자연언어와 인공지능

이 승 종 (연세대 강사)

1. 의미와 자기 인식의 진화

1.1 지상의 척도

인간이 지구상에 처음 출현했을 때 그가 당면한 최대의 문제는 생존의 문제였을 것이다. 자연은 인간에게 종종 적대적이기 때문이다. 생존에 필요한 초보적인 작업은 세상 만물을 인간의 입장에서 갈라 보는 것이다. '나'와 나 아닌 것, 먹을 수 있는 것과 없는 것, 더운 것과 찬 것, 해로운 사람과 이로운 사람 등등. 이러한 여러 가름을 해 나가다 인간은 가름을 좀더 용이하고 체계적으로 하기 위해 '이름'을 고안하게 되었을 것이다. 먹을 수 있는 것은 '음식'으로, 없는 것은 '독'으로, 해로운 사람은 '적'으로, 이로운 사람은 '친구'로 등등. 언어는 바로 이러한 가름의 작업에서 비롯된다. 그리고 언어에 대한 이러한 발생적 가정은 언어를 가름의 체계로 보는 소쉬르와 데리다의 언어 이론과도 부합한다.

다음 단계의 작업은 자연을 인간의 입장에서 이해하는 것이다. 어떻게 해서 이러저러한 현상이 일어나는지, 왜 일어나는지, 그리고 그 의미는 무엇인지에 대해 인간은 물음을 던지고 그것에 어떤 답을 찾거나 구성했을 것이다. 이 작업 역시 그의 생존과 직결되는 것이다. 가령 기상 현상과 생태계의 먹이 사슬을 이해 못했다면 그는 얼어죽거나 굶어 죽거나 다른 동물의 먹이가 되었을 것이다.

인간이 세상 만물을 이해하는 과정에서 그는 지식을 형성하게 된다. 초보적 단계에서 지식은 순수하게 체험적인 것이지만 그것은 곧 언어적인 것으로 변환되었을 것이다. 지식이 전제로 하는 가름이 그러하듯 지식 자체도 마찬가지로 언어의 사용에 의해 보다 명확하고 체계화된 모습을 갖추게 된다. 그리고 무엇보다도 우리가 지식을 언어화했을 때 비로소 다른 사람과 그 지식에 관한 의사 소통이 가능

해 진다. 이처럼 언어는 의사 전달을 가능케 하고 세상 만물의 가름과 이해의 결집으로서의 지식을 창출하는 역할을 동시에 수행한다.

인간이 세상 만물을 갈라 보는 작업은 오직 생존을 위한 동기에서 출발했지만 그것엔 이미 일정한 가치가 개입된다. 애초에 가름의 작업은 인간 중심적인 것이었다. 가령 먹을 수 있는 것과 없는 것, 뜨거운 것과 찬 것의 가름 등은 모두 인간을 중심으로 한 가름이다. 그렇다면 가름의 객관적 근거는 존재하지 않는가? 만일 객관적 근거가 존재하지 않는다면 가름은 어떻게 정당화될 수 있는가? 이 문제에 답하기 위해 다음의 예를 살펴보자.

코페르니쿠스(Copernicus) 이전 시대에는 '운동'이라는 용어는 지구 표면에 대한 상대적 위치 변화의 의미로 사용되었다. 톨레미(Ptolemy)를 비롯한 거의 모든 사람들은 태양과 별들이 지구를 중심으로 운동하고 있으며 지구는 정지해 있다고 생각했다. 그러나 코페르니쿠스는 지구 자체가 태양을 중심으로 운동하고 있으며 태양과 별들은 정지해 있다고 주장했다. 즉 그에게 있어서 '운동'은 태양과 별들에 대한 상대적 위치 변화의 의미를 갖는다. 이러한 관점에서 보았을 때 톨레미의 천동설과 코페르니쿠스의 지동설간의 논쟁은 운동의 의미, 혹은 운동하는 것과 운동하지 않는 것 사이의 가름을 둘러싼 논쟁이었다. 그렇다면 양자는 사실은 하나의 '운동' 개념에 관해 대립하고 있다기 보다는 서로 다른 두 운동 개념, 즉 운동$_P$와 운동$_C$를 각각 사용하고 있다고 보아야 하지 않을까?

이에 대해 코페르니쿠스와 그의 추종자들은 코페르니쿠스가 새로운 '운동' 개념을 자의적으로 도입했다기 보다는 오히려 기존의 운동 개념에 이미 내포되어 있던 의미를 보다 명확하게 분석하였다고 주장한다. 그들에 의하면 과거에 운동의 의미를 지구에 대한 상대적 위치 변화로 보았을 때 이는 사실은 주변의 대부분의 물질의 평균 위치에 대한 상대적 위치 변화를 의미했었다(Barker 1964, p. 13). 요컨대 과거에는 지구의 크기와 질량이 천체보다 훨씬 크다고 생각했으며 따라서 주변의 대부분의 물질의 질량의 평균적 위치에 대해 상대적으로 운동한다는 것은 결국 지구에 대해 상대적으로 운동하는 것을 의미했다. 그러나 16세기에 들어 천체가 지구보다 작지 않음이 알려지게 되었고 이를 바탕으로 코페르니쿠스는 주위에 있는 대부분의 물질의 평균 위치에 대한 지구의 상대적 위치가 변한다는 지동설을 주장한 것이다.

코페르니쿠스 이전의 '운동' 개념이 과연 주변의 대부분의 물질의 평균 위치에 대한 상대적 위치 변화의 의미로 사용되었는가? 만일 그렇다면 천동설과 지동

설의 대립은 주변의 대부분의 물질의 평균 위치에 대한 지구의 상대적 위치가 변하는가 아닌가의 문제, 즉 종합적 문제가 된다. 그렇지 않다면 천동설과 지동설의 대립은 결국 용어의 문제, 즉 분석적 문제가 되고 만다. 콰인(Quine 1951)의 주장처럼 이 경우 역시 문제의 본질이 분석, 혹은 종합의 어느 한 쪽이라고 단정적으로 말하기 어렵다. 그 이유는 '운동'의 개념에 대한 코페르니쿠스 이전의 이해가 코페르니쿠스의 주장처럼 명확하지 않기 때문이다. 16세기를 전후해서 운동의 개념은 의미론적 진화라 부를 수 있는 점진적 변화를 겪어 왔다고 보는 것이 옳을 것 같다. 어느 한 시점에서 분석적 문제가 종합적 문제로 바뀌는 극적인 형태 변화(Gestalt switch)의 뚜렷한 계기나 근거는 발견되지 않는다.

1.2 투사의 역사

자기 인식은 카시러(Cassirer 1944, p. 1)의 주장처럼 철학적 탐구의 최고 목표는 아닐지언정 적어도 하나의 중요한 목표로 여겨져 왔다. 우리의 인식은 우리의 시선이 그러한 것처럼 대상 지향적이다. 내가 보고 인식하는 것은 내게 보여진, 내게 인식된 대상이다. 내가 나 자신을 볼 수 있을까? 문제는 내 시야에 나 자신은 포함되지 않는다는 사실이다. 이를 극복할 수 있는 방법적 도구는 거울이다. 거울은 그 앞에 선 대상의 모습을 형상화한다. 우리는 거울에 투사(project)된 모습을 통해 자신을 확인한다. 인간의 자기 인식의 추구는 이러한 거울의 은유(metaphor)를 통해 전개되었다.

자기 인식의 추구를 중심으로 보았을 때 서양 철학사는 투사의 역사였다. 필자는 시대 구분을 기준으로 투사의 역사를 다음과 같이 네 단계로 나누어 보고자 한다. 첫째, 고대의 자기 인식은 자연, 혹은 존재에의 투사에 의해 형성된다. 자연의 존재론화(Nature Ontologized)가 이 시대에 자기 인식을 반영하는 거울의 구성 방법이었다. 가령 플라톤에 있어서 현상계 너머의 이데아계가 그러한 거울의 대표적 예이다. 둘째, 중세의 자기 인식은 신에의 투사에 의해 형성된다. 믿음의 신학화(Belief Theologized)가 이 시대에 자기 인식을 반영하는 거울의 구성 방법이었다. 셋째, 근대에 이르러 믿음은 신에로 정위되지 않고 인식론적 근거에로 정위된다. 요컨대 투사의 방향이 신으로부터 믿음의 주체, 즉 인간 의식에로 역전된 것이다. 믿음의 인식론화(Belief Epistemologized)가 이 시대에 자기 인식을 반영하는 거울의 구성 방법이 된다. 끝으로 콰인(Quine 1969a)은 인식론을 통하여 믿음의 체계에 궁극적 토대를 마련하려는 근대의 시도가 실패하였음을 선언하면서

인식론이 믿음의 유일한 가능한 체계로서의 자연과학에 포섭되어야 한다고 주장한다. 즉 그에게 있어서는 인식론의 자연화(Epistemology Naturalized)가 거울의 새로운 구성 방법이었다.

인식론의 자연화 시대에서 자기 의식을 찾기 위한 투사의 방향은 어데로 정위되는가? 그것은 자연도, 신도, 인간 의식도 아닌 테크놀로지이다. 인간은 이제 테크놀로지에서 자기 의식의 구체적 모델을 발견할 수 있다고 믿게 되었다. 가령 맥카시(McCarthy 1979)는 다음과 같이 말한다.

> 자동 온도조절장치와 같이 단순한 기계도 믿음을 갖고 있다고 할 수 있으며, 믿음을 갖는 것은 문제 해결의 기능을 갖는 대부분의 기계들의 특징인 것처럼 보인다.

우리의 문제의식과 관련된 테크놀로지의 총아는 컴퓨터이다. AI(Artificial Intelligence)를 탐구하는 많은 인지과학자들은 각각의 AI 프로그램이 저마다 지성(intelligence)을 발현하고 있다고 주장한다. 논의를 더 진행하기에 앞서 우리는 여기서 지성이 무엇인지를 살펴볼 필요가 있다.

2. 삶이란 무엇인가?

2.1 확산되는 삶의 형식

지성은 일부의 생명체가 갖는 특성이다. 따라서 우리가 지성을 살펴보기 위해서는 먼저 생명의 개념을 고려해야 한다. 처치랜드는 생명을 내부 질서를 유지하거나 증가시키기 위해 이미 얻어진 질서와 에너지 흐름을 이용하는 모든 반폐쇄적 물리 체계로 정의한다(Churchland 1988, p. 173). 이러한 정의 하에서는 생명을 지닌 체계와 그렇지 않은 체계 사이의 차이는 종류(kind)의 차이가 아니라 정도(degree)의 차이임이 드러난다. 가령 벌집도 인간과 마찬가지로 생명체로 볼 수 있다. 그리고 벌집으로부터 인간에 이르기까지의 생명체의 범주 내에서 그 어떠한 확고한 불연속적 단절도 발견되지 않는다.

이는 지성에 대해서도 그대로 적용된다. 지성의 일반적 특성은 변화하는 환경에 적절히 반응할 수 있는 능력이다. 그러나 이러한 특성은 감자로부터 고등동물에 이르기까지 나타나고 있으며 이들 사이에는 어떠한 불연속성도 존재하지 않는

다. 처치랜드는 지성을 이미 가지고 있는 정보를 사용할 뿐 아니라 체계를 통과하는 에너지 흐름을 체계가 가지고 있는 정보를 증가시키는 방향으로 이용하는 능력으로 정의한다(Churchland 1988, p. 173) 요컨대 학습을 통한 정보의 증가가 지성의 핵심적 요소라는 것이다.

지성에 대한 처치랜드의 정의는 생명에 대한 그의 정의와의 밀접한 관련하에 내려진 것이다. 즉 지성에 대한 그의 정의는 지성을 생명이 보다 복잡하게 환경과 연계될 경우에 나타나는 특성으로 보는 물리주의에 토대해 있다. 그는 다음과 같이 말한다.

> 지성을 지닌 존재라는 것도 결국 고도의 열역학적인 집약과 내적 질서와 외적 환경 사이의 특별히 밀접한 연결을 지닌 생명체에 불과한 것이다. (Churchland 1988, p 174)

지성을 지닌 존재의 범주에 속하는 존재자들 사이에서도 앞서 생명체의 경우와 마찬가지로 어떠한 극적인 불연속성도 발견되지 않는다. 지성은 인간에게만 고유한 것은 아니다. 지구상에서만도 수많은 다른 종(species)들이 지성에 대한 처치랜드의 정의를 만족시키고 있기 때문이다. 그리고 이 우주에서 생명과 지성이 취할 수 있는 삶의 형식의 종류는 무한하다고 보아야 할 것이다. 우리와 다른 삶의 형식들이 우리를 닮아야 하거나 반드시 우리에게 이해 가능해야 한다고 기대하는 것은 지나친 인간 중심주의적 발상일 것이다.

2.2 컴퓨터의 삶의 형식

일견했을 때 생명과 지성에 대한 처치랜드의 정의는 태양계의 운동에 대한 코페르니쿠스의 정의가 그러했던 것처럼 정의하려는 개념에 대한 인간 중심적, 상식적 태도를 벗어나 있는 것처럼 보인다. 벌집이 생명을 갖고 있다는 사실은 지구가 운동하고 있다는 사실만큼이나 우리의 직관이나 통념에 반하는 주장이기 때문이다. 처치랜드의 정의는 운동에 대한 코페르니쿠스의 정의가 그러했던 것처럼 정당화될 수 있을까?

일반적으로 컴퓨터는 처치랜드가 정의하는 지성을 소유하고 있는 것으로 여겨지고 있다. 컴퓨터의 핵심 기능이 정보처리이기 때문에 이는 어쩌면 당연한 이야기일 것이다. 문제는 컴퓨터가 소유하고 있는 지성이 인간의 지성과 동일한가 하는 점이다. 혹은 달리 말해 컴퓨터가 인간의 마음을 소유하고 있는가 하는 점이다. 이러한 문제의 제기는 그 자체만으로도 유익한 것인지 모른다. 우리는 인간의 지

성, 혹은 인간의 마음에 대해 확실히 알고 있지 못하기 때문이다. 따라서 위의 문제 제기는 컴퓨터에 우리 자신을 투사해 봄으로써 우리 자신에 대해 보다 많은 것을 알 수 있으리라는 일종의 작업 가설의 역할을 하고 있다.

물론 컴퓨터가 지성이나 마음을 가지고 있음을 인정하면서도 컴퓨터의 지성이나 마음이 인간의 그것과 같지 않다고 보는 인지과학자들도 있다. 자동 온도조절장치가 믿음을 가지고 있다고 보는 인지과학자는 자동 온도조절장치가 인간의 마음을 가지고 있다고 생각하지는 않을 것이다. 자동 온도조절장치의 마음이 인간의 마음과 다르다면 그것이 도대체 어떠한 마음인지 필자로선 알 수 없다. 그런데 잠시만 생각해 보면 자동 온도조절장치가 마음을 갖는지 아닌지, 그리고 그 마음이 어떠한 성격을 갖는지의 문제가 사실은 언어의 문제, 좀더 좁혀 말하자면 의미의 문제임이 드러난다. 관건은 자동 온도조절장치가 마음을 갖는다고 주장하는 인지과학자가 사용하고 있는 '마음'의 의미 영역, 그리고 자동 온도조절장치에 그 의미가 귀속될 수 있는 근거의 제시에 달려 있다. 제시된 의미의 테두리와 귀속의 근거가 과연 정당화될 수 있는지, 그리고 어떻게 정당화될 수 있는지의 문제는 정밀한 논의를 필요로 한다. 필자의 소박한 우려는 만일 자동 온도조절장치에 마음을 귀속시킬 수 있다면 그것이 달려 있는 보일러나 공부방에도 마음을 부여해서는 안될 이유가 있느냐는 것이다. 이는 현대판 애니미즘이며 철학적으로 별 가치가 없는 주장이다.

이 글에서 필자는 하나의 집합적 생명체의 지성과 마음을 그 생명체의 삶의 형식이라고 정의하기로 한다. 그렇다면 이제 문제는 컴퓨터의 지성과 마음, 즉 컴퓨터의 삶의 형식이 인간의 삶의 형식과 동일한지를 살펴보는 것으로 좁혀진다. 그런데 컴퓨터의 삶의 형식이란 대체 무엇인가? 그리고 그것이 인간의 삶의 형식과 동일하다, 아니다는 구체적으로 무엇을 말하는 것인가?

3. 의미 해체의 두 양상

3.1 콰인의 면도날

비트겐슈타인에 있어서 삶의 형식의 핵심은 자연언어의 의미 이해와 사용에 있다. 컴퓨터의 삶의 형식이 인간의 삶의 형식과 동일하다고 보는 인지과학자들의 주장도 마찬가지로 결국은 컴퓨터가 자연언어의 의미를 이해하고 사용할 수 있다는 것으로 요약된다. 자연언어의 의미를 이해하고 사용함은 무엇을 말하는 것일까? 언어의 의미 이해와 사용은 전통적으로 다음과 같은 세 단계의 학습에 의해

이루어지는 것으로 여겨져 왔다. 첫째로 구문론(Syntax)의 단계이다. 구문론은 언어적 표현들의 어형의 역할 및 그들 각각의 형식적, 기능적 측면, 그리고 상호간의 내적 관계를 다룬다.

그런데 구문론에서 단어, 문장 등의 제반 언어적 단위와 요소들은 아직 의미 부여가 되지 않은 상태이므로 진리, 내용과 같은 의미론적 개념들을 갖고 있지 않은 형식적 기호에 불과하다. 이 형식적 기호에, 그것에 연관되는 사물, 사건 등을 개입시켜 보자. 이제 기호들은 기호 이외의 것들에 대하여 지칭, 의미, 함의, 대응 등의 관계를 맺게 되며 따라서 의미, 진리, 내용 등과 같은 의미론적 개념들이 정립된다. 확장된 이 단계는 의미론(Semantics)의 단계이다.

다시 의미론적 단계의 언어 이해에 화자(speaker)와 청자(hearer), 행위, 의도, 효과 및 언어가 쓰여지는 상황들을 개입시켜 보자. 언어를 발언하는 것은 언어의 쓰임에 수반되는 일련의 행위와 연관되고 그것은 언어 행위적 문맥 하에서 이해되므로 여기서는 언어와 인간의 행위를 포함하는 삶의 상황과의 관계가 중요시된다. 확장된 이 단계는 화용론(Pragmatics)의 단계이다.

위의 세 단계 중 철학에서 가장 많이 논의된 분야는 의미론이다. 필자는 현대의 철학적 의미론에 결정적 영향을 미친 사람으로 특히 프레게와 브렌타노에 주목한다. 프레게(Frege 1892; 1918)는 의미를 문장이 표현하고 있는 사고나 명제로 보았다. 브렌타노(Brentano 1960)는 의미의 영역에서 지향성을 필요 불가결한 요소로 간주하여 의미론을 곧 지향성의 과학으로 규정하였다.

구문론-의미론-화용론의 구분은 타당한가? 구분의 와해는 프레게/브렌타노의 철학적 의미론에 대한 콰인과 비트겐슈타인의 해체 작업에서 비롯된다. 콰인은 자신이 지지하는 경험론과 전체론의 전제로부터 의미의 불확정성이 도출됨을 원초적 번역 불확정성론을 통하여 논증한다.[1] 이 논증을 통해 콰인은 프레게에 대해 다음과 같이 답변한다; 원초적 번역의 과정에서 문장의 의미로서의 명제는 보존되지 않는다. 보존되어야 할 확정적 명제가 존재하지 않기 때문이다. 그 이유는 첫째, 의미는 개개의 문장이 아니라 전체 문장의 맥락에 귀속되며, 둘째, 그 경우에도 의미는 오직 번역의 틀에 연관해서만 언급될 수 있기 때문이다. 한편 콰인은 브렌타노에 대해 다음과 같이 답변한다; 지향성은 문장의 의미의 토대가 되지 못한다. 문장의 확정적 의미로서의 명제가 부정되는 과정에서 지향성의 소재가 되는 명제적

1) 이에 관한 보다 자세한 논의를 위해서는 다음을 참조할 것. 이승종, "의미와 해석에 관한 콰인/데이빗슨 논쟁," 『철학』, 1993년, 39집.

태도(propositional attitudes)의 확정성 역시 부정되기 때문이다.

콰인에 있어서 의미론과 지향성의 제거는 또한 인식론의 자연화를 함축한다. 그는 다음과 같이 말한다.

> 인식론이 이제 의미론이 된다는 사실은 비엔나 서클의 입장에서는 놀랄 바가 못된다. 인식론이 언제나처럼 증거를 중심으로 전개되고 의미는 언제나처럼 검증을 중심으로 전개되기 때문이다. 증거는 검증이다. 비엔나 서클의 입장에서 더 충격적인 사실은 관찰 문장을 넘어서서는 일반적으로 의미가 개별 문장에 명확히 적용되지 않는다는 점, 그리고 인식론이 언어학 뿐 아니라 심리학에 합병된다는 점이다. (Quine 1969a, pp.89-90)

콰인에 있어서 의미론을 대체하는 것은 그것이 자연화된 뒤의 양상, 즉 언어학, 심리학 등의 과학이다.

한편 프레게/브렌타노의 의미론에 대한 해체는 비트겐슈타인(Wittgenstein 1967)에 있어서는 의미론이 화용론에 환원되는 것으로 귀착된다. 의미가 쓰임에 의해서 드러난다는 그의 주장은 의미가 프레게/브렌타노가 생각한 것처럼 인간에서 독립된 추상적, 혹은 지향적 내용이 아니라 언어를 사용하는 인간의 행위와의 밀접한 연관하에서 이해되어야 한다는 것이다.[2]

3.2 뫼비우스의 띠

콰인과 비트겐슈타인이 전개하는 의미론의 해체 작업은 그러나 의미에 대한 새로운 이론의 구성으로 받아들여져서는 안된다. 이 점은 이 글의 주제와 밀접히 연관되어 있기 때문에 좀더 부연해 논의해 보고자 한다.[3] 의미에 대한 이론, 즉 의미론은 종종 의미를 표면에 직접 명백히 드러나지 않는, 감추어진 어떤 것으로 전제하곤 한다. 이처럼 감추어진 의미를 찾아내고 그것을 규명해야 할 필요성이 의미에 관한 '이론'의 존립 근거가 된다. 이 경우 '이론'의 개념은 과학에서의 이론의 개념과 같은 의미로 사용되고 있지만 전통적 과학의 범위를 넘어서는 영역에

---

2) 이에 관한 보다 자세한 논의를 위해서는 다음을 참조할 것. 이승종, "언어철학의 두 양상," 『철학과 현실』, 1993년 겨울호.

3) 이에 관한 보다 자세한 논의를 위해서는 필자와 뉴턴 가버(Newton Garver)가 같이 쓴 다음의 책을 참조할 것. Newton Garver and Seung-Chong Lee, *Derrida and Wittgenstein*, Philadelphia: Temple University Press, 1994, 7장.

적용되고 있다.

과학에서 이론의 의미는 무엇인가? 물리학자 켐벨(Campbell, 1952)은 이론의 개념에 대해 다음과 같은 정의를 제시하고 있다: 이론은 법칙을 설명한다. 따라서 이론은 법칙과는 구별되어야 한다. 법칙은 현상들 간의 관찰 가능한 규칙성을 기술한다. 일단 규칙성이 우리에게 알려지면 우리는 개별적 사건이나 현상을 설명하기 위해 법칙을 사용할 수 있다. 가령 많은 수의 도시 어린이가 충치를 가지고 있다는 현상을 우리는 도시 어린이들이 단 것을 과다 섭취하고 있다는 사실, 그리고 당분과 충치의 연관 관계에 관한 법칙에 의거하여 설명할 것이다.

법칙과 달리 이론은 경험적 규칙성의 발견과 직접 연관되지 않는다. 이론은 오히려 관찰 불가능한 존재를 상정함으로써 법칙을 설명한다. 가령 뉴턴은 호이겐스에 의해 발견된 원심력의 법칙, 그리고 데카르트와 갈릴레오에 의해 발견된 관성의 법칙이 참이라는 가정 하에서, 행성의 운동에 관한 케플러의 세 법칙을 설명하기 위해 중력의 존재를 상정하였다. 이로써 행성의 관찰 가능한 규칙적 운동은 중력이라는 관찰 불가능한 가정적 존재에 의해 설명된다. 중력은 이처럼 오로지 법칙의 설명을 위해 요청된 개념일 뿐이다. 중력의 존재 그 자체에 대해서는 정작 뉴턴 자신도 확신하지 못한 실정이었다. 그의 이러한 심경은 그가 벤틀리(Richard Bently)에 보낸 다음의 편지에 잘 나타나 있다.

> 한 물체가 다른 어떤 물체의 개입도 받지 않고 진공을 통해서 다른 물체에 작용할 수 있으며 그에 의하여, 그리고 그것을 통하여 그 물체들의 작용과 힘이 서로간에 전달될 수 있다는 생각은 내게는 너무도 황당한 것이어서 철학적 문제에 충분한 사고 기능을 가진 사람이면 누구도 그와 같은 이론에 빠져들 수 없으리라고 나는 믿는다.[4]

중력의 완전한 설명은 아인슈타인의 일반 상대성 이론이 등장한 뒤에야 가능하게 된다.

의미론이 과학적 이론이라면 그것은 케플러의 법칙에 대한 뉴턴의 설명과 같은 어떤 것을 시도해야 할 것이다. 그런데 문제는 의미론이 문제삼는 의미가 자연과학의 경우에서처럼 경험적 규칙성을 갖느냐 하는데 있다. 브렌타노와 콰인의 대립은 바로 이 문제를 둘러싼 것이었다. 콰인은 브렌타노와 후설이 제시하는 의미

4) *Proceedings of the Royal Society of London*, vol. 54, 1893, p. 381.

의 지향성을 "무근거한," 그리고 "과학적 통찰에 아무런 보탬도 되지 않는" 것으로 거부한다(Quine 1960, p. 221). 그러나 콰인은 그렇다고 의미론이 의미에 관한 법칙을 설명할 수 있을 것으로도 보지 않았다. 케플러의 법칙에 대한 뉴턴의 설명은 결코 케플러의 법칙을 해체하지 않는다. 반면 콰인에 의하면 의미론은 오히려 그것이 설명하고자 하는 의미에 관한 법칙의 가능성을 해체하는 결과를 초래한다는 것이다. 그 이유가 의미의 불확정성, 혹은 확정적 의미의 부재에 있음은 앞서 살펴본 바 있다.

한편 의미론의 이념에 대한 비트겐슈타인의 불만은 의미론이 의미를 설명하려 한다는데 있다. 그에 의하면 의미란 결코 설명의 대상이 아니다. 의미를 창출하는 언어의 사용이 인간의 원초적 삶의 현상이기 때문이다. 원초적 삶의 현상은 오직 기술될 수 있을 뿐이다.

구문론-의미론-화용론의 구분에 대한 또다른 해체의 시도는 최근 일부의 기능주의 심리철학자들과 AI 연구가들에 의해 이루어지고 있다. 비트겐슈타인과는 반대로 그들은 의미론이 화용론이 아니라 구문론에로 환원된다고 본다. 이러한 주장의 근거와 그 타당성을 검토하기 전에 우리는 그것이 궁극적으로 의도하는 바를 살펴볼 필요가 있다. 앞서 우리는 컴퓨터의 삶의 형식이 인간의 삶의 형식과 동일하다고 보는 인지과학자들의 주장이 타당한지의 관건이 컴퓨터가 자연언어를 이해할 수 있는지에 달려 있음을 보았다. 만일 구문론-의미론의 구분이 와해된다면 자연언어의 의미에 대한 인간의 이해는 컴퓨터가 수행하는 구문론적 조작으로 환원가능하게 된다. 이는 컴퓨터의 삶의 형식과 인간의 삶의 형식의 동일성 주장의 입지를 강화하는 결정적 논증이라 할 수 있다. 요컨대 구문론-의미론이 뫼비우스의 띠와 같다면 컴퓨터와 인간도 뫼비우스의 띠와 같다는 것이다.

4. 자연언어와 인공지능

4.1 기계 밖의 인간: 튜링

지금까지의 논의를 정리해 보자. 컴퓨터와 인간이 동일한 삶의 형식을 갖는지의 여부는 컴퓨터가 자연언어의 의미를 이해하고 사용할 수 있느냐의 문제로 좁혀졌다. 이는 다시 의미론이 구문론으로 환원될 수 있느냐의 문제로 좁혀졌다. 이제 문제를 다시 소급하여 컴퓨터가 인간과 동일한 삶의 형식을 갖는지의 문제, 컴퓨터가 자연언어의 의미를 이해하고 사용할 수 있는지의 문제, 그리고 의미론이 구

문론으로 환원될 수 있는지의 문제의 순으로 논의를 검토해 보자.

컴퓨터의 삶의 형식에 관한 논의는 튜링에 의해서 최초로 정립되었다. 그는 컴퓨터가 지성과 마음을 갖는지의 여부를 판정할 수 있는 다음과 같은 테스트를 구상하였다(Turing 1950; Cf. 이초식 1993, p. 51). 컴퓨터, 테스트의 판정관, 그리고 다른 한 사람이 각각 독방에 격리 수용된다. 일정한 주어진 시간 동안 판정관은 텔레타이프를 통해 옆방의 사람과 컴퓨터와 각각 교신한다. 그는 어느 방에 컴퓨터가 있고 어느 방에 사람이 있는지 알지 못한다. 그는 주어진 시간 내에 오직 텔레타이프를 통한 교신에 의존해서만 컴퓨터와 사람을 판별하는 수 밖에 없다. 튜링에 의하면 이러한 상황하에서 컴퓨터가 지성과 마음을 가지고 있는지에 대한 판정관의 판정 기준은 다음과 같을 수 밖에 없다.

> 컴퓨터는 판정관이 컴퓨터와 사람을 판별할 수 없을 경우에, 그리고 오직 그 경우에만 지성을 갖는다.

혹은

> 컴퓨터는 판정관이 보기에 마음을 갖고 있는 것처럼 행동할 때, 그리고 오직 그 때에만 마음을 갖는다.

튜링이 제시하는 이러한 기준은 과연 컴퓨터와 인간이 동일한 삶의 형식을 갖는지에 대한 필요 충분한 객관적 기준인가? 우선 그것이 필요 충분한 기준인지를 살펴보자. 튜링의 기준은 튜링 테스트가 전제로 하고 있는 상황, 즉 판정관과 사람, 그리고 컴퓨터가 서로 직접 볼 수 없이 격리된 채 주어진 시간 내에서 텔레타이프에 의한 교신만이 허용되는 상황에서만 필요 충분할 수 있을지 모른다. 테스트에 주어진 시간을 짧게 조절한다면 와이젠바움(Weizenbaum 1976)의 엘리자(ELIZA)나 블럭(Block 1978)의 버블즈 아주머니(Aunt Bubbles) 등의 프로그램은 튜링 테스트를 통과할지 모른다. 그러나 이론상 튜링 테스트의 시간은 무한으로 늘릴 수 있으며 이 경우에 인간과 이들 프로그램의 차이는 결국 식별될 것이다. 그런데 문제는 우리가 아직 인간에 버금가는 인지 능력의 컴퓨터를 갖고 있지 않다는 사실이 아니라 우리가 인간의 인지 능력의 범위에 대해서 확실히 알지 못한다는데 있다.

짧게 한정된 시간 하에 엘리자나 버블즈 아주머니가 튜링 테스트를 통과했다

고 가정해 보자. 이들 프로그램은 과연 인간의 지성을 소유하고 있는가? 가령 심리 상담자를 본뜬 엘리자의 대화는 사실은 환자가 대화 중에 사용한 문장의 간단한 변형, 그리고 환자가 사용한 핵심적 개념으로 조립한 일군의 표준 질문 형식 등으로 구성한 것이다. 엘리자는 이들 개념이 무엇을 의미하는지 전혀 이해하지 못한다. 단지 엘리자는 프로그램 입안자가 구성한 요령에 따라 기계적으로, 그러나 성공적으로 환자와 표준적 형태의 대화를 이끌어 갈 뿐이다. 즉 엘리자는 구문론적 정보처리 프로그램 그 이상도 이하도 아닌 것이다.

버블즈 아주머니의 경우도 사정은 마찬가지이다. 버블즈 아주머니의 제작자인 블럭은 자신의 프로그램이 사실 알고 보면 쥬크박스(jukebox) 수준의 지성에 머물러 있다고 고백한다(Block 1990, p. 252). 이 프로그램이 장시간, 혹은 무한정 진행되는 튜링 테스트에서 탈락할 것임은 말할 나위 없다.

튜링 테스트의 기준은 그렇다면 객관적인가? 문제는 튜링 테스트의 판정관이 인간이라는데서 비롯된다. 튜링 테스트의 관건은 테스트를 치르는 컴퓨터가 테스트의 판정관인 인간으로 하여금 컴퓨터가 정말로 인간의 지성을 갖고 있다고 판단하게 하느냐 못하느냐에 달려 있다. 그러나 이 경우 판정관이 내리는 판단의 객관성은 판단 행위의 객관성과는 구별되어야 한다. 논리적 관점에서 보았을 때 "나는 이 컴퓨터가 정말로 인간의 지성을 갖고 있다고 본다"는 문장은 "이 컴퓨터가 정말로 인간의 지성을 갖고 있다"는 문장을 함축하지 않기 때문이다. 전자를 정당화하는 사실이 곧 후자를 정당화하고 있다고 보는 것은 명백한 논리적 오류이다.

아울러 누가 튜링 테스트의 판정관인가에 따라 판정이 크게 달라질 수 있다. 가령 자동 온도조절장치가 믿음을 갖고 있다고 믿는 맥카시가 판정관일 때와, 버블즈 아주머니가 쥬크박스 수준의 프로그램이라고 혹평하는 블럭이 판정관일 때, 그리고 서양 문명을 전혀 이해 못하는 원시 종족이 판정관일 때 판정의 결과와 근거는 각각 판이하게 다를 것이다.

4.2 기계 안의 인간: 썰

튜링 테스트의 철학적 배경은 행동주의이다. 지성이나 마음의 유무를 외적으로 드러난 행동에 의해 판정하고 있기 때문이다. 과연 심적 상태나 심적 현상에 대한 고려를 배제한 상태에서 지성이나 마음을 정의하는 작업이 타당하다고 볼 수 있을까? 행동주의자들은 심적 상태나 심적 현상도 결국은 행동으로 환원, 설명될 수 밖에 없다고 주장할 것이다. 그러나 환원의 대상이 환원의 과정에서 조금이라도

누락되거나 왜곡된다면 그 환원은 성공했다고 볼 수 없다.

컴퓨터와 인간이 동일한 삶의 형식을 갖는지의 여부를 판정하는데 있어서 행동주의적 환원 방법은 최선의 방법인가? 보다 나은 방법이 있다면 그것은 우리가 비교의 대상인 컴퓨터가 직접 되어 보는 것이다. 그러나 어떻게 우리가 컴퓨터가 될 수 있는가? 그리고 그 상황에서 우리와 컴퓨터가 동일한 하나의 삶의 형식 하에 있다는 기준은 무엇인가? 썰(Searle 1980; Cf. 이초식 1993, pp.52-55)은 우리가 컴퓨터가 될 수 있는 방안을 하나의 사유 실험을 통해 제시한다. '중국어 방'으로 통칭되는 이 상황에서 컴퓨터가 된 우리가 자연언어의 의미를 이해하고 사용하는지의 여부를 컴퓨터가 우리와 삶의 형식을 공유하는지에 관한 테스트로 삼는다. 이제 그의 사유 실험을 살펴보자.

튜링 테스트에 동원된 컴퓨터가 제거되고 대신 내가 그 방에 들어간다. 이제부터는 내가 컴퓨터의 역할을 하게 되는 것이다. 나의 모국어는 한국어이고 나는 중국어를 전혀 알지 못한다. 내 옆방의 한 사람, 그리고 다른 방의 판정관은 모두 중국어가 모국어인 중국인이라고 가정해 보자. 지금 내 방에는 중국어 단어들이 널려 있고 이에 관한 한국어로 된 규정집이 주어져 있다. 그 규정집의 규칙들은 내게 주어진 중국어 단어들을 구문론에 의해 형식적으로 처리하도록 하는 지침이다. 가령 이러이러한 단어를 저러저러한 단어 다음에 놓으라는 식이다. 이제 내 방으로 중국어 단어들이 들어오면 나는 규정집의 규칙에 따라 단어들을 연결하여 방 밖으로 내보낸다.

위의 경우에 내 방으로 들어오는 중국어는 튜링 테스트의 판정관에 의해 입력되는 질문이고 방 밖으로 나가는 중국어는 출력되는 대답이다. 그리고 방 안의 나는 정보를 구문론적으로 처리하는 컴퓨터인 셈이다. 내가 가지고 있는 규정집이 아주 효율적인 것이고 이에 의존하는 나의 정보처리 능력이 빠르고 정확해서 판정관이 튜링 테스트에서 나를 합격시켰다고 해보자. 나는 중국어를 과연 이해하고 사용하고 있는가? 중국어에 대한 나의 이해는 튜링 테스트 이전이나 이후나 똑같다. 즉 나는 중국어에 대해 여전히 아무것도 알지 못하고 있는 상태인 것이다. 그런데 나는 여기서 컴퓨터의 역할을 하고 있으므로 내가 중국어를 이해 못한다는 말은 컴퓨터가 자연언어를 이해하지 못한다는 말과 같다. 자연언어의 이해가 인간의 삶의 형식의 관건이므로 이로부터 컴퓨터가 인간의 삶의 형식을 갖지 않는다는 결론이 도출된다.

썰의 '중국어 방'은 행동주의에 기초한 튜링 테스트에 대한 강력한 반증이다.

그렇다면 행동주의에 반대하는 썰 자신은 어떠한 입장에 서 있는가? 이를 위해 우리는 썰이 제시한 다음과 같은 논증을 살펴볼 필요가 있다(Searle 1984, pp.39-41). 그의 논증은 다음과 같은 네 개의 전제를 갖는다.

1. 두뇌가 마음을 야기한다.
2. 구문론은 의미론의 충분조건이 아니다.
3. 컴퓨터 프로그램은 전적으로 형식적, 혹은 구문론적 구조에 의해 정의된다.
4. 마음은 심리적 내용을 갖는다. 구체적으로 말해 마음은 의미론적 내용을 갖는다.

이로부터 썰은 다음과 같은 네 개의 결론을 연역한다.

결론 1. 어떠한 컴퓨터 프로그램도 그 자체로서는 어느 체계로 하여금 마음을 갖도록 하는 충분조건이 못된다. 요컨대, 프로그램은 마음이 아니며 프로그램만으로는 마음을 갖게 하기 위한 충분조건이 못된다. (전제 2, 3, 4로부터의 연역)

결론 2. 두뇌의 기능이 마음을 야기하는 방식은 단지 컴퓨터 프로그램을 작동시켜서 만으로는 성립될 수 없다. (전제 1과 결론 1로부터의 연역)

결론 3. 마음을 야기하는 다른 어떤 것은 최소한 두뇌의 능력과 동등한 인과적 힘을 가져야 할 것이다. (전제 1로부터의 연역)

결론 4. 컴퓨터 프로그램의 운용만으로는, 우리가 만들 수 있는 어떤 인조물이 인간의 심리적 상태와 동등한 심리적 상태를 갖도록 하는 데 충분하지 않다. 오히려 그 인조물은 인간의 두뇌와 동등한 능력을 가져야만 할 것이다. (결론 1, 3으로부터의 연역)

논증의 전제로부터 우리는 썰이 행동주의 뿐 아니라 기능주의와 강한 AI(strong AI), 제거적 유물론(eliminative materialism)에도 모두 반대하고 있음을 알 수 있다. 좀더 구체적으로 말하자면 전제 1, 2, 3은 기능주의에 대한 반대를, 전제 2, 3, 4는 강한 AI에 대한 반대를, 그리고 전제 4는 제거적 유물론에 대한 반대를 표명하고 있다. 기능주의는 컴퓨터의 내적 체제가 우리의 심리 상태

와 기능적으로 동형적이라면 컴퓨터 역시 마음을 갖는다는 입장을 말한다. 강한 AI는 컴퓨터가 인간의 마음을 가질 수 있다는 입장을 말한다. 제거적 유물론은 마음이 소유하는 것으로 여겨지는 심리적 내용의 개념이 설은 완전히 잘못된 개념이기 때문에 제거되어야 한다는 입장을 말한다. 썰의 전제들은 이러한 입장들과 첨예하게 대립하고 있다.

기능주의와 강한 AI에 대한 썰의 비판의 근거는 앞서 살펴본 '중국어 방'에서 찾아진다. 문제의 실험에서 컴퓨터와 나는 기능적으로는 어느 정도 동형성을 유지하고 있지만, 즉 양자가 다 유능한 정보처리 능력을 보여주고 있지만, 그럼에도 불구하고 컴퓨터와 나 사이의 삶의 형식의 동일성은 확보되지 않는다. 컴퓨터는 마음을 갖지 않기 때문이다. 제거적 유물론에 대한 썰의 반대는 제거적 유물론이 제거하려는 민간 심리학(folk psychology)의 핵심 개념들에 대한 다음과 같은 수용적 태도에서 잘 나타나 있다.

> "이해"는 심적(지향적) 상태의 소유를 함축한다 … 믿음, 바램, 지향 등은 지향적 상태이다. (Searle 1980, p. 208)

썰의 입장은 하만(Harman 1989, p. 838)이 지적하고 있듯이 유형 동일론(type identity theory)에 가장 가깝다. 유형 동일론은 인간의 심적 상태의 유형이 그것을 창출하는 물리적 상태의 유형과 동일하고 또 그 물리적 상태에로 환원된다는 입장이다. 많은 인지과학자들은 유형 동일론이 그릇된 이론이라고 보고 있으며 그 비판의 논증이나 사례도 이미 널리 알려져 있다. 이러한 비판에 대한 응수는 썰에게 남겨진 과제이다.

기능주의, 강한 AI, 제거적 유물론에 대한 썰의 비판은 타당한가? 썰이 제거적 유물론에 반대하는 이유가 무엇인지는 확실하지 않다. 그는 단지 민간 심리학의 개념들을 다른 설명 없이 그대로 수용하고 있을 뿐이다. 따라서 엄밀한 의미에서 썰은 제거적 유물론을 비판하지 못했다. 문제는 기능주의와 강한 AI에 대한 썰의 비판이 타당한가 이다.

4.3 기계 인간: 래퍼포트

래퍼포트는 기능주의와 강한 AI를 썰의 비판으로부터 옹호하기 위해 썰의 '중국어 방'에 대항하는 다음과 같은 '한국어 방'을 구성한다(Rapaport 1988a; Cf. 이초식 1993, pp.55-57). 튜링 테스트에서 썰의 중국어 방 안에 있는 나는

중국어 뿐 아니라 영어도 알지 못한다. 그러나 나는 셰익스피어 작품들을 한국어 번역판으로 읽고 이에 관한 논문들을 역시 한국어로 써낸다. 그리고 이 논문들은 영역되어 세계적인 인정을 받는다.

방 안의 내가 영어로 된 셰익스피어 작품을 이해하지 못하면서도 셰익스피어의 한국어 번역을 이해하고 있음은 사실이다. 그것이 가능한 까닭은 셰익스피어의 작품이 영어로도, 한국어로도, 혹은 그 외의 어떠한 자연언어로도 표현될 수 있다는데 있다. 따라서 엄밀한 의미에서 나는 단지 셰익스피어의 한국어 번역을 이해하고 있는 것이 아니라 '셰익스피어'를 이해하고 있는 셈이다.

래퍼포트는 셰익스피어의 작품이 여러 종류의 자연언어로 표현되고 이해될 수 있다는 사실에서 이해가 인간 뿐 아니라 컴퓨터나 그 밖의 다른 여러 인공물에 의해 구현될 수 있다는 사실을 추론한다. 중국어를 유능하게 구사하는 사람이 중국어를 이해하고 있는 것처럼 중국어를 유능하게 처리하고 있는 컴퓨터도 중국어를 이해하는 것으로 볼 수 있다는 것이다. 따라서 중국어 방의 나도 설령 나 자신이 중국어를 이해 못한다고 믿는다 할지라도 사실은 중국어를 이해하고 있다고 보아야 한다는 것이다. 논리적 관점에서 보았을 때 "나는 중국어를 이해 못한다고 믿는다"는 문장이 "나는 중국어를 이해 못한다"는 문장을 함축하는 것은 아니다. 전자를 정당화하는 사실이 곧 후자를 정당화하고 있다고 보는 것은 논리적 오류이다.

'한국어 방'을 통해 드러나는 래퍼포트의 입장은 기능주의와 강한 AI이다. 그는 다음과 같이 말한다.

> 나는 [이해에 관한] 기능적 개념, 즉 계산적 개념의 가능성을 논증하고 있는 것이다. 이 개념은 인간의 두뇌 뿐 아니라 컴퓨터에서도 운용될 수 있다고 본다. 따라서 두뇌 뿐 아니라 컴퓨터도 이해할 수 있음을 논증하려 한다. (Rapaport 1985, p. 343; cf. 1988a, p. 91)

래퍼포트는 자연언어의 이해와 같은 심적 과정을 인간의 두뇌나 컴퓨터에서 운용될 수 있는 추상으로 보아야 한다고 주장한다. 그에 의하면 이러한 추상의 요체는 컴퓨터가 수행하는 것과 같은 계산적, 구문론적 조작이다. 이러한 관점에서 보았을 때 컴퓨터는 자연언어를 이해하며 따라서 마음과 지성을 소유하는 것으로 간주해야 한다.

래퍼포트의 '한국어 방' 논변은 썰의 '중국어 방' 논변에 대한 타당한 반증인

가? 우선 우리는 래퍼포트와 썰이 이해의 개념에 있어서 상이한 태도를 취하고 있음을 유의할 필요가 있다. 썰에 있어서 이해는 인간의 이해인 반면에 래퍼포트의 이해는 반드시 인간의 이해를 의미하는 것은 아니다. 그는 다음과 같이 말한다.

> 나는 인간의 심성(mentality)이 무엇이며 실제로 그것이 어떻게 운용되는지에 관한 계산 이론보다는 운용의 매개와는 독립적으로 어떻게 심성이 가능한지에 관한 계산 이론에 더 관심이 있다. (Rapaport 1988b, p. 595; cf. 1986, p. 14)

래퍼포트에 의하면 컴퓨터는 자연언어를 이해한다. 그 이유는 그가 자연언어의 이해를 컴퓨터가 수행하는 계산적, 구문론적 조작과 동일시하고 있기 때문이다. 따라서 컴퓨터가 자연언어를 이해한다는 래퍼포트의 주장과 컴퓨터가 자연언어를 이해하지 못한다는 썰의 주장 사이의 대립은 지구가 운동한다는 코페르니쿠스와 지구가 운동하지 않는다는 톨레미 사이의 대립의 양상과 같다. 톨레미와 코페르니쿠스의 대립이 하나의 '운동' 개념에 관한 대립이 아니고 사실은 서로 다른 두 운동 개념, 즉 운동$_P$와 운동$_C$를 각각 사용하는데서 비롯된 것처럼 래퍼포트와 썰의 대립도 하나의 '이해' 개념에 관한 대립이 아니고 서로 다른 두 이해 개념, 즉 이해$_R$과 이해$_S$를 각각 사용하는데서 비롯된 것으로 보아야 할 것이다. 여기서 이해$_R$은 래퍼포트가 말하는 이해이며 이해$_S$는 썰이 말하는 이해를 뜻한다.

래퍼포트의 논증이 썰의 주장에 대한 효과적인 반증이 되기 위해서는 톨레미에 대해 코페르니쿠스가 그러했던 것처럼 이해$_R$이 이해$_S$에 이미 내포되어 있는 의미를 보다 명확하게 하고 있음을 보이거나, 이해$_R$이 이해$_S$에 대해 갖는 그 밖의 다른 이론적, 혹은 경험적 이점을 제시해야 할 것이다. 그러나 래퍼포트는 위의 인용문에서 보듯이 이해$_R$과 이해$_S$를 단지 병렬적으로만 취급하고 있을 뿐이다. 필자는 이러한 방법이 문제의 해결에 별 도움이 되지못함을 맥카시의 경우를 예를 들어 이미 살펴보았다. 코페르니쿠스의 경우와는 달리 래퍼포트의 이해$_R$은 이해$_S$보다 오히려 허약한 개념이다. 이해$_R$의 내용과 양상이 구문론적, 계산적 기능에 머물러 있는데 반해 이해$_S$, 즉 인간의 이해의 내용과 양상은 그보다 훨씬 더 복잡하고 다차원적일 수 있기 때문이다.

래퍼포트의 '한국어 방'의 또 하나의 문제는 그것이 썰의 '중국어 방'과는 달리 컴퓨터의 경우에 구체적으로 어떻게 적용될 수 있는지가 애매하다는 것이다. '한국어 방'에 입력되는 자료는 무엇이고 그 방에서 '출력'되는 것은 무엇인가?

그리고 한국어 방에는 무엇이 구비되어 있으며 어떠한 작업이 그 안에서 일어나고 있는가? 번역과 자연언어의 이해는 정확히 어느 단계에서 이루어지는가? 이와 관련하여 자렌(Jahren 1990, p. 325)은 래퍼포트가 선결 문제 요구의 오류에 빠져 있다고 비판한다. 그는 래퍼포트가 자연언어의 이해와 같은 심적 과정을 인간의 두뇌나 컴퓨터에서 운용될 수 있는 추상으로 보는 점이 문제라고 주장한다. 왜냐하면 그러한 추상이 인간의 두뇌나 컴퓨터에서 운용되었을 때 양자에 모두 마음이 형성된다는 사실은 그 추상적 과정의 컴퓨터에서의 운용이 같은 과정의 인간에서의 운용과 기능적으로 동등하다는 사실을 전제로 하고 있기 때문이다. 그리고 이러한 전제는 바로 래퍼포트가 정당화하려는 기능주의와 강한 AI의 주장이다.

썰의 '중국어 방'의 경우에는 주어진 정보 처리의 업무가 기계적인 구문론적 조작으로 제한되어 있었으므로 컴퓨터와 인간의 기능적 동형성은 어느 정도 전제될 수 있었다. '중국어 방'에서 썰은 이러한 기능적 동형성이 전제되어도 컴퓨터에 인간과 동등한 마음이 형성되지는 않음을 보였다. 반면 래퍼포트는 컴퓨터가 (인간과는 다른 형태의) 마음을 갖고 있음을 주장하면서도 기능주의와 강한 AI의 이념인 컴퓨터와 인간의 기능적 동형성은 무비판적으로 수용하고 있다. 그러나 래퍼포트의 '한국어 방'은 썰의 경우와는 달리 주어진 정보 처리의 범위가 제한되어 있지 않다. '한국어 방'은 셰익스피어 작품의 이해, 그리고 그에 대한 학문적 저술 작업도 수행해야 한다. 이는 분명 '중국어 방'에서의 구문론적 조작의 수준을 넘어서는 복잡하고 어려운 과제이다. 따라서 '한국어 방'의 경우에 컴퓨터와 인간의 기능적 동형성은 유지되기 어렵다.

래퍼포트의 '한국어 방'의 결정적인 문제점은 무엇보다도 그것이 래퍼포트 자신의 가정을 위반하는 지경에 이를 수 있다는 점이다. 잠시 '한국어 방'으로 되돌아가 보자. 방 안의 나는 영어로 된 셰익스피어 작품을 이해하지 못하면서도 셰익스피어의 한국어 번역을 이해한다. 셰익스피어의 작품이 영어로도, 한국어로도, 혹은 그 외의 어떠한 자연언어로도 표현될 수 있기 때문이다. 요컨대 나는 단지 셰익스피어의 한국어 번역을 이해하고 있는 것이 아니라 '셰익스피어'를 이해하고 있는 셈이다.

그러나 여기서 한국어로 번역된 셰익스피어 작품에 대한 이해와 독립된 '셰익스피어'의 이해는 무엇인가? 자연언어로 표현된 셰익스피어를 넘어서는 셰익스피어는 무엇인가? 우리는 앞서 의미, 혹은 명제를 그것을 표현하는 개별적 문장, 즉 자연언어를 넘어서는 제 3의 영역에 위치시키는 프레게의 플라톤주의가 콰인의 번역 불확

정성론에 의해 와해됨을 보았다. 셰익스피어에 대한 래퍼포트의 플라톤주의는 콰인의 비판에 어떻게 대처할 것인가? 이는 래퍼포트가 해결해야 할 과제이다.

필자로선 한국어로 번역된 셰익스피어 작품에 대한 이해와 독립된 '셰익스피어'의 이해가 무엇인지 알 수 없다. 따라서 사정상 논의의 범위를 한국어로 번역된 셰익스피어 작품에 대한 이해로 좁혀 보자. 한국어로 번역된 셰익스피어 작품에 대한 나의 이해는 과연 같은 번역에 대한 컴퓨터의 이해와 동등한가? 래퍼포트에 의하면 컴퓨터의 역할을 하는 내게 주어진 기능은 자연언어에 대한 계산적, 구문론적 조작이 전부다. 그러나 나는 그 기능만 가지고서 한국어로 번역된 셰익스피어 작품을 이해하고 또 이에 관해 논문도 써낸다. 이는 어떻게 가능한가? 컴퓨터의 기능이 자연언어에 대한 계산적, 구문론적 조작에 국한되어 있다는 자신의 가정을 어기지 않고 이에 답하려면 래퍼포트는 구문론적 조작에서 의미 이해가 도출됨을, 혹은 의미론이 구문론으로 환원됨을 먼저 증명해야 할 것이다.

끝으로 썰의 '중국어 방'의 경우 중국어를 이해 못한다고 믿는 나의 신념 상태에도 불구하고 사실 나는 중국어를 이해하고 있다는 래퍼포트의 응답을 검토해 보자. 썰은 이러한 종류의 응답을 '체계 응답'(The Systems Reply)이라고 명명하고 이에 대해 다음과 같이 답변하고 있다(Searle 1980, pp.194-197). 썰에 의하면 래퍼포트는 중국어 방의 '나'를, 중국어를 이해 못한다고 믿는 신념 체계로서의 '나'와 중국어를 이해하는 '나'로 각각 이원화하고 있다. 래퍼포트는 중국어를 이해하는 '나'가 그렇지 않다고 믿는 '나'보다 더 궁극적인, 그리고 총체적인 '나'의 모습이라는 것이다. 그가 그리는 '나'의 모습은 '나' 속에 또 하나의 '나'가 에일리언(alien)처럼 숨어 있는 형국이다. 그러나 만일 총체성이 문제의 관건이라면 중국어를 이해 못한다고 믿는 '나,' 즉 신념 체계로서의 '나'가 총체적 '나'에 관한 모든 것을 익혀 '나'의 것으로 포섭할 수 있을 것이다. 요컨대 어느 것이 궁극적이고 총체적인 나의 모습인지는 상대적인 문제이다.

썰은 '나'를 이원화해서 얻어지는 이점이 무엇인지를 묻는다. 이원화 이후에도 '나'가 어떻게 중국어를 이해할 수 있는지는 여전히 해명되지 않고 있기 때문이다. 그 이유는 이원화된 두 '나' 사이의 관계를 설명할 수 없다는데서 기인한다.

4.4 구문론적 기계 인간: 블럭

래퍼포트가 지금까지의 비판에 맞설 수 있는 길은 의미론이 구문론으로 환원됨을 보이는 것이다. 래퍼포트는 '한국어 방' 논변이 이를 해 내고 있다고 믿고

있지만 필자는 '한국어 방' 논변에서 어떻게 그것이 이루어지고 있는지 알 수 없다. '한국어 방' 논변 자체가 설득력이 없어서가 아니라 그것이 컴퓨터와 어떻게 연관되는지가 모호하기 때문이다. 최근 블럭(Block 1990)은 의미론이 구문론으로 환원됨을 보여주기 위한 보다 구체적인 논증을 시도하였다. 그의 논증의 방법은 포더(Fodor 1968)와 데넷(Dennett 1975) 등이 제시한 다음과 같은 '기능적 분석'에 의거해 있다. 인간의 마음을 머리 속에 존재하는 지성적 소인(homunculus)으로 의인화 해보자. 이 소인이 다시 보다 작고 덜 지성적인 소인들로 구성되어 있으며, 이들 각각의 소인들은 다시 보다 더 작고 보다 덜 지성적인 소인들로 구성되어 있다고 생각하자. 이러한 경로를 밟아 내려가다 보면 우리는 완전히 기계적인 소인들에 도달하게 된다는 것이다.

이제 다음과 같은 자연언어 이해의 모델을 살펴보자. 이 모델은 다음과 같은 세 소인으로 구성되어 있다. 첫 번째 소인은 입력되는 단어를 접수해서 두 번째 소인에게 넘겨준다. 두 번째 소인은 각각의 단어에 대한 구문론적, 의미론적 정보를 기록한 사전을 가지고 있다. 그는 주어진 단어를 사전에서 찾아 이를 세 번째 소인에게 알린다. 세 번째 소인은 두 번째 소인이 찾아 준 사전에 나온 구문론적, 의미론적 정보를 검색한다.

이 과정에서 동원되는 기능은 단어의 접수와 대조, 검색 등 단순하고 기계적인 조작이다. 각각의 세 소인들이 과연 자연언어를 이해하는지에 대해 블럭은 확실한 입장을 취하지 않는다. 그는 최소한 세 소인들이 모여 이루어진 모델만큼은 자연언어를 이해하고 있다고 믿는 것 같다. 그러나 이러한 시도는 앞서 살펴본 '체계 응답'의 범주에 속하는 것이며 썰은 이미 이를 비판한 바 있다.

블럭은 기능적 분석 방법을 바탕으로 계산기의 구조를 분석한다(Block 1990, pp.257-259, 264-267). 그는 이를 통하여 의미론이 구문론에서 도출됨을 입증하려 한다. 계산기는 이진법의 체계이다. 우리가 사용하는 십진법의 2는 계산기의 이진법에 있어서는 10이다. 그러나 우리는 십진법의 2와 이진법의 10이 동일한 수(number)를 표현하고 있음을 안다. 십진법의 2와 이진법의 10이 표현하고 있는 동일한 수가 바로 이들이 공유하는 단일한 의미이다. 하나의 수의 의미는 이처럼 여러 방식으로, 즉 여러 숫자(numerals)로 표현될 수 있다. 우리가 계산을 할 때 우리는 계산에 동원되는 수의 의미가 무엇인지 알고 있다. 그러나 우리의 계산을 대행하는 계산기의 내부에서 우리가 사용하는 수의 의미는 완전히 제거되고 대신 수를 대행하는 숫자들이 기계적 프로그램에 의해 구문론적으로 조작되고 있음

을 알게 된다.

블럭이 의도하는 바는 계산의 기능적 분석을 통하여 의미론적 수 개념이 구문론적 숫자 개념으로 변모하는 구체적 과정을 보이려는 것이다. 블럭에 있어서 궁극적으로 수와 숫자는 각각 의미론적 엔진과 구문론적 엔진에 연관된다. 그에 의하면 구문론적 계산기가 의미론적 수의 계산을 수행하듯, 구문론적 엔진인 뇌가 의미론적 엔진인 마음을 운전한다는 것이다(Block 1990, pp.264-267). 즉 의미론이 구문론에 환원된다는 블럭의 주장은 마음이 뇌에 환원된다는 물리주의에 이르게 된다.

블럭의 이러한 논의는 래퍼포트의 논의가 안고 있는 문제점들을 다시 한번 노정하고 있다. 셰익스피어에 대한 래퍼포트의 플라톤주의는 블럭에 있어서는 수에 대한 플라톤주의로 부활된다. 사실 블럭의 플라톤주의는 프레게의 수리철학의 답습에 불과하다. 콰인의 번역 불확정성론은 프레게와 래퍼포트에 있어서와 마찬가지로 블럭에게도 큰 부담으로 남는다. 어떤 의미에서 블럭이 안고 있는 문제는 래퍼포트의 경우보다 더 심각하다. 래퍼포트가 문제삼고 있는 자연언어의 경우와는 달리 블럭이 문제삼고 있는 수의 경우에는 의미론과 구문론의 구분 자체가 불분명하기 때문이다. 수학은 의미론인가, 아니면 구문론인가? 수의 의미란 구체적으로 무엇인가? 문제는 의미론적 수의 개념이 구문론적 숫자 개념으로 환원될 수 없다는데 있다기 보다는 수의 의미론과 구문론이 무엇인지가 확립되어 있지 않다는데 있다.

블럭은 어느 시점에서 수가 숫자가 되는지, 즉 의미론이 구문론이 되는지에 대해 확신하고 있는 듯하다. 그러나 사실 그가 말하는 시점에서 변화하는 것은 아무것도 없다. 단지 그는 같은 대상을 놓고 이것을 이 시점 이후부터는 수로 보지 말고 숫자로 보아야 한다고 주장하고 있을 뿐이다. 필자는 블럭에 있어서 플라톤주의와 물리주의가 어떻게 조화될 수 있는지 알지 못한다. 블럭을 좇다 보면 수가 숫자가 되는 것, 의미론이 구문론이 되는 것은 결국 같은 대상에 대한 우리의 태도 변화에 달려 있다는 결과에 도달한다.

같은 맥락에서 블럭의 기능적 분석에 의하면 인간은 어느 시점부터는 구문론적 기계로 변화된다. 이 시점이 정확히 어느 시점인지는 앞서 수의 경우보다 훨씬 불분명하다. 더구나 설령 시점을 확정할 수 있다 해도 이 시점에서 역시 변화하는 것은 아무것도 없다. 그러나 블럭은 어느 시점 이후부터는 인간이 명백히 구문론적 기계가 된다고 주장하고 있다. 물론 우리는 블럭의 의도를 알고 있다. 한 편으로 그는 인간을 구문론적 컴퓨터에 투사하여 이해하려 하고 있다. 그는 컴퓨터가 인간의 모델

이라고 믿고 있다. 그러나 다른 한 편으로 블럭은 자신의 논증의 핵심인 의미론의 구문론에의 환원을 설명하는 과정에서 이에 대한 인간의 태도와 판단에 깊이 의존하고 있다. 환원이란 곧 인간의 태도 변경에 지나지 않기 때문이다. 이러한 방법론상의 혼란은 블럭의 논증을 선결 문제 요구의 오류에 직면하게 한다. 블럭에 있어서 선결 문제는 컴퓨터가 아니라 바로 인간에 대한 올바른 해명이다.

필자는 의미론과 구문론의 구분이 절대로 해체되어서는 안된다 고는 보지 않는다. 그러나 해체가 이루어질 경우에는 운동 개념에 대한 코페르니쿠스의 혁명에서처럼 의미론과 구문론의 전통적 정의 역시 개정되어야 할 것이다. 아울러 해체와 개정의 타당한 이유가 확고하게 제시되어야 할 것이다.

5. *Différance*

지금까지 우리는 인식론의 자연화를 배경으로 한 현대의 자기 의식이 컴퓨터에서 자신의 모습을 찾으려는 시도들을 비판적으로 살펴보았다. 컴퓨터와 인지과학은 우리가 알지 못했던 많은 정보와 새로운 경지를 보여주었다. 이러한 점에서 인식론의 자연화는 매우 성공적인 전략이었다. 방법론적 관점에서 보았을 때 AI와 인지과학은 인식론의 자연화에서 직접, 혹은 간접으로 비롯된 결실들이기 때문이다.

이 글에서 우리의 관심은 AI와 인지과학의 찬양이 아니라 컴퓨터의 삶의 형식이 인간의 삶의 형식과 동일하다는 가설의 점검이었다. 그 가설의 진위를 판정하는데 있어서의 관건은 컴퓨터가 자연언어를 이해하느냐의 문제였다. 이를 긍정하는 주장의 약점은 그것이 이해가 무엇인지를 제대로 설명하지 못하고 있다는 점이었다. 요컨대 컴퓨터가 자연언어를 이해한다는 주장의 논거는 자연언어를 이해하는 인간의 입장에서 보았을 때 설득력이 부족하거나 지나치게 이질적인 것이었다.

인식론이 자연화된 뒤에도 이해의 개념이 과학적으로 완전히 정의되지 못하고 있다는 것은 무엇을 의미하는가? 이에 관해 필자는 다음과 같이 생각한다. 첫째, 그것은 이해의 개념이 인식론적 개념도, 과학적 개념도 아님을 시사한다. 필자는 그것이 인간적 개념이라고 본다. 우리는 은유와 상상의 세계에서를 제외하고는 우리 아닌 다른 대상에 이해의 개념을 한 번도 진지하게 적용해 보지 못했다. 물론 우리는 때로 우리와 가까운 개나 고양이에게 이해의 능력을 귀속시켜 보곤 한다. 그러나 이는 단편적인 시도이며 결코 온전하고 지속적인 의사 소통의 수준에 이르

지 못한다. 우리가 개나 고양이를 언제나 인간과 구별하듯 컴퓨터에 대해서도--설령 그것이 아무리 효율적이고 능란한 기교를 구사한다 해도--같은 구별을 유지한다. 우리는 컴퓨터로부터 많은 도움과 배움을 얻는다. 그러나 그것은 역시 테크놀로지에 불과하다. 우리의 이러한 태도는 우리가 인간이라는 하나의 삶의 형식에 속해 있다는 결속력의 확인인 동시에 다른 대상, 다른 삶의 형식에 대한 배타적 편가름의 표명이다.

둘째, 이해의 자연화를 콰인이 제시한 방향이 아니라 비트겐슈타인(Wittgenstein 1967)이 제시한 방향으로 추구하는 것을 제안해 본다. 콰인의 자연화, 혹은 줄여서 자연화$_Q$는 문제되고 있는 주제를 과학적 방법에 포섭시키는 과정을 말한다. 그에 있어서 과학은 유일하게 합법적인 지식의 형태이기 때문이다. 비트겐슈타인에서 자연화라는 용어는 등장하지 않는다. 그러나 그는 인간과 연관된 자연사(natural history)의 일반적 사실을 환기시키고 이를 기술하는 것을 자신의 철학의 방법으로 삼는다.[5] 필자는 문제되는 주제를 이러한 방법에 포섭시키는 과정을 비트겐슈타인에 있어서의 자연화, 혹은 줄여서 자연화$_W$라고 부르고자 한다. 인식론의 자연화$_W$는 자연화$_Q$와는 달리 이해나 인식, 의미 등 과거 인식론의 핵심 개념들이 실제적 삶의 문맥에서 어떻게 사용되고 있는지를, 인간의 삶을 제약하는 자연사의 사실과 연관지어 보여줌으로써 수행된다. 실제로 심리철학에 관한 비트겐슈타인의 최후 작품들은 이러한 개념들의 쓰임의 기술과 분석에 집중되어 있다.

이러한 제안은 자연화$_Q$가, 그것이 가져온 많은 유익한 결실에도 불구하고 인간에 대한 올바른 이해를 도모하지 못하고 있다는 우려에서 비롯된다. 자연화$_Q$의 문제점은 과학과 테크놀로지가 인간의 투사의 한 양상에 불과하다는 사실을 간과하고 있다는 점이다. (이는 물론 과학과 테크놀로지가 객관성을 결여한 주관적이고 상대적인 것이라는 말로 받아 들여져서는 안된다.) 인간은 여러 대상에 자신을 투사하여 이를 거울삼아 자신을 이해하려 해 왔다. 그러므로 인간을 이해하기 위해서는 적어도 인간이 투사하고 또 구성해 본 거울들을 공평하게 살펴볼 필요가 있다. 자연화$_Q$는 과학, 특히 자연과학에 특권적 지위를 부여함으로써 과거에 논리 실증주의가 그러했던 것처럼 공평성을 잃고 있다.

컴퓨터와 AI는 인간이 자신을 이해하기 위해 투사적으로 동일시해 본 대상 중

5) 이에 관한 보다 자세한 논의를 위해서는 다음을 참조할 것. 이승종, "인간의 얼굴을 한 자연주의," 한국 칸트학회 발표논문, 1993년.

가장 최근의 것이다. 그러나 과거의 많은 투사적 동일시의 시도가 그러했던 것처럼 그 결과는 투사적 동일시의 대상과 우리 사이의 다름의 확인이었다. 이 확인은 과거의 경우가 그랬던 것처럼 여러모로 유익한 것이었다. 그럼에도 불구하고 인간이 무엇인가 하는 궁극적 물음에 대한 해답은 이로서 또 한번 연기된다. 어쩌면 이는 이 물음이 갖는 숙명인지도 모른다. 다름과 연기를 하나로 묶는 개념이 있다면 그것이야 말로 인간의 물음에 대한 우리의 처지를 대변할 수 있을 것이다. *différance*가 바로 그러한 개념이다.

참 고 문 헌

이승종(1993a) "의미와 해석에 관한 콰인/데이빗슨 논쟁," 『철학』, 39집.

_____.(1993b) "언어철학의 두 양상," 『철학과 현실』, 겨울호.

_____.(1993c) "인간의 얼굴을 한 자연주의," 한국칸트학회 발표논문.

이초식(1993) 『인공지능의 철학』, 고려대학교 출판부.

Barker, S.(1964) *Philosophy of Mathematics*, Englewood Cliffs: Prentice-Hall.

Block, N.(1978) "Troubles with Functionalism," Lycan 1989에 재수록.

_____.(1990) "The Computer Model of the Mind," Osherson and Smith 1990에 수록.

Boden, M.(ed.)(1990) *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.

Brentano, F.(1960) "The Distinction between Mental and Physical Phenomena," Chisholm 1960에 재수록.

Campbell, N.(1952) *What is Science?* New York: Dover.

Cassirer, E.(1944) *An Essay on Man*, New Haven: Yale University Press.

Chisholm, R.(1960) *Realism and the Background of Phenomenology*, Glencoe: Free Press.

Churchland, P.(1988) *Matter and Consciousness*, revised ed., Cambridge: MIT Press.

Dennett, D.(1975) "Why the Law of Effect Will Not Go Away,"

Dennett 1978에 재수록.

_____.(1978) *Brainstorms*, Cambridge: MIT Press.

Fetzer, J.(ed.)(1988) *Aspects of Artificial Intelligence*, Dordrecht: Kluwer.

Fodor, J.(1968) "The Appeal to Tacit Knowledge in Psychological Explanation," Fodor 1981에 재수록.

_____.(1981) *Representations*, Cambridge: MIT Press.

Frege, G.(1892) "On Sense and Reference," Geach and Black 1952에 재수록.

_____.(1912) "The Thought: A Logical Inquiry," Strawson 1967에 재수록.

Garfield, J.(ed.)(1990) *Foundations of Cognitive Science*, New York: Paragon.

Garver, N. and Seung-Chong Lee.(1994) *Derrida and Wittgenstein*, Philadelphia: Temple University Press.

Geach, P. and M. Black.(eds.)(1952) *Translations from the Philosophical Writings of Gottlob Frege*, Oxford: Basil Blackwell.

Harman, G.(1989) "Some Philosophical Issues in Cognitive Science: Qualia, Intentionality, and the Mind-Body Problem," Posner 1989에 수록.

Jahren, N.(1990) "Can Semantics be Syntactic?" *Synthese*, vol. 82, no. 3.

Lycan, W.(ed.)(1989) *Mind and Cognition*, Oxford: Basil Blackwell.

McCarthy, J.(1979) "Ascribing Mental Qualities to Machines," Ringle 1979에 수록.

Osherson, D. and E. Smith.(eds.)(1990) *Thinking*, Cambridge: MIT Press.

Posner, M.(ed.)(1989) *Foundations of Cognitive Science*, Cambridge: MIT Press.

Quine, W. V.(1951) "Two Dogmas of Empiricism," Quine 1980에 재수록.

_____.(1960) *Word and Object*, Cambridge: MIT Press.

_____.(1969a) "Epistemology Naturalized," Quine 1969b에 수록.
_____.(1969b) *Ontological Relativity and Other Essays*, New York: Columbia University Press.
_____.(1980) *From a Logical Point of View*, 3rd ed., Cambridge: Harvard University Press.
Rapaport, W.(1985) "Machine Understanding and Data Abstraction in Searle's Chinese Room," *Proceedings of the 7th Annual Conference of the Cognitive Science Society, University of California at Irvine*, Hillsdale: Lawrence Erlbaum에 수록.
_____.(1986) "Philosophy, Artificial Intelligence, and the Chinese Room Argument," *Abacus*, vol. 3.
_____.(1988a) "Syntactic Semantics: Foundations of Computational Natural Language Understanding," Fetzer 1988에 수록.
_____.(1988b) "To Think or Not to Think,"*Noûs*, vol. 22.
Ringle, M.(ed.)(1979) *Philosophical Perspectives in Artificial Intelligence*, Atlantic Highlands: Humanities Press.
Searle, J.(1980) "Minds, Brains, and Programs," Garfield 1990에 재수록.
_____.(1984) *Minds, Brains and Science*, Cambridge: Harvard University Press.
Strawson, P.(ed.)(1967) *Philosophical Logic*, Oxford: Oxford University Press.
Turing, A.(1950) "Computing Machinery and Intelligence," Boden 1990에 재수록.
Weizenbaum, J.(1976) *Computer Power and Human Reason*, San Francisco: Freeman.
Wittgenstein, L.(1967) *Philosophical Investigations*. Ed. G. E. M. Anscombe and R. Rhees, trans. G. E. M. Anscombe, 3rd ed. Oxford: Basil Blackwell.