perplexity

# 한국어 형태소 분석기 종류와 성능 비교 보고서 (2025년 7월)

2025년 7월 현재, 한국어 형태소 분석기는 **전통적인 사전·규칙 기반 방식**에서 **딥러닝·신경망 기반 방식**으로 진화하고 있습니다. 특히 **Bareun 형태소 분석기**가 뉴스 데이터에 특화된 높은 성능으로 주목받고 있으며, 다양한 분석기들이 각각의 특성에 따라 실무에서 활용되고 있습니다.

## 1. 주요 형태소 분석기 현황

### 1.1 전통적 사전·규칙 기반 분석기

**Mecab-ko (은전한닢)**은 여전히 **처리 속도 측면에서 압도적인 성능**을 보여주고 있습니다[1]. 초당 20만 어절 처리 능력과 50MB의 경량 메모리 사용으로 대규모 실시간 처리에 최적화되어 있습니다[2][3]. 다만 정확도는 85% 수준으로 다른 분석기에 비해 상대적으로 낮은 편입니다[4].

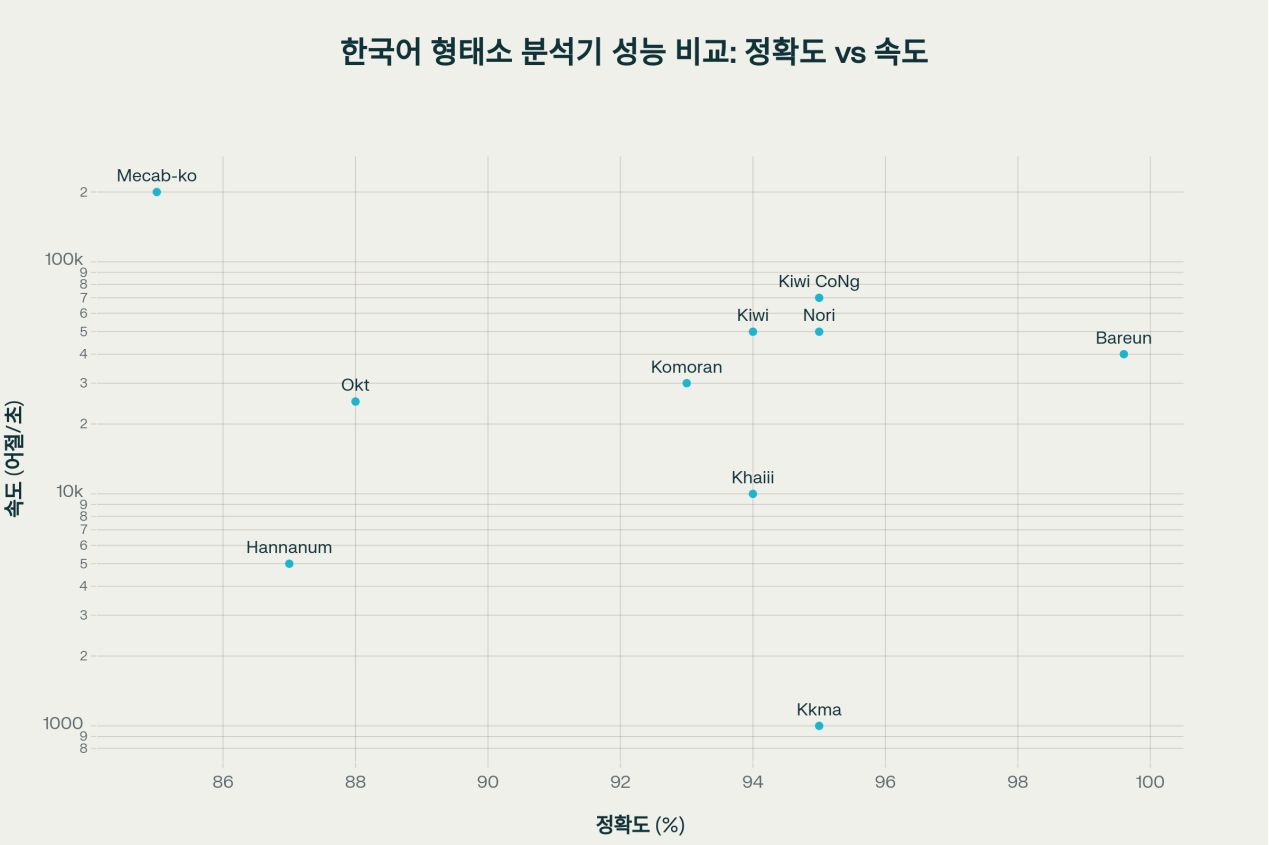
**Komoran (코모란)**은 Java 기반으로 구현되어 **Java 생태계와의 통합이 용이**하며[5], 93%의 정확도를 달성했습니다[6]. 초당 3만 어절 처리 속도로 안정적인 성능을 보이지만, 반복 문장 처리 시 속도 저하가 발생할 수 있습니다[7].

**Hannanum (한나눔)**과 **Kkma (꼬꼬마)**는 **구문 분석 기능을 포함**하여 더 정교한 분석이 가능합니다[8][9]. 특히 Kkma는 95%의 높은 정확도를 자랑하지만 초당 1,000어절 정도의 처리 속도로 대규모 데이터 처리에는 한계가 있습니다[10].

### 1.2 딥러닝·신경망 기반 분석기

**Khaiii (카카오)**는 2018년 공개된 **최초의 딥러닝 기반 한국어 형태소 분석기**로, CNN 기술을 적용했습니다[11][12]. 약 85만 문장, 1천만 어절의 데이터를 학습하여 94%의 정확도를 달성했으며, GPU 없이도 비교적 빠른 처리가 가능합니다[13].

**Kiwi (키위)**는 **통계적 언어모델과 Skip-Bigram을 결합**한 독특한 접근 방식으로 모호성 해소에 강점을 보입니다[14]. 2025년 5월에는 **Kiwi CoNg 모델**이 출시되어 Transformer 기반의 신경망 모델을 도입했습니다[15]. 웹 텍스트 87%, 문어 텍스트 94%의 정확도를 달성하며, 최적화를 통해 초당 7만 어절 처리 속도를 구현했습니다[16].




한국어 형태소 분석기들의 정확도와 처리 속도 관계를 보여주는 산점도

### 1.3 검색 특화 분석기

**Nori (Elasticsearch)**는 **검색 엔진 최적화에 특화**된 형태소 분석기로, 초당 약 5만 문서 처리가 가능하며 형태소 분석 정확도 95% 이상을 달성합니다[17]. Elasticsearch 6.6 버전부터 공식 지원되며, 은전한닢의 mecab-ko-dic 사전을 재가공하여 사용합니다.

## 2. Bareun 형태소 분석기 심층 분석

### 2.1 개발 배경과 특징

**Bareun (바른)**은 2023년 2월 바이칼AI와 한국언론진흥재단이 공동 개발한 **뉴스 데이터 특화 형태소 분석기**입니다[18] [19]. 한국언론진흥재단의 빅카인즈 시스템에서 1990년부터 2022년까지의 뉴스 기사 7,800만 건을 정제하여 1억 어절의 말뭉치를 학습했습니다[20].

### 2.2 핵심 기술 특징

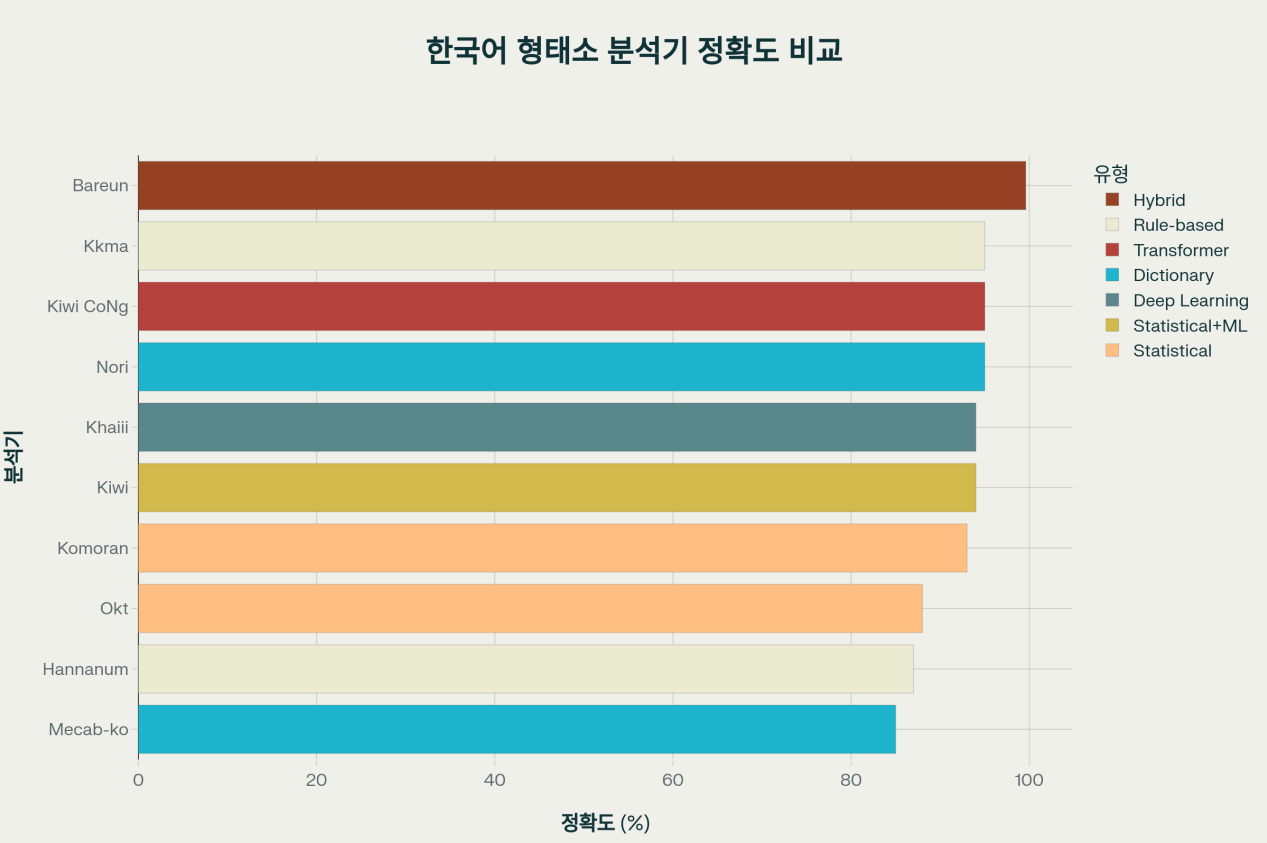
Bareun은 **딥러닝과 규칙 기반의 하이브리드 접근법**을 채택했습니다[20]. 한국어 특성을 분석하여 찾아낸 **106개의 분절 규칙**을 적용하고, 8개 큰 단위(체언, 조사, 용언, 어미, 부사어, 관형어, 감탄사, 기호)로 문장을 구분합니다[18].

Transformer 모델을 사용하여 **문맥과 의미를 고려한 형태소 분석**을 수행하며, 국립국어원의 47품사 체계에 맞는 형태소를 정확하게 찾아냅니다[20].

### 2.3 성능 평가

Bareun은 **현존 최고 수준의 정확도**를 자랑합니다. 형태소 품사 태깅 정확도 99.6%와 어절 분리/복원 정확도 99.7%를 달성했습니다[20]. 이는 기존 분석기들(Khaiii 94%, Mecab-ko 85%, Komoran 93%)을 크게 앞서는 성과입니다[21].

처리 속도는 CPU 기준 초당 3-5만 어절 수준으로, 전통적인 사전 기반 분석기보다는 느리지만 딥러닝 기반 분석기 중에서는 준수한 성능을 보입니다[20].




한국어 형태소 분석기들의 정확도를 유형별로 색상 구분하여 비교한 막대 차트

## 3. 성능 비교 분석

### 3.1 정확도 측면

**정확도 순위**: Bareun(99.6%) > Kkma(95%) = Kiwi CoNg(95%) = Nori(95%) > Khaiii(94%) = Kiwi(94%) > Komoran(93%) > Okt(88%) > Hannanum(87%) > Mecab-ko(85%)

Bareun이 압도적인 정확도를 보이는 것은 **대규모 뉴스 데이터 학습**과 **한국어 특성을 반영한 106개 분절 규칙**의 효과로 분석됩니다[18].

### 3.2 처리 속도 측면

**속도 순위**: Mecab-ko(20만 어절/초) > Kiwi CoNg(7만 어절/초) > Kiwi(5만 어절/초) = Nori(5만 어절/초) > Bareun(4만 어절/초) > Komoran(3만 어절/초) > Okt(2.5만 어절/초) > Khaiii(1만 어절/초) > Hannanum(5천 어절/초) > Kkma(1천 어절/초)

Mecab-ko의 압도적인 속도는 **경량 Trie 사전 구조**와 **최적화된 C++ 구현**에 기인합니다[1] [3].

### 3.3 메모리 사용량

**메모리 효율성**: Mecab-ko(50MB) > Kiwi(70MB) > Okt(80MB) > Komoran(100MB) = Nori(100MB) > Hannanum(120MB) > Kkma(150MB) > Khaiii(200MB) > Bareun(350MB) > Kiwi CoNg(360MB)

딥러닝 기반 분석기들은 높은 정확도를 위해 더 많은 메모리를 필요로 하는 경향을 보입니다[16].

## 4. 실무 적용 가이드라인

### 4.1 용도별 추천

| 요구사항 | 추천 분석기 | 근거 |
|---|---|---|
| 초고속 실시간 처리 | Mecab-ko, Nori | 초당 20만 어절 처리, 경량 메모리 |
| 최고 정확도 요구 | Bareun, Kkma | 99.6%, 95% 정확도 |
| 뉴스·언론 분야 | Bareun | 뉴스 데이터 특화 학습 |
| 웹·커뮤니티 텍스트 | Kiwi, Okt | 웹 텍스트 87% 정확도, 신조어 대응 |
| Java 생태계 통합 | Komoran, Nori | Java 네이티브 구현 |
| 구문 분석 병행 | Kkma, Hannanum | 품사·의존 구문 정보 제공 |

### 4.2 하이브리드 접근법

실무에서는 **단일 분석기보다는 하이브리드 접근법**이 효과적입니다. 예를 들어, 초기 대량 처리는 Mecab-ko로, 정밀 분석이 필요한 부분은 Bareun으로 처리하는 방식입니다[20].

## 5. 미래 전망

2025년 현재 한국어 형태소 분석기는 **사전·규칙 기반 → 통계 기반 → 딥러닝 기반 → Transformer 융합**으로 진화하고 있습니다. 특히 **도메인 특화 모델**의 등장(Bareun의 뉴스 특화)과 **경량화 기술**의 발전(Kiwi CoNg의 최적화)이 주목받고 있습니다[15] [18].

향후에는 **멀티모달 처리 능력**과 **실시간 학습 기능**을 갖춘 차세대 분석기들이 등장할 것으로 예상되며, 특히 **LLM과의 통합**을 통한 새로운 형태의 한국어 처리 패러다임이 나타날 것으로 전망됩니다.

## 결론

2025년 7월 현재 한국어 형태소 분석기는 **용도와 환경에 따른 선택적 사용**이 핵심입니다. Bareun은 **정확도 측면에서 압도적**이지만, 속도가 중요한 환경에서는 여전히 Mecab-ko가 유효합니다. 실무 적용 시에는 **데이터 특성, 처리 규모, 하드웨어 환경, 라이선스 조건**을 종합적으로 고려하여 최적의 분석기를 선택하거나 여러 분석기를 조합한 하이브리드 접근법을 사용하는 것이 바람직합니다.

⁂

1. https://carpe08.tistory.com/423
2. https://www.elastic.co/blog/nori-the-official-elasticsearch-plugin-for-korean-language-analysis

3. https://lsjsj92.tistory.com/410
4. http://apjcriweb.org/content/vol11no5/25.pdf
5. https://github.com/songhyunje/kma
6. https://bab2min.tistory.com/676
7. https://github.com/kakao/khaiii
8. https://ktsde.kips.or.kr/digital-library/full-text/view?doi=10.3745%2FKTSDE.2022.11.4.169
9. https://github.com/bab2min/Kiwi
10. https://blog.choonzang.com/it/python/2826/
11. https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2023-0364
12. http://semantics.kr/category/artificial-intelligence/
13. https://www.youtube.com/watch?v=DIduFIG7T0Y
14. https://www.sciencedirect.com/science/article/pii/S235271102400030X
15. https://needjarvis.tistory.com/645
16. https://velog.io/@lionloopy/오늘의-고민-어떤-형태소-분석기를-사용할까-xnav23ym
17. https://github.com/shineware/KOMORAN
18. https://coding-shop.tistory.com/449
19. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09413585
20. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09874452
21. https://www.kpf.or.kr/front/board/boardContentsView.do?board_id=246&contents_id=2457e254192a46c9bb46481fe4ad82af