123종(11·112·0), 종성군 303종(31·208·63·복합값 /t+s/ 1)이다. 종성 표의 과거 소계가 1종 작았던 것은 /t+s/를 구성 연쇄가 아니라 복합값 ㅈ으로 배당해 세 범주 대조에서 제외했기 때문이다.

재현 결과와 해석 범위

완전 해소에 필요한 새 겹낱자는 RP 516종, GA 320종이며 빈도 상위 60종의 누적 해소율은 각각 90.4%, 91.8%다. 2요소 음절핵은 14종이고, 코드포인트 이름상 후보는 3종이지만 U+118D와 U+1182는 이미 각각 /ʊ/와 GA /ɔ/에 배당돼 자동 동치로 쓸 수 없다. 이 수치는 현 사전과 전처리 규칙의 재현 결과이며 영어 전체의 영구적 하한은 아니다.

어두·어말 추출은 고유명사·차용어·오류·음절자음·구두점과 각 사전의 RP/GA 표기 관행에 영향을 받는다. 따라서 결과에는 사전 파일, 실행 코드, 분모·제외 조건과 실행일을 함께 인용한다. 저장소에 포함된 사전의 원출처·판·재배포 조건은 별도 확인 대상으로 남긴다.

재현 자료: data/dict-rp.json, data/dict-ga.json, js/registry.js, tools/cluster_census.mjs. 실행: node tools/cluster_census.mjs. 확인일: 2026-09-16. 기존 보고: [I4] 25–28·35–40쪽.

22. 두 갈래의 해법과 그 비용

22.1 갈래 A · 필요한 복합 부호를 검토한다

필요한 조합에 독립 부호를 배당하는 길은 문자 부호화 제안, 용례·구별 필요성 검토와 글꼴·엔진 지원을 함께 요구한다. 기존 부호를 활용할 수 있어도 해당 부호의 자형·문자적 정체성과 HMBR 의미의 적합성을 검토해야 한다.

새 부호가 등록되었다고 기존 글꼴이 즉시 그릴 수는 없다. §13의 대조는 이미 있는 일부 복합 부호를 처리한 결과다. 아직 없는 부호에 대한 구현 비용이 없다는 증거로 사용할 수 없다.

22.2 갈래 B · 낱자 연쇄를 보존하는 프로파일

구성 자모열을 기준형으로 유지하고 블록 경계·교환·표시 규칙을 별도로 정하는 길이다. 연쇄의 음운 대응과 직렬화 구조를 명시하므로, 새 언어를 추가할 때 공통 스키마와 처리 도구를 재사용할 수 있다. 다만 새 음가 배당, 언어별 분석과 배치·사용자 시험은 여전히 필요하다.

비교 항목	갈래 A · 복합 부호	갈래 B · 구성 연쇄
추가 절차	문자 부호화·글꼴·엔진 검토	프로파일·배치·상호운용성 구현
확장 비용	새 조합의 문자 식별 검토	새 모듈의 분석·registry·표시 시험
기존 환경	기존 부호의 일부는 지원	지원되지 않는 연쇄는 fallback
의미 동치	자동 인정하지 않음	자동 인정하지 않음
공통 조건	음가·순서·출처와 버전 보존	음가·순서·출처와 버전 보존

22.3 이 연구의 권고

구성 연쇄와 구조형을 보존하는 갈래 B를 우선 연구한다. 갈래 A는 독립된 문자적 필요성과 음가 보존이 입증된 항목에 한해 검토한다. 어느 경로든 단일 복합 부호와 연쇄의 동일성은 명시적 registry 관계로만 허용하며, Unicode 정규화 알고리즘을 바꾸지 않는다.

갈래 비교와 권고는 본 연구의 설계·정책 제안이다. 새 부호의 등록 가능성이나 일정은 확정하지 않는다.

23. 개정 제안

제1본이 제안한 것은 일반 외래어 표기의 기능과 통일성을 유지하면서, 원어 음가를 정밀하게 기록할 분야에 별도의 연구·적용 공간을 마련하자는 것이다. 아래 제안은 그 공간에서 데이터를 저장하고 교환하는 기술적 조건을 다룬다.

과제	제안 방향	보존할 조건
구성 연쇄의 표현	확장 프로파일에서 복수 초·중·종성 연쇄를 명시적으로 허용	원래 음가·순서·블록 경계
기준형의 선택	일반 조합형 교환 프로파일과 수신 정규화 절차를 문서화	Unicode 정준 동등성
의미 동치의 범위	registry가 승인한 경우만 프로파일 별칭 허용	서로 다른 음운 대립
표시·편집 요구	블록 처리·대체 표시·복사·접근성 시험을 포함	미지원 상태의 투명성

23.1 적용 범위

확장 프로파일은 교육·사전·학술·고유명사·접근성·언어기술 등 원어 정보 보존이 필요한 용도를 대상으로 한다. 일반 한국어 문서에 새로운 표기나 조판을 강제하지 않는다. 또한 훈민바름의 특정 자모 배당을 국가표준의 유일한 값으로 채택하도록 요구하지 않는다.

23.2 표준과 응용 규칙을 분리한다

국가표준 또는 부속 프로파일에는 허용 문자열·교환 조건·상호운용성·검증 방법을 중심으로 최소 공통 요구를 둔다. 특정 언어의 음가·자모 대응과 교육용 표시 수치는 언어모듈·응용 프로파일에서 다룬다. 기술중립성을 지키기 위해 특정 글꼴이나 특정 앱을 적합성의 유일한 기준으로 삼지 않는다.

본 제안은 일반 표기를 흔들지 않고 정보 보존이 필요한 영역의 적용 범위를 합리화하는 방향이다. 개정 형식과 대상 조항은 현행 국문판의 확인 및 공공 검토를 거쳐 정한다.

[I1] 23-24·29-33쪽의 공공성·단계적 표준화 원칙을 전산 규칙으로 구체화한 제안.

23. 개정 제안 · 문언 예시

아래 문언은 채택된 표준 조항의 인용이 아니라, 공개 검토를 위한 연구자의 예시안이다. 현행 국문판의 절 번호와 문체에 맞추는 작업은 별도로 남긴다.

확장 표기 프로파일의 자모 연쇄와 처리

- ① 정밀 음가 기록을 위한 확장 표기 프로파일은 해당 프로파일이 정한 자모값의 순서와 음절 경계를 보존하기 위하여, 하나의 음절블록 안에 복수의 초성·중성·종성 조합형 자모를 차례대로 표현할 수 있다.
- ② 프로파일은 허용 자모, 기준 직렬화, 경계 판별, 버전 식별과 수신 시 정규화 절차를 명시하여야 한다. Unicode의 정규화 정의를 변경하는 방식으로 별도의 문자값을 정하여서는 안 된다.
- ③ 단일 복합 부호와 구성 낱자 연쇄의 의미 동치 여부는 부호의 이름이나 자형만으로 정하지 않는다. 해당 프로파일이 두 표현에 같은 값과 역규칙을 명시적으로 부여하고 음운 대립의 보존을 검증한 경우에만 프로파일 내부의 동치로 인정한다.
- ④ 처리계는 하나의 정보 음절블록을 정보처리 단위로 보존하고, 표시 프로파일에 맞춰 구성 요소를 식별 가능하게 나타내야 한다. 압축 합성을 지원하지 못하면 블록 경계·선형 표시·IPA 또는 대체텍스트를 제공하고 지원 상태를 밝혀야 한다.
- ⑤ 정보 교환 형식은 표시 결과와 분리하여 보존하여야 한다. 표시를 위한 치환이나 축약은 원문, 자모값, 음소 순서와 규칙 버전을 덮어써서는 안 된다.
- ⑥ 본 프로파일의 적용은 일반 한국어 문서의 표기와 교환 방식을 일률적으로 변경하지 않으며, 다른 연구 표기체계에도 같은 검토 조건을 적용한다.

문언의 효과와 한계

①은 연쇄 표현의 경로, ②는 교환 조건, ③은 음가 대립의 보호, ④-⑤는 실제 사용자에게 전달되는 정보의 보존을 담당한다. 이 조항들을 제안한다고 해서 글꼴 합성과 음가 배당이 자동으로 완성되는 것은 아니다. 적합성 시험과 지원 범위를 함께 공개해야 한다.

24. 국제표준과의 관계와 이행 경로

24.1 2009년 의견서의 의미

미국과 Unicode Consortium의 2009년 공동 의견서 L2/09-151은 KS X 1026-1:2007을 ISO/IEC 10646에서 참조하려는 제안에 대해, 영문판 제6절의 별도 정규화 권고와 Unicode·ISO/IEC 10646의 정합성 우려를 제기했다. 국문 원문을 검토하지 못했다는 한계와, 정규화가 의무가 아닌 상황의 교환 권고라면 다른 해석이 가능하다는 점도 함께 밝혔다. [E6]

따라서 이 자료는 당시 정규화 논점이 실제로 제기되었다는 역사적 근거다. 그 우려가 2026년에도 동일한 문언으로 남아 있다거나 이미 법적 위반이 확정됐다는 증거로 사용하지 않는다.

24.2 단계적 이행

단계	산출물	다음 단계의 조건
공개 응용 프로파일	정본·registry·스키마·시험 벡터	독립 구현과 오류 제보가 가능함
기술보고서	환경별 계측·왕복·접근성 결과	재현 자료와 부정적 결과 공개
단체·공공 검토	최소 공통 규격과 용도별 적용 범위	음운·교육·문자·정보기술 검토
국가표준 검토	현행 국문판 대조와 개정·부속서안	운용 근거와 상호운용성 검증

제3본은 응용 프로파일과 검증 요구를 세우는 단계다. 제4본은 구현 결과와 그 한계를 기록한다. 연구 규격의 공개는 상위 표준 개정을 기다리지 않고 진행할 수 있으나, 표준 적합성을 선언할 때는 적용 판본·프로파일·예외를 밝혀야 한다.

24.3 공개 자료의 안내

공개 저장소 주소는 <https://github.com/ivanpark/hmbr> 이다. 제3본·제4본, 웹 참조구현과 시험자료의 최신판은 그 공개 경로에서 관리한다. 현재 배포 파일과 웹앱 동작의 감사 결과는 제4본과 2026-10-09 네 본 상태 보고서를 참조한다. 이 전산본은 설계 규격과 문서 내부의 일관성을 다룬다.

[I1] 29-31쪽; [I3] 2-3쪽; [E6] 2009년 공동 의견서 전문. 일정과 현재 배포 상태는 별도 확인 대상.

제6부 · 적합성과 판정

25. 적합성 등급

적합성은 선언한 모듈·프로파일·지원 범위 안에서 판정한다. 다음 세 등급은 시범 분류다. 상위 등급은 하위 요구를 포함하지만, 범위 밖 미결 항목까지 지원한다는 뜻은 아니다.

등급	요구	필수 보고
1 · 보존 적합	source와 구조형 보존, 유효한 참조 관계, 직렬화 가능 여부의 정확한 판정	미배당·미결 항목과 지원 범위
2 · 표시 적합	1등급 + 블록·음가 식별, 미지원 표시와 접근성 정보	압축 합성 범위와 linear/fallback 범위
3 · 왕복 적합	2등급 + 선언 범위의 구조형 왕복·평문 역변환·검색·복사 검증	복원 정보 수준, 후보·실패 및 실제 수신 환경

선형 대체 표시를 정확히 제공하는 구현은 해당 범위의 표시 요구를 충족할 수 있다. 그러나 이를 복합 블록의 압축 합성 성공이라고 부르지 않는다. 보존 적합에서도 미결 레코드를 조용히 버리거나 음소를 바꾸면 실패다.

25.1 적합성 선언의 최소 항목

규격·schema·registry·모듈 버전, 입력 데이터셋과 범위, 지원·제외 항목, 시험 결과, 글꼴·엔진·플랫폼 버전, 검사 일자와 재현 자료를 함께 공개한다. ‘모든 테스트 통과’라는 문구에는 어떤 테스트와 어떤 분모인지를 붙인다.

25.2 서로 다른 검증을 섞지 않는다

이 문서를 작성하면서 JSON과 예시 코드포인트를 검사한 결과는 문서의 내부 일관성 검사다. 그것이 웹앱의 3등급 적합성이나 전체 영어의 왕복 성공을 의미하지 않는다. 음성학적 타당성, 학습 효과와 사회적 수용성 역시 별도의 검증이다.

미결 항목을 처리하지 못한다는 사실보다, 처리하지 못한 범위를 숨기거나 다른 음가로 대체하는 것이 더 큰 보존 실패다. 적합성 보고는 성공과 함께 범위와 실패를 읽을 수 있게 해야 한다.

26. 시험 벡터와 27. 버전·변경 관리

26.1 시험 묶음

묶음	확인할 내용	예시
기본·자음군	L/V/T 순서·블록 수·자모 순서	pick, bank, friend, text
복합 음절핵	1:N 대응과 단모음 대립	boost, boy, food, full/fool 핵
복합값	N:1 대응과 두 음소 회수	cats, boots
일반 초성군	/l/ 한 번 기록	play, blue
모듈 대조	RP·GA의 서로 다른 기준형	pass, car, go
미결·미배당	원문 보존과 거부 상태	/ɜː/, catch의 종성 /tʃ/
정규화	NFC 수신과 NFD 기준형 회수	pat, bank, pass(GA)
구조 무결성	누락·중복·없는 참조·경계 불일치 검출	토큰·mapping·블록 교차 검사

부록 B의 양성 벡터와 음성 벡터를 함께 사용한다. 양성 예시만 통과한 것을 오류 검출의 증거로 삼지 않는다. 아주 짧은 자모열의 성공을 사전 전수조사 성공으로 확대하지 않는다.

27.1 버전의 대상

문서판, schema, registry, 언어모듈, 직렬화 프로파일, 표시 프로파일과 시험셋의 버전을 각각 기록한다. 문서 설명 수정과 음가 배당 변경은 영향이 다르다. 미결이 잠정 고정으로 바뀌더라도 기존 출력·역해석이 달라지면 호환성 변경으로 취급한다.

27.2 이행 원칙

구판 데이터는 source에서 새 분석을 다시 만들 수 있는지 먼저 확인한다. 원 source를 보존하고 이전 규칙과 새 규칙의 차이를 기록한다. 구판의 filler나 중복 위치를 단순 삭제하는 방식으로 이행하지 않는다. source가 없어 음소를 확정할 수 없으면 수동 검토 대상으로 남긴다.

공개 배포에는 변경 이유, 영향받는 레코드, 새 시험과 이전판 읽기 절차를 포함한다. 여기서 제안한 draft-2026-09-08 식별자는 제품의 정식 배포 버전 번호가 아니다.

28. 잠정 고정·시범·미결 목록

28.1 제2본에서 받는 기준

한 영어 음절 = 한 음절블록, 내부 음소 순서, /l/ 한 번, /tʃ/와 /t+s/의 구별, /ɪ/·/ʊ/·/ə/의 독립 대립, source IPA 보존을 따른다. 자모값의 음가를 이번 전산 설계의 편의를 위해 바꾸지 않는다.

상태	이 개정에서의 처리
잠정 고정 원칙	음운 토큰과 자모 요소 분리; 1:1·1:N·N:1 보존; 층 1과 source의 권위; 표시와 저장 분리
시범 전산 프로파일	L+ V+ T* 기준형; NFD 수신·저장 절차; mappingUnits 필드; 오류 코드; 교환 형식; 적합성 등급
시범 표시	복합 핵의 뒤 모음에 읽기용 무음 초성 ㅇ을 붙이는 선형 표시(저장 V 배열에는 미포함), /s/ 위치형 경계, GA /ɔ/ 값, 강세 상대 크기.
미결 자모·분석	/ɜ/ 기저 모음, /j/+V·/w/+V 최종 결합, [ʔ], 음절자음 최종 처리
미결 전산·표시	미배당 종성 부호, 복합 압축 자형, 동치 별칭의 개별 승인, 다중 환경 적합성

이 상태표는 초안의 적용 범위를 밝히는 것이다. 새 전산 필드와 프로파일이 이미 저자의 최종 승인을 받았다는 뜻이 아니다. 제2본이 시범·미결로 둔 항목은 이 문서에서도 해당 상태를 유지한다.

28.2 제4본으로 넘기는 것

제4본은 이 문서의 registry와 schema를 실제 데이터·코드로 옮기고, source 정규화와 음절분석, 직렬화·역변환, 표시·복사·검색·접근성을 시험한다. 사전 전수조사는 출처·라이선스·원파일·실행 버전과 함께 다시 수행한다. 앱 화면과 실제 동작의 차이는 그 구현본에서 보고한다.

그 결과가 현재의 조자법·음가 배당을 다시 검토해야 함을 보이면 제2본의 공개 판정 절차로 되돌린다. 기술 구현이 정본을 몰래 바꾸거나, 정본의 미결사항을 구현 성공으로 덮지 않는다.

[I2] 35–37·58–59쪽; [I5] Master 3, 137–140쪽. 새로운 전산 항목은 검토용 설계안.

맷음말 · 부호는 정본을 섬긴다

훈민바름 전산본의 첫 과제는 더 많은 글자를 화면에 보여 주는 일이 아니다. 어떤 원문의 어떤 소리를 어떤 분석과 규칙으로 기록했는지 잃지 않는 일이다. 정본의 자모값, 원어의 음운 단위, Unicode 부호열과 화면 자형을 나누는 이유도 여기에 있다.

하나의 /u:/가 두 모음 요소로 적히고, /t/와 /s/ 두 음소가 하나의 복합값으로 적힐 수 있다. 이 관계를 데이터에 담지 않으면 화면은 그럴듯해도 원래 소리를 다시 읽는 과정에서 정보가 달라진다. 음운 토큰과 자모 요소를 분리하는 모델은 이런 차이를 감추지 않기 위한 장치다.

한 영어 음절을 한 블록으로 보존하는 일도 같은 문제다. 영어 자음군과 복합 음절핵이 현재 글꼴에서 여러 자리로 나뉘어 보이더라도, 그 표시 사정 때문에 원어에 없는 음절을 만들어서는 안 된다. 압축된 최종 자형이 아직 미결이라면 블록 경계와 내부 순서, IPA를 통해 현재의 한계를 정직하게 보여 주어야 한다.

국가표준의 검토 역시 같은 태도를 요구한다. 국제 문자부호의 구조적 가능성, 국가표준의 적용 범위, 개별 글꼴과 엔진의 구현 한계를 하나의 원인으로 뭉뚱그리지 않는다. 직접 확인한 조항과 측정값을 밝히고, 아직 확보하지 못한 원자료와 미결사항도 함께 남긴다. 그래야 연구가 다른 사람의 검토와 재구현으로 이어질 수 있다.

이 연구가 요청하는 것은 특정 표기체계의 독점적 채택이 아니다. 일반 외래어 표기의 역할을 지키면서, 원어 음가를 정밀하게 보존할 필요가 있는 분야에서 한글의 역사적 자원과 조자 원리를 공개적으로 시험할 수 있는 공간이다. 그 공간에는 훈민바름뿐 아니라 다른 연구와 대안도 함께 들어올 수 있어야 한다.

정본은 사람이 적고 읽는 원리를 세우고, 전산본은 그 원리를 잃지 않게 보존하며, 구현본은 실제로 얼마나 지켰는지 밝힌다. 세 단계가 서로의 한계를 드러내고 수정할 수 있을 때, 가표본은 고정된 문자 목록을 넘어 계속 검증하고 발전시킬 수 있는 연구가 된다.

우리 글로 적고, 기록한 소리 정보를 다시 되찾는다.

부호는 정본을 섬기고, 구현은 그 보존을 증명해야 한다.

부록 A. 복합 음절핵과 코드포인트 식별

아래 표는 음운 단위와 저장 자모 요소열의 대응이다. 표시용 무음 ɔ은 포함하지 않는다.

읽기용 표시는 뒤 요소에 ɔ을 붙이며, 제2본의 음가 대응과 이 표의 부호열은 유지한다.

음운 단위	모듈	자모 요소 (각 항 JV-)	부호열 (각 항 U+)
/i:/	공통	I I	1175 1175
/u:/	공통	U U	116E 116E
/ɑ:/	RP	A A	1161 1161
/ɔ:/	RP	O O	1169 1169
/ɜ:/	RP	EO EO	1165 1165
/ɛ:/	RP	E E	1166 1166
/eɪ/	공통	E KIT	1166 119D
/aɪ/	공통	A KIT	1161 119D
/ɔɪ/	공통	O KIT	1169 119D
/aʊ/	공통	A FOOT	1161 118D
/əʊ/	RP	SCHWA FOOT	119E 118D
/oʊ/	GA	O FOOT	1169 118D
/ɪə/	RP	KIT SCHWA	119D 119E
/ʊə/	RP	FOOT SCHWA	118D 119E

GA NURSE /ɜ/는 이 완성 대응표에 넣지 않는다. /ɝ/는 source에서 normalized /əɪ/로 분석하는 규칙을 따르며, 모음 하나의 복합 nucleus로 등록하지 않는다. /ɪ/은 그 음절 분석에 맞는 자음 위치에 기록한다.

A.1 허용 부호의 이름 검사

registry의 codepoint와 Unicode 공식 이름, Hangul_Syllable_Type을 대조한다. 코드포인트 이름의 문자열 유사성은 음가 대응이나 프로파일 동치의 근거로 사용하지 않는다. 별칭과 역사적 이름이 필요한 문자는 해당 Unicode 판본의 주석도 함께 확인한다.

[I2] 13-14-22-23-35-37-50-53쪽. 구성 요소 ID는 이 개정의 시범 식별자.

부록 A. 핵심 문자 식별표

부호	단독 자모	Unicode 공식 이름
U+119D	ㅣ	HANGUL JUNGSEONG I-ARAEA
U+118D	ㅍ	HANGUL JUNGSEONG U-U
U+1182	ㅍ	HANGUL JUNGSEONG O-O
U+119E	.	HANGUL JUNGSEONG ARAEA
U+1144	ㅐ	HANGUL CHOSEONG IEUNG-PIEUP
U+114B	ㅐ	HANGUL CHOSEONG IEUNG-PHIEUPH
U+A976	ㅐ	HANGUL CHOSEONG IEUNG-RIEUL
U+D7DB	ㅐ	HANGUL JONGSEONG RIEUL-YESIEUNG
U+111B	ㅐ	HANGUL CHOSEONG KAPYEOUNRIEUL
U+11D0	ㅐ	HANGUL JONGSEONG SSANGRIEUL
U+11BE	ㅐ	HANGUL JONGSEONG CHIEUCH
U+1149	ㅐ	HANGUL CHOSEONG IEUNG-CHIEUCH
U+115F		HANGUL CHOSEONG FILLER
U+1160		HANGUL JUNGSEONG FILLER
U+110B	ㅐ	HANGUL CHOSEONG IEUNG

단독 자모 열은 코드 식별용이다. 독립 음절블록이나 새 정본 예시가 아니다. 표에서 글리프를 홀로 보여 주는 일과 기준형 자모열을 직렬화하는 일은 구분한다.

특히 구별할 두 쌍

U+118D (FOOT /ʊ/) ≠ U+116E U+116E (/u:/)

U+1182 (GA /ɔ/) ≠ U+1169 U+1169 (RP /ɔ:/)

어느 쪽도 이름이나 모양만으로 자동 치환하지 않는다.

이름과 코드값은 이번 실행 환경의 Unicode 데이터에서 대조했다. 확인 대상은 오래전부터 배당된 본 표의 문자들이며 최신판 전체의 변경 검사는 아니다.

부록 B. 시험 벡터와 기대 판정

§7의 25개 직렬화 예시는 양성 벡터다. RP·GA 공통 표시는 두 모듈에 같은 예시 부호열을 적용한다는 뜻이다. 아래 항목은 그 예시만으로 드러나지 않는 보존 실패와 미결 처리를 추가로 검사한다.

시험	입력·조건	기대 판정
B-01	boost /u:/의 mapping	음운 토큰 1개, 자모 요소 2개
B-02	cats의 /t+s/	음운 토큰 2개, JV-TS 1개
B-03	play·blue	초성군의 JV-L이 1개
B-04	FOOT 부호와 U U 연쇄	두 값의 동치 비교는 거짓
B-05	GA AW와 RP O O 연쇄	두 값의 동치 비교는 거짓
B-06	about 평문만 역변환	강세·출처를 복원했다고 선언하지 않음
B-07	catch 중성 JV-CH	구조 보존, 평문 unavailable
B-08	GA /3~/	source 보존, 기저 모음 unresolved
B-09	음절자음만 있는 nucleus	모음 삽입 없이 지원 범위 밖으로 보고
B-10	filler가 포함된 기준형	E-CHAR
B-11	T 다음 V 또는 V 없는 L	E-ORDER
B-12	없는 mapping 요소 참조	E-REFERENCE
B-13	같은 입력 음소의 중복 소비	E-COVERAGE
B-14	NFC(bank)를 수신	원입력 보존, NFD 후 원 코드열 회수
B-15	모듈·버전 없는 평문	필요 정보 요청 또는 후보 상태 유지

표의 기대 판정은 구현 요구이며 이 책을 작성한 것만으로 웹앱에서 모두 통과한 것은 아니다. 문서 내부 검사는 부호열·경계·대립·JSON 예시의 일관성을 확인한다. 사용자 인터페이스와 실제 복사·낭독 동작은 제4본의 시험 결과를 필요로 한다.

부록 C. 측정 방법과 재현 절차

C.1 이번 세이핑 계측의 식별 정보

항목	값
측정일	2026-09-08
글꼴	NanumMyeongjoOTF-YetHangul.otf
파일 판	Version 1.0.0.9 Build 20140917
글꼴 좌표	unitsPerEm = 1000
엔진	HarfBuzz 8.3.0
방식	HarfBuzz C API, OpenType font functions
버퍼	UTF-8 입력, segment properties 자동 추정
추가 기능	사용자 지정 feature 없음, 엔진 기본값
폭 기준	가 U+AC00의 advance 합 = 950
정규화	입력 코드포인트 열을 직접 구성

글꼴 파일 SHA-256:

```
c2e3217e63456a8cdcfce5670692d5ec  
78d7b1d1396f6968bef70ddb72dc1bb
```

위 해시는 줄바꿈 없이 이어 붙인 64자리 값이다. 시험 입력의 정확한 코드포인트는 §7과 §13의 대조표에 있다. 동일한 글꼴 이름이어도 파일 해시가 다르면 같은 실험 조건으로 취급하지 않는다.

C.2 재현 단계

동일한 해시의 글꼴을 로드하고 `hb_face_create`와 `hb_font_create`로 font를 만든다.

`hb_ot_font_set_funcs`를 적용하고 font scale을 (1000,1000)으로 둔다. 각 코드포인트 열을 UTF-8로 만들고 새 buffer에 넣어 `hb_buffer_guess_segment_properties`를 호출한다. 사용자 지정 feature 없이 `hb_shape`를 실행한다.

`hb_buffer_get_glyph_positions`에서 `x_advance`를 합하고 기준값 950으로 나눈다. 글리프 수, 코드포인트 열, advance 합과 비율을 함께 저장한다. 자형 판정에는 실제 화면 또는 페이지 렌더를 별도로 확인하며, 1자리 폭만으로 정상 HMBR 표시를 선언하지 않는다.

이번 환경의 직접 계측. Linux 공유 라이브러리 HarfBuzz C API를 Python ctypes에서 호출했다.

부록 C 계속 · 사전 전수조사 재현 기록

C.3 재현 환경과 명령

묶음	기록할 정보
자료 식별	사전명·판·날짜·원파일·해시·표제어 수·변이 처리
이용조건	출처·권리자·라이선스·재배포 허용 범위
발음 모듈	RP·GA 표기 관행, 비권설·rhotic·길이·강세 처리
전처리	기호 치환·IPA 분해·음절자음·제외 조건과 로그
매핑	registry 버전, 위치형, /t+s/ 소비 규칙
범주	배당된 구성값 / 미배당 구성값 / 분석 미결 / 오류
계수	자음군 종류 수·출현 수·표제어 수를 분리
재현성	실행 코드 버전·환경·명령·원 집계표

저장소 루트에서 `node tools/cluster_census.mjs`를 실행한다. 스크립트는 `data/dict-rp.json`과 `data/dict-ga.json`을 읽고 `js/registry.js`의 현재 배당을 적용한다. 결과에는 표제어·해석 수, 조성·종성 군집, 부호 있음·없음·미배당, 복합값 /t+s/, 새 겹낱자 수, 상위 60종 누적률과 2요소 음절핵을 함께 출력한다.

C.4 이번 문서의 내부 검사와 사전 재현 범위

§7의 25개 예시에서 허용 자모와 L/V/T 배열, 블록 경계를 검사한다. 각 자모열의 NFD 안정성과 NFC 뒤 NFD 복원을 검사하고, FOOT/U U 및 AW/O O 대립을 확인한다. §3 JSON의 참조와 boost 직렬화가 대응하는지도 검사한다.

이 검사는 예시의 내부 일관성과 현재 사전 집계 of 재현성을 확인한 것이다. 독립 변환 엔진의 영어 전범위 발음 정확도나 사전 원출처의 권리 상태를 대신하지 않는다. §13의 세이핑 계측과 §21의 사전 집계는 서로 다른 자료·명령·판정 범위를 가진다.

C.5 공개 기록과 인용 조건

제4본의 적합성 보고에는 성공·실패·미결 결과와 전체 분모를 함께 남긴다. §21 수치를 인용할 때에는 저장소 커밋, 사전 파일, `tools/cluster_census.mjs`와 실행일을 함께 적는다. 사전판·라이선스가 확정되기 전에는 이 결과를 영어 전체의 절대적 규모나 법적 결론으로 확대하지 않는다.

부록 D. 용어 대조

이 문서의 용어	이름	뜻
원문 IPA	source IPA	출처의 기호와 문자열을 보존한 IPA
정규화 IPA	normalized IPA	언어모듈이 매핑에 사용하는 분석형
음운 토큰	phoneme token	모듈이 정한 음운 단위
대응 단위	mapping unit	음운 토큰열과 자모 요소열의 관계
자모값 ID	jamo value ID	등록부의 자모 요소·복합값 식별자
음절블록	syllable block	선택된 영어 음절 하나를 담는 정본 단위
초성군	onset[]	블록 앞 자음 요소. 없으면 무음 초성 사용
음절핵	nucleus[]	한 음절을 이루는 모음 구성 요소
종성군	coda[]	블록 뒤 자음 요소의 순서 있는 배열
위치형	positional form	같은 자모값의 자리별 코드·모양
직렬화	serialization	구조를 정해진 순서의 교환 표현으로 변환
셰이핑	shaping	문자열을 글리프 선택·배치로 처리
기준형	profile canonical form	본 응용 프로파일이 선택한 저장형
정준 동등성	canonical equivalence	Unicode가 정한 문자열의 동등성
의미 동치	profile equivalence	모듈·버전이 승인한 의미 관계
대체 표시	fallback	미지원 환경에서 정보 보존을 돕는 출력

‘정본’이라는 연구상의 권위와 Unicode의 canonical이라는 기술 용어를 혼동하지 않는다. 정본의 음가 규칙은 제2본을 따르고, Unicode 정준 동등성은 Unicode 규칙을 따른다. 이 책의 기준형은 응용 프로파일이 선택한 보존 형식이다.

참고문헌 및 작성 근거 · 내부 자료

아래 자료는 이번 개정에서 직접 열람한 근거다. 페이지는 원본 PDF의 물리적 쪽수 기준이며, 본문에 인쇄된 차례 번호가 다른 경우 PDF 쪽수를 우선한다.

[I1] 제1본 선언본

Park Il Hwan, 「훈민바름 영어확장판 · 제1본 선언본」, 제출용, 원표지 날짜 2026-08-14, 36쪽. 2026-09-08 제공본에서 출발했으며 현행 기준은

HMBR_Book1_Seoneonbon_20261009.pdf(2026-10-09 교정판)다. 문서 지위·정밀 표기층·공공성·표준화·한계의 표현 기준.

[I2] 제2본 해례본

Park Il Hwan, 「훈민바름 영어확장판 · 제2본 해례본」, 제출용, 원표지 날짜 2026-08-14, 59쪽. 현행 기준은 HMBR_Book2_Haeryebon_20261009.pdf(2026-10-09 교정판)다. 최초

업로드의 ‘정본’ 대신 문서 내부 명칭인 ‘해례본’을 따른다. 음가 배당과 현 시범 조자법·상태의 기준.

[I3] 제안서와 커버레터

「규제합리화 대국민 공모전 제안서 작성본(1)」, 1쪽; 「심사위원 여러분께(3)」, 4쪽. 정밀 표기층의 적용 범위 합리화와 4본의 역할·공개 계획을 확인했다.

[I4] 이전 전산본과 작성 보고

「훈민바름 영어확장판 · 제3본 전산본 260807」, 12쪽.

HMBR_BOOK3_JEONSANBON_A4.docx와 HMBR_BOOK3_PROOF.pdf, 2026-08-14, 41쪽. HMBR_BOOK3_REPORT.md, Claude, 2026-08-14. 이 가운데 41쪽 후속판의 조직과 판형을 개정 기반으로 삼았다. 과거 사전 집계는 2026-09-16 저장소의 스크립트로 재현했다.

[I5] 이전 대화 기록

「바르미 Master 3 / 공모전 2」, 140쪽; 「바르미 Master 2 / 공모전」, 623쪽. 전산본 관련 결정과 마지막 수정 지시를 검색·대조했다. 특히 Master 3의 132-140쪽에서 후속 원고 평가, mapping unit, /I 한 번과 의미 동치 수정 지시를 확인했다. 대화에 언급된 과거 첨부물 전부를 복원한 것은 아니다.

이번 개정의 JSON 필드, 오류 코드, 교환·역변환 조건과 적합성 요구는 위 자료를 전산적으로 정합화하기 위한 설계안이다. 원 저자의 새로운 음성학 판정을 대신하지 않는다.

참고문헌 및 작성 근거 · 외부 자료

[E1] Unicode 17.0.0, Chapter 3

§3.12 Conjoining Jamo Behavior, D133–D134. 음절블록 정의와 한글 정준 분해·합성.

<https://www.unicode.org/versions/Unicode17.0.0/core-spec/chapter-3/>

[E2] UAX #15: Unicode Normalization Forms

정준·호환 동등성, NFC·NFD·NFKC·NFKD의 구분.

<https://www.unicode.org/reports/tr15/>

[E3] Microsoft, Developing OpenType Fonts for Korean Hangul Script

등록 조합 전처리, Uniscribe 설명, ccmp·ljmo·vjmo·tjmo. 모든 엔진의 현재 동작을 대표하지 않는다.

<https://learn.microsoft.com/en-us/typography/script-development/hangul>

[E4] UAX #29: Unicode Text Segmentation

Unicode 17.0.0, Revision 47. GB6–GB8과 한글 분절.

<https://www.unicode.org/reports/tr29/tr29-47.html>

[E5] N3422, KS X 1026-1:2007 영문 번역본

한국 측 공개 제출 자료. PDF 5쪽의 지위 표기와 12–14쪽 관련 조항. 국문 정식본을 대신하지 않는다.

<https://www.unicode.org/L2/L2008/08225-n3422.pdf>

[E6] L2/09-151 / N3633

US / Unicode Consortium Comments on N3611, 2009-04-21, 2쪽. 당시 정규화 우려와 해석 한계를 확인했다.

<https://www.unicode.org/L2/L2009/09151-us-utc-cmt-n3611.pdf>

외부 자료 열람일: 2026-09-08. 법령의 구속력, 현행 KS 국문판과 저장소의 현재 배포 상태는 이 문서의 확인 범위 밖이다.

원형 토큰과 t 표면 주석의 보존

2026년 10월 12일 개정 부록. 제3.6절과 모델·교환 논의에 연결 발음 연습층의 실제 구현을 보충한다. 기존 Broad 정보, 음가 배당, 블록 구조와 NFD 직렬화는 변경하지 않는다. Narrow 전체 스키마의 완성을 뜻하지 않는다.

서로 다른 세 층

층	보존하거나 생성하는 정보
원형 records	입력 IPA, 출처, phonemeTokens, canonicalModel, 정보 직렬화
파생 records	지원 음소만 달라진 약형·동화·탈락을 기존 엔진으로 변환한 연습형
표면 surface_records	[ʔ], [ŋ], [tʰ] 등을 포함할 수 있는 IPA와 훈민바름 참조 상태

표면 IPA가 기본 엔진에 없는 세부 음을 포함해도 그 기호를 원형 자모열에 끼워 넣지 않는다. button을 [ˈbʌʔŋ]으로 선택하면 원형 /bʌtən/ 레코드를 유지하며 display_status는 broad-reference다. winter의 t 탈락처럼 지원 음소만 남는 경우에는 연습용 파생 훈민바름을 만들되 원형 records를 별도로 보존한다.

t 주석의 필드

필드	의미
word_index	원형 단어 배열의 0부터 시작하는 위치
token_ids	해당 원형 canonicalModel.phonemeTokens의 ID
underlying_ipa	영향받은 원형 토큰의 IPA
surface_ipa	선택한 실현. 탈락이면 빈 문자열
feature_id	ntdrop, glottal 또는 unreleased
environment / source	적용 환경 설명과 참고 자료 URL

button의 /tən/을 [ʔŋ]으로 바꾸는 선택은 원형 tən 토큰 세 개를 가리킨다. 표면 [ŋ]을 새 Broad 음소로 오인하지 않는다. 겨냥한 원형 토큰이 남으므로 선택을 해제할 때 추측으로 복원할 필요가 없다. 이 필드는 이번에 추가한 t 변이에 제공하며 기존 연결 규칙 전체의 토큰 수준 주석을 완성한 것은 아니다.

내보내기와 일관성 규칙

연음화 패널을 연 뒤 전체 기록 저장을 누르면 원형 records에 connected_speech를 추가한다. 일반 변환만 한 결과에는 이 필드를 넣지 않는다. version은 connected-speech-t-variants-20261012다. accent, pace, emphasis, pauses, candidates, applied, 원형과 선택형 IPA, 파생 records 및 surface_annotations와 surface_records를 담는다. GA 단어 하나라도 t 후보가 있으면 이 연습층을 저장할 수 있다.

surface_records 필드	해석
ipa	선택 후 표면 문자열
hmbr_reference_ipa	현재 훈민바름이 참조하는 IPA
display_status	derived 또는 broad-reference
display_notes	미정 자형·미파열 주석 등 화면 표시의 이유

derived는 선택 후의 지원 가능한 IPA로 표시할 수 있다는 뜻이며 항상 원형과 다른 문자열이라는 뜻은 아니다. broad-reference이면 훈민바름만 보고 표면의 [ʔ][n]까지 확정해 읽어서는 안 된다. 표면 문자열에는 원형에서 유지한 음절 점이 남을 수 있으며 완전한 재음절화·운을 분석을 뜻하지 않는다.

충돌과 되돌리기

같은 원형 t의 탈락·성문음화·미파열은 하나의 선택 그룹으로 묶는다. 수동 선택에서는 마지막 선택이 앞 선택을 대체한다. 빠른 회화에서 자음군 탈락과 성문음화가 모두 추천될 수 있는 경우 성문음화의 추천 우선순위를 높인다. 이는 인터페이스의 단순화이며 복합 조음 자체를 금지하는 음성학 규칙이 아니다.

쉽은 해당 단어 경계의 연결 부호를 막는다. 쉽 앞 성문음화·미파열은 별도의 수동 선택이다. 입력이나 발음 기준을 바꾸면 이전 선택을 버린다. 원형 자모열 복사는 기존 원형만 복사한다. 표면 IPA나 연결 부호를 정본의 무손실 직렬화로 바꾸지 않는다.

검증 범위와 음성 연계

새 시험은 원형 불변, 토큰 ID의 존재, 세 토큰을 가리키는 button 주석, 강제 조건, 쉽, 선택 충돌, 실제 파생 자모의 변화를 검사한다. 역변환은 surface_ipa만 보고 원형을 추정해서는 안 된다. [r]에서 t/d를 복구하기 어려운 기존 문제와 동일한 보존 원칙을 따른다.

다음 TTS 단계는 이 저장된 선택을 음성 요청에 연결하고 반환된 발음이 요청과 일치하는지 별도로 확인해야 한다. 현재 내보내기 구조에는 합성된 오디오나 발음 일치 보증을 넣지 않는다. 외부 JSON 가져오기와 완전한 Narrow 변환도 별도 과제로 남긴다.