한글 맞춤법 검사 시스템의 개발

○
박 종 만, 김 영 찬
금성소프트웨어(주) 연구소

A Development of Han-geul Spelling Checker

○
Jong Mahn Park, Young Chan Kim
R&D Institute, GoldStar Software Ltd.

요 약

문자 생활에 있어서 컴퓨터의 도입이 가속화되면서, 한글의 맞춤법 검사를 자동으로 해주는 문안 편집 도움 기구의 필요성이 대두되었다. 교착어인 한국어의 문자인 한글의 맞춤법 검사는 다른 언어에 비해 상대적으로 어렵게 여겨져 왔다. 본 논문에서는 한글 맞춤법 검사 시스템의 개발에 대하여 실용화의 관점에서 논한다. 실용화의 관점에서는 한글 맞춤법 검사뿐만 아니라 문서 편집기를 통한 인터페이스, 사전의 제공, 틀린 경우의 적절한 조치 등이 필요하다.

I. 서론

우리의 문자생활에 컴퓨터의 도입이 가속화되면서 한글 맞춤법 검사기의 개발이 절실히 요구되고 있다. 한글 맞춤법 검사기는 자연언어 처리의 입장에서 기초연구가 이루어져 왔으며, 이미 다수의 논문 및 시제품이 발표된 바 있다[1,4,5]. 이러한 연구자료들을 종합해 볼 때, 이제는 실용화 단계로 접어들었다고 할 수 있다. 이에 금성소프트웨어(주) 연구소에서는 실용화를 목표로 한글 맞춤법 검사 시스템의 개발을 추진해 왔다. 이 논문은 본 연구소의 경험을 바탕으로 한글 맞춤법 검사기를 실용화하는데 있어서의 문제점과 그 해결방안, 앞으로의 과제들에 대하여 논한다.

II. 실용화 조건

2.1 환경 요인

한글 맞춤법 검사기가 쓰이기 위해서는 컴퓨터에 입력된 문서가 있어야 한다. 따라서 광범위한 컴퓨터의 보급과 문자생활에의 이용은 맞춤법 검사기 실용화의 전제조건이다. 현재 한글 문서 편집기의 활발한 개발에 힘입어 이러한 조건은 만족되었다고 평가된다.

문서 편집의 대부분은 PC에서 이루어지므로 맞춤법 검사기는 PC에서 수행되어야 한다. 프로그램 자체가 이식가능한(portable) 것이라면, PC에서 수행될 경우 고려해야 하는 점은 적은 기억장소와 느린 수행속도로 인한 응답시간(response time)의 지연이다. 대부분의 시제품에서는 사전의 구조를 기억장소의 한계에 대한 고려없이 설계했기 때문에 PC에의 직접적용이 어렵다. 특히 사전구조는 수행속도에 가장 큰 영향을 미치는 것으로서 신중히 고려되지 않으면 응답시간의 지연이 심각해지게 된다. 따라서 최적화된 사전의 구조가 요구된다.

2.2 사전

사전은 관리의 관점에서 시스템 제공 사전과 사용자 정의 사전으로 분류된다. 시스템 정의 사전은 맞춤법 검사기의 개발자 측에서 기본적으로 제공하는 것으로서 사전의 정확성을 개발자 측에서 보장하는 특징을 갖게되며, 범용으로 쓰일 수 있는 어휘를 수록하고 있다. 사용자 정의 사전은 사용자가 여러가지 이유로 인하여 추가로 어휘를 수록시킨 사전으로서 그 정확성은 사용자의 책임이며, 사용자의 특수분야를 반영하는 도메인 사전의 역할을 할 수도 있다. 또한 사용자의 편의를 위하여 인명, 지명, 기타 고유어들을 등록시켜 사용할 수도 있다.

사전을 실제로 구현하는 물리적인 관점에서는 다음과 같은 사항들이 고려되어야 한다.

1) 복수의 사용자들이 사전의 병합을 통해 편리하게 그들의 사전을 확장시킬 수 있어야 한다.
2) 상위 버전의 시스템 제공 사전이 기존의 시스템 제공 사전 및 사용자 정의 사전과 문제없이 병합될 수 있어야

한다.

3) 사전의 병합시 사용자 사전의 항목에 대해 검토할 수 있어야 한다.

사용자 사전은 그 특성상 특수 분야나 그 사용자만이 사용하는 고유어들이 등록될 수 있기 때문에, 단순한 병합은 필요없는 이휘들의 누적을 수반하게 된다. 또한 사용자가 사전의 정확성을 책임지기때문에, 틀린 이휘나 틀린 정보가 일단 수록되면 계속 남아있게 된다. 맞춤법 검사기에 일단 수록된 단어가 틀렸다면, 그 단어에 접근하려는 시도 자체가 없을 가능성이 높기 때문에, 찾아내어 수정하기가 매우 곤란하다.

4) 사용자 정의 사전이 수행속도에 큰 영향을 주어서는 안 된다.

시스템 제공 사전과 사용자 정의 사전을 구별하는 방법은 두가지가 있다. 하나는 두개의 사전을 하나의 구조안에 수록하고 플래그(flag)로써 구별하는 것이고, 다른 하나는 별도의 사전을 구성하는 것이다. 플래그로써 구별하는 방법은 사전의 검색이 한번에 이루어지므로 수행속도에 이득이 있는 반면, 사용자 정의 항목을 검토하거나 사전을 병합할 때에 일반적으로 대량인 시스템 사전까지 모두 처리해야 가능하므로 상대적으로 매우 큰 오버헤드(overhead)를 감수해야 한다. 병합이 얼마나 자주 일어나느냐에 그 타당성이 달려있다고 볼 수 있다.

별도의 사전으로 구성하는 방법은 사용자 정의 사전을 위하여 추가로 사전 검색이 필요하다. 사용자 정의 사전에 시스템 제공 사전의 항목과 같은 항목이 있으면서 정보의 내용이 다를 경우가 문제되는데, 먼저 찾는 사전에서 원하는 항목을 발견한 경우에도 나머지 사전을 검색해야 하거나, 한 사전에 중복 수록한 후 그 사전을 항상 먼저 찾도록 해야 한다. 후자의 경우 시스템 제공 사전은 편집을 할 수 없으므로 사용자 정의 사전에 중복 수록해야 한다. 별도의 사전을 구성하면 병합이 사용자 정의 사전에 대해서만 이루어지므로 편리하고, 사용자 정의 항목만을 쉽게 찾아볼 수 있다. 또한 시스템 제공 사전은 더 이상 확장되지 않으므로, 사전을 압축하여 구조를 최적화할 수 있다. 사전의 압축은 블록(block) 단위로 구성되는 사전의 경우에 발생하는 단편화(fragmentation) 현상을 제거하고 재배치를 통하여 가능하다. 이 이득은 PC의 환경에서 매우 큰 것으로 평가될 수 있다.

사전의 항목 수는 실용 가능성의 가장 큰 척도라고 할 수 있다. 한글의 경우 어느 정도의 항목 수가 되어야 사용자가 불편없이 사용할 수 있는지에 대해서는 알 수 없으나, 영어의 경우에는 10만 이상의 항목을 갖추고 있다. 한글의 경우에는 단어 형성의 방법이 다양하여 영어보다 많은 항목 수가 필요할 것으로 예상된다. 접두어에 의한 파생어 및 합성명사가 주된 원인이 되는데, 이러한 과정을 통하여 생기는 단어들은 공백문자(space)없이 조성되면서, 일정한 형태소에 대하여 일관되게 적용되는 일반성을 갖지는 못하기 때문에 사전에 등록시키기를 요구하기 때문이다. 이러한 입장은 맞춤법 검사기의 기능 특성상 맞는 것을 틀리다고 하는 오류보다는 틀린 것을 맞다고 하는 것이 더 나쁜 결과로 평가되기 때문이다[4]. 사전의 항목 수가 큰 경우에는, 항목의 선정 방법과 각 항목의 입력 과정에서 발생할 수 있는 오류의 해결이 중요한 문제이다[6].

사용자 정의 항목이 갖추어야 하는 정보의 양과 복잡도는 최소화되어야 한다. 한글의 맞춤법 검사의 경우에는 첨가어적 특성에 의해, 단어만을 수록하는 영어와는 달리, 형태소간의 접속 가능 여부를 판단하기 위한 정보가 필요하게 된다. 사전의 항목 수가 충분히 큰 경우에는 사용자 정의 사전에는 명사의 추가 수록만이 가능하도록 제한하는 방법도 있으나 고급 사용자에게는 큰 불편이 될 수 있다.

2.3 사용자 인터페이스

맞춤법 검사기의 가장 큰 용도는 문서 편집을 돕는 것이다. 지금까지 발표된 시제품들은 모두 일괄처리(batch)방식을 채택하고 있는데, 실용화를 위해서는 대화식(interactive)으로, 문서 편집기와 통합된 환경을 제공해야 한다.

대화식의 경우 문서 편집을 하다가 사용자가 맞춤법 검사를 요구할 때에 맞춤법에 어긋나는 부분을 지적하며 사용자의 다음 행동을 대기하는 것이 일반적이다. 이때 사용자에게 제공되는 기능들은 다음과 같다[6].

- 무시 : 주로 고유어에 대하여 맞춤법 검사기가 오류로 인정하는 경우에 단지 지나치는 기능.

  이 기능은 1회만 지나치는 기능과 이후에 같은 고유어가 나타나는 경우 자동으로 무시하는 기능으로 세분된다.
- 수정 : 지적된 부분만을 수정하는 기능.

  이 기능은 다시 그 부분만을 수정하는 1회 수정과 이후에 같은 모습으로 틀린 부분이 나타나는 경우 자동으로 수정하는 수정 후 기억 기능으로 세분된다.
- 편집 : 지적된 부분뿐만이 아니라 다른 부분까지 고칠 필요가 있는 경우에 맞춤법 검사를 중지하고 편집 환경으로 전환하는 기능.
- 사전 등록 : 사전에 해당 단어가 없이서 지적된 경우에 그 단어를 등록하는 기능.
- 사전 편집 : 사전을 편집(검색, 삭제, 변경 등)하는 기능.
- 제시 : 음소 치환, 삭제, 추가 등의 방법을 통하여 유사어를 찾아 제시하는 기능.

  이 기능은 사용자가 제시되는 유사어를 검토함으로써, 사전에 틀린 단어가 수록되는 경우를 줄일 수 있다.

한글 맞춤법 검사기는 한글 입력을 처리하는데, 실제에서 사용되는 문서에는 한글 이외에 일반적으로 영어, 숫자, 도표, 포매팅 명령이 등을 포함하게 된다. 따라서 이러한 입력에 대한 적절한 처리가 필수적이다.

III. 시스템의 구현

3.1 개요

본 연구소에서 개발한 한글 맞춤법 검사기의 개요는 다음과 같다.

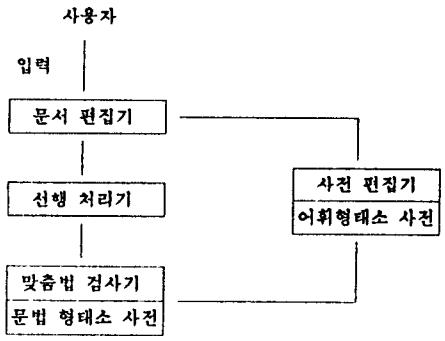

그림 1. 시스템의 개략적인 구성도

사용자는 문서 편집기를 인터페이스로 사용하여 맞춤법 검사를 하게 된다. 사용자의 입장에서 맞춤법 검사기의 대화식 인터페이스는 기존의 문서 편집기에 맞춤법 검사의 기능이 추가된 것이다.

이 시스템은 MS-DOS 환경에서 C 언어로 구성되어 있다. 현재는 완성형 한글 코드에 대해서 지원하고 있으며, 다른 한글 코드의 처리는 약간의 모듈 교체에 해당하는 수정만으로 가능하다.

3.2 선행 처리기

선행 처리기는 문서 편집기와 맞춤법 검사기 사이에서 인터페이스 역할을 한다. 주된 기능은 다음과 같다.

- 문서 편집기로 부터 입력문서를 받아 맞춤법 검사기의 입력 단위인 단위 토큰으로 분리
- 문서 편집기에 종속적인 요소의 처리 (축소, 확대, 특수문자 등)
- 포매팅 명령어를 무시
- 쌍으로 이루어진 문자의 처리 (괄호, 따옴표 등)
- 한자를 한글로 치환 (한자의 독음 처리)
- 한자나 영어로 구성된 어절을 별도의 화일에 저장
- 미완성 한글 검사
- 도표와 같은 그래픽 문자 처리
- 맞춤법 검사에서 틀린 것으로 판단된 어절에 대한 사용자 인터페이스
- 틀린 어절에 대하여 사용자가 원하는 경우 올바른 어절 후보 생성 (재시 기능)
- 한글 코드의 변환 (완성형 -> N byte)

한자와 영어는 그 자체의 정확성을 검사하지는 않지만 사용자의 입장에서는 당연히 그 정확성의 검사까지도 요구하게 된다. 그러나 이러한 요구를 그대로 수용하는 것은 매우 많은 추가작업이 필요하다. 개발된 시스템에서는 그에 대한 일부 해결책으로서 한자와 영어를 사용한 어절들을 모아 별도의 화일로 사용자에게 제공하고 있다. 이러한 해결책은 한글을 기본으로 하고 한자와 영어를 부분적으로 사용하는 문서를 취급하게 된다는 것을 그 전제로 한다. 사용자는 상대적으로 매우 적은 양만을 검토하여 정확성을 확인할 수 있다. 특히 한자의 경우는 여러 문서 편집기에서 단어별 변환 기능을 제공하므로, 그 자체의 정확성을 검사하는 것은 중복 검사일 수도 있다. 또한 한자의 독음을 한글로 치환하여 조사 등과의 결합관계를 검사한다.

3.3 사전 및 사전 편집기

시스템의 사전은 크게 조사, 어미 등을 수록한 문법 형태소 사전과 체언, 용언 등을 수록한 어휘 형태소 사전으로 분류되며, 어휘 형태소 사전은 다시 보조 기억장치에 있는 마스터 사전과 주기억 장치에서 유지되는 캐쉬 사전으로 분류된다. 문법 형태소 사전과 어휘 형태소 사전의 분리로 인한 이득 및 사전의 구조는 [4]를 참고하기 바란다.

캐쉬 사전은 한국어에서 가장 빈번히 사용되는 어휘들을 기본적으로 1000 단어 정도를 보유하며, 문서 처리 시 마스터 사전에서 검색된 어휘가 추가되어 이후에 다시 사용되거나, 재검사를 요구하는 경우에 대비한다. 또한 특정 사용자만이 사용하는 고유어에 대하여 등록을 요구하지는 않지만, 문서내에서 자주 나타나기 때문에 현재 처리되는 문서내에서는 계속 무시되기를 요구하는 경우(무시 후 기억 기능)에 그 고유어를 캐쉬 사전에 등록시키게 된다. 이때 시스템을 종료시킨 후 다시 실행시키면 캐쉬 사전에 등록된 어휘들은 기본 1000 단어 이외에는 모두 사라진다. 캐쉬 사전과 마스터 사전의 일관성은 항상 유지된다.

맞춤법 검사기는 일단 캐쉬 사전을 검색하고 나서 마스터 사전을 검색한다. 마스터 사전은 인덱스가 트라이(tric)구조로 구성되어 주기억 장치에 상주하며, 나머지 부분은 블록단위의 선형구조로 보조 기억장치에 위치한다. 한 번 검색된 블록은 버퍼에 남아 있게 된다.

문법 형태소 사전은 이진 정렬 트리 구조로 구성되어 있다. 문법 형태소 사전은 그 특성상 사용자에게 편집이 허락되지 않으며, 어휘 형태소 사전은 현재 플래그로써 사용자 정의 어휘와 시스템 제공 어휘를 구별하는 전략을 채택하고 있다. 이 전략은 최종 시스템의 응답속도가 만족스럽다면 독립된 사용자 정의 사전을 유지하는 것으로 전환될 것이다.

사용자가 사전을 편집하는 경우, 사전내의 일관성 및 정확성의 유지가 사용자의 책임하에 있게 된다. 따라서 사용자로 하여금 쉽게 사전을 편집하면서 정확하게 처리할 수 있도록 도울 수 있어야 한다. 본 시스템에서는 사전 편집기를 제공하여 사용자가 쉽게 사전을 편집할 수 있도록 하고 있다. 사용자는 시스템 제공 사전에 대해서는 검색만을 할 수 있다. 사용자가 어휘를 추가 수록할 때, 필요한 정보는 품사 및 불규칙 정보 뿐이지만, 사용자가 어렵게 느낄 수도 있기 때문에 도움말을 제공하고, 용언의 불규칙을 입력할 때에는 용언의 특성을 자동으로 조사하여 가능한 활용 형태를 생성, 사용자가 그러한 형태가 사용되는지를 대답만 하면 되도록 구성되어 있

다. 예를 들어 사용자가 "아름답다"라는 형용사를 입력하려고 한다면, 시스템은 "아름답아-라고 쓰입니까" 라는 질문을 하며, "아니오"라는 사용자의 대답에 따라 "아름답다"라는 어휘가 ㅂ 불규칙이라는 것을 자동 입력하는 것이다. 불규칙의 경우 해당 어휘가 한정되이 해당 어휘를 시스템 제공 사전에 모두 수록하고, 사용자는 규칙 용언만을 추가할 수 있는 집근 방법도 가능하다. 그러나 합성어가 용언을 구성할 수 있이서, 그러한 집근 방법은 고급사용자에게 제한이 된다. 제공되는 사전 편집기는 어휘의 추가뿐만 아니라 검색, 수정, 삭제 등도 지원한다.

PC에서 지원되는 사전은 정전 등으로 인한 시스템의 비정상적 종료에 대하여 대처할 수 있어야 한다. 그러한 비정상적 종료가 사전 구조의 붕괴를 야기한다면 PC환경에서는 치명적인 약점이 될 수 있다. 시스템의 사전은 변경이 발생할 경우, 즉시 그 변경을 보조 기억 장치에 반영하여 비정상적 종료에도 사전 구조의 붕괴는 발생하지 않으며, 최악의 경우 추가된 몇 개의 항목을 잃어비릴 수는 있지만 전체에 영향을 미치지는 않는다. 잃어비린 항목들은 누적되지 않는다.

3.4 사용자 인터페이스

사용자의 입장에서 맞춤법 검사 시스템은 기존의 문서 편집기에 맞춤법 검사 기능이 추가된 것이다. 맞춤법 검사는 메뉴에서 선정하여, 커서의 현재 위치와 문서의 처음 위치 중 하나로부터 시작한다. 맞춤법 검사에서 틀린 어절로 판정된 경우에 대한 인터케이스는 2장에서의 논의를 그대로 적용하고 있다. 이러한 인터페이스의 관리는 선행 처리기에서 맡고 있다.

3.5 맞춤법 검사기

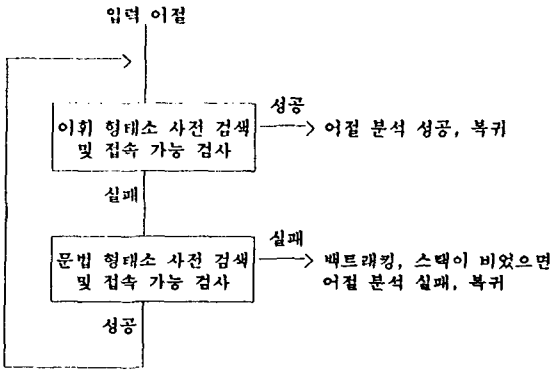

그림 2. 개략적인 처리 알고리즘

맞춤법 검사기의 처리 알고리즘은 [4]를 채택하였다. 입력되는 어절은 뒷부분부터 처리되며, 문법 형태소 사전에서 검색된다. 검색된 문법 형태소들은 스택 구조에 저장된다. 맞춤법 검사에서는 일단 하나의 가능한 형태소 간의 조합이 찾아지면 분석을 중지하고 다음 처리를 하면 되므로, 한글 어절의 특성상 어떠한 순서의 검색이 처리시간을 최적화하는가의 문제가 의미를 갖는다. 개발된 맞춤법 검사기에서는 한글에서는 명사, 관형사, 부사 등이 홀로 쓰이는 경우가 빈번하며, 사전 검색 시 한 번 검색된 블록은 다시 읽어오지 않는 사전 검색 알고리즘의 특성을 이용하여 전체 이절에 대한 검색을 먼저 시도하고, 실패하면 문법 형태소의 검색을 시도한다.

계속하여 찾아지는 문법 형태소들은 매번 이전의 문법 형태소들과 접속 가능 검사를 한다. 보조 기억 장치에 대한 접근을 요구할 가능성이 있는 어휘 형태소의 검색은 문법 형태소들이 모두 접속 가능한 경우에 한해서 시도된다.

3.6 틀린 어절에 대한 처리

틀린 어절은 시스템의 처리 오류, 사전의 미등록, 사용자의 입력 오류, 맞춤법에 대한 지식 부족 등 여러가지일 수 있다. 맞춤법 검사 시스템의 입장에서는 시스템의 오류, 미등록어의 존재를 일단 무시하고, 사용자의 입력 오류나 맞춤법에 대한 지식 부족으로 간주하여 대안을 제시하는 것이 바람직하다. 시스템이 제안하는 대안에 따라 사용자는 오류의 종류를 파악할 수 있을 것이다. 시스템의 대안은 맞춤법이나 표준이에 대한 지식 부족으로 인한 오류에 초점을 맞추었다. 사용자의 입력 오류일 경우, 그 옳은 형태를 사용자가 안다면, 사용자는 맞춤법 검사기에서 틀린 어절로 판정되었다는 사실만으로 옳은 형태로 수정할 수 있을 것이다. 예를 들어, "헵센"이 입력될 경우, "웹센"을 그 대안으로 제시하게 된다.

일반적으로 사용자들이 주로 틀리는 유형이 종합, 분류될 수도 있는데, 그러한 유형들에 대해서는 따로 사전의 정보를 추가하여 사용자에게 옳은 형태를 제시하도록 하였다. 예를 들어, "남비"라는 입력이 들어오면 "냄비"라는 대응이를 준비하여 제시하는 것이다. 앞의 대안 제시와 다른 점은, 사전에 미리 그 대응어를 준비하여 제시한다는 집이다. 이러한 처리는 사용자가 후보제시 기능을 사용해도 그 대응어를 시스템이 생성, 제시하기 어려운 형태(맞상 ==> 겸상)를 처리할 수 있고, 후보제시 기능에서 다른 형태를 제시함으로써 사용자를 혼란스럽게 하는 것을 방지할 수 있다.

IV. 결론

맞춤법 검사는 언어를 그 처리 대상으로 하기 때문에 방대함, 모호함, 많은 예외 현상 등에 직면하고 있다. 한글로 표현 가능한 전체 집합을 대상으로 하는 경우 아직 맞춤법 검사기의 정확성은 완전하지 않지만, 사용빈도가 높은 집합을 대상으로 하면 사용자에게 큰 도움을 줄 수 있다. 개발된 시스템은 사용자가 어려움을 느끼는 부분에 대하여 중점적으로 개발, 지원하고, 틀린 이절에 대한 처리의 질의 향상을 통하여 사용자의 만족도를 높였다. 또한 실제 사용되는 문서의 다양한 입력을 대부분 포괄하는 인터페이스를 제공하고 있다. 맞춤법 검사기의 문제는 많은 부분이 사전에 귀착된다. 그런데 사전은 사용함에 따라 성장과 안정의 속성을 가지므로 한글 맞춤법 검사기의 빠른 실용화가 요구된다.

현재 한글 맞춤법 검사기가 안고 있는 문제는 한글 맞춤법 규정 자체에서 비롯되는 것도 많다. 한글 맞춤법 검사기의 기반 기술은 한국어의 형태소 분석인데, 이 기술은 한국어의 통계자료를 얻기 위한 기술로 사용될 수 있다. 그러한 통계자

료는 다시 보다 정확한 한글 맞춤법의 제정에 기여할 수 있을 것이다.

맞춤법 검사 시스템이 갖추어야 하는 요소중의 하나는 적절한 도움말 기능이다. 적절한 도움말은 어떤 이유로 현재의 어절이 틀렸을 것이라는 조언을 해줄 수 있는 것을 말하는데, 이는 결국 틀린 어절에 대한 정확한 처리와 연관되는 기술에 해당된다. 맞춤법 검사 시스템이 단순한 문서 편집을 돕는 것에서 교육용으로서의 의미를 가지기 위해서는 이러한 기술이 더욱 개발되어야 할 것이다.

참 고 문 헌

[1] 강 재우, 송 춘환, 김 연배, 최 기선, 권 용래, 김 길창, "한글 철자 및 띄어쓰기 검사기의 설계 및 구현," 한글날 기념 학술대회 발표 논문집, 한국 인지 과학회, 한국 정보 과학회, 1989.

[2] 남 기심, 고 영근, 표준 국어 문법론, 탑출판사, 1986.

[3] 미 승우, 새맞춤법과 교정의 실제, 어문각, 1988.

[4] 박 충만, "효율적인 한국어 형태소 분석기 및 철자 검사 교정기의 구현," 서울대학교 공학석사 학위논문, 1990.

[5] 송 춘환, 강 재우, 김 연배, 최 기선, 권 용래, 김 길창, "한글 철자 및 띄어쓰기 검사기," 한국 정보 과학회 가을 학술 발표 논문집, 1989.

[6] Peterson J. L., "Computer Programs for Detecting and Correcting Spelling Errors," CACM Vol. 23, No. 12, Dececber, pp. 676-687, 1980.