논 평

마음S, 마음R, 그리고 마음 (non-S, non-R)

- '자연언어와 인공지능'에 관한 논평

서 정 신 (한양대 강사)

이승종 박사가 그의 논문 '자연언어와 인공지능'에서 다루고 있는 주제는 넓게 말해서 '컴퓨터가 인간의 마음과 같은 마음을 소유하고 있는가?'하는 문제이다. 이 문제는 다시 '컴퓨터가 자연언어를 이해할 수있는가?'하는 더 구체적인 문제로 좁혀져 있다. 이 구체적인 문제에 접근하기 위해 이 승종박사는 다시 '자연언어를 이해한다'는 말이 어떻게 설명되어야하는 지에 촛점을 맞춘다. 전통적으로 자연언어의 이해가 의미론(semantics)의 영역에서 다루어져왔음을 인정하면서 그는 구문론(syntax), 화용론(pragmatics)과 구별되어온 의미론의 독자적인 영역이 와해되고 의미이해의 새로운 모델이 제시되는 세가지 방식에 주목한다. 그 첫째는 콰인의 방식이다. 이 방식은 문장의 의미인 명제가 불확정적이기 때문에 의미론은 심리학이나 언어학등의 과학으로 대체되어야 한다는 주장에 근거한다. 둘째는 비트겐슈타인의 방식이다. 의미는 언어의 사용에서만 드러나기 때문에 의미론은 화용론으로 환원되어야 한다는 주장에 의해 의미론의 독자적 영역이 와해된다. 세째는 AI이론가들의 방식이다. 이는 의미론이 구문론의 영역으로 환원되어야 한다는 주장에 근거하고 있으며 자연언어의 의미에 대한 인간의 이해가 컴퓨터가 수행하는 구문론적 조작으로 환원된다는 입장에 의미론의 독자영역이 와해된다. 여기에서 이박사는 AI이론가들의 의미이해에 대한 모델을 비판하고 비트겐슈타인의 의미이해에 대한 입장에 동조적인 입장에서 컴퓨터가 자연언어를 이해할 수없으며, 따라서 컴퓨터는 인간의 인식을 투사하는 모델이 아니라고 주장하고 있다. 그의 논의는 크게 두 갈래로 나누어 진다. 첫째 논의는 AI이론가들의 입장에 대한 비판적 논의이다. 자연언어의 이해가 구문론적 조작가능성으로 설명될 수 있고 그에 따라 의미론이 구문론으로 환원될 수 있다는 AI이론가입장의 전형인 래퍼포트와 블락의 입장이 그들의 논의에 의해 지지되지 않는다고 주장한다. 반면 이박사의

둘째 논의는 AI 모델이 아닌 삶의 양식을 그 기준으로 삼는 비트겐슈타인의 언어 이해의 모델을 언어 이해에 대한 적절한 모델로 상정하고 컴퓨터와 인간의 마음이 갖는 상이성(difference)를 주장하는 긍정적 논의이다. 이 두 가지 논의를 우선 살펴 보자.

그의 부정적 논의는 다시 두 부분으로 나누어진다. 첫째 부분은 썰과 래퍼포트 간의 논쟁에 대한 비판적 접근이다. 튜링 테스트를 언어이해의 기준으로 받아들이는 AI이론가들의 입장이 지지될 수없다는 썰의 논의가 래퍼포트의 소위 한국어방 논의에 의해 타당하게 반증되어 있지 못하다는 논의가 그것이다. 둘째 부분은 래퍼포트의 반증을 보충해주는 듯 보이는 블락의 입장역시 근본적인 문제가 있기 때문에 의미론의 구문론적 환원이 정당화되어 있지 않다는 주장을 담고 있다.

이박사가 소개한대로 썰의 중국어방논의는 튜링테스트의 행동주의적 해석을 그대로 사용한 사유실험에 기초해서 자연언어를 구문론적으로 조작하는 능력이 자연언어를 이해하는 능력일 수없다고 주장하고 있다. 썰은 컴퓨터 프로그램의 구문론적 조작에서는 인간두뇌가 갖는 인과적 힘이 결여되어 있다고 지적하면서 컴퓨터가 자연언어를 이해하지 못한다고 주장한다. 이에 대해 래퍼포트는 썰의 중국어방과 유사한 한국어방을 상정하고 영어를 이해하지 못하는 그 방안의 사람이 중국어방에서와 유사한 구문론적 조작을 통해 셰익스피어 작품을 번역하고 논문을 써내는등 셰익스피어를 이해하는 듯 보이는 상황을 구성하고 있다. 래퍼포트에 의하면 한국어방의 사람은 그가 영어를 이해하지 못한다고 생각한다 하더라도 사실은 영어를 이해하고 있다는 것이다. 여기에서 이박사는 래퍼포트의 한국어방논의가 썰의 중국어방 논의에 대한 타당한 반증인가를 묻고 있다. 그리고 그의 답은 부정적이다. 이박사는 그 이유를 세가지로 내세운다. 그 첫째는 래퍼포트가 생각하는 의미이해R와 썰이 생각하는 의미이해S는 마치 천동설을 받아들인 톨레미의 운동개념P와 지동설을 주장한 코페르니쿠스의 운동개념C처럼 이질적이기 때문이라는 것이다. 썰과 래퍼포트가 하나의 의미이해에 대해 얘기하고 있지 않고 서로 다른 개념을 사용하고 있어서, "이해R 이 이해S에 이미 내포되어 있는 의미를 보다 명확하게 하고 있음을 보이거나, 이해R이 이해S에 대해 갖는 그밖의 다른 이론적, 혹은 경험적 이점을 제시"(13페이지)하고 있지 않는 한 래퍼포트의 한국어방논의는 썰의 중국어방논의에 대한 타당한 비판이 아니라는 추론을 따르고 있는 것이다. 그런데 정말 이승종 박사의 지적대로 썰과 래퍼포트의 '이해' 개념이 톨레미와 코페르니쿠스의 '운동' 개념처럼 이질적인가? 이박사의 지적대로 래퍼포트는 "자

연언어의 이해를 컴퓨터가 수행하는 계산적,구문론적 조작과 동일시하고" 있어서 "컴퓨터가 자연언어를 이해한다"고 주장하는가? 아니면 래퍼포트가 상정하는 자연언어이해의 기준자체가 오히려 컴퓨터의 계산기능을 포함하는, 그리고 어느 인지기능체에 의해서도 발현이 가능한 계산적 인식기능(computational cognitive function)에 의한 기준이기 때문에 결과적으로 그 기능체들중의 하나인 인간처럼 또 다른 인지기능체인 "컴퓨터가 자연언어를 이해한다"고 래퍼포트가 주장하고 있는 것이 아닌가? 이는 이승종박사자신의 관찰, 즉 이해S가 자연언어에 대한 인간의 이해인반면 이해R는 "반드시 인간의 이해를 의미하는 것은 아니다"(12 페이지)라는 관찰과도 일관된다. 래퍼포트 역시 "인간의 심성(mentality)가 무엇이며 그것이 어떻게 운용되는지에 관한 계산이론보다는 **운용**의 매개와는 독립적으로 어떻게 심성이 가능한지에 관한 계산이론에 더 관심이 있다." (래퍼포트의 논문, 이승종박사의 글 13페이지에 인용, 진한 부분은 본인의 강조)라고 말하고 있다. 다시말해, 이승종박사가 이해S, 이해R구분을 하면서 래퍼포트의 자연언어이해개념을 컴퓨터가 수행하는 '운용의 매개에 의존적인' 계산기능에만 연결시키는 것은 래퍼포트에게 공정하지않은 게 아닌가 생각된다. 인간의 자연언어이해나 컴퓨터의 자연언어이해를 각각의 운용매개에 의존적인 방식으로 설명할 수도 있고 각각의 운용매개에 독립적인 방식으로 설명할 수도 있다. 썰이 전자의 방식을 취하고 있는 것은 그가 인간의 두뇌라는 운용매개를 근원적으로 사용하고 있다는 면에서 분명히 나타난다. 그러나 래퍼포트는 그 스스로도 강조하듯이 후자의 방식을 사용하기 때문에 썰처럼 운용매개의존적인 설명을 한다고 말할 수 없다.

래퍼포트의 논의에 대한 이박사의 두번째 문제제기는 다음의 물음으로 요약된다. 즉, 한국어라는 자연언어에 의한 셰익스피어와 영어라는 자연언어에 의한 셰익스피어를 넘어서는, 즉 자연언어로 표현된 셰익스피어를 넘어서는 셰익스피어는 존재하는가라는 질문이 그것이다. 이 물음에 대한 그의 답은 부정적이다. 즉, 콰인의 번역불확정성론에 의해 프레게가 말하는 의미의 제3영역귀속이 와해되었기 때문에 자연언어로 표현된 셰익스피어를 넘어서는 셰익스피어의 작품은 존재하지 않는다는 입장을 갖고 있는 것이다. 정말 그러한가? 그렇지 않다. 콰인의 불확정성 논의는 프레게가 문장의 의미로서의 사고내용체(the thought)를 제 3의 영역에 설정했다는 사실을 또 하나의 플라톤주의로 몰아부치는 입장에서만 프레게에게 위협적이다. 프레게가 말했던 문장의미의 객관성은 플라톤의 이데아처럼 형이상학적으로만 해석되어지는 것은 아니다. 더밋(Dummett)등의 프레게학파들은 문장의

의미가 갖는 객관성을 간주관성(intersubjectivity)으로 설명하고 있으며 그에 따라 문장의 의미가 상정되었던 제3영역에 대한 설득력이 확장될 수있다고 보았다. 한 자연언어를 사용하는 우리가 다른 자연언어로 쓰여진 간 자연언어적(interlingual)인 셰익스피어를 이해할 수 없는가?

나는 이것이 가능하다고 생각한다. 우리가 셰익스피어를 이해한다는 것은 결국 영어로 쓰여진 셰익스피어에 담긴 정보의 합을 얻는 것을 말한다. 예를 들어 햄릿은 햄릿이라는 가공인물과 그 인물을 둘러싼 상황에 대한 관계적 정보의 합으로 이루어져 있기 때문에 이 정보가 자연언어에 독립적으로 다른 자연언어를 사용하는 셰익스피어학자에게 사용될 수있다는 것이다. 제 3의 영역안에 존재하는 형이상학적 실체로서의 사고내용체는 동일성조건이 문제가 될 수있다. 그러나 간 자연언어적 정보는 자연언어간의 차이점을 인정하는 동시에 그 자연언어라는 수단(medium)을 넘어서는 내용적 유사성을 기준으로 동일화 될 수 있을 것이기 때문에 획일적인 동일성기준이라는 제한조건에서 벗어날 수 있다. 그것은 여러사람이 둔, 여러 경우의 특정한 형태의 바둑의 수(手)가 모두 동일한 것이라고 말할 수있다는 이야기와 같다. 어떤 바둑의 수를 둔다는 행위는 그 구체적인 바둑돌들과는 떼어질 수 없지만 그럼에도 불구하고 구체적인 바둑돌과는 독립적으로 어떤 바둑의 수를 이야기할 수 있다는 것이다. 여기에는 다음과 같은 유추또한 도움이 될 수 있다. 문학비평론가들은 뮤지컬 '웨스트 사이드 스토리'가 셰익스피어의 비극 '로미오와 줄리엣'의 내용을 담고 있다고 한다. 구체적인 등장인물이나 배경등의 차이점에도 불구하고 두 작품의 내용의 동일성이 주장될 수있었던 근거는 두 개의 작품모두 주인공들간의 관계설정이나 작품내의 갈등의 구조등이 같다는 관찰에 있다. 즉 구체적 작품들이 갖고 있는 표현수단(medium)과는 독립적으로 추상화된 정보들이 갖고 있는 동일성이 이야기될 수있다는 것이다. 가상진실(virtual reality)과 그것이 모사하는 현실사이에 존재하는 대응적 관계성에 의한 동일성 상정도 위와 같은 맥락에서 가능할 것이다.

위의 두 가지 문제제기에 비해 세번째 문제제기는 훨씬 더 설득력을 갖고 있다. 이박사는 썰이 래퍼포트의 응답을 '체계응답(The Systems Reply)이라고 비판하는 내용을 검토한다. 즉, 썰은 래퍼포트가 중국어방, 혹은 한국어방에서 자연언어를 이해하는 '나'와 자연언어를 이해못한다고 믿을 수 있는 신념체계로서의 '나'로 이원화하고 있다고 비판한다는 것이다. 이박사의 이해S/이해R구분에 상응해서 후자의 '나'와 전자의 '나'를 '나S'/'나R'의 구분에 의해 상정해보자. 각각의

'나'는 각각의 이해를 하는 당사자들이다. 이박사는 이해S와 이해R을 다음과 같이 비교한다.

> "래퍼포트의 이해R은 이해S보다 오히려 허약한 경험이다. 이해R의 내용과 양상이 구문론적, 계산적 기능에 머물러 있는데 반해 이해S, 즉 인간의 이해의 내용과 양상은 그보다 훨씬 더 복잡하고 다차원적일 수있기 때문이다."

하지만 이박사는 이해S의 상대적 우위성에 대해 더 구체적으로 부연설명하지는 않고 있다. 썰이 거의 암묵적으로 전제하고 있는 그 상대적 우위성을 그대로 받아들이는 듯하다. 인간의 이해가 그 내용이나 양상에서 "훨씬 더 복잡하고 다차원적"인 이유는 무엇이며 그 구조는 어떠한가? 이 구조는 컴퓨터가 실현할 수없는 구조인가? 이해 S가 이해R보다 더 복잡하고 다차원적이라면 '나S'역시 '나R'보다 더 복잡하고 다차원적이라는 말이 되는데 그 설명은 어떻게 가능할 까? 불행히도 우리는 더 이상의 설명을 듣지못한다. 래퍼포트의 썰비판에 대한 세번째의 문제제기는 그러므로 그 직관적 설득력에도 불구하고 불완전하다. 블락에 대한 이박사의 비판은 래퍼포트에 대한 비판과 상당부분 중복되므로 생략한다.

여태까지는 이승종박사의 부정적 논의부분을 살펴보았다. 이제는 그의 긍정적 논의부분을 살펴보자. 부정적 논의에 비해 이 부분은 매우 축약되어 있다. 그는 우리가 컴퓨터와 다른 삶의 형식을 따르고 있으므로 우리가 자연언어를 이해하는 내용이나 양식을 설명하려면 우리의 실제적 삶의 문맥안에서 이해, 인식, 의미등의 개념들이 사용되는 방식을 기술해야 한다고 한다. 후기 비트겐슈타인의 입장에 근거해서 그는 투사의 대상으로서의 컴퓨터나 컴퓨터의 인식은 인간, 인간인식과 다를 수밖에 없다고 결론내리고 있다. 여기에서 우리는 다시한번 부연설명에 대한 아쉬움을 느낄 수밖에 없다. 무엇보다 우리는 다른 삶의 형식을 따른다는 사실이 곧 다른 마음을 갖고 있을 수 밖에 없게 하는지를 - 그 반 비트겐슈타인적인 함축에도 불구하고- 다시 물을 수있다. 우리의 삶의 형식의 특수성이 설명되거나 이해S, '나S'의 구조가 어떤 방식으로 더 고차원적이고 복잡한지가 밝혀 질때까지 우리는 이승종박사의 매우 호소력있는 결론에 선뜻 동의하기가 어려울 것이다.

이박사는 우리에게 컴퓨터가 자연언어를 이해못한다고, 그리고 따라서 컴퓨터가 사람의 마음과 같은 마음을 갖고 있지 못한다고 말해준다. 그리고 그 이유는 인간과 컴퓨터의 삶의 형식이 다르기 때문이라고 말한다. 그런데 우리는 이 모든

주장들에도 불구하고 다시 이렇게 물을 수 있다. 비록 컴퓨터가 사람의 마음을 갖지 못한다고 인정한다고 치자. 그렇다면 컴퓨터는 사람의 마음이 그 한 부분이 될 수 있는 마음은 갖고 있는가? 이박사의 문제제기에도 불구하고 아직 래퍼포트가 말한 매개수단 독립적인 이해를 하는 마음이 상정 가능 하다면 우리는 컴퓨터가 그 마음, 즉 마음S도 마음R도 아닌 마음(non-S, non-R)을 갖는지 물어볼 수있을 것이다.