

機械翻訳システムの間言言語

The interlingua for a machine translation system

村木一至 亀井真一郎 野村直之
MURAKI Kazunori KAMEI Shinichiro NOMURA Naoyuki

NEC Corporation

1] はじめに

言語理解システムや、機械翻訳システムに於てテキストの情報内容を盛る器としてどのような機構を持った形式言語を用意すべきかについて定説はない。ところが最近、CICCの多言語間翻訳プロジェクトやEDR日英電子化辞書プロジェクトなどの目的、目標を達成するために、中間言語ないし概念記述言語が必須の道具と認識されるようになり、徐々に具体的な提案がなされるようになってきた。理想的な中間言語が一朝一夕にできあがるとは考え難く、それを得るためには多方面からの研究が継続的に具体的検証と平行して続けられなければならないと考える。

筆者らも、中間言語方式による多言語翻訳システム開発のため中間言語を開発・利用してきたが、その設計法ないし中間言語自身の内容に関してはいままで報告する機会がなかった。

本稿では実際に現在稼働中のシステムに於ける中間言語(PIVOT)の基本構成を説明し理想的な中間言語設計・開発の資料の一つに供したい。

2] 何を目指したか

中間言語ないし概念構造記述言語と呼ぶものは人為的に作る情報表現の器である。また、目標とするのは翻訳という情報伝達が可能な文内容表現を担う言語の開発である。形式論理(例えば述語論理、様相論理)は意味内容をきっちり表現

する唯一の道具であると考えられてきた。しかしながら、例えば「ペンギンは飛ぶ」という文の言語翻訳に於て、中間言語が事実の真偽を明確に記述できる必要があるかといえば、必ずしもそうではない。与えられた入力文の記述内容を伝えることが必要であっても、「そんなばかな?冗談でしょ」といった内容の解釈結果を伝達する必要性はない。つまり、ある状況(モデル)において言明の真偽を判定する枠組みとしての形式論理以外に、機械的に翻訳を行うための文内容を盛る形式言語でより扱いやすい言語が設計できないかと考えた。

中間言語の機能としては、一方の言語の単語表現が他方の言語の単語表現にうまく移し換える為の情報の橋渡し機能を持つ必要がある。言語の文化を担う部分を端的に著す単語表現は、一般に複数言語間で単語レベルの一对一対応を持たないことがある。これを解決するためには、lexical-decomposition、exhaustive-listingと言った戦略が考えられる。

Schankの動作の意味記述法における14 primitivesモデルは前者の戦略の立場である。14種の独立な意味(意義)素の組合せによって全ての動作の意味を記述できるというものである。それらの意義素がすべて言語独立で且つ動作に関する諸々の意味を表現できるときに各言語の文化差異をも吸収する規範表現を与える期待がある。しかし翻訳システムにとっては異なった言語間で容易に双方の単語表現に対応つけられるものをたった14種の語義に分解してその組合せで表現しなければならない積極的な理由はない。

3),4)

岡田らは後者の戦略に従って動作の基本的

なプリミティブを数え上げた。語彙のリストアップとその分析的研究から、約2000の動作意義素(単純動作概念)を抽出し、それに直接対応しない動作単語表現はその組合せにて表現することを提案している。筆者らの中間言語「PIVOT」の語彙設計はこの戦略を採用している。^{1),2)}

後者は存在する各単語表現そのものにそれ固有の独立した意味が対応していると考えられる立場である。筆者らは、他と差別的に認識すべき、且つ認識できる意味を担う単語表現は、各文化に於てそれが如何に頻繁に情報伝達行為中に意識されることがあるかに依存して作り出されてきたと考える。また、既に存在する単語表現が担う意味とその差異が識別可能な意味が存在する時、

- 1) もともと存在する意味を含み、新たな差と認識できる意味を加えた複合意味を担う単語表現をあらたに作るか、
- 2) 既存意味に対立する意味を担う新たな単語表現を作り出す、

のではないかと考えている。そう考えると、中間言語の語彙(CONCEPTUAL-PRIMITIVE CP)は例えば14種の意義素である必要はなく、またそのように決まった数の物であることは不可能である可能性がある。そこで、単語表現のリストから、少なくとも各言語で認識できる単語の担う意味に対応してCPを設計することで最終的に望むものが得られると考えた。さらに、そこで抽出したCPを用いて複合的概念の分析的PARAPHRASEが可能であると考えた。

さらに、複合的概念を素な語彙を用いて表現する構造記述機能が必要であり、それは内容を両言語表現間に橋渡しする機能でもある。その方法としては、仮に名付けるとsemi-translation、およびparaphrasingという2手法が考えられる。

前者はある単語表現があった場合にその直接対応する単語表現が他の言語にないとき、上位概念、関連概念等による次善対応付けで対処する方法である。情報伝達という目的から考えると、伝えたい内容をその読み手の身近な言葉で表現できるので、ある場合には伝達目的をより効果的に達成できる可能性もある。

後者はあくまでその直接対応がない表現を目標言語において言い替え説明するという方法である。どちらの方策を採用しても、中間言語としての語彙を網羅的に数え上げるならば、両手法は選択的に組み合わせる用いることができる。PIVOT設計では、paraphrasingによる構造的な対応付けを基本として採用した。インプリメンテーションでは両者を効果的に組み合わせている。

PIVOTは当初日本語・英語と言う2カ国語の言語データを基に設計・開発し、日・英翻訳システムに利用しながら強化してきた。最近スペイン語や韓国語やフランス語などに関しても同言語の有用性の検証を開始した。それは、緒についたばかりであるが、語彙セットに関しては殆ど手を加える事なく利用している。

以下の節ではまず、PIVOTの言語仕様における語彙と文法について説明し、その後で、中間言語にとって最も基本と思われる特約語彙を具体的に例示し解説する。

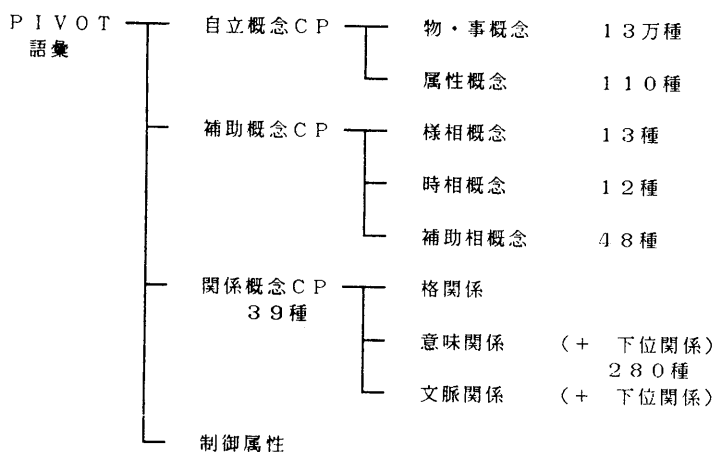


図1 PIVOT構成

3] 構成要素はなにか

P I V O T は言語のリソースとしてその文法(書式)と図1に示す各種の語彙(CP)とから成る。各種語彙は原則辞書に定義される物で、ほとんどの場合表層に対応表現を持つ。

語彙は大きく分けて4種類ある。以下の説明中、“大文字アルファベット+#”はP I V O Tの語彙CPを現すものとし、“用語”語彙”と”CP”は同一意味とする。

第一は自立概念CPと呼ばれる物で”A P P L E #”、“W A L K #”等の事、物に関する概念と”F R E Q U E N C Y #”、“C O L O R #”、といった属性概念、“Y E L L O W #”、“H I G H #”といった属性値概念を表現する語彙である。現在、日本語・英語単語見出し各々9万に対して総数約13万種類ある。

第二は文章中に出現するCP間に文章中で指定される意味の関係を記述する語彙であり関係概念CPと呼ぶ。“A G E N T #”、“T I M E

表1 関係概念リスト

NO	CP	Full-spell	BRIEF DESCRIPTION	EXAMPLE SENTENCE
1	OBJ	object	Object of a Predicate	John <u>hit</u> the desk with his fist.
2	AGT	agent	Subject of an action	He <u>gave</u> me his books.
3	CAU	causer	causer	She <u>let</u> him leave.
4	EXP	experiencer	Subject of feeling, sense	They <u>suspect</u> that he's the murderer.
5	INS	instrument	tool, instrument	The computer <u>solved</u> the problem.
6	MEA	means	means, method	She <u>persuaded</u> him to stay with a kiss.
7	BEN	beneficiary	beneficiary	They are <u>working</u> for me.
8	LOC	location	place	I saw <u>it</u> under the table.
9	TIM	time	time	We haven't <u>seen</u> land in 20 days.
10	SOR	source	source, starting point	John <u>left</u> town.
11	TAR	target	target, destination point	John <u>spent</u> all his money on clothes.
12	PRT	participant	companion, accompanied thing	They <u>married</u> Taro to Hanako.
13	CAP	capacity	role, function	He <u>attended</u> meeting as a leader.
14	FCS	focus	focus	He wrote a <u>book</u> about Japan.
15	MAT	material	material	Their <u>house</u> is built of wood.
16	ELM	element	element	<u>Japan</u> consists of four islands.
17	POS	possessor	possessor	This is my <u>house</u> .
18	POF	part of	part of the whole	The front of <u>the car</u> was destroyed.
19	NUM	number	number	<u>five meter</u> ...
20	NAM	name	name	The Yamanote <u>line</u> ...
21	NMOD	noun modifier	succession of noun	The magnetic <u>discs</u> ...
22	ATT	attribute	attribute	What is your <u>shoe size</u> ?
23	VAL	value	value of attribute	<u>The length</u> of the axis is 30 meter.
24	MOD	modifier	adverbial modifier	People struggle to <u>live</u> better.
25	QUNT	quantity	quantity	She bought two <u>bedding sets</u> .
26	EQ	equality	equal relationship	That must be a <u>whale</u> .
27	APP	apposition	apposition	<u>Memory devices</u> , such as magnetic discs
28	REF	reference	determination of reference	<u>Which man</u> did John say saw him?
29	CPQT	quantity of comparative	quantity of comparative	He is 5 cm. <u>taller</u> than I.
30	MDQT	quantity of modification	quantity of change, continuation	He <u>walks</u> 5 km.
31	RLBS	base of relativity	base of relative concept	He looks at <u>the outside</u> of the window.
32	DCMP	dummy case for comparative	than of comparative	He is <u>taller</u> than I.
33	MODS	sentential- modifier	sentential modifier	He is, in a word, <u>an idiot</u> .
34	PAR	parallel	parallel	<u>John</u> and Mary heard the news.
35	GOA	goal	goal, aim, purposive	They <u>stopped</u> in order to rest.
36	CON	connection	connection of affairs (in a sequence of time)	Returning to my office, I <u>slept</u> .
37	REA	reason	reason	The rain <u>made</u> us seek shelter.
38	CAS	case	establishment of case	In <u>case of</u> the accident, <u>call</u> me.
39	CASA	case assumption	assumption	If he comes, <u>call</u> me.

＃”、“REASON＃”等の格関係、意味関係、文脈関係等を表現する2項関係記述子からなる。また、意味関係および文脈関係はその下位分類を持ち、例えば、“UNDER＃”、“BEFORE＃”等は時刻、場所関係を表現する下位関係CPの例である。この中には関係概念“PARALLEL＃”の下位関係“AND＃”、“OR”等の論理的CPも含まれる。関係CPは現在39種類、同下位関係CPは約280ある。

第三は補助概念CPと呼ぶ様相、時相、補助相からなり、話者態度、動作様態等を表現する語彙であり総数73種である。

第四は制御属性と呼ばれ、主題、焦点等の区別、意志、無意志の別、状態、動作の別、並列の範囲のマーク、分野・語用の別と言った属性からなり、これらは、原則的にそれらの値に対応する表層単語がない。入力文章のスタイルによって話者が述べたい事実内容の表現の仕方（言い回し）を出力言語に伝達する機能を担っている。

PIVOTの基本文法は非常に簡素である。それは以下の3つの約束からなっている。

中間言語表現書式

1) 構造

- 1.1 中間表現は有効グラフでありLOOPを含まない。
- 1.2 関係概念CPと補助概念CP、自立概念CPはグラフパス上で交互に出現する。

2) ノード

- 2.1 必ず一つのCPないし複合CPを含み、そうでないものはノードでない。制御属性はそれだけでノードとは見なされない。
- 2.2 複合CPは関係CP1=関係CP2=下位関係CPないし関係CP=下位関係CPで与えられる。
- 2.3 ノードはCPの他に種々の情報を持ち、それは制御属性、下位分類属性CP、補助属性CPによって記述される。

3) 有向ポインタ

- 3.1 有効ポインタの方向は、格関係概念CP毎に定義されている。それを満たさないものは有向ポインタでない。

複合概念CPは、例えば次のような書式で記述する。書式例1の意味は、目標を表現する格関係“TARGET＃”は意味関係“LOCAT

ION＃”の下位関係“ON＃”に格を割り当てる。つまり表層表現「。。。の上に」に対応する。書式例2は英語では表現「until」に対応する複合関係CPである。

例1.

TARGET＃=LOCATION＃=ON＃

例2.

TARGET＃=TIME＃=BEFORE＃

3.1] 自立概念CP

自立概念CPは各言語の単語表現が持つ各意味を表現し、関係概念CPが乗り物を記述すると考えるとそこに乗る人に対応する。自立概念CPは大きく分けて2種類あり、物・事概念を表現する語彙と属性を表現する語彙がある。属性概念はさらに2種類に細分類され、1)属性名概念、2)属性値概念から成る。属性名としては“COLOR＃”、その値としては“RED＃”、“YELLOW”といった離散的値を採るものと、程度・度合を現す属性名、例えば“FREQUENCY＃”とその値“ALL＃”、“MANY＃”、“NEUTRAL＃”、“#A-FEW＃”、“NOTHING＃”といった決まった連続的値を取るものがある。現在それらをあわせて約110種ある。

書式上、属性名は他の自立概念CP、補助概念CPと次節で述べる関係概念CP“ATTRIBUTE＃”によって結合され、属性値とは関係概念CP“VALUE＃”によって結合されなければならない。

3.2] 関係概念CP

関係概念CPは表1に示すように39種類の格関係、意味関係、文脈関係CPと表2にその一部を示す下位関係CPとからなる。関係概念CPは、大きく分けて3群に分類できるが、その境界は必ずしも明確に決めていない。表1は左から関係CP、そのSPELL、簡単な説明、例文からなる。

第1群は格関係と呼ぶものであり、“SOURCE＃：始点”、“TARGET＃：目標点”を代表とする同表1~14の14種から成る。

また、第2群は意味関係と呼び、“LOCATION＃”、“TIME＃”、“ELEMENT＃”等を代表とする同表15~30の15種類から成る。

第3群は"CONNECTION#：動作連鎖"、"CASE#：場合条件"等を代表とする同表31～39の9種類から成る。

実際、意味関係、文脈関係はかなり広い意味をカバーしており、各言語は表1で定義する語彙以上に時間、場所、場合条件などを細かく表現する能力を持っている。そこで、意味関係及び文脈関係には、その下位属性として下位関係CPを設定することにより、関係表現の多様さを表現できるようにした。表2はその一部を示している。

を補助的に記述する意味を担うもの13種からなる。同表中、英語対応表現が空白であるのは、英語表層として独立且つ適切な単語表現がない場合を意味し、翻訳においては何等かのParaphrasingの必要性を示唆する。

時相表現は現在12種類のCPを設定しており、その中は3種類に細分類される。表4はおのおのの語彙リストを示す。また、対応英語が空欄であるのは表3と同じ意味である。

表2 下位関係CP「0***」の例

文脈関係CP： PAR	
OADDITION	He wanted many things <u>besides</u> a new car.
OAVSPAR	He tried very, <u>albeit</u> not very hard.
OINSTEAD	They planted tulips <u>in place of</u> roses.
OAND, OOR, OXOR	
文脈関係CP： REA	
ORESULT	He managed to survive <u>as the result of</u> his effort.
ODUETO	日程の <u>関係から</u> 、前途を断念した。
意味関係CP： TIM	
ONEXT	論文を書き上げた <u>その次に</u>
OEACH	彼は論文を書き上げると <u>その都度</u> 休暇をとる。
OEACH	その話は、 <u>毎回</u> 聞かされる。
ONEXT	明くる日曜日、 <u>帰国</u> した。
OBEOFRE	彼より <u>前に</u> そこへ着いた。
OAFTER	They stayed <u>beyond</u> the summer.
OSOONAFTER	そこへ着くや <u>否や</u> 直ちに指揮をとり始めた。
OSIMULTANEOUS	<u>At the same time as</u> Mary left, John arrived.
OIN	<u>3時間</u> でできる仕事。
ODURING	彼がここに <u>いるうちに</u> 仕事を片付けよう。
OFUTURE	He will speak <u>in</u> a few minutes.

3.3] 補助概念CP

補助概念CPは様相、時相、および補助相概念からなっている。この中には話者の態度(Propositional-Attitude)、物事の起こる様、ある様(Temporal, Aspect)を表現する語彙を含む。表3は様相(話者態度および論理的評価を表現する)と補助相概念として数え上げた64種のCPをリストしたものである。同表に於てMDLは様相と呼ばれるにふさわしい話者判断にかかわる内容を表現するもので、13種類ある。C S H Cは、出来事に関する表現の客観的判断記述、客観的叙述

4] 評価、および検討

中間言語PIVOTはその開発から約5年間を経過した。全てが一時に揃ったのではなく順次に3度の大きな改良を加えた。その改良に於ては無計画に設定されたCPの統合が主なものであった。自立概念、関係概念はその種の改良対象の最たるものであった。関係概念においてはシステム辞書の意味構造記述言語としても利用されるので、大きな改版は膨大な関連改版作業を発生させる。それを少しでも軽減するために当初より関係概念は下位関係概念との組合せによって表現す

表3 様相概念と補助相概念

MDL 13種

CP	Japanese surface samples	English surface samples
XLOOK	そうだ、*ようだ、らしい	SEEM, LOOK LIKE ING
XSEEM	かのようにだ、*ようだ	SEEM AS IF, APPEAR AS IF
XINFE	だろう、であろう、*う	WILL, MAY, MIGHT
XPRED	ことは言うまでもない	MUST
XPROP	はずだ、はずである	SHOULD
XGUES	かもしれない、かねない	MAY, MIGHT, COULD
XNEED	なければならない、べきだ	NEED, MUST, HAVE TO
XPROTH	てはならない、べきでない	MUST NOT, SHOULD NOT
XPERM	てよい、*てください	MAY, CAN, COULD
XRECO	ほうがよい、たほうがよい	SHOULD, BETTER, HAD BETTE
XADVI	とよい、のがよい、ばよい	SHOULD
XSUFF	ば十分だ、さえすればよい	ONLY HAVE TO
XCONCLU	のである、*ことになる	0

C S H C 48種

CP	Japanese surface samples	English surface samples
XTEND	がちだ、*やすい	TEND TO, BE APT TO
XESTIM	にたる、価値がある	BE WORTHY OF, DESERVE
XPOSS	*られる、える、うる	CAN, BE ABLE TO, COULD
XINVI	*よう、*う	let's
XCAUS	せる、させる、*ようにする	MAKE, HAVE, GET, let's
XBENE	てもらう、でもらう	HAVE
XPLAN	*ことにしている、考えだ	INTEND, PLAN
XWILL	*よう、てみせる、*まい	WILL
XWISH	たい、てほしい、でほしい	WANT, WISH, DESIRE, HOPE
XNECE	ないではいられない	CANNOT BUT, C.F.CANNOT
XPASS	*られる、*れる	BE-ed
XPOLI	*られる、*れる	0
XREPOR	そうだ、ということだ(未)	It is set, that
XHONO	てさしあげる、てあげる	
SOME^OF^	ことがある、こともある	
FREQ	ことはある、場合がある	SOMETIMES
MANY^OF^	ことが多い、ことも多い	
FREQ	場合が多い、場合も多い	OFTEN

ALL^OF^	てばかりいる	
CONSI	でばかりいる	ALWAYS
NO^OF^	ことがない、ことはない	
FREQ	たことがない	
XONLY	にすぎない、にとどまる (の)でしかない、ばかりだ	ONLY
XEXCES	すぎる	TOO MUCH
XENOUGH	たりる、たりない	ENOUGH
XAGAIN	なおす	AGAIN
XBACK	かえす	BACK
XVERY	かえる、てしかたがない	VERY
XDIFF	*にくい、がたい、づらい	HARD
XEASE	*やすい、よい	EASY
XINTE	*うとする、うとさえる *ようにする *ようとする	ATTEMPT
XPULE	*ことにしている、たものだ *ようにしている	MAKE IT A RULE TO, USE TO
XPULED	*よくなってる	
XDECIDE	*ことにする	DECIDE
XDECIDE	*ことになる	BE DECIDED
XREGARD	*ことにする *ことになっている	REGARD
XREGARD^	*つもりだ、つもりである	
ED		BE REGARD TO
XBETO	*よくなっている	BE TO
XSHOW	*てみせる	
XSUPPOSE	*ことにする	SUPPOSE
XTRY	てみる、でみる	TRY-ING
XUNWILL-		
ING	しぶる	BE UNWILLING
XEAGER	たがる	BE EAGER TO
XFORGET	忘れる	FORGET
XFAIL	そびれる、そこなう、のがす	FAIL
XCHAN	となる、になる、よくなる	BECOME, GET
XACCUS-		
TOMED	なれる、つける	BE ACCUSTOMED TO -ING
XMISTAKE	違える、間違う、まちがえる	MISTAKE
XFEEL	がる	
XREQ	てくれ、でくれ、	
XEACH^		
OTHER	あう	EACH OTHER
XNULL	する、を行う	
XDARE	(あえて)	DARE

ることとした。下位関係概念は恒常的に設計・定義し追加・変更を行い得るが、上位関係概念は原則的に変更不能とした。これにより、統配合を下位関係にのみ押えながら、発見される新しい関係記述に関する意味を取り入れ、拡張することができた。

また、補助概念はそれに対応する日本語表層表現が叙述補助部（助動詞）におおく現れることより、日本語叙述部を網羅的に収集・分析し、それに基づいて各概念の設計・開発を行った。他方、英語に於ける対応表層表現を動詞、副詞、助動詞を対象として網羅的に調べることによって、最終的に語彙を決定した。この開発において、補助概念もそれが明確な内部意味構造を持つ事が容易に推定できた。そうした、補助概念の構造指定機能も同概念CPが中間言語中で自立概念CPと同一な扱いができることにより、無理なく中間言語に親和させることができた。

現在も保守改良を行っており、中心は依然自立概念、下位関係概念に関してである。自立概念は新たな単語表現が新たな意味をもたらすときに、新たに中間言語に語彙を追加するという作業

表4 時相概念

TEMPORAL 4種

CP	Japanese sample	English sample
XBEGIN	始める、始まる、	BEGIN, START
XABOUT	かかる、かける、	BE ABOUT TO,
XCONT2	続ける、とおす	CONTINUE
XFINISH	終わる、きる	FINISH

ASPECT 5種

CP	Japanese sample	English sample
PROG	つつある、なかだ	BE + ING
PERF	ている、てある	HAVE + EN
STAT	ている、ておる	
CONTI	ていく、ゆく	
PSTAT	*た	

TIME 3種

CP	Japanese sample	English sample
PAST	*た	(past tense)
PRES		(present)
FUTU		(future)

と、それに先立つ単語表現の意味を担うCPがすでにあるかどうかの判断を行う作業を含む。この手法と問題点については本予稿集中の「他言語間機械翻訳用辞書の開発手法」にて論じる。下位関係概念は、関係概念CPの細分を担っており、その統配合に関しては、日常的な管理を必要としている。

これらの管理はシステムエンジニアによって行っている。しかし、翻訳システム利用者が辞書登録を行う際にはCP管理作業が発生する。そこで、利用者が登録する単語の意味は自立概念に分類される物のみとし、辞書登録の影響をユーザ専用辞書のみで反映させることにしている。それによって、利用者を煩雑且つ一般利用者には実行が困難なCP管理から解放している。

他方、この枠組はスペイン語、韓国語、フランス語について適用検証をしている。その過程で、いくつかの改案が必要になっている。ただし、自立概念、関係概念、補助概念に関しては既存の物に手を加える必要を感じてない。

本稿では、中間言語の4種の語彙の内、制御属性を除く自立、関係、補助の3種の概念CPを具体的に与えた。これらを利用して翻訳システムを構築するためには、1) 制御属性、2) 書式の詳細な意味記述が必要である。これらに関しては別の機会に報告する。

また、中間言語PIVOTが多言語間翻訳システムを真に支えられる枠組みであるかどうかについては、今しばらくの研究と検証が必要である。CICCプロジェクト、EDRプロジェクトでもたらされる中間言語ないし概念記述言語の成果報告をも利用して、検証を行なっていきたい。

[参考文献]

- 1) 村木、「知識ベースと言語に独立の中間表現とを用いた日英機械翻訳システム」、日経エレクトロニクス1984, DEC.
- 2) Muraki, 「Augmented dependency grammar for language comprehension」、Second European AI Conference, 1986, FEB.
- 3) 岡田、他「自然語および図形解釈のための単純事象概念の分析及び分類」、信学論D、1973、OCT.
- 4) 岡田、他「自然語および図形解釈のための非単純事象概念の分析及び分類」、信学論D、1973、OCT.