

SERI Test Suites '97 :
한국어 구문분석기 성능 평가용 문장 모음

성원경*, 장명길, 박재득, 류범모, 이현아, 박동인
시스템공학연구소, 자연어정보처리연구부

SERI Test Suites '97 :
Test Sentences for Korean Syntactic Analyser

Won-Kyung Sung*, Myung-Gil Jang, Jae-Deuk Park, PumMo Ryu, Hyun-A Lee, Dong-In Park
Systems Engineering Research Institute (SERI), Natural Language Information Processing Dept.
e-mail : {wksung, mgjang, jdpark, ryupm, halee, dipark}@seri.re.kr

요약

자연어 정보처리 분야의 거듭된 발전은 다양한 언어처리 도구들의 출현을 가져왔다. 그러나 객관적인 성능 평가 기준의 부재로 인해, 개발된 도구들은 임의의 기준에 따라 평가될 수 밖에 없었다. 그 결과 성능 평가 결과는 평가자와 평가자가 제안한 기준에 따라 다를 수 밖에 없었고 따라서 평가 결과 자체 역시 설득력을 갖을 수가 없었다. 이와 같은 문제에 대한 해결책을 찾고자 하는 노력의 일환으로, 본 연구에서는 한국어처리 도구들 중 특히 구문분석기의 체계적이고도 객관적인 성능 평가를 목적으로 제작된 문장들과 관련 주석 정보들로 구성된 SERI Test Suites '97 을 소개한다.

1. 서론

오늘날 우리 사회의 정보화에 대한 요구는 자연어처리 기술에 대한 수요를 크게 증가시키고 있다. 이와 같은 수요에 부응하기 위해 몇몇 자동번역 시스템들과 정보검색 시스템들이 이미 상품화된 바 있고 또한 여러 기업들과 연구 기관들에서도 다양한 언어처리 도구들을 계속적으로 개발 및 상용화하는 움직임을 보이고 있다. 그 결과 오늘날의 사용자들은 그 어느 때보다도 쉽게 자연어처리 도구들을 접할 수 있게 되었으며 개발된 도구들의 성능 역시 날로 향상되는 추세이다. 그러나 개발된 도구들은, 객관적인 성능 평가 기준의 부재로 인해, 개발자들이 임의로 설정한 기준에 따라 평가될 수 밖에 없었다. 당연히 성능 평가 결과는 평가자와 평가자의 기준에 따라 다를 수 밖에 없었고 그로 인한 평가 결과 자체에 대한 신뢰도의 저하는 불가피한 것이었다.

평가 결과의 신뢰도의 보장은 크게 다음의 두 차원에서 그 의미를 갖는다. 그 하나는

개발의 차원으로서, 개발자들의 개발 중 혹은 개발 완료된 도구들에 대한 정확한 진단 정보에 대한 요구이다. 객관적이고도 체계적인 검증 기준에 근거한 진단 자료는 개발자들에게 도구의 성능을 개발 각 단계마다 효율적으로 비교 검토할 수 있게 할 뿐만 아니라 내재된 결함을 체계적으로 포착할 수 있게 하기 때문이다[Balkan 94], [Lehmann 96]. 평가 결과의 신뢰도 보장의 또 다른 의미는 사용자의 차원에서 찾을 수 있다. 개발 완료된 도구들에 대한 평가 결과가 객관적이고도 신뢰도가 높다면, 사용자는 그 평가 결과에 근거하여 도구들을 비교함으로써 보다 성능이 좋은 그리고 사용 목적에 부합하는 도구를 선택할 수 있을 것이다.

본 연구에서는, 이상과 같은 의미를 갖는 객관적인 평가 기준을 설정하고자 하는 노력의 일환으로, 한국어처리 도구들 중 특히 구문분석기의 성능 평가를 목적으로 제작된 평가문들과 관련 주석 정보들(annotations)로 구성된 SERI Test Suites '97 을 소개한다.

2. SERI Test Suites '97 의 구축 원칙들

SERI Test Suites '97 은 시스템공학연구소의 한국어 구문분석기(KOSA (Korean Syntactic Analyser))의 평가를 목적으로 구축된 평가문 모음인 SERISET96[장명 96]의 확장된 version 인 동시에 현재 동연구소에서 개발 중인 범용 한국어처리 도구 평가문 모음인 KONTEST(KOREAN TEST SuiTes)¹의 사전 작업 결과의 일부에 해당된다. KONTEST 의 일부로서의 SERI Test Suites '97 은 다음과 같은 KONTEST 구축 방법론을 공유한다².

- 평가문의 특성에 대한 체계적 통제 :

SERI Test Suites '97 은 평가 결과의 명확한 해석을 그 기본적인 목표로 한다. 이와 같은 목표 달성을 위해, SERI Test Suites '97 의 각 평가문들은 반드시 필요한 경우 이외에는 다음과 같은 조건들에 따라 제작하였다:

- ① 각 평가문에는 오직 하나의 평가 항목만이, 예를 들어 하나의 문법 현상만을 그리고 하나의 술어구문만이, 포함되도록 할 것.
 - ② 중의성(ambiguity)이 없게 할 것.
 - ③ 수식어 사용을 피할 것.
 - ④ 표준 어순을 사용할 것.
 - ⑤ 의문문이나 명령문보다는 평서문 형태를 사용할 것.
 - ⑥ 부정문보다는 긍정문 형태를 사용할 것.
- 여러 문법현상들과 여러 구문을 포함하는 테스트 데이터 혹은 중의성이 있는 테스트 데이터에 기초한 평가는, 혼재한 문법 현상들 간의 상호 작용이나 중의성으로 인해, 그 결과의 해석이 쉽지 않을 것임은 물론 말할 필요조차 없다. 또한 수식어를 포함하지 않은, 표준 어순에 따른 긍정적 평서문으로도

¹ KONTEST 는 EC 4 개국(독일의 DFki, 스위스의 ISSCO, 영국의 Essex 대학, 프랑스의 Aerospatial)이 공동 참여한 TSNLP(Test Suites for Natural Language Processing) 프로젝트(1993 년 12 월 - 1996 년 3 월)의 연구 결과에 기초한다. TSNLP 프로젝트에 관한 보다 상세한 정보는 다음의 URL 에서 얻을 수 있다 : <http://tsnlp.dfki.uni-sb.de/tsnlp>

² KONTEST 는 위에 열거된 다섯 가지 항목 이외에 “비문 데이터(ungrammatical data 혹은 negative data)”를 포함하는 것으로도 특징지워진다. 비문 데이터는 특히 문서되고 시스템의 성능 평가를 위해서는 필수적이며 구문분석기의 경우에도 특히 시스템의 강건성(robustness)에 대한 진단 시에 아주 유용하다. 그러나 이와 같은 유용성에도 불구하고, 비문 데이터는 그 제작 방법론의 미체계화로 인해 SERI Test Suites '97 에는 아직 포함되지 않았다.

충분히 평가가 가능한데도 일부러 그렇지 않은 형태의 평가문을 사용할 필요도 없기 때문이다[Balkan 94].

- 평가 항목의 체계화 :

성능 평가 결과는 반드시 평가 대상인 구문 분석기가 우리말을 전반적으로 잘 처리하고 있는지의 여부를 드러내 주어야 한다. 우리말의 몇몇 부분에 대한 처리 면에서는 뛰어나다 해도 다른 부분은 잘 처리하지 못하는 도구들은 그 성능 면에서 결코 좋은 도구가 되지 못함은 자명하기 때문이다. 이와 같은 취지에서 SERI Test Suites '97 은 우리말의 여러 문법현상을 종합적으로 평가할 수 있도록 제작되었다.

- 평가 항목의 중복 출현 회피

SERI Test Suites '97 의 각 평가문들은 앞에서 언급한 바처럼 하나의 평가 항목만을 포함할 뿐만 아니라 그 역으로 가능한 한 하나의 평가 항목은 단 하나의 평가문에만 나타나도록 제작되었다[Balkan 94], [Estival 95]. 평가 항목의 중복 출현 회피에 대한 이와 같은 배려는 SERI Test Suites '97 에 의한 평가 과정의 효율성과 평가 결과의 최적화를 위한 것이다.

- 제한된 수의 어휘 사용

SERI Test Suites '97 의 평가문들은 총 1937 어절로 구성되어 있으며 총 440 개의 비교적 제한된 수의 어휘를 포함한다³. 이와 같이 사용 어휘 수를 가능한 한 제한하고자 하는 것은 구문분석기의 성능 평가를 목적으로 하는 SERI Test Suites '97 에 의한 평가가 자칫 평가 대상 도구의 lexical coverage 자체에 대한 평가가 되는 잘못을 피하고 오직 구문 분석에 관련된 성능에 대한 평가가 될 수 있도록 하기 위해서 이다[Balkan 94].

- 주석 정보

SERI Test Suites '97 의 각 평가문은 주석 정보들과 함께 제시된다. 주석 정보들은 각 평가문의 평가 목표를 명확하게 함으로써 평가자와 그 이외의 사람들에게도 평가 결과에 대한 정확하고도 객관적인 이해를 가능

³ SERI Test Suites '97 에 포함된 440 개의 어휘는 다음과 같이 구성된다 : 명사(의존명사 및 대명사 포함) 269 개, 술어 125 개, (과생)부사 35 개, 한정사(관형사 포함) 11 개.

하게 할 뿐만 아니라 평가문의 재활용 가능성을 높여준다[Balkan 95a], [Fouvry 96]. 참고로, KONTEST는 주석 정보의 입력, 관리 및 활용을 보다 체계화하기 위해 다음과 같은 schema를 사용한다.

[표 1] KONTEST에서 사용하는 schema의 예

Test Item				
item id: 24001	author: seri	date: 05-97		
format: none	origin: invented	difficulty: 1		
wellformedness: 1	category: S	length: 3		
input: 철수가 밥을 먹는다				
comment:				
position	instance	category	function	domain
0:1	철수	NP_sg	subj	2:3
1:2	밥	NP_sg	obj	2:3
2:3	먹는다	V	func	0:3
Phenomenon				
phenomenon id: 24	author: seri	date: 05-97		
name: C_Complementation				
supertype: Complementation				
presupposition: C_Agreement				
restriction: neutral	interactions: none	purpose: test		
comment:				

3. SERI Test Suites '97의 구성

SERI Test Suites '97은 총 472문장으로 구성되었으며, 그 평가 대상 항목에 따라 다시 180개의 술어구문 평가문들과 292개의 문법현상 평가문들로 대별된다. 술어구문 평가문들은 구문분석기가 입력된 문장을 분석하여 출력한 구성 성분들의 문법적 기능과 관계의 정확성에 대한 평가를 그 목적으로 하며 문법현상 평가문들은 입력문에 나타난 문법현상을 구문분석기가 정확하게 포착하는가를 평가하기 위한 것이다. SERI Test Suites '97은 문장당 평균 어절의 수가 약 4.10 어절인 비교적 복잡도가 낮은 문장들로 구성되었으며, 앞에서 언급한 제작의 기본 목표 달성을 위해, 평가문들을 말뭉치에서 추출하기 보다는 우리말을 모국어로 하는 제작자들의 언어적 직관에 근거하여 작성하였다. 물론 제작자들에 의해 만들어진 평가문들은 기존의 평가 작업에서 빈번히 사용되던 말뭉치에 비해 그 대표성이 떨어질 수는 있다. 그러나 일반적으로 말뭉치의 각 개별 문장들에는 여러 문법현상들이 혼재하고 있을 뿐만 아니라 동일 현상의 반복적 출현, 누락된 문법현상 등과 같은 문제점들이 있기 때문에 말뭉치를 평가 자료로 삼았을 경우, 대표성은 어느 정도 확보될 수는 있다 하더라도, 평가 과정의 가장 중요한 요소인 평가의 효율성과 체계성을 보장하기가 어렵다[Balkan 94], [Dauphin 95a]. 더욱이 주석 정

보의 부재는 평가 결과의 객관적인 이해와 평가문으로서의 재활용을 불가능하게 한다. 물론 평가문의 대표성과 평가 작업의 효율성 및 체계성 모두를 확보하기 위해서는 말뭉치에서 전술한 바와 같은 요건들을 충족시키는 평가문들을 추출하여 활용할 수도 있을 것이다[Balkan 95b]. 아래의 [표 2]는 SERI Test Suites '97의 전체적 구성에 대한 통계 자료를 보여준다.

[표 2] SERI Test Suites '97의 통계 자료표

	문장 수	어절 수	문장당 평균 어절의 수
술어구문 평가문	180 문	768 어절	4.27 어절
문법현상 평가문	292 문	1169 어절	4.00 어절
누계	472 문	1937 어절	4.10 어절

3.1. 술어구문 평가문

하위 범주화 정보에 대한 평가를 목표로 하는 180개의 술어구문 평가문들은 아래의 [표 3]에 열거된 20개 술어들을 기준으로 작성되었다. 이 술어들은 STEP2000 과제의 일환으로 KAIST에서 구축한 약 10,000문장(약 100,000 어절) 규모의 구문 정보 부착 말뭉치에서 추출된 것으로 통계적으로 그 출현 빈도수가 높은 상위 20개 술어이다.

[표 3] 술어구문 평가문에 사용된 술어

술어	10,000 문장 중 해당 술어를 포함하는 문장의 수	술어	10,000 문장 중 해당 술어를 포함하는 문장의 수
받다	232	이르다	81
나타나다	188	먹다	78
보이다	169	들다	74
살다	159	삼다	61
주다	158	나다	60
파르다	145	부르다	54
쓰다	128	만나다	50
가다	103	타다	47
오다	100	나가다	41
나오다	84	돌다	22

앞의 표에 열거된 20개 술어들을 중심으로 해서 작성된 술어구문 평가문들은 다음의 [표 4]와 같이 구성되었다. 각 술어에 해당되는 평가문의 수는 그 술어가 허용하는 구문

의 수에 비례하여 결정되었다.

[표 4] 상위 고빈도 20 개 술어를 포함하는 평가문의 수

술어	해당 평가문의 수	술어	해당 평가문의 수
가다	13	받다	6
나가다	13	보이다	15
나다	19	부르다	4
나오다	10	살다	6
나타나다	5	삼다	5
돌다	10	쓰다	17
들다	4	오다	10
따르다	4	이르다	5
만나다	5	주다	4
먹다	14	타다	11

SERI Test Suites '97에서는 평가문 구성에 사용된 동형어를 구분짓기 위해 꼬리 번호를 사용하였다. 예를 들어, 아래의 [표 5]에서의 “가다 1”과 “가다 2”의 꼬리 번호 “1”과 “2”는 이 두 “가다”가 형태는 동일하지만 그 쓰임에 있어서는 상호 구분되는 동형어임을 의미한다. 참고로, 제시된 술어들의 하위 범주화에 대한 정보는 다음에 열거된 기존의 연구 결과물들에 근거하였음을 밝혀둔다.

- Classification syntaxique des constructions adjectivales en coréen [Nam 94]
- 한국어 동사 구문 사전 [홍재 96]
- 한국어 문장 분석을 위한 용언의 하위 범주화 [시스템 96]
- 한국어 문장 분석을 위한 용언의 하위 범주화에 관한 연구 [시스템 97]

다음의 [표 5]는 술어구문 평가문의 예이다.

[표 5] 술어구문 평가문의 예

술어	평가문	구문 형태
가다 1	(1) 봄이 다 갔다.	N0-가 V
	(2) 이 차가 서울로 간다.	N0-가 N1-로 V
	(3) 철수가 학교에 갔다.	N0-가 N1-에 V
	(4) 철수가 책을 사러 서점에 갔다.	N0-가 S-러 N1-에 Vmvt
가다 2	(5) 철수가 영수를 만나러 밥집을 혼자서 간다.	N0-가 S-러 N1-을 Vmvt
	(6) 철수가 제주도도로 신혼여행을 갔다.	N0-가 N2-에게/로 N1-을 V

3.2. 문법현상 평가문

SERI Test Suites '97의 문법현상 평가문들에 의해 검증되어질 사항들은 특정 언어들 내에서 그 중요성이 부각된 현상들 또는

특정 구문분석기를 위한 특정 현상들이라기 보다는 우리말 문법서들에서 집중적으로 지적되는 문법현상들로서 우리말 구문분석기라면 당연히 포착해내야 할 기본적인고도 핵심적인 문법현상들만을 포함함을 밝혀둔다. SERI Test Suites '97의 문법현상 평가문들이 다루는 문법현상들과 각 현상별 평가문의 수는 아래의 [표 6]과 같다. 괄호 안의 수치는 해당 문법현상 평가문들 중 일부가 다른 현상을 평가하기 위한 항목에 포함되었음을 의미한다. 예를 들어, 부정문의 경우, “(2/명령문, 청유문)”는 2개의 평가문이 “명령문, 청유문” 현상 평가문에 포함되었음을 의미한다.

[표 6] SERI Test Suites '97에서 다루어진 문법현상들과 해당 평가문의 수

문법 현상	문장 수	문법 현상	문장 수
의문문 Question	21	이중 주어 Double Subject	5
명령문, 청유문, Imperative, Propositive	6	이중 목적어 Double Object	6
시제, 시상, 양상 Tense, Aspect, Modality	18	특수조사 Delimiters	56
부정문 Negation	6 + (2/명령문, 청유문)	수식 Modification	17 + (2/의문문)
태 Voice	19	관계화 Relativization	(5/수식)
사역문 Causative	16	등위 접속 Coordination	37
주제화 Topicalization	34	공백화 현상 Gapping	13 + (2/주제화) + (5/관계화)
비명사구 주 어 Non-NP Subject	3	중의성 Ambiguity	29
격조사 중출 Case Stacking	6		

다음은 문법현상 평가문의 한 예이다.

[표 7] 문법현상 평가문의 예

NP/PP 등위 접속		
-(7)	철수와 영수가 학교에서 영희를 만났다.	(N Conj N)-가
	(8) 철수가 학교에서 영수와 영희를 만났다.	(N Conj N)-들
	(9) 철수가 영수와 영희에게 책을 주었다.	(N Conj N)-에게
	(10) 철수가 영수에게와 영희에게 책을 주었다.	N-에게 Conj N-에게
	(11) 철수가 학교와 집에서 영수를 만났다.	(N Conj N)-에서
	(12) 철수가 학교에서의 집에서 영수를 만났다.	N-에서 Conj N-에서

4. SERI Test Suites '97 의 평가표 작성 및 평가 요령

SERI Test Suites '97 은 문장 모습들에 근거한 구문분석기의 성능 평가를 보다 체계적이고도 용이하게 하기 위해 각각의 평가 결과들을 일목요연하게 요약하고 시각화해 주는

다음의 [표 8]과 같은 평가표를 사용할 것을 권고한다. 참고로, 평가표의 “문 번호”는 SERI Test Suites '97 의 평가문들의 일련 번호를 의미하며 “어절 수”는 해당 평가문에 포함된 어절의 수를 의미한다.

[표 8] SERI Test Suites '97 평가표의 일부

문 번 호	어 절 수	형태소 분석 결과	구문 분석 결과					비고		
			생성된 구문 tree 개수		평가					
			wft	ift	1	2	3		4	5
001	3									
002	4									
003	3									

* wft : well-formed tree / ift : ill-formed tree

다음은 평가표의 각 항목에 대한 설명이다.

형태소 분석 결과

해당 평가문을 구성하는 형태소에 대한 형태소분석 결과에 오류가 있을 경우에 분석이 잘못된 형태소를 기입한다. 그 외의 경우에는 빈 칸으로 남겨둔다.

형태소분석 단계와 구문분석 단계는 개념적으로 구분된 두 과정에 해당된다. 따라서 형태소분석의 결과에 대한 평가를 구문분석기 평가표에 포함시키는 것이 무의미한 것처럼 보일 수도 있을 것이다. 그러나 일반적으로 잘못된 형태소분석의 결과에 근거한 구문분석은, 그 구문분석기의 실제의 성능과는 무관하게, 옳지 못한 결과를 낼 수 있기 때문에, 보다 공정한 평가를 위해서는 구문분석기로 입력되는 형태소분석의 결과에 대한 검증이 반드시 필요하다.

생성된 구문 tree 개수

구문분석의 결과 생성된 구문 tree 의 개수를 “올바른 구문 tree(well-formed tree (wft))”와 “잘못된 구문 tree(ill-formed tree (ift))”로 나누어서 기입한다.

평가 (1, 2, 3, 4, 5)

- 구문분석의 결과가 해당 평가문의 주석 정보란에 제시된 모든 wft 만을 포함한 경우에는 “1”항에 표시한다.
- 구문분석의 결과가 모든 wft 뿐만 아니라 ift 도 포함한 경우에는 “2”항에 표시한다.

- 구문분석의 결과가 일부의 wft 만을 포함한 경우에는 “3”항에 표시한다.
- 구문분석의 결과가 일부의 wft 와 ift 를 포함한 경우에는 “4”항에 표시한다.
- 구문분석의 결과 모두가 ift 인 경우에는 “5”항에 표시한다.

비고

평가 시 고려해야 할 구문분석기의 특수성으로 인한 사항들을 기입한다.

위와 같은 요령으로 작성된 평가표에 근거하여 구문분석기의 성능에 대한 평가를 다음에 기술된 절차에 따라 실시할 수 있다.

성능 혹은 정확도 평가

구문분석기의 성능은 크게 두 측면에서 측정될 수 있다. 그 하나는 처리 시간이며 또 다른 하나는 분석 결과의 정확도이다. 그러나 하드웨어에 특히 민감한 처리 시간은 하드웨어의 조건에 따라 가변적일 수 있으므로 평가 시에 분석 결과의 정확도 만큼 중요한 평가 기준으로 사용되기는 어렵다⁴. 따라서 SERI Test Suites '97 에서는 분석의 정확도를 성능 평가 기준의 유일한 척도로 사용한다.

구문분석기의 정확도는 구문 분석기가 입력

⁴ 물론 물적 조건이 동일한 경우라면 처리 시간 역시 성능 평가의 중요한 기준이 될 수도 있을 것이다.

된 문장의 구문을 얼마나 정확하게 인식하는가에 대한 척도로서 다음의 기준에 의해 산출한다.

- A. 평가 결과가 "1"인 문장의 수
- B. 평가 결과가 "2"인 문장의 수
- C. 평가 결과가 "3"인 문장의 수
- D. 평가 결과가 "4"인 문장의 수
- E. 평가 결과가 "5"인 문장의 수

라고 할 때,

분석의 정확도 δ 는 다음과 같이 가중치를 적용하여 산출한다. (A는 "5", B, C, D는 "3", E는 "0"이라는 가중치를 부여한다.)

$$\delta = (A*5 + B*3 + C*3 + D*3 + E*0) / (\text{총 평가문 수} * 5)$$

이때 형태소분석의 결과가 잘못된 평가문들은 총 평가문의 수에서 제외시킨다.

구문분석 결과의 관점에서 B, C, D는 서로 구분이 되는 경우이다. B는 구문 tree가 과잉 생성된 경우이며 (overgeneration), C는 과소 생성의 경우에 해당된다 (undergeneration). D는 과소생성된 wft와 과잉생성된 ift가 혼재하는 경우이다. 그러나 앞에서 제시된 구문분석기의 정확도 산출 방식은 이 세가지 경우를 구별하지 않으며 동일하게 가중치 "3"을 임의로 적용하였다. B, C, D 모두 분석 결과가 충분하지 못하다는 면에서는 동일하기 때문이다. 물론 개발자는 구문분석기를 포함하는 상위 시스템의 작업 목표 및 특수성으로 인해 내포된 구문분석기로 하여금 B와 같이 과잉생성을 하게 하거나 혹은 C와 같이 과소생성하게 하는 특수한 전략적 입장을 선택할 수 있을 것이다. 그럴 경우, 개발자의 입장에서는 B, C 그리고 D는 동일한 의미를 갖을 수가 없을 것이며 이와 같은 특수성들을 고려하지 않는 범용 평가도구로서의 SERI Test Suites '97의 정확도에 대한 평가 결과 역시 큰 의미를 갖지 않을 수도 있을 것이다. 그러나, 그러한 경우에도, SERI Test Suites '97은 평가자에게 해당 구문분석기의 전체적인 분석의 정확도에 대한 일별을 가능하게 할 뿐만 아니라 생성된 wft와 ift의 구분을 함으로써 해당 구문분석기의 특수성에 따른 필요 정보를 체계적으로 제공할 수 있다는 의의는 갖는다.

5. 참고문헌

- [시스템 96] 한국어 문장 분석을 위한 용언의 하위 범주화, 시스템공학연구소, 1996.
- [시스템 97] 한국어 문장 분석을 위한 용언의 하위 범주화에 관한 연구, 시스템공학연구소, 1997.
- [장명 96] 장명길, "한국어 구문분석기 (KOSA)의 평가용 문장 모음 SERISET96 및 시스템 평가", *JMG Technical Report TR-NLP-02*, 시스템공학연구소, 1996.
- [홍재 96] 홍재성 외, 한국어 동사 구문사전, 두산동아, 1996.
- [Balkan 94] Lorna Balkan et al., "Test Suite Design: Guidelines and Methodology", *D-WP2.1a*, University of Essex, 1994.
- [Balkan 95a] Lorna Balkan et al., "Test Suites for NLP", *Fourth International Conference on Cognitive Science of Natural Language Processing*, Dublin, 1995.
- [Balkan 95b] Lorna Balkan et al. "Corpus-Based Test Suite Generation", *D-WP5.2*, University of Essex, 1995.
- [Dauphin 95a] Eva Dauphin et al., "Checking Coverage Against Corpora", *D-WP3.2*, University of Essex, 1995.
- [Dauphin 95b] Eva Dauphin et al., "Testing and Customisation of Test Items", *D-WP4*, University of Essex, 1995.
- [Estival 95] Dominique Estival et al., "The Construction of Test Material", *D-WP3.1*, ISSCO, Université de Genève, 1995.
- [Fouvry 96] Frederik Fouvry et al., "Test Suites for Quality Evaluation of NLP Products", *Le traitement automatique du langage et les applications industrielles (Natural Language Processing and Industrial Applications)*, Moncton, New-Brunswick, Canada, 1996.
- [Lehmann 96] Sabine Lehmann et al., "TSNLP --- Des jeux de phrases-test pour l'évaluation d'applications dans le domaine TALN", *Traitement Automatique du Langage Naturel (TALN) Conference*,

[Nam 94] Marseille, 1996.
Jee-Sun Nam, *Classification
syntaxique des constructions
adjectivales en coréen*, Doctoral
thesis, Univ. Paris 7, 1994.