토 론

사회: 저희가 본래 한 시간 가량 토론을 할 생각이었는데, 주제발표해 주시는 분들께서 아주 좋은 내용을 자세하게 말씀해 주시느라고 시간이 많이 지났습니다. 불과 40분 정도밖에 남지 않았기 때문에 효과적인 방법으로 이야기를 진행해야 할 것 같습니다. 오늘 정희성 선생님께서는 권혁철 선생님이 발표해 주신 내용의 끝부분을 도와주신 셈이고, 결국 세 분께서 발제를 해주셨는데, '형태론과 어휘부'라는 제목 아래서 우리 언어학이 이러한 문제를 어떻게 다루며, 요즈음 전산학이 자연언어처리를 하는데, parsing이라든가 여러 가지 다른 부분이 있겠지만, 특히 lexicon의 구성을 어떻게 하는 것이 현재로서 좋은 방안이 될 것이냐를 보겠습니다. 앞으로 전산학이 발달하면 물론 다른 이론까지 수용할 수 있겠지만, 현재로서 언어학에서 전산학에 대한 아무런 고려 없이 개발된 이론들을 볼 때, 전산쪽에서는 어떠한 다른 생각을 가지고 있으며, 서로 어떻게 접점이 형성되어서 협조를 해나갈 것인가 하는 문제를 토론해 보는 것이 오늘의 과제라 할 수 있겠습니다.

발제를 해주신 순서에 따라서 여기 나오신 세 분께 한 분씩 좋은 참고하실 의견을 이야기해 주시면 좋겠는데요, 먼저 아까 김영석 선생님께서 발표하신 내용 중에서, Mohanan은 파생(derivation)은 level 1, 2(Mohanan은 stratum이라는 용어를 썼습니다)에서 다루고, 복합(compounding)은 level 3에서 다루며, level 4에 가서 inflection이 있다고 하였습니다. 그런데, 권혁철 선생님처럼 전산쪽의 의견은 파생 같은 것은 있는 대로 받아들이고, inflection은 syntactic한 문제와 함께 처리를 할 수 있는 것으로 본다고 하고 있습니다. 그래서 lexical phonology의 입장에서처럼 lexicon안에 level이 네 개이건 두 개이건 몇 개가 있을 때, 앞부분은 불문에 붙이는 것이 전산쪽에 유리하기 때문에, 언어학에서 제시하는 이러한 모델을 전산학에서 어떻게 달리 수용해야 할 것인가 하는 문제점이 있을 것으로 생각합니다. 이러한 이야기부터 시작하기로 하죠. 이기동 선생님께서 먼저 이야기 해주실까요?

이기동: 김영석 선생님께서 소개해 주신 것은 소위 말하는 formal syntax에서 lexicon을 어떻게 생각하고 있느냐 하는 것이라 할 수 있겠는데, 저는 조금 관점을 달리 하기 때문에 formal syntax의 범위를 벗어나서 던져보고 싶은 질문들이 몇 가지 있습니다. 우선 Chomsky의 문법에서 lexicon을 이야기하는 사람들이 아마 어휘를 제대로 살펴보지도 않고 하는 사람들이 많은 것 같다는 느낌입니다. 그래서 어떤 모형이든지 자료를 중심으로 해서 모형이 나와야 하는데, 먼저 모형을 정해놓고 거기에 맞는 데이타를 집어 넣으려고 하는 데서 아주 거북스러운 문제들이 생겨나지 않나 생각합니다. 그리고 자료집 74페이지에 Halle의 모형이 나오

는데, 이게 model이라고 하지만 과연 무슨 model이냐 하는 문제가 있습니다. 언어학이라고 하는 것은 native speaker가 무의식적으로 가지고 있는 지식을 의식화하는 것이기 때문에, Chomsky도 말했듯이 언어모형은 psychological reality가 있어야 합니다. 그래서 이러한 model에 과연 어떤 psychological reality가 있느냐 하는 의심을 던져볼 수 있겠습니다. 그 다음에 모형이 자꾸 바뀌어 나가는데, 이 모형이 왜 바뀌어 나가느냐, 즉, native speaker가 가지고 있는 언어지식을 좀더 충실히 반영하기 위해 바뀌어 나가는 것인지 아니면 어떤 이론적인 틀을 먼저 제시해 놓고 이 틀을 좀더 일관성있게 만들어 가기 위한 것인지 하는 문제도 있습니다. 이 두가지는 전혀 다른 문제라 생각됩니다. 이론적으로 일관성이 있다고 해서 그것이 바로 언어지식을 보다 충실히 반영하는 것은 아니라고 생각합니다.

다음에 여기에 나와 있는 모형을 가지고 몇가지 생각해 보겠습니다. 우선 형태소 목록이 있고 그 다음에 word formation rule이 있고 filter가 있고 dictionary가 있는데, 이런 경우 간단한 예를들어 'push'라는 동사와 '-er'이라는 형태소가 있고, '동사 + -er'이라는 word formation rule에 의해 'pusher'가 만들어져 사전에 들어갑니다. 그런데, 'pusher'의 뜻을 생각할 때, 과연 'push'와 '-er'만 합쳐 가지고 일반 사람들이 사용하는 'pusher'의 의미를 나타낼 수가 있겠느냐는 것이 문제입니다. 'pusher'는 '마약같은 것을 강매하는 사람'을 뜻하는데, 이러한 규칙을 가지고 그러한 의미를 어떻게 포착할 수 있겠느냐 하는 것입니다. 만약 그 뜻이 모든 사람이 다 쓰는 것인데도 포함이 안된다면, 이러한 모형을 만들어 가지고 오히려 언어를 빈약하게 만드는 결과를 초래하고, 자연언어를 제대로 파악하지 못하는 절름발이식 연구가 되지 않겠는가 하는 생각입니다. 그 다음에 여기에 보면 word formation rule에서 dictionary로 가서 다시 돌아오게 되어 있는데, 이것은 사전에 있는 것만… 글쎄요 이것을 어떻게 풀이해야 좋을지 모르겠습니다만, 만약에 이런 식으로 된다면 이런 모형을 가지고는 native speaker가 갖는 낱말형성에 대한 창의성 같은 것은 도저히 포착할 수 없는 게 아닌가 생각되는군요. 그러니까 syntax가 창의성이 있다고 한다면, 낱말쪽에서도 의미라든지 합성을 만들어 낸다는 점에서 굉장한 창의적인 면이 있는데 이러한 것을 어떻게 cover할 수 있겠느냐는 것입니다.

그리고 또 한가지 세부적인 문제로, 홍선생님께서 비슷한 문제를 이야기 해주셨지만, 언어분석을 하는데 있어서 economy를 굉장히 중요하게 생각하고 있습니다. 예를 들어 Chomsky 문법에서는 규칙을 만드는데 feature를 10개로 할 것을 5개로 하면 그만큼 economic하니까 그 rule이 가치가 있다 하는 식입니다. 그런데, 과연 그러한 argument를 언어분석에 적용해도 되는지 하는 생각이 듭니다. 말을 이해한다는 것이 규칙도 알고 그 규칙에 맞는 실례도 아는 것을 의미한다고 보면, 언어의 규칙이냐 목록이냐 하여 둘 중의 하나를 택할 것이 아니라, 두가지가 다 포함되어야 할 것이 아니겠느냐 하는 생각입니다. 예를들어 영어에서 명사의 복수형을 만드는 것은 '명사+-s'인데, 만약 그 규칙만 알고 과연 영어를 안다고 할 수 있을지 의문입니다. 그

래서 언어기술에 있어서는 규칙과 실례가 공존해야 하지 않겠는가 하고 생각합니다. 물론 economic하지는 못하지만, 그것이 자연언어를 자연 그대로 보여주는 것이라고 봅니다.

끝으로 오늘 모임에서 느낀 바는 언어학자들이 어떤 이론적인 것에만 관심을 갖고, 실제 자료를 처리해 보지 않은 상태에서 일반화만 추구하는데, 이러한 것이 조금 지양되어야 한다고 생각합니다. 그리고 언어학자들이 사전편찬에 관하여 별다른 가치를 부여하고 있지 않고 있는데, 제 생각으로는 이러한 작업을 해가지고 거기서 나온 많은 자료를 살펴서 어떤 일반화를 얻는 것이 바람직한 방향이 아닌가 합니다.

사회: 네, 사실 김영석 선생님의 자료에 보이듯이 우리 phonologist나 morphologist들이 내어놓은 model은 다양합니다. 그 다양한 것을 받아보는 입장에서는 그러한 이야기가 나올 수 있겠지요. 이 점에 대해서 김영석 선생님께서 답변해주시지요.

김영석: 말씀하신 순서에 따라서 답변해 드리겠습니다. 일반적으로 lexicon이라는 용어를 사용할 때에, 전에는 전혀 예측할 수 없는 idiosyncratic한 것을 쓰레기 모아놓듯 쳐박아 놓는 창고와 같은 것을 lexicon이라고 생각하였기 때문에 그러한 의미에서 어휘목록이라는 뜻도 됩니다. 여기서는 그야말로 어휘부라는 의미로 사용하고 있는데, 어떤 경우는 두 용어 사이에 혼동을 가져올 가능성도 있어서, 우선 그 점을 말씀드립니다. 후자의 의미, 즉 어휘부로 보았을 때의, lexicon은 우리가 보통 말하는 dictionary와는 차이가 있습니다. 어휘부라고 하는 lexicon은 실제로 눈에 보이는 것도 아니고 추상화되어 있으며, 다분히 mental한 그러한 의미로써의 구성이라 할 수 있습니다. 이것을 어떻게 보았을 때, 영어 혹은 국어의 언어현상을 자연스럽게, 무리없이 설명할 수 있느냐에 따라서 여러가지 모델이 나오는 것입니다. 그러니까 조금전에 말씀하신 것처럼 무조건 그려놓고 여기에 끼워 맞추자는 그러한 뜻이라기보다는 주어진 기본 데이타를 설명하는데 이와같은 모델이 적합하지 않겠느냐 하여 나온 것이지요. Halle가 첫번째 시도를 했었는데, Halle의 모델은 아까 지적한 것처럼 문제점이 많습니다.

그리고, 두번째 이야기하신 이러한 모델의 변화가 언어직관이라든가 사실에 바탕을 둔 것이냐 하는 문제에 대해, 예를 들어 'push'에 '-er'을 붙여 'pusher'가 나왔을 때, 그 자체의 semantic compositionality가 생길 것으로 기대될 법한데 뜻밖의 의미가 나오는 경우를 보기로 하지요. 그와 비슷한 예를 두번째 말씀드린 Aronoff는 'transmission'을 들어 설명하고 있습니다. 'transmission'은 'transmit'에 '-ion'이 붙은 경우인데, 'transmission'이 자동차의 한 부분을 의미한다고 하면, 이것을 'transmit'의 명사형으로 보기에는 너무 의미가 달라진 것이지요. 이와 관련하여, Aronoff는, 아까 권혁철 선생님께서 말씀하신 것과도 관련이 있는데, derivation이 lexicon에 속해있다고 하는 것을 다음과 같이 설명하였습니다. 이론적 측면에서 derivation의 과정 즉 morphological rule은 syntactic rule과 같이 어떤 단어를 말할 때 마다 적용시키는 것이 아니라, 한번 적용시켜 단어가 만들어지면 그것으로 끝인 rule이라는 것

입니다. 그래서 Aronoff는 이것을 once-only rule이라고 했지요. 이러한 의미에서 우리가 흔히 말하는 syntactic rule과는 다른 것입니다. 그러니까 word-formation rule이 갖고 있는 기능에는 새로운 낱말을 만들어 내는 기능도 있고, 낱말을 분석하는 기능도 아울러 있는데, 이렇게 once-only rule로 lexicon에 기재되고 나면, 이것은 결국 처음부터 lexicon에 있다고 하는 이야기와 일맥상통하는 것이 됩니다. 그래서 lexicon에 있으면서 의미적으로도 분화가 생길 수 있는데, 이것을 semantic drift라고 설명하기도 합니다. 다른 문제가 언급되면 그 때 말씀드리기로 하고 이것으로 이선생님의 질문에 대한 답변을 대신할까 합니다.

사회: 김선생님이 말씀하신 내용에 대해 다른 토론자께서 거론하실 부분이 있으면 말씀해 주십시오. 특히 어휘부의 구조를 어떻게 형성해야 하는가 하는 점부터 이야기해 주시면 좋겠습니다.

박병수: 예, 제가 말씀드리겠습니다. 아까 사회자께서 시간을 걱정하셨는데, 한 문제씩 따로따로 이야기 하면 너무 시간이 걸릴 것 같으니까, 주로 김영석 선생님께서 발표하신 부분에 대해 초점을 두되 다른 문제와도 연결시켜볼까 합니다.

첫째 발표에서 lexicon의 설계를 어떻게 할 것이냐를 말씀해 주셨는데, 지금 우리 토론회의 주제가 자연언어 처리와 관련된 것이라면, 생성음운론 또는 생성언어학에서 제시해 놓은 lexicon의 설계는 별로 도움이되지 않는다고 말씀드릴 수 있습니다. 이것이 언어 이론상으로는 흥미로운 것이겠지만, 자연언어를 처리하는 데에는 부분적으로 참고될 수 있을지언정 결정적인 도움은 못된다는 것입니다. 그 이유로서는 그 이론들이 모두 절차적(procedural)이고, non-monotonic하고, 또 information-based가 아니라는 점을 들 수 있습니다. 대체로 이 세가지 조건에 맞아야 아까 권혁철 선생님과 정희성 선생님께서 제시하신 자연언어 처리의 모형으로 적절하다고 할 수 있는데, 하나도 맞지 않는 것이지요. 그러한 면에서 보면 국내의 음운론 하시는 분들께서 computational morphology에 대해 별로 관심을 갖지 않고 계시는 것 같다는 생각입니다. 그러한 부분을 좀더 연구해야 하지 않을까 하는 생각이 드는데, 예를들면 L. Karttunen이 하는 computational morphology 같은 것이지요. 아까 정희성 선생님께서 말씀하신 Kimmo와도 관련이 되는 것인데, 그런 이론을 참조하면 보다 실용적인 lexicon의 구성이 나올 수 있을 것 같습니다.

아마도 제일 중요한 것은 information-based라는 조건을 지키는 것이라고 생각됩니다. 지금까지 생성언어학에서 볼 수 있는 규칙이나, 어휘부는 크게 보아 constituent, 작게는 morpheme 또는 phoneme과 같은 단위들을 조작하여 이동하기도 하고 없애기도 하지요. 이런식으로 하는 것을 procedural 또는 derivational하다고 하는데, 그렇게 하지 않고 전체 unit은 그대로 둔채 그 unit이 각기 가지고 있는 음운정보, 통사정보, 의미정보, 또 화용론적 정보 등 모든 정보들을 부분적으로 조작할 수 있는 방법이 있습니다. 그러니까 그러한 정보를 가지는 단위들에는 아무런 변동이 가해지지 않고, 다만 그 단위가 가지고 있는 여러가지 아주 복잡한 복합적 정보들을 한꺼번에가

아니라 부분적으로 조작을 가하는 그러한 방법을 쓸 때에 information-based theory라고 하는데, 그런 면에서 볼 때에는 기존의 생성음운론 또는 생성문법론에서 제시되는 방법들은 그와는 맞지 않는 것들입니다. 그래서, 앞으로 자연언어 처리를 할 때에는 이런 방법으로 정보기반의 이론을 개발해야 할 것이라고 생각합니다.

이러한 맥락에서 실제적인 자료분석을 많이 해야 한다는 이기동 선생님의 말씀도 좋은 지적이고, 또 홍재성 선생님께서 하신 것과 같은 노력도 좋다고 보여집니다. 사전이 순수 언어적 정보만 다루어서는 안되며, world knowledge까지 다루어야 한다는 것이 Richard Hudson의 지적인데요, 실제 우리가 상식적으로 사용하는 사전은 더러 그렇게 하고 있습니다. 언어학에서 사용하는 사전은 그래서는 안된다고 우리가 늘 배워 왔는데 최근에 와서는 그것이 한계가 있다고 생각하고 있고, world knowledge의 바탕이 언어적인 문제를 처리하는 데 있어서도 결정적인 역할을 할 때가 있다는 것이 많이 알려졌습니다. 따라서 Richard Hudson의 그러한 지적도 lexicon 문제를 생각하는 데 대단히 중요한 방향을 시사해 준다고 할 수 있겠습니다.

김영석: 한마디만 comment하겠습니다. 지금까지 나온 이론들, 특히 lexical phonology와 같은 이론은 lexicon 안에 상당한 규칙성이 있어서 그것을 그냥 팽개쳐 놓아서는 안된다는 점을 강조하는 frame work입니다. 지금까지 제가 말씀드린 이론들은 어떤 이론이건 자연언어의 기계처리를 전제로 하고 그의 선행작업으로 model을 제시한 것은 아닙니다. 그러니까 어떤 의미에서는 상당히 sophisticated한 이론이라 할 수 있겠는데, 예컨데 영어의 강세규칙을 예측하기 위해 metrical tree를 그려가는 것은 Hayes의 주장대로라면 우리가 갖는 mental representation을 나타내는 것이라는 것입니다. metrical tree를 그린 것이 어떻게 stress pattern에 대한 우리의 mental representation이 될 수 있냐고 반문할 수도 있겠지만, 이러한 것은 약간 현학적임에도 불구하고 이론적인 측면에서 그것의 설명력을 강조하고 있다고 볼 수도 있습니다.

아까 주제발표에서 Semitic language와 같은 비연쇄 형태론을 설명하기 위한 McCarthy의 prosodic theory를 말씀드린 바 있습니다. 권선생님께서 이야기하신 것처럼 IA model 같은 데서는 예를 들어 '책장'이면 '책'부터 차례대로 가면서 잘라가는 방법을 사용합니다. 그런데 McCarthy가 다룬 이런 non-concatenative morphology에 있어서의 morpheme 분석은 그렇게는 안됩니다. 이와 같이 discontinuous한 자료를 처리하기 위해서는 조금 더 sophisticated한 model이 필요하다고 할 수 있겠습니다. 기계처리를 위해서는 어려울지 모르지만 나중에 더 훌륭한 자연언어 처리의 기술이 개발되면 이런 경우도 처리될 수 있을지 모른다고 볼 때, 당장 이것이 이용될 수 없다고 하여 쓸모가 없다고 해서는 곤란하다고 봅니다. 예를 들어 GB가 LFG보다 기계번역에 도움이 안된다 하여 GB를 틀렸다고 결론짓는다면 마치 사람이 기계에 얽매여서 기계가 틀렸다고 하면 틀린 것으로 인정하고 고개를 숙이는 것과 같다고 생각합니다.

사회: 네 감사합니다. 정선생님께서 말

씀해 주시겠습니까?

정희성: 저도 김영석 선생님의 말씀에 동의합니다. 저희가 어떤 computer system을 권하는 데 있어서 가장 중요하게 생각하는 것은 전체성입니다. 그 전체성 안에서 이와 같은 현상도 어떤 rule이라든가 설명에 의해 나타내어질 수 있으면, 그 이상 좋은 것이 없겠지요. GB이론을 이용한 system도 지난 8월에 Carnegie Melon 대학에서 발표된 바 있는데, 아주 훌륭합니다. GB 안에 있는 여러 principle 등을 전부 programming하여 실제로 잘 돌고 있습니다. 이러한 것을 어디에 활용할 것인가가 문제가 될 수 있는데, 예를 들어 기계번역 system에 활용한다든가 하는 생각을 해 볼 수 있겠지요. 그러나 이러한 것들은 어디까지나 결과론에 불과합니다. 지금 우리가 여기서 lexicon이나, phonology의 여러가지 model을 보았는데, 제 생각으로는 이러한 이론에서 수용할 수 있는 것은 전적으로 받아들이는 입장이 되어야 한다고 생각합니다. 예를들면, lexicon이란 무엇인가?, 어떻게 정의를 할 것인가?, 또 이 안에 음성, 음운적인 정보, 형태적인 정보, 의미적인 정보 등 이질적인 여러 정보들이 들어 있어야 한다고 생각할 수 있는데, 이들을 어떻게 하여 동일한 frame에 넣을 것인가?… 등등 여러 문제들이 많이 있습니다. 이들을 한 frame에 넣을 때, 서로 계층을 나눈다거나, 여러 rule이 서로 엉켜서는 안되겠지요. 이렇게 되면, 다루기가 힘들 뿐만 아니라, 언어학적으로 설득력이 약화된다고 할 수 있습니다. 또한 computer science에서 가장 중요한 개념은 *topology*입니다. 예를들어 음운론적인 문제를 다룰 때에 그것이 hierarchical하게 tree 구조가 되면 search나 sorting을 할 때 상당히 도움이 됩니다. 다시말해서 computer science에서는 그러한 data structure가 요구되는 것입니다. 김영석 선생님께서 survey해 주신 여러 이론 가운데에서도 필요한 것은 전적으로 받아들여야겠지요. 실제로 잘 아시겠지만, MIT에서 MIT Talk라는 음성합성 system을 만든 적이 있는데, 언어학의 많은 이론들을 도입하고 있습니다.

사회: 이정민 선생님께서 말씀해주시지요.

이정민: 김영석 선생님께서 McCarthy의 prosodic theory에 관해서 Semitic 언어를 예로하여 consonantal tier와 vocalic tier를 구별하여 말씀해주셨는데, 우리말의 vowel harmony는 거기에 해당한다고 봅니다. '먹었다'와 '잡았다'에서 과거가 '었'이 되느냐 혹은 '았'이 되느냐 하는 것이 vocalic tier와 관련되고, 이때에는 consonant가 영향을 미치지 않기 때문에 이들을 나누어야 할 필요가 있을 것 같습니다. 그러나 그 밖의 경우는 Semitic 어처럼 consonant를 따로 쪼개낸다고 하여, 어떠한 이득이 있을 것인지는 의문스럽습니다.

그리고 저는 형태론 또는 lexicon 안에서 morpheme이나, 단어를 정리할 때에 syntactic structure 내지 semantic structure의 information이 얼마나 중요한 것이냐 하는 문제를 강조하고 싶습니다. 예컨대, 대체로 derivational한 것으로 간주하는 '고기잡이'는 '고기를 잡다'에 '-이'가 결합하여 된 것이지만, '고기잡이'는 하나의 단어로 쓰이고 있기 때문에 '고기를 잡다'라는 개념은 들어 있지만 '고기'를 따로 수식할 수는 없습니다.

'거창한 고기잡이'라 하면 '거창한'은 '고기잡이' 전체에 걸리지, '거창한 고기를 잡는다'는 의미로는 절대로 해석될 수 없는 특징을 가지는 것이지요. 반면에 '거창한 고기를 잡기가 쉽지 않다'의 경우의 '-기'는 infectional한 것이어서 syntactic information이 그대로 들어갈 수 있다는 차이를 볼 수 있습니다. 여기에서 재미있는 것은 syntactic, semantic한 것과 phonogical한 concatenation과의 bracketing 내지 match의 paradox 문제가 생겨날 수 있다는 것입니다. 영어의 penhandler의 경우, 원래에는 handle과 pen에서 「동사-목적어」의 관계인데, switching이 되어서 개념적으로 pen-handle이 된 것이지요. 음운론적으로 -er은 끝의 handle에 붙는 것이지 전체에 걸리는 것은 아닙니다. 그러니까 phonological level에서는 끝에 -er을 선형적으로 붙여주지만, 개념적으로는 syntactic, semantic hierarchy가 드러나야만 의미를 파악할 수가 있다는 점이 중요하다는 것을 알 수 있습니다.

다음에 아까 권혁철 선생님께서 '겠'과 '다'가 sentence를 subcategorize 해 주는 방식으로 처리하셨는데, 상당히 개념적인 면에서는 일리가 있다고 봅니다. GPSG에서는 VP의 lexical head인 verb가 subject도 subcategorize하고 sentence까지 subcategorize하는 방식을 취하고 있습니다. 그러니까, verb가 전부라는 것인데, verb의 stem 다음에 오는 이 꼬리 같은 것이 문장 전체를 쥐고 흔드는 역할을 한다는 것이지요. 우리말의 경우는 modal이 중요하다고 생각되는데, 이 modal이 문장을 subcategorize하는 이러한 현상을 어떻게 포착할 것이냐 하는 문제는 앞으로 풀어야 할 과제이며, 외국이론만 가지고는 충분치 않다고 봅니다. GB이론에서는 INFL이 subject를 subcategorize하고, subject의 θ-marking은 간접적으로 이루어 진다는 방식으로 처리하고 있습니다. 그래서 Levin, Rappaport는 Stowell 등의 방법을 개선하여, lexicon에 표기하는 방법을 제시한 바 있습니다. 그것이 얼마나 plausible한 것이냐에는 문제가 있지만, syntactic, semantic information이 담겨지도록 노력을 하고 있다는 점은 분명합니다. MIT의 lexicon project group에 속해 있는 이들 Levin, Rappaport는 연구를 계속하여, load와 같은 동사의 경우 'loaded the truck with hay'일 때에는 'full'이라는 implication이 성립하지만, 'loaded hay onto the truck'인 경우는 그러한 implication이 성립하지 않아서 causative analysis로 나아갈 수 밖에 없다는 결론에 이른 것으로 알고 있습니다. 아까 decomposition이 난관에 봉착하여 meaning postulate 방법으로 하고 있는 것이 경향이라고 말씀하셨지만, 깊이 파고 들어가면 자연언어의 처리에 있어서는 decomposition이 여전히 필요한 부분으로 대두된다고 하는 점을 지적할 수 있겠습니다.

그리고, 화용론적인 것이 문제가 되는데, 아까 modal과 관련하여 '-르 것이다, 겠'등이 언급되었지만, 이들은 volitional modal로 취급이 될 수 있겠지요. 이 외에 speech act form이 무엇이냐가 결정적일 때가 있습니다. 예를들어 '나 가겠다', '너 가겠니?', '나 갈래', '너 갈래?' 등 모두 성립하고 의문문도 가능합니다. 그런데, '나 갈께', '*너 갈께?'인 경우는 안됩니다. speech act로서의 promise

의 marker인 '-르께'를 포착해 주지 않으면 다른 volitional과의 관계가 포착되지 않으며, 따라서 volitional만으로 그치는 것이 아니라는 것을 알 수 있습니다. 이런 문제는, 약간은 차원 문제라 할 수 있겠지만, computer AI에서 처리하기에는 아직 어렵다고 하겠습니다. 그런 의미에서 볼 때, 아까 ETRI에서 시와 같은 상당히 context-dependent한 자료를 분석하려고 시도하고 있다고 하셨는데, 조금은 ambitious한 것이 아닌가, 그러니까 순서상으로 더 나중에 해야할 일이 아닌가 생각됩니다.

사회: 네, 한꺼번에 여러가지 언급을 해주셨는데, 우선 권선생님 답변하실 말씀이 있으십니까?

권혁철: 대충 저의 approach에 대해서는 괜찮다고 이야기하시는 것 같아서 특별히 말씀드릴 것은 없을 것 같습니다. 그래서 저의 경험을 한가지만 말씀드리겠습니다. 제가 우리말 '고'를 분석해 보았는데, '고'가 대등접속이냐, 아니면, 종속접속이냐 하는 문제에 대해서 다룬 논문이 있어서 이를 바탕으로 했습니다. 그런데, '고'에 관하여 대등이냐 종속이냐 하는 것의 syntactic structure 자체가 해석이 거의 동일하게 나왔습니다. 그러니까 comp ıter로 '고'를 분석하면서, '고'가 두 요소를 접속시킬 때, subcategorization이 같은 두 요소인 경우이면 대등이고 그렇지 않은 경우이면 종속이라고 할 수 있는데, 대등일 때에도 종속일 수 있기 때문에 그런 식으로 해보았더니 상상외로 비슷한 결과가 나와서 흥미로웠습니다.

사회: 네 감사합니다. 처음 예정으로는 주제발표한 순서대로 문제를 하나씩 거론해 볼까 했었지만, 이야기가 자연히 번져서 그러한 절차는 밟지 않았습니다. 이제 토론자로 나오신 분들께서 대개 한번씩 말씀하셨는데, 그동안 청중들께서도 생각해 놓으신 질의나 comment가 있으실 것 같아 마이크를 잠깐 객석으로 돌리겠습니다. 어떤 문제이든 주저없이 말씀해 주시면 감사하겠습니다.

박대현(한국교원대학원 교육학과): 아까 정희성 선생님께서 Kimmo system이 한국어를 처리하는데 약간 장애가 있다고 하셨는데, Kimmo system을 보완하는 문제에 대해 말씀해 주시면 감사하겠습니다.

정희성: 보완하는 것이 아닙니다. 그러니까 Kimmo system자체의 결함이 아니고, 우리말의 morphology에 대한 충분한 검토없이 그냥 맞추어 놓았으니 output이 나올 리가 없었겠지요.

박대현: 권혁철 선생님께 한가지 더 질문드리겠습니다. 우리나라에서 문자인식 computer는 이미 나온 것으로 알고 있는데, 음성인식 computer의 개발이 어떻게 진행되고 있는지 궁금하고, 또 인공지능에 관한 연구가 과학기술원, 전자통신 연구소, 및 각 대학에서 분산되어 이루어지고 있어서 그러한 것을 통합해서 하면 여러 측면에서 더 좋지 않을까 생각되는데 이에 대해 말씀해 주십시요.

권혁철: 문자인식 system이 나왔다 하여도 그것은 인쇄가 아주 잘된 것만 가능하고 또 글자체가 바뀌면 안되는 그 정도의 수준이어서 아직 실용화는 안되어 있는 상황입니다. 음성인식 system의 경우는 잡음도 많고하여 생각보다 훨신 어렵습니다. 외국의 경우는 음성인식 system이나, 문자인식 system이 실용화되어

있는 예가 많지만, 우리의 경우는 한국어에 대한 연구가 충분치 못하고, 투자가 적었기 때문에 아직 못하고 있다고 할 수 있겠습니다. 그리고 AI에 관한 연구가 여러 단체들이 통합하여 이루어 진다면 잘 되지 않겠느냐고 하셨는데, 그럴 가능성이 없지는 않습니다. 하지만 AI를 연구하는 사람들이 모두 자연언어 처리에 뛰어들지도 않을 것이고, 전세계적으로 보아도 가까운 시일 내에 실용화된다는 것보다는 어떤 domain을 잡아서 조금씩 실용화하는 단계에서 시간이 흘러 경험이 쌓이다 보면 실용화가 되지 않겠느냐 하는 것이 기본적인 태도라 할 수 있습니다. 다시 말씀드려서, 자연언어 처리에 대해 아직 저희들이 경험이 없습니다. 이렇게 경험도 없는 상태에서 모두 뛰어든다 하여 되는 것은 아니겠지요. 조금씩 조금씩 난관을 극복하다 보면, 어느 순간에는 실용화가 가능하지 않겠나 생각합니다.

박대현: 마지막으로 한가지만 더 묻겠습니다. 자동번역의 경우는 요즈음 어떻게 되어가고 있는지 말씀해 주십시요.

정희성: 거기에 대해 제가 말씀드리겠습니다. 기계번역 쪽에서 가장 발전을 하고 있는 나라가 일본입니다. 상당히 실용화되어서 영·일 혹은 일·영의 번역을 하는데 도움이 될 수 있는 system을 아주 싼 가격으로 구입할 수 있습니다. 이것이 어느정도 수준이냐 하면, 없는 것보다 낫다는 정도라 할 수 있습니다. 사전을 일일히 펴보는 것보다는 좋더라는 정도이지요. 그리고 일본의 대기업들이 기계번역을 많이 하고 있어서, main frame을 팔 때에 도움을 받는, 그러니까 computer-aided machine translation system이라고 하는데, 이에 대한 연구가 이루어지고 있습니다. 미국 쪽은 Carnegie Melon대학에서 음성 interface를 붙여 영/독, 영/일의 번역이 상당한 수준까지 이루어지고 있습니다. 한 예로 speech decoding system이 300 sentences를 화자 구분없이 받아들이는 정도입니다. 상당한 기술수준이라 할 수 있습니다. 외국에서는 이러한데, 우리는 왜 못하고 있느냐?라고 물으실 수 있겠지만, 조금만 시간을 주시고 기다려주시기 바랍니다.(웃음)

사회: 잠시 일반적인 질문이었습니다. 다시 오늘 주제에 관련되는 말씀이 있으시면 해주시기 바랍니다.

김한곤: 아까 논의되었던 여러가지 문제 가운데에서 우리가 짚고 넘어갔으면 하는 philosophical한 문제가 있는 것 같습니다. 그리고 lexicon의 구조에 있어서 어떤 model이 좋은가의 문제가 거론되었는데, 이에 대한 저의 견해까지 두 가지를 말씀드리도록 하겠습니다.

먼저 philosophical한 문제에 관하여 언급하면, 언어학자가 기술한 것을 computer가 못한다는 것은 사실입니다. 현재로서는 언어학자가 기술한 것을 computer가 따라갈 수 없습니다. 그렇다면 기계의 수준에 머물도록 언어학자가 기술을 포기할 것이냐 하는 문제가 생겨납니다. 저는 본래 언어학을 전공했고 computer를 어깨넘어 보았는데, 그런 사람의 입장에서 생각을 해보았습니다. NLP(자연언어 처리)를 하는 사람들은 자신들이 추구하는데 있어서 목표의 수준(levels of ambition), 그러니까 어느정도까지 달성하는 것을 목표로 하느냐 하는 것을 반드시 전제하고 시작합니다. 아까 그런 문제가 전혀 언급되지 않은 상황에서 이야기가 진행되어서, 청중들의 입장에서 다소 con-

fusion이 생긴 것 같습니다. 기계번역을 예로 들 때, 인간처럼 완벽하게 하느냐, 또는 인간이 도움을 주어서 기계가 주가 되도록하느냐, 아니면, 기계의 도움을 받아서 인간이 주가 되는 것이냐, 또, 가장 낮은 수준에서, 기계를 인간이 하는 것에 대한 단지 utility 정도로만 사용할 것이냐 등등 여러 수준의 목표를 생각할 수 있습니다. 저로서는, 언어학을 전공하는 사람으로서, 기계가 언어에 대해 인간이 연구하는 것을 그대로 다 implement해 주는 그러한 수준이 이상적이겠지만, 조금은 포기하는 심리가 생겼습니다. 다시 말해서 우선 우리의 목표로서는 그렇게까지 하는 것은 지나친 것이 아닌가 하여 타협하는 입장이 되었다고 할 수 있습니다. 반대로 기계번역하는 사람의 입장에서도 그럴 것 같은 생각이 듭니다. 그래서 그러한 여러가지 level을 생각해 보면, 김영석 선생님께서 하신 것처럼 아주 이상적인 것도 자신의 필요에 따라 목표로 해 볼 수 있는 것이고, 홍선생님이나 박병수 선생님이 목표로 하신 것도 해 볼 수 있는 것이라는 생각입니다. 결국 각자의 목표에 따라서, 또한 자신에게 주어지는 경제적, 시간적 여건에 따라서 마음대로 해볼 수 있겠지요. 그러니까 결론적으로 제 이야기의 philosophical background는 기계란 인간을 위해서 쓰는 도구이기 때문에, 그 기계를 무슨 목적으로 쓰고자 하느냐 하는 목표에 따라서, 그리고 시간과 공간, 경제적 여건이 허락하는 범위 안에서 각자가 여러가지 목표를 세울 수 있을 것이고, 따라서 그것 가지고 너무 옳고 그른 것을 이야기할 필요는 없지 않을까 하는 것입니다.

다음에 lexicon에 대한 이야기인데, 아까 언어학자들이 쓰는 lexicon의 model을 그대로 computer에 implement할 수 있느냐 없느냐 하는 문제가 거론되었었지요. 이 역시 처음에 말씀드린 philosophy에 관계되는 문제인데, 만일에 이상적인 언어습득의 model이나 description을 목표로 한다든지, 아니면 예를들어 GB theory specific하게 그것을 만들어서 GB theory가 과연 물리적으로 proof가 되느냐 하는 것을 알아보기 위한 그런 식의 목표가 아니라면, 우선은 수준을 조금 낮추는 것이 좋을 것 같은데, 이 때에는 역시 unification-based grammar가 편리할 것 같습니다. 여기 계시는 전문가분들께서 더 잘 아시겠지만, 요즈음은 GB parser도 나와 있습니다. 하지만 그런 것을 다 감안한다 하여도, unification-based grammar가 더 비용도 적게들이고 여러모로 편리하다고 봅니다. Lexicon의 구조가 사실은 굉장히 복잡합니다. 제가 사용해 본 경험이 있고, 개발하는 데 기여도 한 model에 대해서 말씀을 드리면, lexicon은 categorial concepts는 물론이고, 그 밖에 아까 박병수 선생님께서 이야기하신 바와 마찬가지로, world knowledge의 부분도 가지고 있어야 합니다. 예를 들어, 명사의 경우 Chomsky가 *Aspects*에서 inherent feature라 불렀던, semantic feature도 다 들어있고, 심지어 어떤 명사가 technical field로 보아 astronomy의 술어이냐, 또 medical field라면 그 중에서 surgical term이냐 하는 것까지 다 들어 있습니다. 그렇게 하는 이유는, 이것이 machine translation system이기 때문에, ambiguity를 resolve할 때에는, 그 분야를 맞추어야 translation 하는 시간을 절약할 수 있다는 데 있습니다. 다시말해서 world

knowledge를 넣어야 performance가 좋아지고, 오역도 줄어들며, ambiguity resolution의 한 방법도 된다는 것이지요. 동사의 경우도, form, stativity/non-stativity,... 등등을 다 넣어 주어야 하고, 게다가 영어만 보더라도 complement structure가 보통 복잡한 것이 아닙니다. 뿐만 아니라 phrasal verb에 있어서 동사에 따르는 전치사에 대한 정보도 필요합니다. semantic interpretation도 물론 넣어주어야 하지만, 이것을 syntactic하게 규정하는 자체가 대단히 어렵습니다. 저의 경우는 transfer model을 사용한 것이었기 때문에 transfer portion도 넣어야 했는데, 이 경우 transfer portion에서 pickup한 lexicon이 한·영 사전과 영·한 사전의 양쪽에서 어떻게 추리가 가능하도록 하느냐, 또한 양쪽의 feature attributes를 implement 한 것이 어떻게 서로 interact하게 할 것이냐 등등 복잡한 문제가 엄청나게 많습니다. 이런 의미에서 볼 때, lexicon의 structure가 transformationalist라든지 우리 언어학자가 만든 neat한 lexicon과는 매우 거리가 멀다고 말씀드릴 수 있습니다. 형용사의 경우도 마찬가지인데, 이러한 모든 것을 고려해 보면, model construction도 문제이지만, 이것을 input하는 자체가 굉장히 큰 일이라는 것을 알 수 있습니다. 즉, research하는 것이 문제가 아니고, 이론적인 것보다 housekeeping쪽이 어떻게 복잡한지 그 자체만으로도 너무 어려워서 언어학자가 만들어 놓은 이상적 model을 가져다가 바로 쓰는 것은 거의 불가능하다고 할 수 있습니다. 그러나 philosophical하게는 어느 목표라도 잡을 수가 있겠지요. 그러니까 우리가 이론적으로는 싸울 필요가 없을 것 같습니다. 양쪽 주장이 다 옳고, 단지 정해놓은 목표에 따라서 택하는, decision-making의 문제만 남아 있다고 생각됩니다.

사회: 오늘 토론회의 매듭을 지어가는 방향의 말씀이라 생각되는군요. 저희가 현재 수준의 여건에서 lexicon의 구조에 관해 한번 논해보자고 했기 때문에 일부러 쟁점화한 것이었지만, 사실 전산쪽에서 처리할 수 있는 능력에 비례해서 우리 언어학에서 설계해 놓았던 것을 수용할 수 있는 미래가 오리라는 것은 이미 예측되었던 것이었습니다. 객석에서 질문이 더 있으신가 본데 말씀해 주십시요.

신경구(전남대 영문과): 아까 박병수 선생님께서 computer로 적용가능한 morphology를 했으면 좋겠다고 말씀하셨고, 또 이에 대한 대답으로 꼭 computer로 implementation이 되어야만 좋은 이론인 것은 아니다는 말씀도 있었는데, 제가 말씀드리고 싶은 것은 computer로 implementation이 될 때에만 좋은 이론인 것이 아니기도 하지만, 거꾸로 computer에 implementation이 된다고 하여 꼭 좋은 이론이라고 규정할 수는 없다는 것입니다. 예를들어 과거에 한자가 좋다고 하는 사람들이 한자도 computer로 처리할 수 있다고 하여 그것이 한자와 한글을 혼용해야 한다는 주장에 대한 중요한 근거라도 되는 양 강조한 바 있습니다. 그러나 그럼에도 불구하고 우리가 실용적으로 쓰기에 한자는 불편한 글자임은 분명합니다. 그러듯이 GB theory가 computer로 처리된다 하여 그것도 받아들일 수 있는 이론이라고 주장할 수는 없을 것 같습니다. 물론 제가 GB theory가 나쁘다고 주장하는 것은 아닙니다만, computer로 처

리가 된다고 하더라도 그 이론이 좋은 것이냐 아니냐는 더 많은 문제를 검토해 보아야 할 문제라고 생각됩니다.

사회: 네, 감사합니다. 저희가 본래에는 lexicon 내에 담길 information이라든가 lexicon을 표현하는 formalism에 대한 것도 오늘 상세하게 다루어 볼까 했었는데, 그동안 이야기하신 중에 직접, 간접으로 언급이 되었고 하여, 더 깊이 들어가지는 않겠습니다. 종결을 짓기 전에 마지막으로 하고 싶은 말씀이 있으시면 발언해 주시기 바랍니다.

정희성: 저는 사실 오늘 morphology와 lexicon에 대한 정의랄까 그런 것을 기대했었습니다. 저희 computer하는 사람들은 순수한 언어학자가 아니거든요. 예를 들어 낱말, 이은말… 등 여러 개념에 대해 언어학하시는 분들께서는 어떻게 생각하고 계시는지 하는 점이 궁금했었습니다. 그리고 어떤 문법의 우월론을 가지고 논의하는 것은 조금은 이상하다는 생각이 듭니다. 각자의 취향이나 학문적인 배경에 따라 추구하고 있기 때문이지요. 하여튼 여러 방향에서 여러 이론이 나오면 좋겠다는 생각입니다. computer에 태우는 것은 어떤 목적이 있어서 태우는 것입니다. 따라서 이런 것을 너무 의식하지 마시고 여러가지 재미있는 연구를 내어 주셨으면 하는 것이 computer science를 하는 저희 입장에서의 바람입니다.

홍재성: 제가 아까 시간을 많이 쓰면서도 이야기하지 못한 것이 많은데, 몇가지만 말씀드리겠습니다. 다른 분들께서는 형태론과 관련된 정보나, 형태론적인 연관성을 주로 말씀해 주셨는데, 그것도 물론 중요합니다. 제가 여러분께 소개해드린 자료들은 대개 단문구조 내에서 구조자체가 보여주는 통사적인 속성들인데, 과거의 변형생성문법에서는 PS-rule이나 transformation rule과 같은 syntactic rule로서 다루던 것들입니다. 사실은 그것이 동사에 매인 속성이기 때문에, 생성문법쪽에서도 이것을 뒤늦게 인식하여 lexicon에서 다루고 projection principle에 의해 syntactic representation에 이어지는 것으로 보는데, 방향은 옳다고 생각이 됩니다. 그런데 그 테두리 안에서도 정보들이 상당히 많이 있습니다. 이 정보들을 기계번역 장치의 일부로 보는 경우에 어떻게 표상하고 입력하느냐 하는 문제는 제 능력 밖의 문제이고, 제가 강조하고 싶은 것은 이 정보들은 모국어 화자가 알고 있는 동사어휘에 관한 지식으로서, 예를들어 한국어 문장을 정확히 만들어 내려면, 이러한 지식이 다 동원되어야 한다는 점에서 필요한 정보이고, 최소한 lexicon에 이러한 정보들이 수록되어야 하지 않을까 하는 것입니다.

그리고 아까 의미표상의 문제는 거론만 했는데, 이정민 선생님께서 말씀하셔서 한마디만 더 말씀드리기로 하지요. Rappaport, Levin, Laughren이 1988년에 쓴 최신 논문에 보면, Stowell이니, Marantz이니 이런 사람들을 모두 비판하고, 결국 two level representation이 필요한데, 하나는 predicate-argument structure이고, 또 하나는 lexical conception structure —사실 이것이 의미표상입니다—라고 하고 있습니다. 또, predicate는 decomposition을 해야 하고, θ-role의 level을 가지고 무엇을 열거하든지, 혹은 linking rule을 고안하는 것은 문제가 많이 있으므로 variable을 사용해야 한다는 등 여러가지 재미있는 주장을 하고 있습니다.

그리고 GB의 틀을 받아 들이면서도, 동일한 한 동사가 두 가지 구조에 규칙적으로 대응되어 나타나면서 의미해석이 규칙적으로 차이를 보이는 그러한 현상을 어휘부에 통사구조와 의미표상을 대응시켜 적절히 나타내어 주어야겠다는 생각인데, 저도 상당히 중요하다고 생각합니다. 제가 보충자료의 (43)번과 (45)에 그것을 소개해 드리려고 적어 놓았는데, load의 lexical conceptual structure를, 장소보어를 직접목적보어로 하는 경우와 장소보어가 비목적보어인 경우의 의미차이를 나타내기 위해 (45)와 같이 표상합니다. 어휘해체를 하여 공통성이 있으면서도 차이가 있는 것을 그렇게 나타내고 variable을 가지고 대응시키면, predicate argument structure와 대응이 잘 된다는 이야기이지요. 그런데 결국 load를 이렇게 해 놓으면 put과 같은 동사와 구별되지 않게 됩니다. 사실 이런 동사와 구별하기 위해 어떻게 분석하고 나타내어야 할지를 마저 제시했어야 했는데 그렇게 하지 않았습니다. 한국어의 경우도 마찬가지인데, '신다, 놓다, 넣다, 얹다, 지다'가 있으면서도, 조금씩 달라집니다. 이렇게 달라지는 데에 따라서 의미역할의 contact도 달라지는데, 그렇게 하자면, 결국 (43), (45)의 표기방식 가지고는 어렵지 않겠는가 하는 생각입니다. 그렇다 하여 초기의 생성문법처럼 무슨 feature를 열거하기도 곤란하겠지요. 그래서 Melčuk의 정의방식을 제시해 보았는데, 대안으로 한번 고려해 볼 만한 것이 아닌가 하는 생각이 듭니다.

사회: 오늘 토론을 매듭짓는 이야기들이 이미 나왔기 때문에 이 정도로 마무리를 할까 합니다. 제가 한마디 덧붙이자면, 물론 순수 이론도 존재할 수 있고 중요하지만, 우리 언어학자와 전산학을 하시는 분들 사이에는 앞으로도 긴밀한 협조가 있어야 할 것이고, 또 한 쪽에서 이론을 구성할 때에는 그것을 구현하는 면이 어떻게 될 것인가에 대한 고려도 하면서 일을 해 나간다면, 경우에 따라서는 우리 문화, 특히 언어문화의 발전의 첩경이 될 수 있는 길이라고 생각합니다. 이 symposium이 그러한 면에서 의의가 있었다고 생각되고, 앞으로도 어학연구소로서는 이러한 자리를 많이 만들도록 하겠습니다. 감사합니다.

(정리 : 최 동 주)