(제 10회 한글 및 한국어 정보처리 학술대회)

북한 문화어 형태소 분석기(NKMA)의 어절 구조

최운호
서울대학교 언어학과
서울시 관악구 신림동 산56-1
우: 151-742
whchoi@clepsi.co.kr

정희선
(주) 언어과학
서울시관악구 봉천7동 1595-1
우: 151-057
chsunny@clepsi.co.kr

The Word Structure of the North Korean Morphological Analyzer

Choi, Woon-Ho
Department of Linguistics
Seoul National University

Chung, Hoi-Sun
Parole Science, Inc.

요약

분단 이후 북한은 우리와는 다른 언어정책을 추진해 왔고, 그 결과로 지금은 남북한 언어 정책에서 많은 차이를 드러내게 되었다. 본 논문은 북한 문화어 형태소 분석 시스템(NKMA)의 구축을 위한 어절 구조를 제시한다. 북한 문화어의 형태소 분절 및 분석을 위해 사용된 어절 구조는 대체로 말토막 단위와 일치하므로, 음성언어의 인식을 위한 분절 방법에 응용될 수도 있으리라 기대한다.

1. 들어가는 말

우리말 표준어를 대상으로 한 형태소 분석기에 대해서는 많은 연구와 개발이 진행되어 왔으며, 우수한 성능의 시스템도 공개되어 있다. 그러나, 북한 문화어를 대상으로 한 형태소 분석기에 대한 연구 발표는 없었다.

분단 이후, 북한은 우리 나라와는 다른 언어 정책을 펼쳐 왔으며, 그 결과 북한 문서, 신문 등에서 준수하는 북한의 문화어 규정은 우리말 표준어 규정과는 상당한 차이를 보이고 있다.

본 연구는 북한 문화어를 대상으로 한 형태소 분석기(NKMA) 구현을 위해 북한 문화어 규정에 따른 어절 구조를 설계하는 것을 목적으로 한다.

북한의 어문 규정(조선말규범집)은 "한글 맞춤법 규정"과는 달리 다수의 형태론적 구성(의존 구문 구성, 합성 용언 구성, 용언 연쇄 구성)에 대해서 붙여 쓰기를 규정하고 있으며, 이 점은 일부 보조용언 구성에 대해서 붙여 쓰기를 허용하고 있는 "한글 맞춤법 규정"과 큰 차이를 보이고 있으며, 일부 어형 변화에서 음운론적 변이형의 목록에도 차이를 보인다. 따라서, "보조적동사이기 때문이였다", "갈라내는것도" 등과 같은 통사적 구성도 하나의 붙여쓰기 단위로 쓰도록 규정한다.

이러한 북한 문서에 대한 형태소 분석을 위해서 북한어의 어절 구조를 SEGMENT라는 계층 구조로 설정하였으며, 통사적 구성을 형태론적 구성으로 분절하는 것과 각각의 형태론적 구성에 대한 형태소 분석 작업을 동시에 진행한다.

북한 문화어에 대한 형태소 분석기는 2가지 의미를 지닐 수 있다.

첫째, 북한의 문서(신문 등 기타 자료)에 대한 대량 분석을 수행할 수 있는 형태소 분석 시스템 구현에 의미가 있으며, 이렇게 구성된 형태소 분석기는 정보 검색, 언어 처리 등에 응용될 수 있다.

둘째, 북한의 어문 규정에서 규정하는 "붙여쓰기"는 대체로 우리말 말토막 단위와 일치한다고 볼 수 있다. 따라서 음성언어의 분석을 위한 형태소 분절 시스템에 응용될 수 있다. 물론, 음성언어와 문자언어의 자료는 표기법 문제 등을 비롯해서 완전히 다른 모습을 보이지만, 분절 방법은 대체로 약간의 변용을 통해서 음성언어에도 적용할 수 있다고 본다.

2. 형태소 분석의 관점에서 본 표준어와 문화어의 차이

우리나라의 공용어는 표준어이며, 표준어의 정의 및 맞춤법에 대한 규정이 한글 맞춤법 및 표준어 규정이다. 북한에서는 표준어라고 하면 서울말을 표준으로 하는 것으로 이해될 수 있다고 하여 김일성의 1966년 교시에 의해 평양말을 중심으로 표준말을 정의하고 이를 "문화어"라고 불러왔다([6, p. 40, pp. 294-306]).

2.1 언어학적 차이

문자자료만을 대상으로 하는 형태소 분석기의 설계, 제작에 문화어와 표준어의 언어학적인 차이는 큰 장애가 되지 않는다. 왜냐하면 문자 표기 방법에 반영된 언어학적 차이만이 형태소 분석기를 설계할 때 고려 대상이 되기 때문이다. 음성, 음운론적인 차이(장단, 성조 등)가 문자 생활에도 세밀히 반영된다면 각 방언권마다 다른 형태소 분석기를 설계해야 하겠지만 문자생활에는 이러한 요소가 잘 나타나지 않는다. 이러한 차이가 나타난 단순한 예로는 "예절, 노동" 등의 어휘를 북한에서는 "례절, 로동" 등으로 표기한다는 것인데, 이것은 북한의 사전을 가지고 어휘사전을 구축하면 될 뿐, 문제가 되지 않는다.

언어학적인 차이가 문자생활에 반영되어서 표준어와의 차이점을 보이는 예 중의 하나로는 표준어에서는 어미/선어말어미에서 "아/어"만이 음운환경에 따라 이형태의 일부로 대립하고 '-하다'와 같은 여-불규칙 용언에서만 "여"형의 어미가 나타나는데, 북한의 문화어에서는 "아/어/여"가 음운조건에 따라 이형태의 일부로 대립한다는 것이다. 이것은 말음이 개음절이 특정한 부류의 어간이 이러한 어미와 결합할 때, 표준어에서는 축약이 우세한 현상으로 나타나지만 북한 문화어에서는 매개자음 /j/의 삽입이 우세한 현상으로 나타나는 것으로 설명할 수 있는데, 물론 규범집에서는 일부 어휘에 대해 언어현실을 고려하여 축약도 문화어로 인정한다.

2.2 언어정책적 차이

한글 맞춤법 규정은 조사의 경우를 제외하고는 단어를 기준으로 띄어 쓰는 것을 원칙으로 하고 붙여 쓰는 것이 가능한 허용 규정을 제시하고 있다. 반면, 북한의 조선말규범집에는 붙여 쓰는 것에 대한 허용 규정은 당연 규정만이 있다. 북한 문화어를 형태소 단위로 분석할 때 표준어의 어절 구조와 가장 큰 차이를 유발하는 것은 바로 띄어 쓰기 규정의 차이인데, 북한 문화어의 경우 모든 불완전명사와 이에 준하는 단위들은 원칙적으로 붙여 쓰도록 하며 일부 경우에만 띄어 쓰는 것으로 조절하고 있다. 뿐만 아니라 모든 보조용언도 붙여 쓰며 일부 어미가 개입되어 자립적인 용언이 어울리는 경우에도 붙여 쓰도록 규정하고 있는데, 이 부분에서는 규범집에서 설명을 위해 예로 든 어휘들과 조선말대사전의 어휘목록이 일치하지 않거나 불균형을 이루는 것을 파악할 수 있다. 이러한 불균형으로 인해 북한 문화어 형태소 분석기의 분석 단계가 표준어를 대상으로 한 한국어 형태소 분석의 단계보다 많아지게 된다. 예를 들면, [8, p. 4]에는 "받아들이고있는것이"라는 어절이 등장하는데, 우리 표준어의 맞춤법 규정에 따르면 "받아들이고 있는 것이"라고 띄어 쓰는 3개의 어절이 모두 1개의 어절로 붙어 있다. 규범집에 복합어의 예로 등장하는 "캐고들다"라는 어휘는 조선말대사전에서 복합어로 취급하지 않기 때문에, "캐고들어갔던것을"이라는 어절을 분석해야 한다면 "캐고들어가+았+던+것+을" → "캐고들+어+가+았+던+것+을" → "캐+고+들+어+가+았+던+것+을"의 과정을 거쳐야만 한다.

이러한 어절 구조의 차이점은, 표준어의 어절은 기본적으로 형태론적 구성과 거의 일치하는 반면, 문화어에서는 형태론적 구성 뿐만 아니라 통사론적 구성도 하나의 어절로 표현하도록 규정한다는 것으로 요약할 수 있다.

2.3 표준어와 문화어의 어절구조 비교

북한 문화어에서는 통사론적 구성까지도 하나의 어절로 표현하도록 규정하고 있다. 따라서, 문화어 어절 유형의 가지수는 표준어보다 많다.

강승식은 [4]에서 표준어 단어(각주-1)의 유형을 52개로 분류해 놓았다. 이 분류에서 기술하고 있는 것은 형태론적 구성으로 이루어진 어절이며, 여기에서 예외가 되는 것은 다음과 같이 통사론적 구성이 하나의 어절로 사용되는 경우이다.

(1) [4](강승식, 1996)의 한국어 단어의 유형 분류 중 통사론적 구성 유형

ㄱ. 체언 + 용언접미사 + '아/어' + 보조용언 + 어미
ㄴ. 체언 + 용언접미사 + '아/어' + 보조용언 + 선어미 + 어미
ㄷ. 용언 + '아/어' + 보조용언 + 어미
ㄹ. 용언 + '아/어' + 보조용언 + 'ㅁ/기' + 조사

ㅁ. 용언 + '아/어' + 보조용언 + 선어미 + 'ㅁ/기' + 조사

문화어에서 보조용언이 사용된 통사론적 구성의 띄어 쓰기는 한글 맞춤법 규정처럼 붙여 쓰는 것을 허용하는 허용 규정이 아니라 당연히 붙여 쓰도록 규정하는 당연 규정이다. 그리고 띄어 쓰는 예외를 인정하지 않는다. 뿐만 아니라, 모든 불완전 명사는 원칙적으로 붙여 쓰도록 규정하고 있다. 따라서, 문화어의 어절 유형은 당연히 그 가지수가 표준어의 어절 유형보다 많을 수밖에 없다.

용언 어간이 포함된 가능한 어절의 유형 중 일부를 나열식으로 기술하면 다음과 같다.

(2) 북한 문화어 어절의 유형 예

ㄱ. 어간 + 어미{고/아} + 어간 + 어미

ㄴ. 어간 + 어미[보조용언어미] + 보조용언 1형 어간 + 어미

ㄷ. 어간 + 어미[보조용언어미] + 보조용언 2형 어간 + 어미[보조용언어미] + 보조용언 2형 어간 + 어미

ㄹ. 어간 + 어미[보조용언어미] + 보조용언 1형 어간 + 어미[관형형] + 불완전명사 + 지정사 + 어미

ㅁ. 어간 + 선어말어미 + 어미[관형형] + 불완전명사 + 지정사 + 선어말어미 + {ㅁ/기} + 조사

위에서 제시한 어절 유형은 문화어에서 가능한 어절의 극히 일부에 지나지 않는다. 문화어의 어절 구조를 위와 같이 단층적으로 기술하는 것은 바람직하지 않다. 단층적으로 어절의 유형을 분류하는 것이 가능하기는 하지만, 우선 그 가지수가 너무 많아서, 형태소 분석 시스템에서 문화어의 어절 유형을 단층적으로 보고 그에 대한 자료 구조를 관리하는 것이 비효율적이다. 그리고 문화어에서는 통사론적 구성도 하나의 어절료 표현한다는 것을 충분히 고려하여 그 특성을 이용해야 하는데, 그러기 위해서는 어절의 유형도 계층적 구조에 기반해서 기술되고 정의되어야 하며, 형태소 분석 시스템에도 이러한 어절의 계층적 구조가 반영된 자료 구조를 사용해야 한다.

3. 문화어 어절 유형의 계층적 기술

문화어 어절을 계층적으로 기술하기 위해서 어절을 분절(Segment)의 구성으로 본다. 최상위의 분절을 Segment, 하위의 분절을 SegmentX로 기술하며, X는 각 세그먼트를 구분하는 일련 번호를 나타낸다. 각각의 형태소는 종단 분절(Terminal Segment: TS)이라고 정의하고, 형태소들이 결합되어서 이루어진 분절을 비종단 분절(Non-Terminal Segment: NTS)이라고 정의한다. 그러면 분절의 기본 구조는 개념적으로 다음과 같이 정의된다.

(3) 분절의 기본 구조

NTS → (NTS)(TS)* | TS+

NTS: Non-Terminal Segment

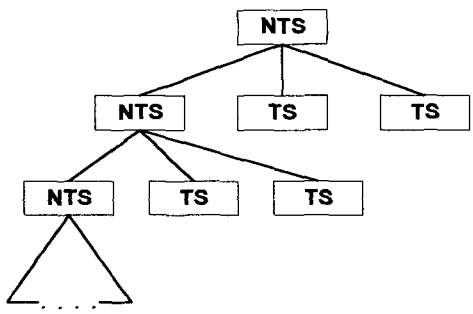
TS : Terminal Segment

* : Kleene Closure

+ : Positive Closure

NKMA에서 분절은 통사론적 구성을 반영하는 것으로 정의하고 어절의 유형에 대한 자료 구조를 설계하는 데 사용되었지만, 한 어절 내에서 나타나는 통사론적 구성을 모두 포함하지는 않았다. 이것은 NKMA가 기본적으로 통사 분석을 위한 것이 아니라 형태소 분석을 위한 시스템이기 때문인데, 모든 통사론적 구성을 어절 유형의 계층적 구성에 반영할 경우, 어절 유형의 계층의 깊이가 너무 깊어지기 때문이다.

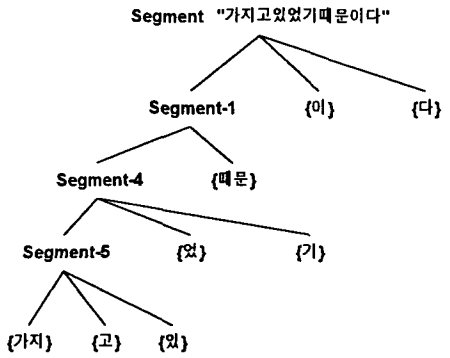
NKMA에서는 하나의 어절을 Segment로, 그리고 한 어절 안에 나타나는 통사론적 구성을 Segment-1, Segment-2 등으로 정의하여 어절의 구조를 계층적으로 구성하였다. NKMA에서 정의된 Segment와 SegmentX는 (3)의 정의에서 NTS가 구현된 것이며, 각 형태소들은 TS가 구현된 것이다. (3)의 정의를 도식화하면 다음 <그림 1>과 같다.



<그림 1> NTS 구조도의 예

(3) 및 <그림 1>과 같은 구조가 NKMA에서

는 어절 구조의 계층적 기술에 기반해서 각각의 Segment CLASS록 구현되었는데, 먼저 한 어절이 분석되었을 때 NKMA의 자료 구조에 저장되는 예를 구조도로 보이면 <그림 2>와 같다.



<그림 2> NKMA 어절 구조의 예

<그림 2>에서 "가지고있었기때문이다"라는 어절은 "본용언+보조용언" 구성, "용언형+불완전명사"의 구성과 같은 통사적 구성을 하나로 붙여서 사용한 예인데, 이러한 통사적 구성은 NKMA에서 분석과 동시에 계층적인 어절 자료 구조에 저장하게 된다. Segment-5는 "용언", "용언+보조용언어미+보조용언"과 같은 구성을 저장하는 분절로 (3)의 NTS에 해당한다. Segment-5의 {가지}, {고}, {있}은 각각 NTS인 Segment-5를 구성하는 TS 유형의 형태소가 된다.

4. 문화어의 어절 구성 전이도

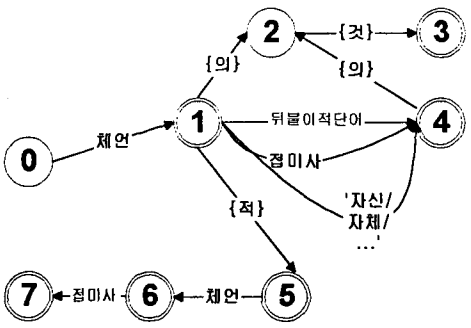
문화어 어절의 계층적 유형에 따라 어절 전이도를 구성하였다.

현재 전체 어절 구성인 Segment에서 Segment-7까지 모두 8개의 분절이 구성되어 있다.

4.1 Segment-7

Segment-7은 명사구 구성과 관련된 단위이며, <그림 3>과 같다.

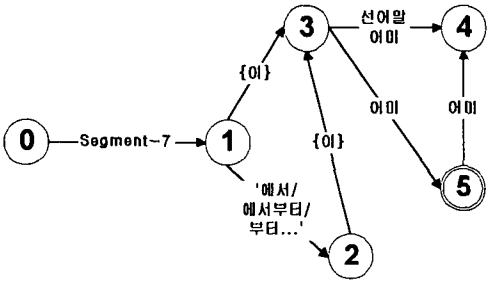
(4) Segment-7의 예
ㄱ. 문법적의미
ㄴ. 사회주의적과업들
ㄷ. 학생의것, 학생들의것
ㄹ. 려행자일행, 기사장자신, 아들딸모두



<그림 3> Segment-7

4.2 Segment-6

Segment-6은 Segment-7에 지정사가 결합하고 여기에 다시 어미가 첨가되는 구성으로 <그림 4>와 같다.



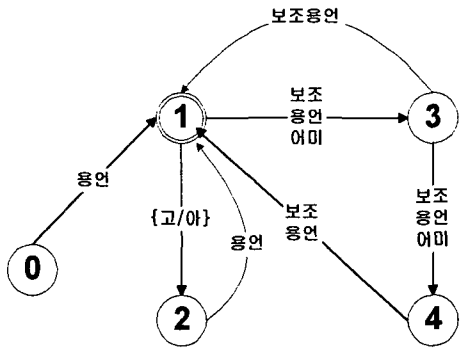
<그림 4> Segment-6

(5) Segment-6의 예
ㄱ. 학생들이였기때문이다
ㄴ. 학생들의것일
ㄷ. 지방에서부터이였을

4.3 Segment-5

Segment-5는 보조용언구성과 같은 의존구문구성, 복합용언구성 등과 같으며, <그림 5>와 같다.

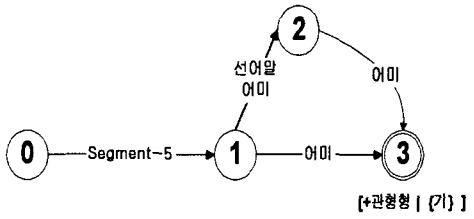
(6) Segment-5의 예
ㄱ. 먹고떨어지다
ㄴ. 기여넘어가다
ㄷ. 되고있다
ㄹ. 되여있다
ㅁ. 가고싶다
ㅂ. 되여가고있다



<그림 5> Segment-5

4.4 Segment-4

Segment-4는 Segment-5에 관형형 어미 또는 명사형 어미 {기}가 첨가된 형태로, <그림 6>과 같다.

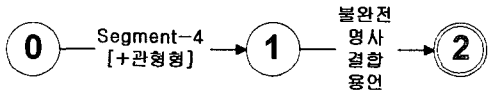


<그림 6> Segment-4

(7) Segment-4의 예
ㄱ. 되여가고있는것이다
ㄴ. 돌아가고싶었던것이다
ㄷ. 되여가고있을듯하였기때문에

4.5 Segment-3

Segment-3는 Segment-4에서 관형형 어미가 첨가된 형태에 불완전명사 결합용언 '-듯하다/만하다' 등이 결합된 형태로 <그림 7>과 같다.

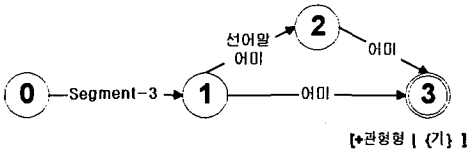


<그림 7> Segment-3

(8) Segment-3의 예
ㄱ. 되여가고있을듯하다
ㄴ. 했을듯싶다

4.6 Segment-2

Segment-2는 Segment-3에 관형형 어미 또는 명사형 어미 {기}가 첨가된 형태로, <그림 8>과 같다.

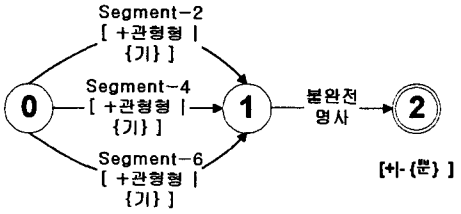


<그림 8> Segment-2

(9) Segment-2의 예
ㄱ. 되여가고있을듯하기때문에
ㄴ. 하였을듯싶기때문에

4.7 Segment-1

Segment-1은 Segment-2, Segment-4, Segment-6에 불완전명사가 결합된 형태로, <그림 9>와 같다.



<그림 9> Segment-1

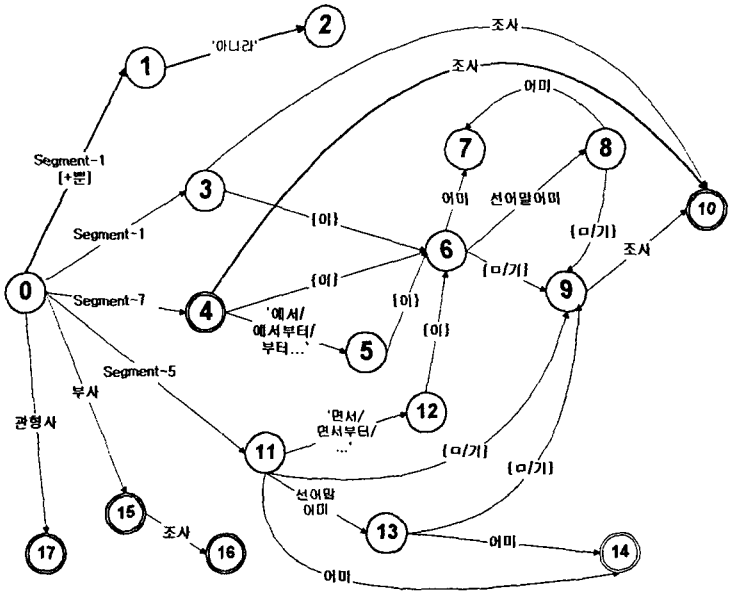
(10) Segment-1의 예
ㄱ. 학생들의것이기때문에
ㄴ. 되여가고있을듯하기때문에

4.8 Segment

Segment는 전체 어절의 구조로, NKMA에서 북한 문화어의 어절 구조는 <그림 10>과 같다.

<그림 10>은 이미 정의된 하위 분절(Segment)과 문법형태소들이 결합하는 관계를 나타낸 것으로, NKMA에서 사용되고 있는 어절의 구성이다.

(11) Segment: 문화어 어절의 예
ㄱ. 학생들의것이였기에
ㄴ. 되여가고있을수
ㄷ. 개발할만하였으므로
ㄹ. 인식되여있었고
ㅁ. 되여가면서부터이였기에



<그림 10> Segment: 문화어의 어절 구조

5. 분석 예

NKMA를 이용하여 분석한 예는 다음과 같다.

● 학생들이였기때문이다

▶Segment-1(Segment-6(Segment-7(학생[MT_N]+들[MT_NSUFFIX:DEUL])+이[MT_COP]+였[MT_T]+기[MT_LINK_GI])+때문[MT_BNOUN])+이[MT_COP]+다[MT_E]

● 학생들의것일수

▶Segment-1(Segment-6(Segment-7(학생[MT_N]+들[MT_NSUFFIX:DEUL]+의[MT_J]+것[MT_BNOUN])+이[MT_COP]+ㄹ[MT_E:DETAIL_ADN])+수[MT_BNOUN])

● 지방에서부터이였을

▶Segment-7(지방[MT_N])+에서부터[MT_JC]+이[MT_COP]+였[MT_T]+을[MT_E]

● 먹고떨어졌다

▶Segment-5(먹[MT_VA]+고[MT_E]+떨어지[MT_VA])+었[MT_T]+다[MT_E]

● 기여넘어가면서부터이였다

▶Segment-5(기[MT_VA]+여[MT_E]+넘어가[MT_VA])+면서부터[MT_EC]+이[MT_COP]+었[MT_T]+다[MT_E]

● 되여가고있었을것이다

▶Segment-1(Segment-4(Segment-5(되[MT_VA]+여[MT_E:DETAIL_E_AUX]+가[AUX_VA]+고[MT_E:DETAIL_E_AUX]+있[AUX_VA])+었[MT_T]+을[MT_E:DETAIL_ADN])+것[MT_BNOUN])+이[MT_COP]+다[MT_E]

● 돌아가고싶었던것이다

▶Segment-1(Segment-4(Segment-5(돌아가[MT_VA]+고[MT_E]+싶[MT_VA])+었[MT_T]+던[MT_E:DETAIL_ADN])+것[MT_BNOUN])+이[MT_COP]+다[MT_E]

● 되여가고있을듯하였기때문에

▶Segment-1(Segment-2(Segment-3(Segment-4(Segment-5(되[MT_VA]+여[MT_E:DETAIL_E_AUX]+가[AUX_VA]+고[MT_E:DETAIL_E_AUX]+있[AUX_VA])+을[MT_E:DETAIL_ADN])+듯하[MT_VA:DETAIL_BN_VA])+었[MT_T]+기[MT_LINK_GI])+때문[MT_BNOUN])+에[MT_J]

● 개발할만하였으므로

▶Segment-3(Segment-4(Segment-5(개발하[MT_VA])+ㄹ[MT_E:DETAIL_ADN])+만하[MT_VA:DETAIL_BN_VA])+였[MT_T]+으므로[MT_E]

6. 결론

북한의 언어정책과 우리말 표준어 규정 및 한글 맞춤법 규정의 차이로 인해 북한의 공용어인 문화어에 대한 형태소 분석기를 북한의 언어정책에 맞게 구성하였다.

북한 문화어 형태소 분석기는 북한의 언어자료에 대한 검색, 분석에 이용될 수 있을 뿐만 아니라, 우리말 음성언어의 분석을 위한 기초 기술로 적용될 수 도 있으리라고 본다. 북한의 띄어쓰기 규정이 대체로 말토막 단위와 일치하는 경

향이 있기 때문에, NKMA에서 사용한 분절 방법은 우리말 음성 언어의 처리에도 적용될 수 있으리라고 본다.

감사의 글

본 연구는 1997년도 문화관광부(구 문화체육부)의 지원으로 수행한 연구 용역 사업의 결과를 수정, 보완한 것이며, 문화관광부의 지원에 감사드립니다.

참고 문헌

[1] Silberztein, M. *Dictionnaire électroniques et analyse automatique de textes*, Masson, 1993

[2] 고신숙, 조선어리론문법-품사론, 과학백과사전출판사, 1987

[3] 국어사정위원회, 조선말규범집, 사회과학출판사, 1988

[3] 권종성, 조선어정보처리, 과학백과사전종합출판사, 1994

[4] 강승식, 한국어 형태소 분석을 위한 단어 유형 분류와 자료구조, 한글 및 한국어 정보처리 학술발표 논문집, pp 241-245, 1996

[5] 김동찬, 조선어리론문법-단어조성론, 고등교육도서출판사, 1986

[6] 김민수 편저, 김정일 시대의 북한 언어, 태학사, 1997

[7] 김용구, 조선어리론문법-문장론, 과학백과사전출판사, 1986

[8] 문영호, 조선어정보처리, 1993

[9] 문영호 외, 조선어빈도수사전, 과학백과사전종합출판사, 1993

[10] 사회과학원 언어연구소 편, 조선말 대사전, 사회과학출판사, 1992

[11] 서울대학교 인문정보연구회, 북한 문화어 형태소 분석기 구현 연구, 문화체육부, 1997

[12] 신현숙, 북한 언어의 실제 분석-<로동신문>을 대상으로-, 북한의 말과 글, 고영근편, 을유문화사, 1989

[12] 연규동, 통일시대의 한글 맞춤법, 박이정, 1998.

[13] 최창호, 조선어맞춤법편람, 과학백과사전종합출판사, 1994