한글 매크로시소러스 구축의 실제

Practical Construction of Hangul Macro-Thesaurus

최석두, 이화여자대학교 문헌정보학과

Suk-Doo Choi, Dept. of Lib. & Inf. Sci., Ewha Womans University

우리나라에도 여러 가지 분야별 시소러스가 연구되고 있으나 여러 분야를 망라하는 대규모 한글매크로시소러스는 아직 없다고 보아야 할 것이다. 또한 분야별 시소러스를 통합하여 매크로시소러스를 구축하는 것은 거의 불가능하며, 통합할 만큼의 분야별 한글 시소러스도 없다. 본 연구에서는 처음부터 매크로시소러스 구축을 위하여 개발한 범용 시소러스관리시스템의 내용과 이 시스템을 이용하여 전 분야를 대상으로 개발하고 있는 한글 매크로시소러스의 개발현황에 대하여 논하고자 한다.

1 서론

시소러스는 의미적으로 관련이 있는 용어들을 모아 조직하고 자연언어를 보다 통제된 시스템언어로 변환할 수 있기 때문에 색인과 검색분야에서의 시소러스 활용방법은 다양하다. 그 중에도 매크로시소러스는 그 자체의 효용성뿐만 아니라 마이크로시소러스의 개발에도 크게 기여하게 될 것이다.

본고에서는 매크로시소러스개발을 목표로 개발된 시소러스구축시스템과, 아직은 매크로시소러스라고 할 수 없지만 그 시스템을 이용하여 개발 중에 있는 한글 매크로시소러스의 개발현황에 대하여 논하고자 한다.

2 시소러스관리시스템

시스템의 환경은 *Digital Server*(주기억 96MB, 보조기억 9GB), *Microsoft SQL Server 6.5, Windows NT 4.0*에서 *PowerBuilder*로 개발되었다. 시소러스관리시스템의 주요 기능과 화면으로 사용자등록, 용어관계의 정의, 패싯보기, 시소러스검색, 시소러스출력, 시소러스편집 등이 있다(그림 1 참조). 이들은 각각 별도의 아이콘을 가지고 있으며, 패러미터화 되어 있어서 주프로그램과 별도로 갱신할 수 있다.

2.1 사용자등록

사용자등록은 등록으로 끝나는 것이 아니라 로컬프로그램을 가져야 한다. 로컬프로그램은 화면의 아이콘을 선택함으로써 바로 센터의 시소러스관리시스템에 링크되도록 짜여 있으며, 하나의 시소러스데이터베이스를 여러 사람이 공동으로 갱신할 수 있다.

2.2 용어관계의 정의

용어관계의 정의에는 '관계의 정의'와 '심볼의 정의'가 있다. 용어관계는 얼마든지 추가확장이 가능하다. 또한 대응되는 심볼을 정의하면 시소러스

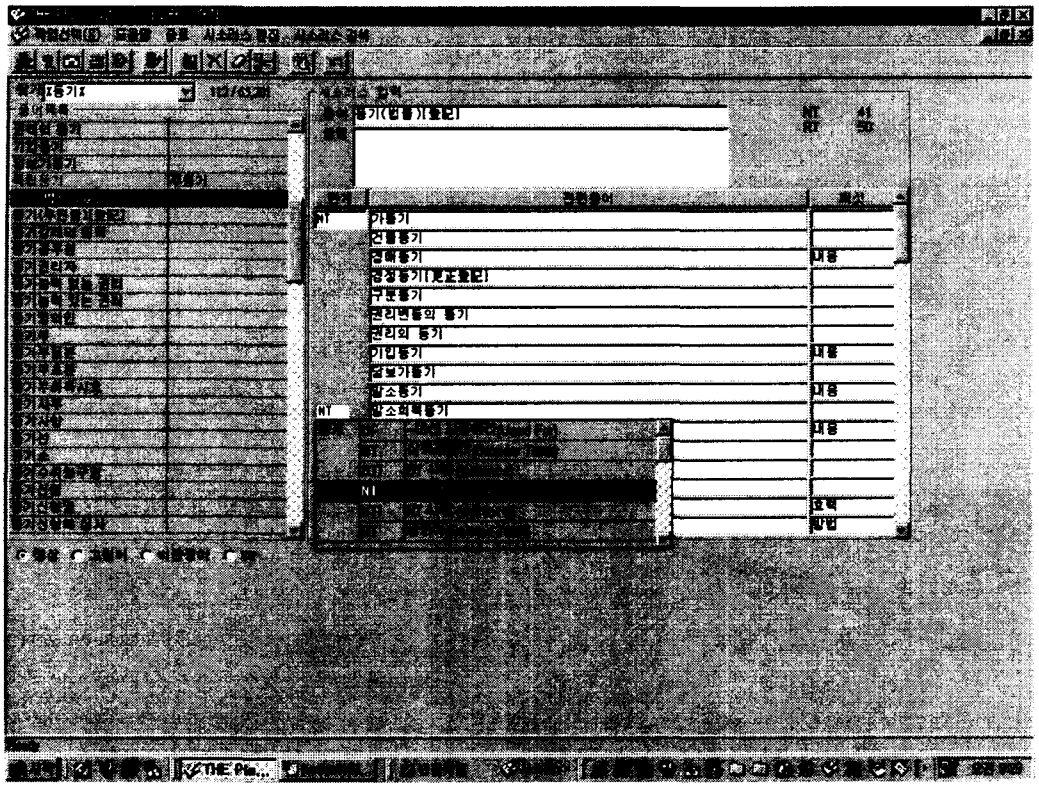

그림 1 시소러스 편집화면

검색기능에서 대신 그 심볼을 사용하게 되며, 그 심볼의 크기를 정수로 부여하면 각 관계의 배열순서가 결정된다. 예를 들면, *nt*에 30, *bt*에 20, *rt*에 40, *uf*에 10의 크기를 부여하면 관계의 배열순서는 *uf, bt, nt, rt* 의 순서가 된다.

2.3 패싯보기

패싯보기는 지금까지 사용된 패싯을 자모순으로 그 사용횟수와 함께 보여준다. 새로운 패싯을 결정할 때 참조하게 된다.

2.4 시소러스검색

시소러스검색은 현재까지 만들어진 시소러스를 검색하는 기능으로 검색용어가 포함되어 있는 트리구조 전체를 보여준다. *folder* 개념을 사용하여 트리구조의 각 노드를 보이거나 감출 수 있으며 모든 용어관계를 망라한다.

2.5 시소러스출력

시소러스출력은 자모순시소러스와 계층시소러스를 출력할 수 있다. 자모순시소러스에서는 각 디스크립터를 중심으로 상하위어관계 1계층만을 갖는 자모순시소러스와, 각 디스크립터를 중심으로 모든 상위개념어와 모든 하위개념어를 자모순으로 정렬한 자모순시소러스의 두 가지를 출력할 수 있다. 후자의 자모순시소러스는 계층관계에 대하여 명확한 개념을 갖고 있지 않은 일반사용자에게 편리하며, 전자의 자모순시소러스는 명확한 계층관계를 정의해야 하는 시소러스개발자에게 편리하다(최석두, 1994). 본 시스템에서는 어느 쪽도 하드카피형식과 파일형식 양쪽의 출력이 가능하다.

2.6 시소러스편집

시소러스편집은 시소러스구축시 개발자가 가장 빈번하게 사용하는 기능으로, 용어의 검색, 수정, 삭제(일괄삭제 포함), 추가, 입력(반복입력, 기존입

력이용, 참조입력, 패싯입력 등을 포함) 등을 가지고 있다. 그림 1은 편집화면을 중심으로 보인 것이다. 참조입력이란 고립어나 현 시점에서 관계를 판정할 수 없는 미판정어를 입력하여 두고 후에 이를 정규용어로 처리하기 위한 것이다. 그림 1의 왼쪽 하단의 해당항목을 선택하여 입력하거나 참조한다.

시소러스편집기능에서는 복수 개의 시소러스를 편집할 수 있으며, 현재 세 종류의 전혀 다른 시소러스를 편집하고 있다.

그림 1에서 왼쪽 상단부의 숫자는 등록된 총용어수와 검색된 용어수를 보여주고 있다. 화면에서의 총용어수는 65,201어이며, '등기'를 양쪽절단으로 검색한 결과가 112가 된다. 화면은 왼쪽의 '검색부분'과 오른쪽의 '편집부분'으로 나뉘어져 있다. 검색부분을 클릭하면 내용이 편집부분으로 표시되지만, 편집부분 내에서의 변화는 검색부분에 영향을 미치지 않는다. 즉, 편집부분의 각 항목을 더블클릭하면 검색부분은 변하지 않고 편집부분만 계속 바뀌게 된다. 따라서 편집시 참조데이터를 적절히 이용할 수 있으므로 매우 편리하다.

오른쪽 상단에는 편집부분에서 참조하고 있는 용어와 관련된 각종 관계어의 수를 보이고 있다. 그림 1에서는 참조하고 있는 용어 '등기(법률)'의 *nt, rt* 등의 각 수가 '41, 50'이라는 것을 보이고 있다.

왼쪽 상단부의 '찾기'에서는 현재 등록된 용어를 한글식별부분으로 검색할 수 있으며, 왼쪽절단검색, 오른쪽절단검색, 양쪽절단검색이 가능하다. 또한 외국어로도 검색할 수 있다. 외국어로 검색하면 단일 외국어용어가 복수의 한글용어로 대응되는 내용도 볼 수 있다. 검색부분은 두 칼럼으로 나뉘어져 있으며, 대부분 왼쪽이 디스크립터이지만, 왼쪽이 희미하게 표시되고 오른쪽에 용어가 있는 것은 오른쪽이 디스크립터이며 왼쪽은 *uf*임을 나타낸다(예에서 '독립등기'는 이며 '주등기'가 디스크립터라는 것을 나타낸다). 이 찾기란의 오른쪽 스크롤심볼을 누르면 지금까지 처리한 이력을 900회까지 볼 수 있으며 그중 하나를 선택하면 해당 용어와 용어관계가 편집부분에 표시된다.

가장 오른쪽의 패싯은 *nt*인 경우에만 입력할 수 있도록 하였다. 이 패싯은 시소러스검색시 출력에서 사용된다. 이 패싯은 출력시 옵션에 따라 다음과 같이 사용된다.

```
등기(법률)[登記]
        (내용)
    nt  경매등기
    nt  기입등기
        .....
        (방법)
    nt  부기등기
    nt  주등기
        .....
```

설명란은 *sn*이며 길이에 관계없이 텍스트데이터를 입력할 수 있다. 편집부분의 하단에 보이는 윈도우가 용어관계기호를 선택하는 부분이 되며, 마우스로 선택하거나 관계기호의 두문자를 입력함으로써 입력할 수 있다.

또한 각종 에러검사기능과 관계의 논리적인 오류를 검사하고 알려주는 기능들을 가지고 있다. 특히 두 용어에 대하여 각각 용어관계를 정의한 후 이 두 용어가 동의어인 것을 알게 되면 두 개의 디스크립터를 하나의 디스크립터로 모으면서 그 중 하나를 *uf*로 결정하는 방법이 매우 간단하다. 즉, 어느 쪽이든지 현재 상황에서 상대쪽 용어에 대하여 *use* 혹은 *uf*관계를 만들고 저장하기만 하면 된다.

3 매크로시소러스의 개발

3.1 개발방향

한글 매크로시소러스를 개발할 때는 전통적인 시소러스에 더하여 다음과 같은 부분이 보완되어야 한다고 생각된다(최석두, 1996; 최석두외, 1996).

1) 용어를 망라해야 한다: 해당 분야에서 사용되고 있는 모든 일반용어, 전문용어, 이들의 복합어, 이들에 대응되는 속어, 방언과 외국어, 필요하다면 고유명사까지도 포함되어야 한다.

2) 관계를 확장하여야 한다: 전통적으로 사용되어온 관계는 단순하게 추상화하여 기본적으로 *nt, bt, rt, use, uf, sn*을 중심으로 대응되고 있다. 용어관계를 다른 측면에서 보아 구별하거나 세분하여 구별할 필요가 있다. 또한 *nt*의 계층수

는 제한이 없어야 한다.

3) 각종 언어정보를 가져야 한다: 구문, 의미, 문맥, 共起, 사례, 통계 등과 같이 언어처리에 필요한 각종 언어정보를 가져야 한다.

4) 해설을 가져야 한다: 시소러스에 포함되어 있는 용어에 대한 해설이 필요하다. 사전 및 백과사전들을 링크할 수도 있으며, 그래픽스, 동화상, 소리 등이 포함될 수도 있을 것이다.

본 시소러스의 구축과정에서는 1)과 2)를 바탕으로 하고 있다. 3)과 4)의 부분은 외부정보를 링크할 수도 있으므로 차후의 과제로 삼고자 한다.

3.2 개발내용

그림 1에서 보는 바와 같이 현재 시스템에 등록된 용어수는 65,201어이다. 이 수는 디스크립터와 비디스크립터를 포함한 것이다. 현재 시소러스의 규모와 다루고 있는 분야를 보면 매크로시소러스를 지향하고 있는 시소러스라 할 수 있으나, 예상으로는 등록용어수가 20만어 정도가 되면 일단 매크로시소러스의 형식을 갖추게 될 것이며, 포괄적인 분야에서의 사용도 가능하리라 사료된다. 궁극적으로는 100만 단위 규모의 용어를 가질 때 명실공히 매크로시소러스가 되리라 생각된다.

어떤 의미와 관련된 용어는 표현의 다양성과 다의성, 계층관계의 다중성, 개념의 부분공유, 중첩구조, 관계의 상대성 등의 특성을 갖고 있다. 특히 전문용어는 정의가 명확하고 다의성을 줄이기 위하여 복합어가 많으며 조어규칙이 단순한 면이 있다. 이와 같은 특징을 살리면서 전체적으로 구조화되도록 시소러스를 설계하는 것이 가장 어려운 일이었다.

용어의 관계는 *nt, bt, rt, nti/bti*(상하관계 중 사례관계: 주로 고유명사에 사용한다), *ntp/btp*(상하관계 중 전체-부분관계), *ntg/btg*(상하관계 중 속관계), *use, uf, sn, tt*(최상위개념어), *dc*(디스크립터 식별기호), *cc*(분류기호), 고립어, 미판정어, *eng/ken*(영어/한글용어), *jap/kja*(일본어/한글용어), *ger/kge*(독일어/한글용어), *fra/kfr*(불어/대응한글용어), ... 등으로 세분하여 사용하고 있으며, 차후 *rt* 중에서 '반의어'를 별도로 분리할 것을 검토하고 있다. 방언, 외국어, 고유명사도 포함되어 있다. 방언과 고유명사의 예를 보면 다음과 같다.

아버지
- *uf* 아바이(아버지: 함경도방언)

독립운동단체
- *nti* 경학사[耕學社]
- 고려혁명군
- 고려혁명당

디스크립터/비디스크립터의 서로 혹은 각각 동형이의어가 있는 경우에는 한글/외국어를 불문하고 한글한정어로 구별하고 있다. 그림 1에서 '등기'라는 용어는 '법률, 우편물'이라는 한정어로 구분되고 있다. 시스템 상에서의 디스크립터 식별은 한글 부분(한정어 포함)만으로 한다. 즉, 부기하는 한자는 단지 참조정보일 뿐, 식별정보는 아니다(한글한정어의 부기원칙에 대해서는 문헌 '김태수, 최석두, 1997' 참조). 현재까지 한정어로 구별된 용어는 2,300개 정도이지만 용어가 늘어날수록 한정어 부가용어의 비율이 높아지고 있다.

4 결론

현재 구축중인 한글매크로시소러스에 대하여 관리시스템과 구축결과를 중심으로 논하였다. 현재 본 시소러스에는 용어의 분류부분이 없으며 그 분류방법에 대한 案도 명확하게 가지고 있지 않다. 그러나 차후 대단위 용어를 가진 매크로시소러스에서 마이크로시소러스를 만들기 위해서는 용어의 분류가 필수적이다. 용어의 분류에 대한 연구가 필요할 것이다.

참고문헌

김태수, 최석두. 1997. 동형이의어의 구별을 위한 한글한정어 사용에 관한 연구. 「情報管理學會誌」, 14(1):107-124.

최석두. 1994. 시소러스의 표시형식에 관한 연구. 「1994年度 韓國情報管理學會 全國論文大會(第1會) 論文集」, 105-108.

최석두. 1996. 한글 매크로시소러스의 개발방향. 「한국어 정보검색 기술세미나」, 35-50.

최석두, 김영환, 남영준. 1996. 하이텔 메뉴검색용 시소러스의 개발에 관한 연구. 「情報管理學會誌」, 13(1):227-241.