특집

한국어 태그 집합

안 동 언
전북대학교 컴퓨터·정보통신공학부
영상·정보신기술연구소

I. 서 론

한국어를 체계적으로 연구하기 위하여 대용량 언어 자료를 축적하고 기반 기술을 연구하고 개발하고자 국어정보베이스를 구축하고 있다.[1] 국어정보베이스에서 가장 중요한 지식베이스는 한국어 코퍼스이다. 한국어 코퍼스를 통하여 문법 정보, 의미 정보, 용례 등을 수집하여 다양한 언어 현상을 발견하고 한국어 처리 시스템을 개발하고 평가를 통해 성능을 향상시킬 수 있다.

코퍼스를 분석하여 언어를 연구하는 코퍼스 언어학은 19세기말부터 시작되었으나 현재에 이르러 컴퓨터와 코퍼스를 처리할 수 있는 소프트웨어의 등장으로 많은 각광을 받고 있다.[2] 가공되지 않은 코퍼스(raw corpus)를 이용하여 얻을 수 있는 자료와 언어 현상 등에는 한계가 있다. 코퍼스로 좀 더 가치 있는 정보를 획득하기 위해서는 품사 정보, 구문 정보, 의미 정보 등을 가공되지 않은 코퍼스에 부가한다. 코퍼스에 정보를 부가하기 위한 기호를 태그(tag)라고 하고 태그들의 모음을 태그 집합(tag set)이라고 한다. 코퍼스에 태그를 붙이는 작업인 태깅(tagging)을 통하여 태깅이 된 코퍼스(tagged corpus)를 얻게 된다.

태깅이 된 코퍼스가 유용한 가치를 가지기 위해서는 우선 태그 집합(tag set)이 잘 설정되어야 한다. 태그 집합은 각 태그가 변별력이 있고 풍부하고 다양한 정보를 나타낼 수 있어야 한다. 따라서, 태그 집합을 정하기 위해서는 한국어 및 한국어 정보처리에 관심이 있는 모든 연구자들의 중지를 모아야 하며 표준화를 유도하고 자료 및 처리 시스템의 호환성을 높여 국어공학의 발전을 촉진하고 이용 편의를 도모하여야 한다.

본 논문에서는 태그 집합에 대해 알아보고 국어정보베이스에서 제시한 한국어 태그 집합을 소개하며 이 태그 집합의 문제점과 개선 방안을 제시하고 한국어 태그 집합의 표준화 동향에 대해서 알아본다.

II. 태그 집합이란

태깅이 되지 않은 코퍼스에 품사 정보, 구문 정보, 의미 정보 등을 부가하기 위해 약속된 기호들을 태그 집합이라 한다. 태그 집합의 종류에는 다음과 같은 것들이 있다.[2]

- 품사 태그 집합
- 구문 태그 집합
- 의미 태그 집합
- 담화 태그 집합
- 음성 태그 집합

국내외 연구를 통하여 각 태그 집합에 대하여 설명한다.

1. 품사 태그 집합

코퍼스 태깅에 있어서 가장 기본이 되는 것은 각 단어에 품사 태그를 부착하는 품사 태깅이다. 품사 태그 집합은 코퍼스에서 언어 현상을 발견하는데에 있어서 가장 기본이 되며 더 나아가 구문이나 의미 정보를 찾거나 태깅하는 바탕이 된다. 자연언어의 분석과정은 형태소 분석과 구문 분석, 의미 분석, 화용 분석의 네 단계로 구분할 수 있고 하위 수준의 해석 결과가 바로 상위 단계로 전달되어 결국에 문장 전체의 의미를 파악하는데 필요한 정보를 제공하기 때문이다.

한국어 정보처리를 하는 연구자들은 대부분 나름대로의 한국어 품사 태그 집합을 설정하여 사용하고 있다. 이러한 품사 태그 집합들이 논문을 통하여 소개는 되었지만, 대량의 태깅된 코퍼스와 같이 공개되고 검증을 받지는 못하였다. 다만, 국어 정보베이스를 구축하면서 만든 품사 태그인 "국어 통사·형태 태그 규격"이 1996년도 제1회 우리말 정보처리 규격 심포지움에서 발표되고, 1997년도 제2회 우리말 정보처리 규격 심포지움에서는 품사 태깅이 된 코퍼스가 제공되었다.[3,4] "국어 통사·형태 태그"에 대해서는 4.1절에서 자세히 다루기로 한다. 한국어 품사 집합의 설정에 있어서 국어학의 품사론 연구가 바탕이 되고 있다.[5]

외국의 품사 태그 집합으로는 LOB 코퍼스(the Lancaster/Oslo-Bergen Corpus of British English)를 태깅하기 위한 CLAWS(the Constituent Likelihood Automatic Word-tagging System) 품사 태그 집합과 BNC(the British National Corpus, http : //info.ox.ac.uk/bnc/)를 태깅하기 위한 C5, C7 품사 태그 집합이 대표적이다.

CLAWS는 UCREL(University Centre for Computer Corpus Research on Language, http : //www.comp.lancs.ac.uk/computing/research/ucrel/)에서 개발한 자동 태깅 시스템으로 품사 태그 집합은 CLAWS1에서 출발하여 CLAWS2로 확장되었고 현재는 CLAWS7을 사용하고 있다.

Lancaster 대학의 UCREL, OUP(Oxford University Press), OUCS(Oxford University Computing Service), 영국국립도서관 등이 참여하여 만든 BNC는 1억 개의 단어로 구성되어 있으며 C5 품사 태그 집합을 가지고 CLAWS에 의해서 자동 태깅되었다. 후에 C5 품사 태그 집합은 C7 품사 태그 집합으로 확장되었다. 품사 태그를 코퍼스에 태깅하는 형식은 SGML을 사용하여 TEI (Text Encoding Initiative)의 지침에 따라 이루어졌다.

Penn Treebank는 구문 태깅 코퍼스로 유명하지만 구문 태깅을 하기 위해서는 품사 태깅이 바탕이 되어야 한다. Penn Treebank 품사 태그 집합은 Church의 PARTS 태깅 시스템을 이용하여 코퍼스를 태깅하는데 사용되었다.(ftp : //ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz/)

2. 구문 태그 집합

구문 트리가 태깅된 코퍼스를 구축하기 위해서는 구문 태그 집합이 설정되어야 한다. 구문 태깅 트리 코퍼스는 각 문장에 해당하는 구문 구조를 구문 트리 형태로 표현하고 있는 코퍼스를 의미한다.

한국어 구문 태그 집합도 품사 태그 집합과 마찬가지로 한국어 정보처리 연구자들이 나름대로의 태그 집합을 사용하고 있다. 그렇지만, 주로 구문

규칙을 기술하기 위하여 사용하고 있을 뿐 대량의 구문 트리 코퍼스를 구축하고 있지는 못하다. 국어정보베이스에서는 구문 트리 코퍼스를 구축하고 있으며 구문 태깅에 사용된 구문 태그 집합을 1996년도 제1회 우리말 정보처리 심포지움에서 발표하였다.[3] 4.2절에서 자세히 다룬다.

외국의 구문 태그 집합으로는 Penn Treebank과 SUSANNE(Surface and Underlying Structural Analysis of Natural English)에서 사용된 구문 태그 집합이 유명하다. Penn Treebank는 Pennsylvania 대학에서 만든 것으로 450만 단어로 이루어진 구문 트리 코퍼스로써 많은 연구자들이 이용하고 참고하고 있다. SUSANNE은 Brown 코퍼스의 일부분에 구문 태깅을 한 것으로 12만 8천 단어로 이루어져 있다.

3. 의미 태그 집합

문장에 의미 태깅을 하는 방법에는 두 가지 방법이 있다. 하나는 문장에서의 문장 요소들간의 의미적 관계를 표시하는 것이고, 또 다른 하나는 문장의 각 단어에 의미 자질을 표시하는 것이다. 아직은 두 번째 방법을 많이 사용하고 있으나 품사 태깅이나 구문 태깅과는 다르게 아직 보편적인 의미 태그 집합이 없다. 따라서, 국내외에 대표적인 의미 태깅이 된 코퍼스가 없다.

의미 태그 집합의 원조는 Roget's Thesaurus라고 할 수 있으며 이를 이용한 의미 태깅 작업도 있었다. 현재 의미 태깅을 하기 위해서 WordNet의 의미 분류를 많이 사용한다. 이 분류는 계층적 의미 분류로 명사는 25개, 동사는 15개의 최상위 분류로 되어 있다. (ftp://clarity.princeton.edu/pub/wordnet/5papers.tar)

4. 담화 태그 집합

London-Lund spoken corpus에 16가지 종류의 담화 태그 집합을 사용하여 담화 태깅을 한 것이 있다.[2] 예를 들어 담화 태그 집합에는 'apologies', "greetings', "hedges', "politeness', "responses' 등이 있다. 담화 태그 집합은 담화 분석에서의 잠재적 역할에 비해서 널리 사용되지는 못하였다. 이러한 분류는 문맥에 의존적이고 또한 다른 태깅에 비해서 논란의 소지가 많기 때문이다.

5. 음성 태그 집합

음성 코퍼스에는 음성 태깅을 할 수 있다. 그러나, 이 작업은 전문가가 직접 해야 하므로 매우 시간이 걸리기 때문에 태깅된 구어체 코퍼스가 매우 드물다. 또한 음성 신호는 하나의 신호로 명확하게 나누어지지 않아 같은 소리가 문맥에 따라서 다른 기호로 표현될 수 있으므로 태깅 작업이 매우 어렵다. 대표적인 음성 태깅된 코퍼스로는 MARSEC 코퍼스가 있으며 이 코퍼스의 출처는 Lancaster/IBM spoken English Corpus이다.[2]

III. 한국어 태그 집합의 표준화 동향

대량의 한국어 코퍼스의 구축과 더불어 코퍼스에 정보를 부가하는 한국어 태그 집합에 대한 논의가 활발하게 이루어지고 있다. 국어정보베이스의 구축과 우리말 정보처리 규격 심포지움을 통하여 품사 태그 집합과 구문 태그 집합이 제시되었다. 그런데 태그 집합에 대한 논의는 아직 품사 태그 집합에 집중되어 있다. 품사 태그 집합이 모든 태깅 작업의 기본이 되고 있으며 이를 바탕으로 구문 태깅이 이루어지므로 우선 품사 태그 집합에 대한 논의가 벌어지고 있는 것이다. 따라서, 품사 태그 집합을 중심으로 표준화 진행 과정을 이야기하고자 한다.

앞에서 언급하였듯이, 한국어 정보처리 연구자들은 나름대로의 품사 태그 집합을 가지고 연구를 해왔으나 대량의 코퍼스를 대상으로 품사 태그를 설정하고 품사 태깅을 한 것은 국어정보베이스를 구축하는 연구과제가 처음이다.[1] 1995년 1차년도에 품사 태그 집합에 대한 논의가 없이 과제를 진행하다 보니 세부 과제에서 사용하고 있는 품사 태그가 서로 달랐다. 그러므로, 세부 과제에서 개발된 시스템, 지식베이스, 결과물 등을 서로 이용하고 공유할 수 없었다. 따라서, 1996년 2차년도

에는 품사 태그의 표준화에 대한 논의가 시작되었다.[6]

1996년 3월 14일에 10명으로 이루어진 연구 소위원회가 구성되어 품사 태그 집합에 대한 연구를 시작하였다. 이 연구를 바탕으로 1차 기준안을 만들어서 "한국어 형태·통사 태그 규격"이라는 이름으로 1996년 7월 11일에 열린 1996년도 제1회 우리말 정보처리 규격 심포지움에서 발표되었다.[3] 이 심포지움에서는 "한국어 구문 태그 규격"도 같이 제시되었다.[3]

품사 태그 집합에 대한 다양한 의견을 수렴하고 보급을 위하여 한국어 품사 태그 세트 검토 및 보완을 위한 전문가 회의가 1997년 5월 30일에 개최되었다. 진지한 토의를 통하여 1차 기준안에 대해서 대분류와 중분류의 품사 태그 집합에 대해서는 대체적인 의견 접근을 보았지만 소분류에 대해서는 다양한 의견이 제시되고 문제점이 도출되었다. 이러한 의견과 문제점은 1997년 6월 28일에 열린 1997년도 제2회 우리말 정보처리 규격 심포지움에서 다시 한 번 제시되고 검토되었다.[6] 그리고, 품사 태깅과 구문 태깅이 된 100만 어절의 코퍼스가 공개되고 배포되었다.

위와 같은 일련의 과정을 거쳐서 한국어 품사 태그 집합에 대하여 지속적으로 연구하고 표준화를 유도할 수 있는 위원회의 필요성이 제기되어 1997년 8월 20일에 "한국어 품사 태그 집합 표준화 위원회"(가칭)가 열렸다. 이 위원회에서 최기선 교수(KAIST)는 "한국어 공학 표준을 위한 제언"을 통하여 한국어 품사 태그 집합을 포함한 코퍼스, 사전, 전문용어를 기술하기 위한 언어, 구문 태그 집합, 구문 구조, 의미 태그 집합, 의미 구조 등의 한국어 전반에 걸친 표준화 대상을 제시하고 표준화의 필요성을 역설하였다.[7]

표준은 모든 응용에 적용되어야 하므로 모든 응용의 공통 분모이다. 또한 표준은 활용에 입각해서 정해지고 표준의 재정립과 확대도 활용에 바탕을 두어야 한다. 협동의 원칙, 협력 개발의 원칙, 표준 준수의 원칙, 품질 관리의 원칙은 표준안 책정의 필요 원칙들이다.[7] 이러한 원칙에 따라서 위원회는 우선 품사 태그 집합의 표준화를 추진해 나가고자 한다. 추후에 구문 태그 집합과 의미 태그 집합 등 한국어 전반에 걸친 표준화에 대한 논의를 진행할 예정이다.

IV. 한국어 태그 집합

우리말 정보 처리 규격 심포지움에서 발표된 품사 태그 집합과 구문 태그 집합에 대하여 자세히 설명하고 논의되었던 문제점을 제시하고자 한다.

1. 한국어 형태·통사 태그 규격

컴퓨터를 통하여 한국어를 처리하기 위해서는 애매성이 없고 풍부하고 다양한 정보를 가진 품사 태그 집합이 필요하다. 따라서, 이 품사 분류는 당연히 세분되고 기능, 형태, 의미의 기준이 상호 보완적으로 적용되어야 한다. 한국어와 한국어 정보처리에 관심 있는 모든 연구자들이 두루 사용할 수 있도록 품사 태그에 대한 분류 기준이 명확히 제시되어야 한다. "우리말 정보처리 규격 심포지움"에서 제시된 "한국어 형태·통사 태그 규격"은 다음과 같은 원칙에 바탕을 두고 개발하였다.[3, 8]

- 한국어를 대상으로 하나, 한국어 문장에서 두루 사용하는 외국어나 특수기호들도 대상으로 삼는다.
- 품사태그 집합의 설정은 학교문법을 최대한 반영하며, 통사론적 분석 위주보다는 형태론적 분석을 위주로 한다.
- 품사태그는 여러 분야(형태소 레벨과 통사적 레벨)에서 다양한 용도로 사용할 수 있도록 계층적으로 분류한다.

국어공학의 현재 기술 수준에 구애받지 않고 국어학에서 현재 논란 중인 항목들을 피하지 않는 규격을 구축하여 국어학 및 국어공학 발전에 기여하고 협동적 국어 형태론 연구를 위한 기초 자료를 조기에 마련하는데 목표가 있다.

<표 1>은 위와 같은 원칙에 의해 형태론에 기준을 두고 구축된 한국어 형태·통사 태그 규격이다. 대분류는 9개 항목으로 나누었고, 소분류로 총

〈표 1〉 한국어 형태·통사 태그 규격

대분류	중분류	소분류			
기호(s)		sp	쉼표	sf	마침표
		sl	여는 따옴표	sr	닫는 따옴표
		sd	이음표	se	줄임표
		su	단위기호	sy	기타기호
외국어(f)		f	외국어		
체언(n)	보통명사(nc)	ncpa	동작성 명사	ncps	상태성 명사
		ncn	비서술성 명사		
	고유명사(nq)	nq	고유명사		
	의존명사(nb)	nbu	단위성 의존명사	nbn	비단위성 의존명사
	대명사(np)	npp	인칭대명사	npd	지시대명사
	수사(nn)	nnc	양수사	nno	서수사
용언(p)	동사(pv)	pvd	지시동사	pvg	일반동사
	형용사(pa)	pad	지시형용사	paa	성상형용사
	보조용언(px)	px	보조용언		
수식언(m)	관형사(mm)	mmd	지시관형사	mma	성상관형사
	부사(ma)	mad	지시부사	maj	접속부사
		mag	일반부사		
독립언(i)	감탄사(ii)	ii	감탄사		
관계언(j)	격조사(jc)	jcs	주격조사	jco	목적격조사
		jcc	보격조사	jcm	관형격조사
		jcv	호격조사	jca	부사격조사
		jcj	접속격조사	jct	공동격조사
		jcr	인용격조사		
	서술격조사(jp)	jp	서술격조사		
	보조사(jx)	jxc	통용보조사	jxf	종결보조사
어미(e)	선어말어미(ep)	ep	선어말어미		
	연결어미(ec)	ecc	대등적 연결어미	ecs	종속적 연결어미
		ecx	보조적 연결어미		
	전성어미(et)	etn	명사형 어미	etm	관형사형 어미
	종결어미(ef)	ef	종결어미		
접사(x)	접두사(xp)	xp	접두사		
	접미사(xs)	xsn	명사파생 접미사	xsv	동사파생 접미사
		xsm	형용사파생 접미사	xsa	부사파생 접미사

54개의 품사 태그로 나누어져 있다.[3, 8]

이 기본 품사 태그 집합의 기본 구조는 응용 연구 분야에 따라 더욱 확장된 태그 집합에 변화 없이 활용될 수 있다.[9]

한국어 형태·통사 태그 규격에 관하여 심포지움과 전문가 회의를 통하여 여러 문제점들이 제기되었다. 국어학의 입장에서는 연구 관점에 따라서, 한국어 정보처리 분야에서는 처리 분야에 따라서 다양한 의견이 제시되었다.[4] 대분류와 중분류에 대해서는 별다른 이견이 없었으며 주로 소분류에 대해서 여러 문제점이 도출되었다. 다음은 소분류의 문제점을 포함하여 전반적인 태그 설정의 원칙 및 과정과 태깅 작업에서의 문제점을 나열하고 있다.

● 품사 태그 집합을 설정하는데 있어서 다양한

의견 수렴이 부족하였다.
- 태깅된 코퍼스의 미보급으로 한국어 형태·통사 태그의 보급이 부진하였다.
- 규격 제정의 기준이 명확하지 못하여 태그 집합의 일관성이 결여되어 있다.
- 태깅을 위한 지침서가 명확하게 마련되지 못했다.
- 품사 태그 집합과 코퍼스를 이용한 실험과 그에 따른 평가가 없었다.
- 기호는 자체가 태그의 역할을 하므로 소분류가 불필요하다.
- 자동사와 타동사의 분류가 필요하다.
- "상태성"과 "동작성"의 구분이 명확하지 않다.
- 지시사의 분류가 불필요하다.
- "수관형사" 태그를 추가할 필요가 있다.
- 격조사 소분류의 분류기준이 모호하다.
- 보통명사 소분류와의 일관성 유지를 위하여 서술성 부사와 비서술성 부사의 구분이 필요하다.
- 접두사의 세분류가 필요하다.
- 동일한 내용이 표기에 따라 태그가 달라진다.(예 : 100$/su, 100dollar/f, 100불/nbu)
- 제목이나 복합어를 하나의 태그로 표기할 수 있는 다어절 태깅 방안이 마련되어야 한다.

위와 같은 문제점들을 해결하기 위하여 표준화 위원회가 결성된 것이며 의견 수렴을 거쳐서 표준안을 도출하고 문서화를 통해 지침서를 배포하고 응용 분야에 따른 태그 집합을 제공하여야 할 것이다. 물론, 표준 품사 태그 집합으로 태깅된 코퍼스도 보급되어야 할 것이다. 또한, 다양한 실험과 활용 분야에의 적용을 통하여 검증이 이루어지고 지속적인 연구가 이루어져야 한다.

2. 한국어 구문 태그 규격

한국어 구문 태그 규격은 한국어 구문 트리 태깅 코퍼스를 작성하기 위한 것이다. 한국어 구문 트리 태깅 코퍼스는 각각의 한국어 입력 문장에 대해, 그 문장에 해당하는 구문 구조를 구문 트리 형태로 표현하고 있는 코퍼스를 의미한다. 이러한 대량의 구문 트리 태깅 코퍼스는 자연언어 처리 전반에 걸친 주요한 정보원일 뿐만 아니라, 한국어 구문 분석기를 위한 구문 규칙을 추출할 수 있고, 한국어 용언의 하위 범주화 정보 및 한국어 구문 유형이나 한국어 단어의 사용례에 대한 연구의 근원이 될 수 있다.

한국어 구문 태그는 문장 분석 단위인 구, 절, 문장을 중심으로 만들어졌다. 또한, 구문 트리 태깅을 위해서는 우선 기본적인 문법 형식이 있어야 하며 여기에서는 구구조 문법을 사용하였다. 한국어는 부분 자유 어순을 가지며, 주어나 목적어와 같은 필수적 성분의 생략이 빈번하다. 이와 같은 특성으로 인하여 한국어 구문 분석에 관한 많은 연구들에서는 주로 의존 문법을 선호하였다. 이는 의존 문법이 어순의 자유성에 의한 문제점을 해결할 수 있으며 구성 요소의 불연속성이나 생략 등과 같은 현상에 큰 영향을 받지 않을 수 있기 때문이다. 그러나, 한국어는 완전한 자유 어순이 아니며, 부분 자유 어순으로서 어순의 제약이 반드시 필요한 부분이 존재한다. 그렇기 때문에 우리말 정보처리 심포지움에서 제시된 구문 태그 집합은 구구조 문법을 기반으로 하여 설정되었다.

〈표 2〉는 한국어 구문 태그 규격이며 〈표 3〉은 문장 성분과 구문 태그간의 관계이다.[3,10] 문법적 관계의 품사 태그는 한국어 형태·통사 태그 규격이다.

〈표 2〉 한국어 구문 태그 규격

구문 태그	설명
S	문장
NP	명사구절
VP	동사구절
ADJP	형용사구절
ADVP	부사구절
MODP	관형사구절
IP	독립구절
AUXP	보조용언구절

〈표 3〉 문장 성분과 한국어 구문 태그간의 관계

문장의 성분	구문태그+문법적 관계
문장	S
주어	NP+jcs
서술어	VP
	ADJP
목적어	NP+jco
보어	NP+jcc
관형어	MODP
	NP+Jcm
	VP+etm
	ADJP+etm
부사어	ADVP
	NP+jca
	NP+jct
	NP+jcr
독립어	IP
기타	AUXP

한국어 구문 태그 규격은 한국어 형태·통사 태그 규격에 비해서 논의가 활발하지 않다. 품사 태그 집합과 구문 태그 집합이 서로 독립된 것이 아니라 형태소의 하위 수준의 해석 결과가 바로 구문의 상위 단계로 전달되어 처리되어야 하므로 품사 태그 집합에 대한 논의가 끝나야 비로소 구문 태그 집합에 대한 논의가 전개될 것이다. 구구조 문법을 표현하는 구문 태그 집합과 더불어 의존 문법을 표현할 수 있는 구문 태그 집합이 마련되어야 하며 당연히 구문 태깅이 된 코퍼스도 구축되고 보급되어야 한다.

V. 결 론

태깅이 되지 않은 코퍼스에 품사 정보, 구문 정보, 의미 정보 등을 부가하기 위해 약속된 기호들인 태그 집합에 대해서 알아보았다. 태깅이 된 코퍼스가 유용한 가치를 가지기 위해서는 우선 태그 집합이 잘 설정되어야 한다. 태그 집합은 각 태그가 변별력이 있고 풍부하고 다양한 정보를 나타낼 수 있어야 한다.

본 논문에서는 국어정보베이스를 구축하면서 설정되고 우리말 정보처리 심포지움을 통해서 발표된 "한국어 형태·통사 태그 규격"과 "한국어 구문 태그 규격"을 중심으로 한국어 태그 집합에 대해서 설명하였다. 그 동안 심포지움과 전문가 회의를 통해서 논의되었던 한국어 태그 집합의 문제점을 제시하였다. 또한, 품사 태그 집합, 구문 태그 집합, 구문 구조, 의미 태그 집합, 의미 구조, 코퍼스, 사전, 전문용어를 기술하기 위한 언어 등 한국어 전반에 걸친 표준화 대상을 제시하고 표준화를 위해 노력하고 있는 학계의 동향을 소개하였다.

훌륭한 한국어 태그 집합을 설정하고 더 나아가 한국어 정보처리의 발전을 위해서는 모든 연구자들의 지속적인 관심과 연구가 필요하며 다양한 실험과 평가를 통한 검토와 대안이 제시되어야 할 것이다.

참 고 문 헌

〔1〕 한국과학기술원, "국어정보베이스," 국어정보처리기술 개발에 관한 연구 3차년도 최종보고서, 1997년 8월

〔2〕 Tony McEntry and Andrew Wilson, "Corpus Linguistics," Edinburg University Press, 1998

〔3〕 한국과학기술원, 시스템공학연구소, "1996년도 제1회 우리말 정보처리 규격 심포지움 발표집,", 1996년 7월

〔4〕 한국과학기술원, 시스템공학연구소, "1997년도 제2회 우리말 정보처리 규격 심포지움 발표집,", 1997년 6월

〔5〕 고영근, "국어 문법의 연구 -그 어제와 오늘-," 탑출판사, pp.31-98, 1987년

〔6〕 안동언, "품사 사전 규칙과 시범 패키지," 국어정보처리기술 개발에 관한 연구 제2차년도 최종보고서, 전북대학교, 1996년 7월

〔7〕 최기선, "한국어 공학 표준을 위한 제언," 한

국어 품사 태그 세트 표준화 위원회 제1차 회의, 1997년 8월
〔8〕 최기선, 남영준, 김진규, 한영균, 박석문, 김진수, 이춘택, 김덕봉, 김재훈, 최병진, "한국어정보베이스를 위한 형태·통사 태그 표준에 관한 연구," 인지과학 제7권, 제4호, 1996년
〔9〕 안동언, "확장 품사 사전 규칙과 보급 패키지," 국어정보처리기술 개발에 관한 연구 제2차년도 최종보고서, 전북대학교, 1997년 5월
〔10〕 이공주, 김재훈, 최기선, 김길창, "구문 트리 부착 코퍼스 구축을 위한 한국어 구문 태그," 인지과학, 제7권, 제4호, 1996년

저 자 소 개

安 東 彦

1958年 5月 25日生
1981年 2月 한양대학교 전자공학과 졸업(공학사)
1987年 2月 한국과학기술원 전산학과 졸업(공학석사)
1995年 2月 한국과학기술원 전산학과 졸업(공학박사)

1988年 3月～1991年 8月 한국외국어대학교 시간강사
1991年 8月～1995年 2月 충남대학교 시간강사
1995年 3月～현재 전북대학교 컴퓨터·정보통신공학부 조교수

주관심 분야 : 한국어정보처리. 기계번역, 정보검색