言語処理技術の現状と将来

長尾 真

長尾でございます．佐和先生の今朝の非常にわかりやすくてすばらしい具体的な話からいっぺんに泥くさい話になりましてちょっと恐縮です．私こちらで今日何をお話しするのがいいかについてはずいぶん迷いました．幾つかお話したいことがありました．一つは何年か前から「情報の生態学」という言葉を作りまして，情報というものが世の中でどういう風に作られて，それがどういう風に使われて，どういう風に摩耗して無くなっていくか，それが人間の善悪とかいろんなbehaviorに対してどういう風に働いていくかとか，そういうことをいろんな観点から調べて明らかにしていかないと情報社会が健全にいかないんじゃないかというわけです．

佐和先生の場合は社会を大きなmassとして御覧になって経済的な観点から考えておられたわけですけれども，私の場合はどちらかというと情報というものが個人に対してどういうインパクトを与えていっているかというようなことを考えないといけないんじゃないかという立場です．そういう話をしようかとも思いましたけれども，過去三回のこのシンポジウムを拝見いたしますと，名古屋大学の福村先生，東大の大須賀先生，それから京大の堂下先生のお話なんかがすでにございましてそういった面がある程度触れられている．したがって，それをやっても重なるかと思いまして，4回目ですからもう少し突っ込んだある程度専門的な分野の限られた範囲を少し詳しくお話しした方がいいんじゃないかと，そういう風に思って言語情報処理の話をさせていただくことに致しました．

佐和先生と違いましてわたくしはOHPを使って，さぼって原稿を作ってきていないという悪いところがありますが……．まず，言語と言いました場合に，それは当然人間の言語であるというのが常識なんですけれども，情報科学をやっている連中からしますと言語というのは機械言語，計算機言語というのがある．それから人間の言語があると，こういう風に対立的に考えるわけですが，そこには大きなギャップがある．機械言語というのはもちろんはっきり定義された言語であるのに対して，人間の言語というのは定義されてない言語である，まあflexibleな言語であるということになります．

で，そういうものをもうちょっと別の見方をすると，機械の言語というのはローカルにdefineされている．定義がローカルな格好で，ある表現は文法的に正しいとかまちがってるとか，そういうことはすべてローカルなルールで決められている．人間は物事を厳密に定義しようとするとローカルなところに

Makoto NAGAO 京都大学

もっていかざるを得ない．つまり，divide and conquerっていいますか，これは，まあ西洋の科学の真髄というか，エッセンスじゃないかと思いますけれども，物事を分解していってユニットに分けてそのユニットごとに解決していく．そしてそれらが解決できれば全体が解決されたんだという風に考えるという世界であります．機械言語がそういう風になっているのは当然なわけでありますが，人間の言語はどうもそうじゃないんじゃないか．つまり，ユニットに分けられないものである．従って，グローバルなことを考えないといけないというわけであります．そういった観点から，今までの自然言語処理，計算機で言葉を扱ってきたのをずっと見て参りますと，文法理論とかいろんなものがあるわけなんですけれども，これはやはり西洋の近代科学の考え方でやってきておりますから，分解していってそこのところで物事を決定して，それで全体が解釈できたとする．そういう文法理論になっている．チョムスキーの文法理論にしても大なり小なりそういうものであるということなんですけれども，どうも，現在の言語処理技術の段階からするとそこのところに大きな行き詰まりがあるのではないかという風に考えられる．

例えば，私どもいろいろやってきております機械翻訳といったものを考えましても，普通のいわゆる文法理論で分析して，そして日本語から英語に単語を入れ換えて英語の文を作り出していくというようなやり方では，どうしてもある程度以上のクオリティーのものは出せないし，しばしば間違う．もう少しいいクオリティーで翻訳というものをやるためにはどうしたらいいかということを考えないといけないわけですけれども，それには今までのような，自然科学が取ってきたようなやり方では駄目なんじゃないかというふうにだいぶ前から思っておりました．そのためにはグローバルなものの見方を持ってこないといけないんじゃないかと考えられる．今日はそれについてお話をしたいと思っているわけであります．

そこで，まず，言語は曖昧です．曖昧であるものを理論言語学のような硬い枠組みでやるということは根本的に問題がある．硬い枠組み，つまり言語の文法理論というようなものは硬すぎてそれでは解釈できない言語現象っていうのは山ほどある．だから，なんらかの形で柔らかい枠組みを考えないといけない．あるいは，曖昧さというものを何らかの形でうまく包含するようなプロセスというようなものを考えないといけないと思われるわけであります．そういうプロセスの研究を私個人は考えて来たわけでして，柔らかい枠組みというか，曖昧な枠組みで何ができるかという例をこれからお示ししようと思っておりますが，ともかく局所的に見たら非常にいい加減なことをやっているけども，大局的にみたらかなり信頼性のあることをやっているというふうなやり方を積極的に開発していく必要があるんじゃないか，ということであります．

これはまったく今までの自然科学的な手法とは逆であります．自然科学的な手法っていうのは物事を分けられるだけ分けていって，局所的に厳密なことをやったらそれが全体をおおうという信念に基づいているわけですけれども，自然言語のように人間にまつわるような内容というのはどうも自然科学の物理的現象というものとはずいぶん違っていて，局所的に厳密なことをやると間違うことが多いんではないか，局所的にはいい加減なことをやるんだけども，大局的にみるとかなり安定した，信頼性のある内容をもってるというふうなことを考えようというわけで，これは非常にむずかしいことなんですけれども，そういうことを試みる．

こういうことについての一般的手法があるかどうかっていうことは今のところわかって

おりませんので，今日お話ししますことは，ある意味では個別的な方法です．それが一般の科学方法論として成り立ち得るのかどうかっていうことについてはちょっとまだまだ疑問な点があるわけですけれども，そういう態度で物事を見ていくことをこれからやらなければならないということであります．

それの一つの例としまして，日本語の文解析をとりあげましょう．現在の言語処理技術ですと大体日本語の文字数にして一つの文が50文字とか60文字ぐらいの文，仮名漢字混じり文で文字の数を数えるわけですけれども，その程度の長さの文ですとまあある程度解析ができる，ある程度成功するわけですけれども，一つの文の文字数が70，80以上の文になるとほとんどの場合解析がうまくいません．失敗いたします．

それで，この問題を一つ解決しなければいけないんですけれども，いろいろよく調べてみると，文字数が多い長い文というのはいろんな意味で文が並列した構造を持っている．ここ（図1）にありますようにいろんな並列構造を持っている．つまり名詞が並列しているとか日本語の場合は「何々して何々して何々した結果何々した」というふうに，いくらでも連用中止法で文を続けていく．そこで，そういう並列している構造をうまく見つける方法はないかということを考える．で，並列しているものを徹底的に調べますと，人間はどういう頭の構造をしているのかわかりませんけれども，似たような構造をしたものが並列をしている．文を書くときに似たような構造で並列させてものを書いている．例えば『解析して，生成する』とかですね．いろいろありますけれども，比較的似たような構造や似たような単語を使う．あるいは同義語じゃなくて反義語みたいなものを使うとかですね．そういう意味において似ているものを使う．だから，似ていることを発見する．つまり，長い文の中のこの辺の部分とこの辺の部分はどうも似てるんじゃないかと考える．それをいかにして機械的に発見するかということを考えるというわけです．

その発見の仕方にはいろんな方法が有り得ますけれども，ここでとりました方法はどんなのかというと，まず並列のキーになるようなポイントは比較的簡単にわかる．『何々して何々する』という場合は『何々して』というような連用中止法の部分がありますから，そ

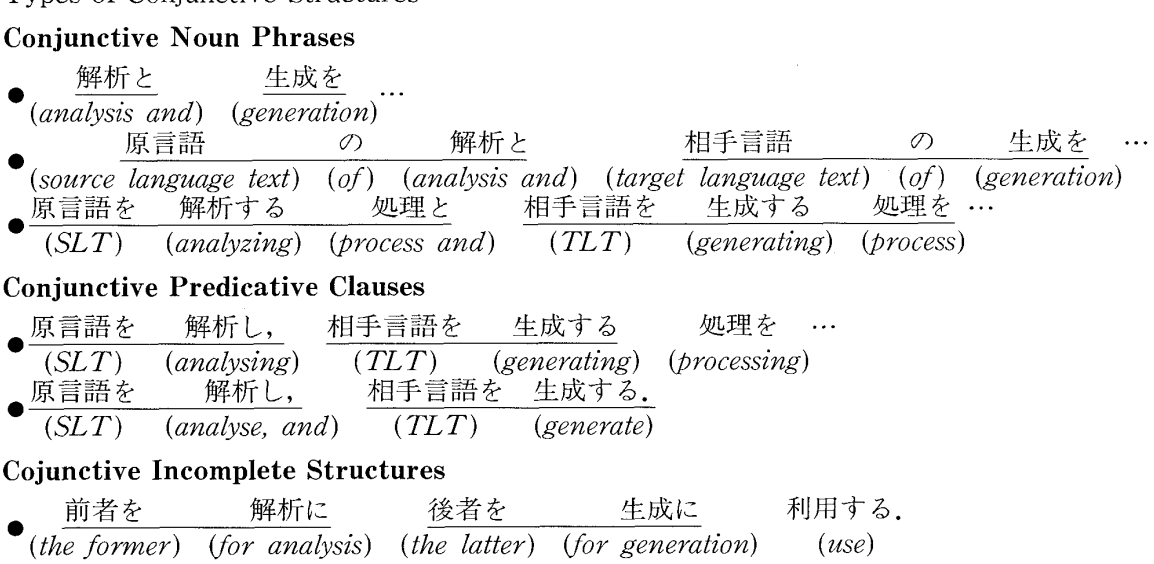

図1 いろいろな並列構造

れは比較的簡単にわかる．あるいは，『何々，何々，何々』というふうに名詞が連なっている場合もそうですし，『何々と何々』という場合は『と』というような並列の助詞が伴っている．そういうことで並列の構造を作り出すキーになるような単語とか，そういうものがわかる．一つの文の中の並列を導き出すようなキーになるものがわかると，そこから手前のほうに何単語かとり，そこから先のほうに何単語かとって，その単語列と単語列が大体似ているかどうかということを調べる．

今までの文法理論ですと，並列の構造をきちっと見つける文法規則はほとんどうまく書けません．同じ単語どうしが並んでいるとかそんなふうなことしか書いてなかったんですけれども，一つの単語どうしは似てないんだけれども，単語のグループとして見た場合，ある単語列とある単語列が大体似ているということを発見したい，そういうことをやるのにダイナミックプログラミングの方法を使うことを考えました．それの話をし出すと細かいんですけれども，一つだけ例を挙げます．

これ（図2）は簡単な例ですが，ここにこう文がありまして，『図に示すように，電流源にpnpトランジスター，スイッチング……』とこうありますと，ここのところにコンマがありまして，これは一つの並列構造を作り出す要素である．この三角形の部分に与えてある数字は何かというと，これは単語どうしの似てさかげんです．例えば，『pnpトランジスタ』というのと『npnトランジスタ』というのは単語どうしが非常によく似ているので，点が高い．ところが，『何とかトランジスタ』というのと，『スイッチング』というのは，それほど単語は似てないから類似度は低い．あるいは，『示すように』というような，そういう単語，というか文節ですけれども，それと，例えば『スイッチング』というようなものはまったく単語が似てないのでここは0点を与えるというふうに，あらゆる単語のペアについてそのペアがどの程度似ているかというようなことで点を与えます．それから，例えば，並列構造を作り出すこのコンマのところから左方向に何単語かとってきて，また右方向に何単語とってきて，それで，一つ一つの単語の似てさかげんの最大値のところをつないでいって，並んだ列全体としてどの程度似ているかというのを調べる，つまり単語の類似性の値を足すということをやる．そういうことを，あらゆる可能な2つの単語列の組について調べるわけですね．

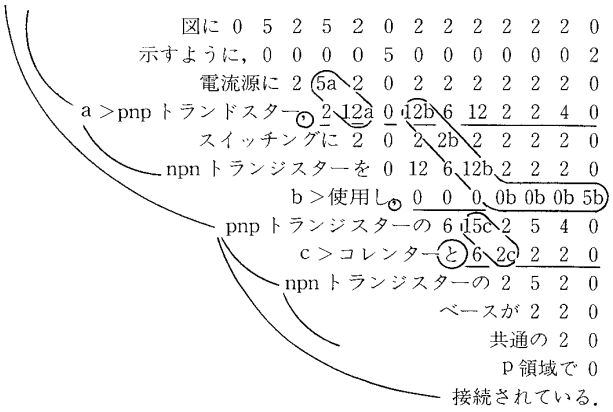

図2　ダイナミックプログラミングの一例

これは非常にややこしい方法ですけれども，最近の計算機ですとこういうことは比較的簡単にできます．それをやりますと，ここにありますように，『電流源にpnpトランジスタ』，『スイッチングにnpnトランジスタ』という2つの単語列どうしがペアになっているということがわかる．あるいは，ここにありますように『トランジスターのコレクター』と『トランジスターのベース』というのがペアになっている，並列しているということがわかる．それで，そういうもの全体を含んだ動詞の構造どうしが並列になっているというようなことがわかる．これは，『文は主語と述語からなる』とか，『主語は名詞か，形容詞と名詞からなる』とか，あるいは，『述部というのは動詞と目的語からなる』とか，『(述部は)動詞と補語からなる』とかいった文法規則という概念とはまったく違っていまして，とに

かく，ある単語列とある単語列がだいたい似てそうだとか，これは似ていそうにないとか，そういうふうなことを見つけるやり方ということになります．

そういう方法でいろいろやりますと，いろんな複雑なものがかなりうまく解析できます．従来は大体80文字以上の文だとほとんど解析がうまくできなかったんですけども，このやりかたを導入すると，大体一つの文が200文字ぐらいからできている文，そういう長い文というのは余程探してもないんですけれども，そのぐらいのところまではいろいろテストしましてだいたいうまくいくということがわかりました．

並列の構造がわかるとまず，一つの並列の構造の中だけを解析してそれ全体を一つの要素で置き換える．並列語句が一つの要素であるということになりますと長い文が短くなる．文が短くなると解析の精度が上がって，結果の信頼性が上がります．しかし，このやり方ですべての場合に成功するかと言うとやっぱり間違う場合がありまして，並列句の構造で9割ぐらいまではうまくいくんですけど1割ぐらいについては失敗する．その失敗したものをよく見ると，人間も間違った読みをするというものが非常に多いということがわかってきた．例えば，一回も読んでない文を人に見せると，スーッと読んで，あれっとおもってもう一回読むといったことをする．2度読むとこの文はこういう構造をしているのだなとわかる．そういうプロセスが人間のreadingにおいて起こるような文がたまにありますけども，そういう場合はこのやりかたでも間違う．それは多くは，並列の構造に極端なアンバランスのあるときです．だけど，そうでない場合はほとんど成功するということがわかりまして，言語の解析についてかなりの進歩をもたらしたんですけれども，それは今さっき言いましたような，なんかふにゃふにゃっとした方法をとっているんだけれども，全体的にはうまくいっているという一つの例であります．

次に，現在の計算機で単語の持っている性質をどこまでくわしく解析できるかというと，なかなかうまくはいっていません．例えば，『彼は学生です』というような文があるとき，日本語の場合は『学生』っていうのは単数であるか複数であるかわからない．英語の場合はstudentであるとか，studentsとなっているので単数か複数かわかるというわけですが，日本語の場合わからない．そこでもっと広い範囲を見ないといけません．『彼は学生です』という場合は，『彼』っていうのは単数ですから，当然『学生』っていうのは単数でないといけないというふうに，つまり，ある程度広い範囲を見ればわかる．一つの単語をみてるだけではわからない．『彼らは学生です』ですと，当然この『学生』は複数にならなきゃならない．それからもう一つは，学生って言いましても一般的な学生と，aとか複数形がつく学生とtheのつく学生がありますね．『彼は昨日一等賞をもらった学生です』という場合，昨日一等賞をもらったというような非常にspecificな修飾語がついている場合の学生というのは特別な個人を指しているわけですので，それはtheがつく．そういうspecificなもの，つまり，definiteなnounなのか，indefiniteなnounなのかということがなかなか今まではわからなかったのですけども，それを文の広い範囲の構造を調べることによってわからせるようにする．あるいは，単数か複数かわからなかったんだけれどもそれを文中の他の単語との関係からわからせる方法を考える．これも，今までの文法ではまったくやられていなかったことです．

こういったことは人間だと当たり前じゃないかということですが計算機ではなかなかできなかったわけです．我々がどういうやりかたをしたかというと，例えば，『犬は役にたつ

動物です』というのと，『その犬は役にたつ』，あるいは『犬が3匹います』というようないろいろな文がありますけれども，『役にたつ動物である』という場合の『犬』というのは，一般的な意味での一般名詞としての犬なんですね．それに対して『その犬は……』のほうは，具体名詞としての犬なんですね．そうすると同じ犬という名詞なんですけれども，それが抽象的に使われている場合と具体的に使われている場合が区別できなきゃいけない．『犬は役にたつ動物です』，『その犬は役にたつ』とか，『犬が3匹います』とかいう文全体を見る．『その』とかがついている場合ですと，これはspecificに『犬』をmodifyしてますから『犬』はそういうspecificなものです．『3匹』というように書いてある場合は，concreteなnounであって複数であるということがわかる．それに対して最初のは『AはBだ』という形の一般的な叙述をしているというような区別をする．そういう区別をするルールを，ヒューリスティックスルールとして書くということをいろいろやりました．

そうしますと，要するに，definiteな名詞か，indefiniteな名詞か，あるいはgenericな汎用な使い方かがある程度わかります．どの程度の精度で判断ができるかというと，だいたい80%ぐらいはいくというようなことがわかってきた．これも，今までのようなdefiniteな文法理論の考え方ではなくて，ある種の修飾語がこの種の名詞についている場合はその名詞の抽象度が高くなってきているんじゃないかとか，過去形でしゃべられている場合ですね，例えば，『学校へいきました』というように，動詞が過去形の場合は，これは，その動作が事実過去に起こったということがわかるわけですから，過去に起こった動作に関しては，これは事実関係であるに違いないから，この場合の『学校』は，実際に存在している学校である．『学校は生徒を処罰した』とかいう場合じゃないとかですね．一つの文の中に現れるいろんな表現によって，ある名詞の抽象性を高めるような表現もあるし，具体性を高めるような表現もある．そういうものをず一っと足し算していくんですね．ある表現があったら抽象度の可能性が高い．ある表現があったら具体度の可能性が高い．種々の表現をチェックすることによって両方両方の可能性が上がって行きますけれども，最終

Referential Property

Generic noun phrase

It denotes all members of the class of the noun phrase or the class itself of the noun phrase.

犬は 役に立つ 動物です

INU(dog)-WA YAKUNITATU(useful) DOUBUTU-DESU(is).

Dogs are useful.

Definite noun phrase

It denotes a contextually non-ambiguous member of the class of the noun phrase.

その 犬は 役にたつ

SONO(the)-INU-WA YAKUNITATIMASU(useful).

The dog is useful.

Indefinite noun phrase

An indefinite noun phrase deenotes an arbitrary member of the class of the noun phrase.

犬が 3匹 います

INU-GA SANBIKI(three) IMASU(there is).

There are three dogs.

図3 犬を例に

value	singular	plural	uncountable	others	total
Usage of the English Articles (140 sentences, 380 nouns)					
correct	274	32	18	25	349
reasonable	1	1	1	0	3
partially correct	0	0	0	11	11
incorrect	3	10	0	4	17
% of correct	98.6	74.4	94.7	62.5	91.8
The Old Man with a Wen (104 sentences, 267 nouns)					
correct	205	24	5	0	234
reasonable	2	0	0	0	2
partially correct	0	0	0	7	7
incorrect	1	22	1	0	24
% of correct	98.7	52.2	83.3	0.0	87.6
An essay "TENSEI JINGO" (23 sentences, 98 nouns)					
correct	64	13	0	3	80
reasonable	2	1	0	0	3
partially correct	0	0	0	6	6
incorrect	1	6	1	1	9
% of correct	95.5	65.0	0.0	30.0	81.6
average					
% of appearence	74.2	14.6	3.5	7.7	100.0
% of correct	98.2	63.3	88.5	49.1	89.0

図4 Number (learning sample)

的に文全体を見た場合には，どっちが優勢しているかということによって，どうもこの名詞は具体性が抽象性よりもちょっと強いようだなというんでこの名詞は具体名詞として判断しておきましょうとか，そういうようなやり方をする．このように，やり方は曖昧なんですけれども，それでもってけっこういい結果が出る．単数か複数かとか，あるいは uncountable かというようなものの推定も，90%近く正しく出てくる．

(質問：佐和) others というのは何ですか．

『こと』とか『もの』とか，そういう種類の名詞です．これは何かのおとぎ話です．こっちは天声人語の文を取ってきたんですね．これは，何か評論みたいなものですね．そこに現れてくる名詞を全部やったんです．ですから『こと』や『もの』などもはいってきます．

(質問：佐和) 天声人語でいうと correct が64で，全部で67個 singular として使われている名詞があるわけですね．そのうち64個が正しいとして判定しているということですね．では reasonable というのは(何ですか)．

絶対的に自信をもっていえるっていう場合は correct で，どうも怪しいけれども，たぶんそうだろうというのが reasonable と分類している．

(質問：田中) 名詞の具体度はあらかじめ(与えているのですか)

それは与えておりません．

(質問：田中) それはどのようにして

例えば，『この店には』などのように(図5)，『この』とかそういう修飾語句がついていると，『店』という単語の具体度に点を10点やるとかするんですね．『店は物を売るところで

す』とか，『AはBです』とか，そういう表現の場合にはこの『店』の抽象性は5点ぐらいあるとかですね，そういう点数をやる．そういうルールを70から80作ったんですね．そうすると，一つの文の中のある単語に対して適用されるルールが5つ6つあるとすると，その一つ一つのルールによって3点与えられたり，5点与えられたりとか，そういうようなことをやって，抽象度の点数と具体度の点数がルールがapplyできる度に増えていく．それで最後にどっちのほうが点数が多いかということで判断するということをやっているんですね．そういうことをやることによって，文中のほかのところの単語の使われかたが，ある単語の使われかたに影響していることを数値的に表している．人間の場合にそういうことをやっているかということはちょっとわかりませんが，先ほど言いましたように，やっぱり人間の場合でも，こういう修飾語句があった場合には具体的な名詞をあらわすんじゃないか，とか，そういういろんなことを考えているに違いないので，それでいいんじゃないかっていうようなことでやっています．

（質問：佐和）『銅はよく熱を伝導します』の銅がuncountableだというのはどこでわかるんですか．

これは，物質名詞かどうかを調べてますね．辞書は持っています．ただ，場合によって物質名詞でもcountableで出てくる場合が有り得るんですね．例えば，cakeというのはa piece of cakeで物質名詞なんじゃないかと思うんですがねえ．『たくさんの』という場合は，結局，複数形にしてしまってますね．だから，辞書では物質名詞か普通名詞か固有名詞かということは調べるんですけれども，その名詞が文の中であらわれたときに本当にuncountableで働いているか，countableで働いているかというのはまた別になる．というのでルールをいろいろ作ってやりますとこの程度のことは出来るということが分かってきた．これはそれほど素晴らしいかどうかわからないやり方なんですけれども，そういうやり方でやりますと，いままで誰も手をつけていなかった名詞の抽象度とか具体度とかそういったことについてある程度の判断ができるようになってきた．

Number
Singular noun phrase
Its referent is singular.
彼女は ケーキを 1個
KANOJO(she)-WA KEIKI(cake)-WO IKKO(one)
MOTTE-IKIMASHITA(take).
持って いきました
She took a cake.

Plural noun phrase
Its referent is plural.
この 店には
KONO(this)-MISE(shop)-NIWA
TAKUSANNO(many)-KEIKI(cake)-GA ARIMASU(there is).
沢山の ケーキが あります
There are many cakes in this shop.

Uncountable noun phrase
Its referent is uncountable.
銅は よく 熱を
DOU(copper)-WA YOKU(well) NETU(heat)-WO
DENDOUSIMASU(conduct).
伝導します
Copper conducts heat well.

図5

さて，文を解析するステップですが，日本語の文が入ってきますと形態素解析というのをまずやります．これは日本語の文を単語に区切りまして辞書を引きまして品詞を付けます．それからさきほど申し上げましたどこからどこまでが並列しているかという並列構造を調べて，その中を係り受け解析する．日本語の場合は語順が任意ですから，句構造解析がうまくできせんので，係り受け解析をやるのが一番いいと今のところ思っております．日本語の場合は省略が非常に多いし，語順が自由であるというその両方に対してこの係り受け解析と言うのは比較的うまく行く．係り受け解析ができた後，格構造解析を行ないますと，どれが主語でどれが目的語かというようなことがわかるようになります．そういう解析をやる．というようなかたちで日本語を解析して行きまして，あとは省略語句の推定とかですね，『それ』とか『あれ』とか『この』とかいろんな代名詞語句が何を指しているかというようなことを調べる．それについてもいろんなことをやりましたけれども，ちょっときょうはやめておきます．そういうことをやって解析をしますと一応日本語の解析ができます．その解析の精度は形態素解析で現在98から99%近くになってきました．ある語句がどの語句を修飾しているかということを文節単位で勘定して，正しく解析できた率ですね．ところが，日本語の文単位で考えるとあまりよくありません．30から50文字の文と，50から80文字の文と，80から150文字の文でやりますと，表（図6）に示したぐらいしか今のところできませんね．

（質問：佐和）これもdependencyの話ですか．

そうですね．dependency解析の話ですね．文節単位でやりまして，例えば，『何々がほしい』とかいう場合は，『何々が』というのが『ほしい』というに係っていっているという2語の関係で，それが正しく係り受け関係になっているかどうかというのを文節単位で勘定すると97%は正しい．ところがこれを文単位でどこか一箇所でも間違いがあるとだめだとなる．だいたい30文字から50文字の文という

KN Parser

★ *Dependency analysis of Japanese sentences:*
97% by BUNSETSU unit
(noun + suffix)

★ *Analysis of long Japanese sentences:*

(by sentence unit)

	30～50ch.	*50～80ch.*	*80～150ch.*	*average*
relative frequency	78% (0.65)	70% (0.30)	46% (0.05)	65% 74% (*weighted average*)
current commercial systems	74% (0.65)	54% (0.30)	20% (0.05)	49% 64% (*weighted average*)

★ *Case structure analysis:*
KN Parser: 93%
commercial systems: 60～70%

図6

と10文節ぐらいは少なくともあるわけですね. 10文節あると, 97%の10文節というわけですから, 30%ぐらいはエラーが出てくるんですね. 50から80文字だと15文節ぐらいありますから, 3%の15倍ぐらいは間違ってくるわけですね, そんなことになって, 精度はなかなかうまく上がらないんですけど, 文単位で60%近くの解析精度のものを80%ぐらいに持ってこようと思うと, dependency analysisを文節単位で99%～99.2%にしなければならない. そのためには, 形態素解析を99.9%以上にしなければならないということになって, 今徐々にやってますけれどなかなかうまくいかない.(最近99.7%まで来た).

(質問：田中) それはたとえば30文字から50文字の文の中に一つも誤りがなくちゃんと出たのが78%ですね. そうすると, 間違っていた場合の間違いの数はどのくらいなんでしょう.

ほとんどは一文中一箇所です.

(田中)そうだとすると, 非常にいいような気がするんですが.

ええ, それが97%ぐらいですね, 文節だけで見ていますから.

(田中) 人が直接見て係りかたを見いだして, それに対してどのくらい間違っているかあっているかをみるのですか.

そうですね.

(田中)その場合人間が見つける係りかたの正しさは, 人によって違いますね.

違いますが, 係り受け解析の場合は我々素人がやってもほとんど正しいんじゃないかと思います.

(田中)それは複数の学生にやらせてみて結果を比べる, なんてことはされてますか.

それはやらなかったですね, 一人の学生がやってますね.

(田中) それは充分時間をかけているんですね.

それは充分時間をかけて, 木構造表現になった係り受けの図を見て, ほんとにこういう修飾関係で正しいかどうかというのをみています.

(田中)私は前に文の分類というのをやりましてね. あの時は独立に5人ぐらいでやりまして, かなりあうんですね.

文によってはこういう係り受け構造も許される, こういうのも許されるといくつかの可能性がありますね. いくつかの可能性がある中のどれかというのは非常に難しいわけですけども, この場合は可能な係り受けのどれかであればOKだとしているのですね.

(田中)係りかたの曖昧な文は, 例としては除いて試みられたんじゃなくていろいろな多意的に見えるものも含めて……そのうちの一つがもっともな答であればよしとする.

そうです. 人間として解釈可能で, ほかの解釈もあるかもしれないけれども, その解釈は間違ってない, と, そういうのはOKだとするわけですね.

(田中)非常にうるさくいいますと, 人間としてもっともだと思われる係りかたは全部出すということは危険ですね.

かもしれませんね. それはそうです. しかし一つの文だけ見ていては曖昧なもののうちのどれが正しいかというのは決定できないのですね. いくつかの文の並びを見ないとどれがいいかっていうのはわからないので, そこは今のところできていませんね.

さて, そういうことをやりまして, 省略語句なんかもかなり正確に復元できるということがこれでわかりました. 例えば図(図7)に書いてあります( )の部分は省略されているんですね.『コンピューターアーキテクチャー』と同じものがここにくるはずなんだということが構造上わかる. だからこれを埋めておく. 英語の文を作るときはひょっとしたら必要になるかもしれないので, 省略語句も埋められるところは全部埋めるというようなことをやっております. 今までのシステムですと20%ぐらいしかうまくいかなかったのがかなりよくなった. 短い文ではあまり変

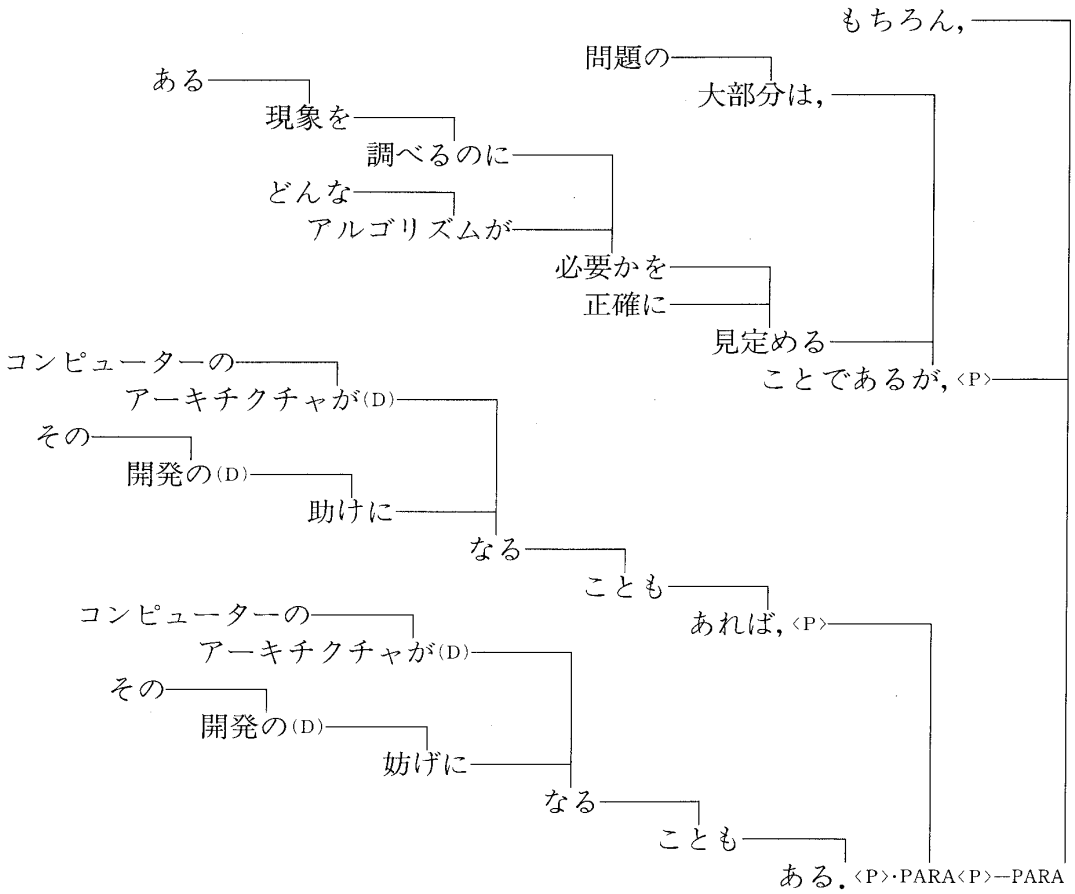

図7 係り受け解析の例

わらないのですね．だけど長い文になってくるとかなりよくなってくる．

(質問：佐和)それは先ほどの並列を発見することで違いが？

そうですね．ほとんどはそれがきいている．ということで，これは希望があるということになります．

話題を変えまして，翻訳のことをお話いたします．アナロジーによる翻訳というのをいろいろやっております．この例（図8）はわたくしがやったんじゃないんですけれども，私が教えた隅田君という人がうまくやりまして，成功した例なんですけれども，『の』という単語はどういうふうに英語に翻訳したらいいか，というのは非常に難しいんですね．いろんな意味の『の』がありますので．そこで，どういうふうにやったかというと，『の』を含む日本語の文をたくさん集めてくる．そして，それに対する適切な英語の訳を並べる．そういうものを100ぐらい集めたんですね．実際は300ぐらい集めたのですが，同じパターンは1つにしたわけです．例えば，『今日の新聞』：『today's paper』，これがベストな訳かどうかわかりませんけれども一応訳がわかるというようなのを与えてありますね．こういうexampleをたくさん集めまして計算機のなかにいれておきまして，そして，例えば，『札幌の夏』というような名詞句が出てくると，これが例文の中のどの表現に一番近いかというのを捜すのです．一番近い表現がどれだというのがわかりますと，その表現の英語訳はこういう訳だというのがわかってますから，それとおんなじスタイルで『札幌の夏』を英語の世界に持っていく．そういうやり方をする．つまり，なぞり翻訳みたいなことを

パターン	出現頻度	%	日本語表現	英語表現
'A[2]	920	36.1	香港の方	Hong Kong
'A'B	692	27.1	ホテルの予約	hotel reservation
'B of 'A	482	18.9	スピーチの内容	the content of the speech
'B in 'A	73	2.9	会議の参加者	the participants in the conference
'A-'B	70	2.7	1の2[3]	1－2
'B for 'A	64	2.5	書類の代金	the charge for the documents
'B at 'A	29	1.1	会議のスピーカー	a speaker at the conference
'A's 'B	28	1.1	今日の新聞	today's paper
'B on 'A	25	1.0	5日の開会式	the opening session on the 5th
'B from 'A	24	0.9	企業からの援助	support from industries
'B, 'A	22	0.9	所長のX教授	Prof. X, the director
'B 'A	20	0.8	5番のバス	the bus number 5
'B before 'A	13	0.5	会議の5日前	5days before the conference
'B about 'A	12	0.5	会議についての情報	info. about the conference
'A of 'B	10	0.4	1枚の紙	a sheet of paper
'B to 'A	8	0.3	香港行きのフライト	a flight to Hong Kong
'A, 'B	7	0.3	最終日の30日	the last day, the 30th
'B prior to 'A	7	0.3	会議の3日前	3days prior to the conference
'B concerning 'A	6	0.2	会社の詳細	details concerning the company

図8 'の'を英語に表現する

やるというやり方です.

(質問)日本語の表現どうしの近いとか遠いとかというのはどうやって判断するのですか.

シソーラス,つまり類義語集を使う.同義語と類義語と反義語とか,そういうものが辞書の中にいれてありまして,ある単語とある単語は本当に同義語の類に入っているか,同義語じゃないんだけれども一つ上の階層を見る.例えば,人間と動物,それから動物には人間のほかにサルとかイヌとかもいる.そうすると,人間とサルは動物という下位概念として別々なんだけれども,一つステップが上がると同じ類にはいるというのである程度近い.例えば,人間と机とか建物とかというものよりはるかに近いということがわかるので,そういう単語のツリー構造がだいたい5～6万単語について入れてありまて,…….

(質問:土屋)ということは,今のケースでしたら『札幌の夏』だったら札幌も香港も同じ地名だから…….

そういう感じでやるわけですね.そういうやり方でベストな例を探してきてそれになぞって翻訳するということをやる.それをやりますと,文法規則というものは必要なくなってくる(笑).極端にいうとそういうことになる.たとえば短い文の典型的なもの,I am a boy.とかThis is a pen.とかを計算機の中にいれておいて,それの訳文も入れておいて,ある文が与えられたときにそれがどれに近いかというのをみてそれになぞらえて翻訳する,ということをやる.複文の場合は単文にうまく分解してからなぞり翻訳をする.しかしその分解が難しいんですね.そこのところを文法規則を使わずに,典型的なフレーズを辞書の中にたくさん入れておくわけですけれども,そういうフレーズを使ってここからここまでが一つのフレーズ単位として取り出せるよ,とかいうようなことを発見する方法というのが今一生懸命研究されていまして,だんだんとうまい方法が見つかってきて,だいたいいくんじゃないかというような感じになってきているんですけれども,まだ完ぺきにはいきません.

(質問:佐和)『の』の意味というのはまさに

ローカルにと言いますか，AのBというのだけで判断できますか．

できないかもしれませんね．そこまでは今のところまだちょっと研究が進んでいませんね．それで，この方法はなぜいいかっていうと，比較的クオリティーの高い翻訳が出せるということなんですね．文法規則でやっていると，『の』なら『の』のbehaviorを文法規則で書くといったって，10種類ぐらいのルールは区別して書けるかもしれないけれども，それ以上は難しいのですね．文法規則の場合はどういうふうに書くかというと，『AのB』とあった場合にこのAはどういう種類の名詞かということを考えるのです．人間をあらわす名詞なのか，物質をあらわす名詞なのか，自然現象をあらわす名詞なのか．簡単にいいますと，名詞の意味を考えてspecifyするのですね．そして，Bについてもそういうことをやる．『の』をはさんで，ある意味の名詞とある意味の名詞があると英語の世界ではどういう表現にするべきかというようなことでルールを書くわけですけども，いくらがんばっても20種類ぐらいしかうまくは書けないのですね．書いても，例外というか誤りがいくらでもでてくるのですね．それに対して文を頼りにやりますと，なんとはなしに似たものが見つかって，なんとはなしにある程度うまく翻訳できる．とともに，もしその翻訳結果が，不満足であるという場合の処置が簡単です．例えば『札幌の夏』という表現をこれにしたがって訳したけれど，これは駄目だという場合には，『札幌の夏』というのを例としてつけ加えまして，それにベストな英語の訳を与えるわけです．つまり，一つ新たな典型例としてこのexampleをつけ加えるのですね．そうすると『札幌の夏』は翻訳できるし，『沖縄の夏』も翻訳できるし，『東京の夏』も翻訳できるっていうふうになっていく．

文法規則でやっている場合は，今までの文法規則で説明できない現象が出てきたときに，それを説明するために文法規則をどう変えたらいいかっていうことは非常に難しい．

文法規則は通常500とか1,000とかそれぐらい入れておりますので，ほかのルールに悪影響を及ぼさないように，現在あるこの文法規則をこういうふうに変えようということを考えないといけないのです．これは非常に難しくて時間がかかるんですが，用例翻訳の場合は翻訳できなかったらそれを一つ例として辞書につけ加えるだけでその類の表現は以後ずっと翻訳できる．そして，ほかのところにあまり悪影響を与えないというのでprogressiveに賢くなっていくというプロセスにおいては断然楽だというわけです．ただ問題は，どういう用例をどれだけ集めなければならないかというところに非常な労力がいる．

それをもうちょっと抽象的な形で説明いたしますとどういうことになるかというと，一つの単語だけを見ていますと，先ほどもいいましたように，抽象的に使われたり具体名詞として使われたりいろいろするし，いろんな意味を持っております．一方，一つの文というのは構造的にいろんなambiguityを持っている．川端康成の有名な例だと，『美しい日本の私』というときに，『美しい』というのは『日本』を修飾しているのか，『私』を修飾しているのか何を修飾しているんだというふうなことをrigorousに解析しないと計算機は動きませんのでどっちかに係けるというわけですね．ところが川端康成は，漠然と『美しい日本の私』とどっちに係ってもいいような調子できっと表現したんじゃないかと思いますけれども，とにかく文にするとそういう構造的な曖昧さが沢山でるのですね．一方で，一つの単語を見るとそれにはそれで意味的な曖昧さが沢山ある．そうすると，単語をいくつか組み合わせたあるレベルのフレーズの時に，曖昧さミニマムというところがあるんじゃないかと考えられるわけですね．曖昧な

単語と曖昧な単語を2つ組み合わせるとかけ算によって曖昧さがもっと増えるかというとそうではなくて，曖昧なものが消えていくわけですね．確定する方向にいくわけですね．ユニークに確定するかどうかはわかりません．だけど，2つあるいは3つの単語を組み合わせるとそこで意味がdefiniteに決まるということが非常に多くなるので，そういうミニマムな曖昧性の世界というのをうまく何かの方法で見つけて，それを辞書の中にフレーズ辞書としていれておいて，それを使って先ほど言いましたような翻訳をやるというようなやり方をする．『美しい日本の私』はこれを1つの単位として英語訳を与え，辞書に登録することによって，その内部に含まれる曖昧さの議論をのがれるわけです．こういう考え方の方が文法規則でやっている翻訳に比べてはるかによさそうだというのがだんだんわかってきました．

これは12～13年前に言い出したんですけれども，最近になってようやく世界中の人がなるほどおもしろそうだなというふうになってきて，今流行っています．最終的にそれが本当にいいのかどうかはちょっとわかりませんけれども，そういう世界になってきている．結局これは文法というような厳密な方法論じゃないんですけれども，なんとはなしにいい．で，そのフレーズっていうのは，ある意味でグローバルな世界を見ている．本当はもっとグローバルな世界を見ないといけないんでしょうけれども，例えば3単語，4単語のフレーズの世界を見るぐらいだと，現在の計算機でやれる範囲ですけれども，20単語ぐらいのもの全部のあらゆる組み合わせについて辞書の中に入れておくということは現在の計算機ではできない．広い範囲を見ていくほどいいんですけれども，今のところはフレーズの単位でとどまっているわけですね．それでも厳密な言語理論とか何とか称されているやり方とくらべて結果がよい．

そういう意味では，どうも人間の頭というのは何か似ているということを基本的に発見して，それをなぞって何かアクションをしているのではないかという気がする．私はそういう根本的な考え方を持っておりまして，それをできるだけ言語現象のいろんなところに使って，ある意味では人間的な方法で翻訳をしていけるんじゃないかと思っています．

ですからそういう意味では，いままで考えられておりますような学問理論というものとは相当アンチな形になっておりまして，そんなのは学問じゃないんじゃないか，という非難も結構あると思うのですけれども，エンジニアリングというのは学問だけではないわけです．最後に言いたかったことですが，田中先生の前でちょっとまずいんですけれども，理学と工学は違うんじゃないか．工学というのは，新聞とかいろんなもので見ると，理学のおこぼれを頂戴して，理学が上流にあって工学あるいは技術が応用として下流にあるというふうに言われる場合が多いんですけれども，実は私はそうは思ってなくて，理学というのは物事のエッセンシャルな部分を取り出す，つまり，abstractionのプロセスである．だから，皮肉な言い方をすると微妙なところは全部捨てる．複雑な現象のエッセンスだけを取り出す，ある意味では気楽なプロセスではないか．それに対して，工学というのはそういうことをふまえた上で，なおかつダーティーな現実世界のきめの細かいところをきちっと説明するとか，それを再現するとか，再構築のプロセスを伴うので理学よりも工学のほうが実ははるかに難しいことをやっている．だから，工学をやっている連中は，理学に対して，インフェリオリティ・コンプレックスを抱かずに，胸を張ってがんばらなきゃならないのです（笑）．

(田中)最後におっしゃったことと，ぼくも実は同じようなことを前から違った言葉で言っ

ているんです．理学は，先ほども言いましたように，ことがどうあるかということについての本当か本当でないか，ということをみればいいんで，いつもチェックできるものですから，気楽に好きなことを言っていてときたま当たればいいというんで，確かに気楽な学問だ．その証拠に理学部の人は服装も気楽な服装の人が多いようです．工学は，自然にないものをつくりださなくちゃいけない．そういう意味では全責任があるんだと思うんですね．理学は自然のほうに責任をかぶせることができますけれども，工学は全責任を自分で負わなくちゃいけないから，その点では，きちんと行儀よくまじめにやっていなきゃいけないんだ，というふうなことを前からよく言ったことがあるものですから．それはそれとして，ちょっと伺いたいのですが，さきほど，これは学問じゃないとそういうことをいろいろおっしゃいましたけれども，そうじゃなくて，言語というものを一つのパターンとして，言語はパターンであると，そういうふうにとらえて議論されているような気がするんです．よくは勉強していませんが，私もチョムスキーやソシュールの言語学をいろいろ見てみますと，一番大きな違いは，ソシュールは言語を一つのパターンであるとみなした上で，例えば音の連続というのはまさしくパターンであるとそういう立場にたっていると私には思えるんです．言語学に非常に詳しい人のお話で，こういうことを伺ったことがあります．エスキモー語というのは非常に面白い言語で，普通言語は何か単語からできていると思っているようであるが，そうじゃなしに，一番最初はひとつずつ何かを序列する叫び声から始まった．その叫び声と叫び声の交わるところに単語は生成した．だから，単語を知った上で言語ができたんじゃなしに，そのような一連の叫び声的ないわばパターンとしての一連のなにか文的なものから単語が生まれたというそういうのが言語の一つの生まれかたの型であって，エスキモー語というのは実はそのような意味で単語でできるんじゃなしに，一連の叫び声のようなもので文として我々が述べることを物語っている．そういうところがあるんだということが頭にあるものですから，先ほどの文法からではなくて一連のフレーズから始めるっていうのは非常に面白い．何年か前に伺ったお話とはまた違って非常に面白い．

（質問：佐和）よく日本語で『は』と『が』というのがありますね．主語が『私は』とか『私が』とか．それが，日本語がちゃんとできる人ならばですね，これはあるポジションをみて，本来『は』であるべきところが『が』になっているとこれはおかしいと思うわけですね．しかし何でおかしいのかと．ルールというのはないんですね．そういう例は日本語の場合いっぱいあるんでしょうかね．

あると思います．文と文の関係には非常に多くのファクターがあって，それを厳密に解析できていないのですね．ある一つの文があってその次にどういう文がくるかという文のつながり，つまり文脈解析と僕らは言っていますけれども，そういう，文と文とのつながり関係によって『が』とか『は』とかが微妙に使い分けられていて，それは簡単になかなかルール化できません．先生のおっしゃるレベルに行く前に，文と文がどういう関係にあるかということを機械的にどこまで認識できるかというようなことを少しずつ研究していまして，ある文とある文が，これは机だ，これは椅子だ，というように並列関係にある，とか，あるいは対比の関係にあるとか，彼はどこどこへ行って誰に会った，その人は何だったというように主題が連鎖していっているのかとか，いろいろな文のつながり関係があって，そういう関係をうまく見つけて一つ前の文の中でどれが話題としてしゃべられているかを推定する．

また，『私は札幌へ来ました』という場合には，『私』が文の中心にあるのではなくて，『札幌へ』ということに焦点が当たっていて，『札幌へ来た』ということを言いたいがために『私が』とかあるいは『飛行機で』とかいうのがくっついている．札幌にくるのにいろんな来かたがあるけれども飛行機で来た，という場合には『札幌へは飛行機で来ました』というふうに言うとかですね，そういう表現の違いがあって，語順とか『は』とか『が』とかの微妙な使い方によって，ある文の中において一番焦点の当たっている概念は何かということを引っ張り出す．それができると，その次の文で『は』が出てくるか『が』をつかうべきかというのがある程度わかってくる，というようなステップでやっておりまして，そのために文と文の関係をまず調べて，文の中で何が話題になっていて何にフォーカスが当たっているかということを見つける．それがだいたい今のところ6割ぐらいはできるんですけど，そこから先はなかなかむずかしい．技術の世界では6割というのは何でもちょちょっとやればすぐできるパーセントで(笑)．8割から9割に持っていくのは大変ですね．9割から9割7分ぐらいに持っていくのは死ぬほど大変です（笑)．

(田中) 100 m 競争みたいなものですね．

(皆川)報告要旨の中に今お話いただいたことのほかにも興味ある話題が書かれておりますので，それに関して先生のほうでご説明いただければと思いますが．

それじゃあ，あまりspecificな話をし過ぎましたので，ご質問もしにくかったんじゃないかと思いますので，書いておきましたことの話題をちょっとお話したいと思います．

4年ほど前からわたくしどもは電子図書館というものの研究をやっておりまして，これは言語処理技術とか画像処理技術とか，要するに情報処理技術の総合技術として情報を扱う場合にやってみると非常に面白い技術だと思っております．4年ほど前からやっておりましたが，最近，情報インフラストラクチャを何とかしなければいけないというゴア副大統領のあの話から，日本でも情報ハイウエイという問題に焦点が当たってきて，情報ハイウエイに何十億円というお金を投資して，NTTの光ファイバーは何兆か何十兆かは知りませんけども，投資してゆくといったことになっているわけです．けれども，じゃあそれに何をのせるかというときに，今朝の佐和先生のお話にありましたように，ビデオ・オン・デマンドとか，あんなのをのせてやる，というのが今話題になっているのですけれども，本当に使われるかどうか私も非常に疑問に思っております．

それに対してこの電子図書館というのは，使う人は少ないんでしょうけれども，それが実現するといろんな意味で便利なことになるんじゃないかと思ってやっています．今，インターネットというネットワークがあって，それはある種の電子図書館の一つの初期的な姿であると言うことができます．それから，電子図書館を考える場合には，テキストというのは多次元的な構造を持っているということを認識しなければならない．つまり，1つの本や論文は章とか節とか項とかに分かれているとか，あるいは索引がついているとか，いろんな意味で立体的なストラクチャを持ったものであって，それをうまく計算機の中で表現できなければならない．

なぜかというと，電子図書館においては出し入れする情報の基本ユニットというのは，今までの図書館とはまったく概念が違うということを認識しないといけないからだと思っております．つまり，今までの図書館というのは本一冊が情報の出し入れの単位だったわけですね．つまり，図書番号をつけるとか図書カードを作るとかやっているのはすべて一冊のものを単位にして，それをどこへしまっ

ておいて取り出してくるかということを考えたのに対して，そういう本とか雑誌とかあるいは美術全集とか，そういうものが計算機の中に入れられた場合は，本を単位にして情報を引っ張り出してくるということは必要ない．もっと細かい単位のものも取り出せる必要がある．最近の本は，論文集のように，5人も6人もの著者が勝手に書いた論文が一つの本にまとめられているというのがあって，そういう場合は本の題名というのは何にも役に立たないわけですね．中の，それぞれの著者が書いた論文題名と著者の名前というのが実は一番大事なのです．本にまとめてしまうと著者の名前も出てこないし，その著者の論文の題も出てこない．そういう図書館ではだめなわけであります．これから電子図書館では中のそういったものが引っ張り出せてこないといけない．そうすると，電子図書館における情報の出し入れのユニットは何かということを真剣に考え直さないといけないということになります．その時にテキストあるいは，本一冊というものはどういう情報構造を持っているかと言うことが非常に基本的な問題になるということです．で，その基本ユニットにアクセスしていくということが情報検索になるわけですけれども，それにはいくつものマルチプル・パスでアクセスできる必要がある．

それから，もう一つは今まで言われておりませんでしたが，本とか雑誌とか論文の目次，つまり第一章何なに，第二章何なに，第一章第何節何なにとか，目次に書いてある情報というのが，少なくとも学術論文の場合はテキストのその部分の内容をある意味で代表しているわけですから，目次というものが非常に重要な役目を占めてくるんではないか．今までは，本の表題か，本の表題だけで不満足な場合はキーワードというのをつけていたわけですけれども，キーワードというのは著者がつける場合と図書館司書がつける場合があるわけですけれども，両方ともそれほど満足のいくキーワードをつけることができていなかったのに対して，目次のタイトルはキーワード以上に役に立つということがわかってきたので，目次検索というのをやるのがいいんじゃないかというのがわたくしの主張でありまして，それで今いろいろ実験システムを作っております．

電子図書館の場合はマルチメディアになっている．つまり，絵とかいろんなものがはいっている．今朝は佐和先生のお話にマルチメディアとはどういうものであるかというのがありました．それは社会学的にみたお話だったと思うのですが，わたくしは技術的にみてマルチメディアの本質というのは何かというのを考えております．その場合は，まずいろいろな情報のタイプがある．絵なら絵，音声なら音声，音楽なら音楽，あるいは本とかいろいろありますね．それぞれの情報にはそれぞれのベストな表現メディア，表現形式がある．ただし，それはそれぞれの環境におけるベストな表現形式である．つまり，利用者がどういう環境にいるかによって，ある情報がこういうメディアで表現されたほうがいい，ああいうメディアで表現されたほうがいいというふうに変わり得るということですね．

例えば，電子読書というシステムを作っているわけですけれども，電子図書館の場合は，図書館の中に情報をいれて検索をするというサービスの他に電子図書館のもう一つの大きな機能として電子読書という部分を考えなければならない．今までの図書館は，本を単位にして貸し出したら，それをどういう読み方をするかというのは図書館の責任の範囲外だったわけですけれども，電子図書館の場合は，読む端末装置が計算機的なものですから，そういうものもきちっとユーザに提供しなければならない．そういう時に電子読書としてどういう機能を提供すべきか，ということが

あるわけですが，その提供するスタイルとして，読者のほうが何をどういうスタイルで読みたいかということによって，表現のメディアをそれぞれ変えていかなければいけないということを考える．

一番簡単には，本を読むというときに目で文字を読んでいく場合と，くたびれたから耳で聞かせてほしいというわけで，音声合成装置を使って発話してもらって文を朗読して聞かせてもらうという読み方をする，ということがあるわけですね．あるいは，カー・ナビゲーションのシステムでは地図が出てくるわけですけれども，地図を見ながら運転するのはあぶないわけですから，その地図を出すとともにそれを音声になおして，地図を解釈して，次の角を左に曲がりなさいとか，もう100mほどいくと何とかがあるからそこを右に曲がりなさいとか，そういうふうに図から文章を合成して出す．そのためには図を認識しないといけませんけれども，地図を認識してそれを文章にして音声になおして出す，というようなこともシチュエーションによっては必要なわけでありまして，それはユーザあるいは利用者あるいは読者のいろいろな環境によってベストなメディアというのがかわる，そういう変換技術が必要である．そういうことにマルチメディアの本質があるのではないか．音とか画像とか何とかというのを多数ごちゃまぜに寄せ集めて同時に出せばいいとか，そういうものではない．一般にはそういうものをマルチメディアというように理解されているんじゃないかと思うんですけれども，わたくしの理解はそうじゃなくて，いまここで言いましたような理解というのがマルチメディアの本質じゃないかと，こういうふうに思っていまして，この考え方で電子図書館の実験システムを作っております．

電子読書については，電子ブックというのを最近売っていますが，これはICカードに小説なんかがはいっていてそれをカチャッと入れると小説が出てきて読めるというようなものです．これから出てくるものに，電子新聞というのがある．電子図書館と関係があるかないかわかりませんけれども，電子新聞というのを大いにやらなければいけないんじゃないか．最近アメリカでも出て来つつあるようですけれども，まず紙を使わないようにしないといけません．地球の森林資源が毎日毎日ものすごく減っていくというわけですから，紙に印刷しない新聞，それを電話線でもって各家庭に配達する．配達人はいま非常にいなくなってきていますので．要するに，各家庭に自分のほしい新聞記事を送ってもらう．それはまあ，やればできる話．

それから，複数テキストの同時並行読書というのをやらなければいけないのです．つまり，参考図書を何冊も広げて拾い読みをしながらいろんな判断をする．あるいは辞書を引きながら本を読むのも二つの書物を同時に読んでいるということになりますので，そういうことができないといけないということとか，それをうまく実現するためにはいろんなテキストの中の単語どうしをダイナミックにリンクしないといけないとか，しおりを挿入できないといけないとか，メモ帳がつけれなきゃいけない，切り抜きができないといけないとか，そういうふうな機能をソフトウェア的に実現する．それから，「本の表題は忘れたけれども，大きさはA4の本で表紙が赤くてぶ厚かったなとか，あの辺に入っていたなとか」，そういう覚え方をしていることが非常に多いわけですね．本や著者の正式な名前を覚えていることはあまりありませんので．そうすると，今の図書館でも電子図書館でも，タイトルをきちっと入れてくれないと動かないというのでは全然役に立ちません．だいたいこんな種類の本だったとか，こんな表紙の本だったとかいうので引っ張り出せないといけないので，そういうfacilityを提供する．だか

ら，「赤い表紙だったと思うよ」とか言うと赤い表紙の本ばかりを出して，この中のどれですかというと，これだということになってそれが出てくるとか，そういうふうな機能を持たせる．

本を読んでいてこの単語がわからないというとき，この単語をポイントすると英和辞典を引きましょうか，百科事典を見ましょうかというのを聞いてくる．この場合は英和辞典を引いたらよろしいというと，この単語の説明がぱっとでてくるとかですね．百科事典のこの項目が見たいというときは百科事典のところをピッと押すと，百科事典の説明がばっと出てくるとか，そういうふうなことができないといけない．そういうシステムを今作っておりますが，そこでの技術的内容として一番面白いのは，やっぱりわたしとしましては，メディアをいかにうまく変換して活用するかということと，10万冊とか100万冊ぐらいの電子化された図書館を考えたときにどういうフレキシブルなretrieval（情報検索）をやることができるか．つまり，さきほど言いましたように，著者名とか本の題名とかがきっちりわかってないと引けないというんじゃ全然だめなので，過去に見たものについては「だいたいこういうふうな感じのものだったよ」といった質問を許す．これからのものについては，例えば「最近の計算機の進歩の状況はどうでしょうね」と聞くと，進歩の状況といっても，いろいろな進歩があるから，それを計算機のほうからユーザーに問い合わせてきて，あなたは素子に関する進歩を考えているのか，ネットワークに関する進歩を考えているのか，どういう意味で進歩という言葉を使っているのか，というのを聞いてくる．といっても，進歩の意味は何ですかといって計算機が人間に聞いたんじゃ人間のほうがどう答えていいかわからなくなるのでガイドしていくわけですね．こういう意味で聞いているんですか，ああいう意味で聞いているんですかというふうに，選択肢を与える．その選択肢がどれだけリッチなものに作れるかというのが一番の問題なんですけれども，そういう漠然としたことを取り扱う．

これも曖昧アプローチみたいなものですけれども，要するに人間に機械が対応していく場合には，人間は曖昧である．その曖昧というものを前提にしてそれをうまくクリアな格好に持っていく，そういうプロセスを入れて電子図書館のシステムを作ろうというんでやっております．ちょっと雑な話になりましたけれども．

(田中)こういう試みをそれに加えていただくことはできますか．前に私ちょっと見たことがあるのですけれども，目次検索というのが出てきましたが，目次のかわりに本の索引が入ると非常に良いと思うんです．これは大変ですけれども，300とか500の専門書に限って，それを索引で検索できるようにしますとよく似た索引が得られるわけです．それを先ほどおっしゃったようなシステムを使ってある索引かそれに近い表現を入れるとその索引とそれに似た索引が出てきて，それに関する事項がどの本の何ページにあるということが出てくると，それは仕事の上で非常に便利ですね．

『岩波情報科学辞典』というのを作ったんですけれども，そこでそういうことを実現しております．この辞典が計算機にハイパーテキストの形で入れてあります．この辞典では，用語の木と称して，見出し語がツリー構造的に作ってあって，この単語の意味を知りたいと思ってここをクリックしますとその単語の意味や説明が出てきます．それからこの文章の中を読んでてこの単語の意味がわからないと言ったら，また別のウィンドウが開いてその単語の説明がでてきます．その単語の説明を『岩波情報科学辞典』の項目でみて，もうちょっと詳しく知りたいという場合には他の本の中の説明部分が指示されます．例えばス

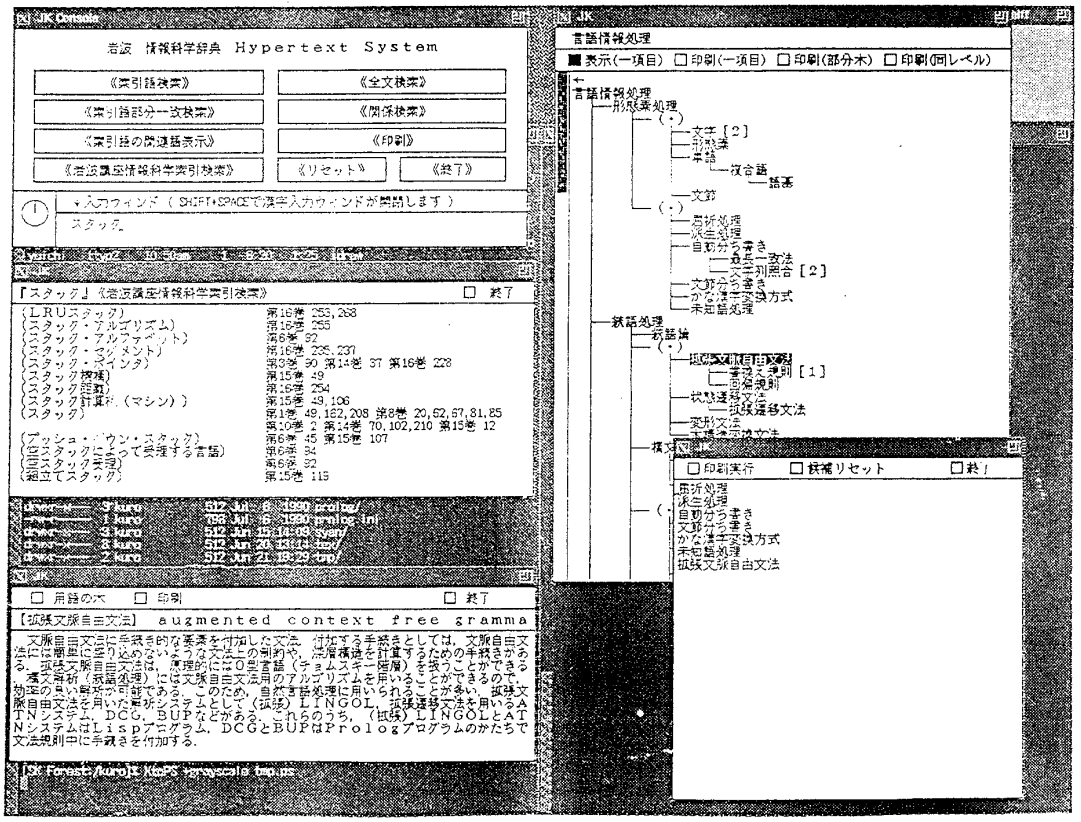

図9

タックという単語を指示しますと，それは『岩波講座情報科学』というのが24巻出ておりまして，その第何巻の何ページにスタックという単語が詳しく説明されているというのがばっと出てきます．これは索引から引くようになっていまして，スタックといいましてもいろんなスタックがありますので，求めている場合のスタックは『岩波情報科学講座』の第何巻の何ページに載っているというわけです．(質問：田中)そのシステムは市販されているのですか．

わたしのところで3，4年前に作りまして，ほしい人には岩波のOKをとってお分けしていますが，CD-ROMにして売ったらどうですかと言っているんですけれども，なかなかそれを岩波がやりません．もしお入り用の方がありましたら僕らのところからSUNワークステーションの上で動きますのですぐお渡しします．かなりの人にお分けしております．こういうようなのをいくつもの本について自由に相互参照できるように電子図書館の場合はしないといけない．

長尾講演に対するコメントと質疑

司会：ご質問，ご意見等ございませんでしょうか．

(1)それぞれの情報にはそれぞれの環境における最良の表現(メディア)のし方がある.
(2)幾つかの情報の表現(メディア)の組合せによって最良の表現が実現される.
(3)情報が最良の表現メディアで表現されていないときは最良の表現(メディア)に変換する必要がある.

図10 マルチメディアの本質

佐和：さっきのOHP(図10)を見せてくれますか．2番目の文章は『の』が何度も出てきますね.『いくつかの情報の表現の組み合わせによって最良の表現』，これですね，ローカルに見ていくと『いくつかの情報』の『いくつか』というのは『情報』に係ることににになってしまいますよね．これが何に係っているのかというのは非常にわかりにくいんですよ．どうも正しいのは，たぶん『表現』に係っているんですね．とにかく，にわかにも人間にもにわかにはなかなかわかりにくいんですよ．

長尾：これどういうふうになりますかね(笑)．

佐和：うるさくいえば，『表現』が複数個であって，情報のいろんな表現の組み合わせによって最良の表現が実現される，というんですね．そうすると，ベストな表現というのが，そこは当然単数ですね．表現の組み合わせも一つの表現と呼ぶということですね．

長尾：この文を解析してみないといけませんね(笑)．

原：『情報の表現』というのが一つのフレーズで，複数の情報の表現があってそれの組み合わせというふうに解釈すべきでは.『いくつかの』も『情報の』も『表現』に係っていて，『情報の表現』でワンフレーズですよね．例えば画像情報で表すとか，音声で表すとか，それが複数あって，それの組み合わせ．

佐和：だから『表現の組み合わせ』というのも表現と呼ぶというわけですね．そこでナイーブに考えればですね，『表現の組み合わせ』というのは本来ならば別の名詞になってほしい感じですね．

土屋：ただ，今日の長尾先生のお話をそのまま使えば，これでだいたい何を言いたいかわかるわけですよ．そのときに，今ナイーブに考えればというのは全然ナイーブでなくて．

佐和：いや，コンピュータですから．コンピュータがこの文章を意味を理解しようとしたときの話ですよ．

土屋：今の技術では非常に困りますけれども，人間は，ナイーブに考えればだいたいわかるんです．で，さっき先生も後になってからもう一回出すとおっしゃったように，その場の時は，「あ，そんなものかな」，と思ってさっと通してしまうわけですよね.ですから，そういうようなことを可能にすることが重要なんだろうと思います．つまり，これがいったいどういう意味かというようなことを，丁寧に分析して，細かく解析し尽くしたところに必ず何かあって，それをコンピュータは最終的に分析できるのだという前提に立ってやっていたのが今までの研究のスタイルだと.それに対して,今日長尾先生のお話にあった研究の手法の一つの非常に大ざっぱなまとめとしては，もしもこれでコミュニケーションが通じているのであればその解析のレベルというのがうまく表現できればいいんだろうな，ということではないかというふうに伺ったんですが．

長尾：そういう感じなんですね．つまり，それ以上細かくanalyticにやると非常に問題が起こって，あっているとかあってないとかいうことになるけれども，そこへ入らずにわかるんだったら，そこのところで分析を止めておく，そこのところで次のアクションのほうに行くとかそういうふうなことも．

田中：ランゲージというのは聞きたいように

聞くというところがありますから，これは確かに6つか7つぐらいの受け取りかたがあると思うんですけれども，そのそれぞれに応じてそれぞれ処理をしているというそういう現状にはなっているような気がします．で，機械処理というのは人間のまねを別にしなくてもいいわけですから，そのかわりは可能な解釈を全部出すというのも一つの方法だと思うんですけれども．

長尾：可能な解釈を全部出すとすると，それは膨大な数になってしまって，ほとんど意味をなさなくなってしまうわけですね．

土屋：でも機械にやらせると，我々が考えるよりも曖昧さは増しますよね．我々にとっては同形だけども機械にとっては別の形というのが出てきちゃうんで．

田中：ただ，普通は考えない独自な理解をする人が，100人の内に1人か2人はいるわけですね．100人のうち1人か2人はその場その場の局面でありうるのではないかと．

佐和：例えば，霞が関用語辞典なんてものがあって，つまり，「検討する」というのは何もしないということであるとかですね，そんなつもりで言ったんじゃないということがよくありますね．

皆川：まだ翻訳業者が失業したという話もあまり聞いたことがないんですけれども翻訳システムというのは，実用化の段階でいうと実際にはどの程度までいっているのでしょうか．8年ぐらい前に雑誌で読んだ内容では，その当時は，大ざっぱに逐語的に翻訳をしておいて，翻訳業者の方が細かい部分を修正して文を仕上げるという感じで使われていたことを記憶しています．それで，現在はどのような状況でしょうか？

長尾：現在も同じ状況ですね．多分現在は10社ぐらいのメーカーが機械翻訳のシステムを売っていまして，多分1，2万セット売れていて，その中で，ほこりをかぶって使われていないというのが山ほどあって，ちゃんと使っているというのはたぶん数百セットから千セットぐらい日本中ではあるんじゃないかなと思います．その中で，翻訳会社が機械翻訳システム，英日あるいは日英ですけれども，を使っているのはかなりあると思います．かなり使われているいうのは漠然とした言い方ですが，まあ100セットから200セットぐらいはあるのではないでしょうか．どういう使われ方をしているかといいますと，典型的には技術文書であるとか，何かのかたい報告書であるとかそういうふうなものを翻訳するのに向いています．しかし，その場合でも専門用語辞典あるいは専門用語集を別途用意して機械翻訳システムに入れる必要があります．たぶん1万単語とか2万単語ぐらい入れる．例えば，医学のある分野のテキストを翻訳しないといけないという場合はそういう辞書を買ってくるか，自力で作って入れる．そういうふうにしないと，失敗の翻訳がたくさんでます．

それでも，機械で翻訳させますとめちゃくちゃな翻訳結果がでますので，そこで人間がpost-editingと称して修正するというプロセスをやる．これはどうしてもやらざるを得ません．それをやりまして納入するということをやっています．うまく使っている翻訳会社は少なくとも何社かあっていろんな会合で発表してますけれども，翻訳スピードは，翻訳するべきものが来てから翻訳をしてcustomerにちゃんと印刷した物を返すまでの時間ですけれども，人間だけで翻訳して返す場合に比べて機械翻訳を使ってやるとpost editingを入れても，スピードは少なくとも2倍は上がっていると言っています．コストはすべてを考えて，これは使い方によっていろいろ違うらしいんですけれども，うまくやると4割ぐらいになる．つまり6割減るという人もいるし，逆でだいたい6割ぐらいになって4割のコストダウンになるという人もいま

す．つまり，速度は2倍ぐらいになってコストは半分ぐらいになる．うまく使うとそうなる．だけど，下手に使うとだめだということになります．

機械から出てきたものを人間の手で直さないといけないのではだめじゃないかという質問がよくありますが，それでは人間が翻訳している場合はどうかと考えると，ヨーロッパ共同体とか，どこの翻訳会社でもそうですけれども，だいたい一番最初は粗訳(あらやく)というのをやるわけですね，だれか比較的新入りみたいな人がですね．で，その粗訳した文章をベテランの翻訳者の人が読んで直すというプロセスを人間の翻訳の場合も必ずやっているわけですね．このように，人間の場合にも2度やっているわけですので，機械翻訳の場合に post editing と称して2回目を人間にお願いするというのは今の段階ではやむを得ない．そんな状況です．

さきほど述べましたような用例を使ってうまくアナロジカルな形で翻訳するという実用システムは今のところありませんが，今メーカーはその方式で一生懸命実用の翻訳システムを開発中です．この方式のシステムはおそらくあと2年ぐらいしたら出てきて，そしてそれをいろいろと改良して，20世紀の終わりか21世紀のはじめには今までのやり方とはまったく違った形の実用システムとなるでしょう．しかも先ほど言いました単数複数とか，あるいは the をつけるとかつけないとかいうことも6，7割ぐらい解決できてという，ちょっと新しいスタイルの翻訳システムが21世紀のはじめには出てくるんじゃないか．そうすると，クオリティーはいちおうワンステップ上がるんだろうと思いますね．

現在の機械翻訳システムは，どんなきたない翻訳を出しているのかという疑問が生じるでしょうが，そもそも原文自身が非常に曖昧で，今，佐和先生からご指摘があったように人間でも「はてな」というのがあるので，原文がいかにクリアに書かれているかというのがキーポイントです．比較的短い文でクリアな書き方がされている場合は post editing も非常に簡単に済む．そういう場合ですと，例えば，わたくしどもの大学の学生に翻訳させるのとそう違わないのが出てくる場合が多い．といいますのは，例えば原子力に関するテキスト，あるいは医学に関するテキストがあった場合に，それを電気工学の学生に訳させたら，これはもうまったくチンプンカンプンで，機械がやるのと同じように読んで機械と同じように単語の変換をやって翻訳しているだけで意味内容はわからないのですからね．あとは，『てにをは』をうまくつける．『は』と『が』は人間の場合工夫してつける．そんな程度しかできませんので，内容を知らない学生に翻訳させると機械とそうかわらない．

佐和：英語から日本語と，日本語から英語とだと英語から日本語のほうがやさしいのですか．

長尾：ええ，はるかにやさしい．それはやっぱり日本語の文体というのが省略が多いし，だらだらと長いし非常に微妙ですからね．

佐和：二つの言語間の翻訳で，片方はほとんど英語なんでしょうけれども，自動翻訳機が一番よくできているのは英語と何語ですか．

長尾：英語とフランス語の間とか英語とドイツ語の間の翻訳はかなりいいようですね．

田中：英露は？

長尾：英露は，最近あまりヨーロッパでも使われておりませんので，なんとも言えないのですけれども，過去にいろいろ行われました．やっぱり英仏とか英独に比べると悪いようですね．ですからそれと同じように，日本語と朝鮮語の間の翻訳システムが作られつつあるのですけれども，これはかなり楽なようですね．

佐和：しかし文章の翻訳というのが難しいと思うのは，これは僕自身の経験なんですけれども，ある翻訳をする人間について下訳を高

校の英語の先生とか，出版社の下請けみたいなのがあって映画の台詞の翻訳とかをやるんですね．そういう高校の英語の先生とか，英語学校の先生とかがある経済に関する本をやったんですね．もうむっちゃくちゃですね．バックグラウンドの知識があるかないかでものすごいですね．だからそれはちょっと計算機には無理ですから難しいですね．

長尾：ですから計算機の場合は，経済用語などの辞書データを沢山入れるとか，用語だけじゃなくて先ほどちょっと言いましたように，経済用語のある種の短いフレーズをできるだけたくさんいれてそういうような特有の表現を扱えるようにしないとめっちゃくちゃですね．

田中：配っていただいたプリントに情報の価値について言及しておられますね．私は情報の価値というのはどこでどのようにして生成するのかという点に大変興味を持っています．実はそれを考える上で岩井さんの貨幣論が役に立つんじゃないかと思っていたんですが，たまたま，岩井さんの貨幣論の書評を朝日新聞で読みまして，岩井さんのその本を購入しまして何度か勉強しましたけれども，そもそも情報の価値というのはどこからどのようにして生まれ，あるいは何に基づいて生成してきているのか．そういうことについて明日の追加の報告の時に言及していただければ，教えていただければと思います．