物語理解シミュレーションの試み：物語テキストから アニメーション自動生成を通して

A trial of story comprehension simulation: Through automatic generation of animation from story text

星名　研吾[†]，野口 武紘[‡]，杉本　徹[†‡]榎津　秀次[†‡]
Hoshina Kengo, Noguchi Takehiro, Sugimoto Toru, Enokidzu Hideji

[†]芝浦工業大学大学院理工学研究科，[‡]株式会社ナビタイムジャパン,[‡]芝浦工業大学工学部情報工学科
Graduate School of Shibaura Institute of Technology, Navitime Japan Co., Ltd, Shibaura Institute of Technology
ma12094@shibaura-it.ac.jp, m110107@shibaura-it.ac.jp,enokizu@shibaura-it.ac.jp

Abstract

In order to investigate the mechanism that underlies the mental simulation during comprehension of the, story, we designed the system that can generate a series of animation from the story text. Mental simulation is analogous to mental reproduction of the microworld depicted by each sentence. Two kinds of information were needed to generate animation: information composing the microworld and information determining how to shoot the microworld, Based on these findings, we indicated some mental computations necessary to mental simulation and the mental representation of the situation model.

Keywords — situation model, mental simulation, event segmentation theory, automatic generation of animation

1. 研究概要

認知科学の分野において, 人間の理解過程は現在でも大きな研究テーマである. 近年でも言語, 特に物語文の理解過程に関する研究が盛んである.

Kintsch ら[1]や Zwaan ら[2]によると，人間は物語を読む際，まず記述されている状況についての言語的な手掛かりを抽出する.そして,これらの情報と過去の知識や経験を頭の中で結びつけて活性化させ, 状況モデルと呼ばれる心的小世界(イメージ)を作ることで物語文に記述された内容を理解しているとされる.

また Zacks らの EST(event Segmentation theory)[3]によれば，人間は状況モデル構築の過程で，物語内で起きる特徴的な次元の変化によって, 物語の分割を行っているとされる. 例えば, 時間/空間の次元に関連した変化によるセグメンテーション, その下での人物の次元に関わる変化によるセグメンテーションというように, いくつかの階層でセグメントを行い, それぞれを意味の単位としてまとめ, 文の内容の統合, 理解をしているとされる.

これら状況モデルや EST の理論を用いることで, 状況モデルの構造や言語的な手掛かりを決定することが出来る. しかし, これらの情報から視覚的イメージを生成するためには, 心的小世界の範囲や視野, それらを定義する視点の情報を補完するための知識が必要であると考えられる.

これに関し,Zwaan ら[4]は，状況モデルを構築する際の視点の存在についても言及している．人間は心的小世界の中に視点を置き, あたかもその状況の中に自分自身が没入しているような, 経験的なシミュレーションを行うとされる.

このことから，状況モデル構築の際には,状況を俯瞰するような固定された視点ではなく,映画のように注目する情報によって様々に変化する視点が,人々が体験によって得た知識から与えられているのではないかとも考えられる.

そこで本研究では,状況や行為に対する視点を与える知識が集積されたものとして,映画に注目する.そして映画をカメラが切り替わるまでの単位であるショットに区切り,ショットを構成する要素を解析,抽出する.これを基に視覚的イメージ化のための構図情報を導出し，人間がイメージを生成するための知識とすることを試みる.

以上より, 本研究では, 物語テキストからアニメーションを自動生成するシステムを構築することを通して, 物語理解, すなわち物語テキストから状況モデル(心的小世界)が構築される過程を検討す

る. そして, 状況モデルが構築されるのに必要な言語的手掛りとイメージ生成に必要な知識を明らかにする. 最後に, 構築したシステムを実際に物語テキストに適用することで, 物語理解についてのメンタルシミュレーションの実験を行うことを研究目的とする.

2. システム概要

全体のシステム概要は図１のようになる.

本システムの入力は物語テキストである．ただし自然言語処理による意味解析を容易にするため1文１述語からなるように単文化したテキストを用いる. 図中番号1では，入力テキストを一文ごとに解析し，構文解析を利用してそれぞれの文の状況についての情報を導出する．これらの情報を一旦状況テーブルとして格納, 文ごとの逐次的な処理が終わったのち，これらの情報にESTに基づいた微分的，積分的計算処理を施す．ここまでの全ての情報は，CSV形式のセグメントテーブルというファイルに格納する．次に図中番号2では, セグメントテーブルの情報を基にして, 構図決定に必要な情報のif-thenルールが記述された構図情報ファイルを参照し，各文でのアニメーションの構図を決定する．セグメントテーブルの情報に構図情報を加えたファイルをアニメーションテーブルと呼ぶ. 図中番号3では，決定した構図をアニメーションで表現するために，人物情報や空間情報，構図情報等を基に，空間上での人物やカメラの配置を決定する. そしてアニメーションテーブルの情報をアニメーションの生成に用いるTVML playerでの出力できるよう, TVMLファイルに変換する.

図 1 システム概要

最後に, TVML ファイルを TVML Player に読み込ませることで，アニメーションを生成する.

3. 言語的手がかり

3.1. 状況モデル

3.1.1. 状況モデルとは

本研究では,物語テキストから言語的手がかりを抽出する過程のモデルとして，状況モデルという考え方を導入する．状況モデルとは，人間が物語理解の際に頭の中で構成する心的小世界，すなわちイメージとその構成過程のモデルである.

Zwaan[2]らによると，状況モデルを構築するためには 5 つの重要な状況的次元，時間(time)，空間(space), 主人公(protagonist), 意図(motivation), 原因及び因果(causation)があり,読者が物語を読む際には, これらの要素に注目しているとされる. 本研究では, Zwaan の主張する状況モデルを構成する要素のうち，主人公を登場人物とし，そして新たに状況に登場する物あるいは対象（object）を人物と分ける.また文脈解析の困難さから意図を除き，主人公の目標に沿って物語が進むという仮定から，意図を潜在的に含むと考えられる行為/状態の次元を加え，時間，空間，人物，物及び行為/状態の 5 つの次元を, 状況モデルを構築する要素とした.

3.1.2. 状況テーブルの生成

入力テキストには形態素解析ツール Mecab を用いて解析を施す．形態素解析結果が出力されると，単語に本研究で必要な情報を付加した辞書を適用し意味解析に相当する処理を行う．その情報に基づき，格文法を利用した構文解析を行い，一文ごとに情報を導出する．ここまでの情報は状況テーブルとして一時的に保存される．状況テーブルに格納する情報は，主に状況モデルに必要な 5 つの次元に関連する情報である.

3.2. セグメンテーション

Zacks ら[3]は，人間が物語を読み、または映画を見ているとき，そこに描かれたイベントを理解するために,ある連続した表現を構成するとして

いる．これは状況モデルの論にも見られる、心的小世界の構成のことである.またこの表現について、現実において行動の連続する流れを知覚するのと同じように，文脈のある特定の境界によってユニットに分割していることを指摘している．

また,これらの分割は知覚システムが次に与えられる情報を予測しようとする副作用として自発的に発生するものだとされている．現実世界での生活のように，読者は常に物語の次の展開を予測しながら読み進めており，予測した活動の知覚的，または概念的特徴が変化するとき，予測はより複雑になり，エラーが一時的に増大する．その時点で，読者は「今起きている状況」についての記憶表現を更新するのである．これは読者には，一つの新しいイベントが生じたという主観的経験として知覚される．

本研究では，Zwaan らが特に重要としている時間及び空間，それに主人公（人物）の変化によってイベントの大きな分割が起こっていると考える．そこで，時間や空間によって分割されるユニットの単位をステージ，人物の変化によって分割されるユニットの単位をシーン，そしてその下の 1 文 1 述語からなる最小単位をイベントと定義し,物語のセグメンテーションを行う．このセグメンテーションは階層的な構造を持ち，情報はセグメントのユニットごとに算出,保存される．

3.3. セグメントテーブルの生成

ここでは，各分の状況についての 5 次元の情報が格納された状況テーブルに,Zacks の EST に基づき案出された差分(微分)計算によって算出した情報を付加する．

本研究では，時間及び空間で分割するステージというユニット，人物の入退場で分割するシーンというユニットが存在する．ステージの変化は，時間及び空間の変化，すなわち解析している文において時間か空間の情報が新しく入ってきたときに判断される．シーンの変化は，新しい人物の登場，あるいは状況に存在している人物が退場して画面から消える場合に判断される．このような人物の入退場の判断は本研究では,移動を伴う動作に，入場もしくは退場の情報を含め，その動作の動作主に対して入場もしくは退場の判断をする形で行う．以上の情報を状況テーブルに追加し，セグメントテーブルとして出力する.セグメントテーブルに含まれる情報を表 1 に示す.

表 1 セグメントテーブルを構成する情報

文番号	ステージ番号	時間	登場人物
地の文	シーン番号	空間	既存登場人物
会話文		動作主	新規登場人物
		被動作主	既存登場人数
		物	新規登場人数
		動作	主人公
		動作タイプ	

4. 視覚的イメージ生成のための情報

4.1. 構図情報

本研究では，状況モデルにおける視点にあたるものとして,アニメーションの構図を決定する．

一般的な（特にスクリーンでの）映像制作において，構図は絵コンテによって決められている．

絵コンテは,シナリオに描かれている情報を土台にしてどのくらいのサイズでどのような向きで登場人物を画面に映し出すかといった情報を描くものであり,原則として 1 ショットにつき 1 枚の絵を描くものとされている．本研究においても言語的手がかりをもとにして,各イベントごとに 1 ショットとして構図を決定していく．

また,構図の要素を特に重要と思われる(ショットサイズ,ショット内人数,コンポジション)の三つに絞り,本研究で扱う構図情報と定義する．

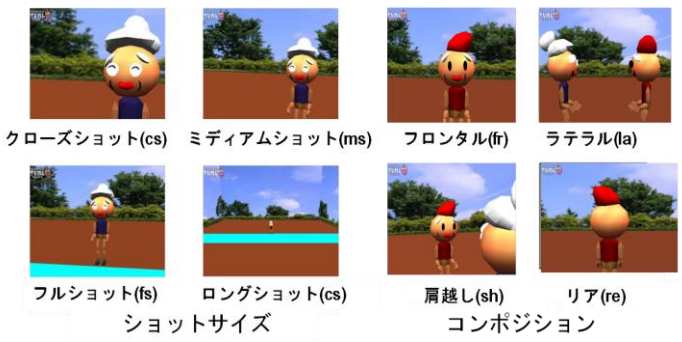

図 2 ショットサイズとコンポジション

アニメーションの出力に用いる TVML は，仮

想空間を持ち，映画と同様のカメラ操作が可能であるため,実際の映画に用いられている撮影技法,構図を解析しアニメーションに反映させることができる.そこで映画における構図情報に注目し構図を決定し,物語の視覚的イメージ化に利用する.

4.2. 構図決定ルールの作成

4.2.1. ショット解析

映画文法[5]において，映像の最小単位はショットであるとされる．本研究では，映画をショット単位に区切り，ショットを構成する様々な要素を抽出する作業を行った．この作業をショット解析と呼ぶ.

表2 ショット解析結果の一部

シーン開始時間	大連辞	人数	ショット	人数	ショットサイズ	コンポジション	登場人物	アクション	アクション(客体)
0:21:48:26~0:22:41:03	sc	3	228	2	id	fr	ディナ	移動	ne
								発話	ne
							レイモンド	見る	ディナ
			229	3	cs	fr	ディナ	発話	ジャニーン
							ジャニーン	発話	ディナ
							ピーター	立ち上がる	ne
								移動	ne
			230	2	cu	la	ピーター	発話	ディナ
							ディナ	発話	ピーター
			231	1	cs	fr	ジャニーン	振り向く	ne

構図や撮影カメラの決定を目的とする既存の研究においては,構図決定のための知識として,映画文法が多く用いられている.しかし,映画文法は実際に撮影ルールとして用いるには状況に限定される技法が多く，物語テキストに記述されるさまざまな状況の全てにおいて,カメラを一意に決定することが難しい．そこで映画に対してショット解析を行い，それぞれの状況で実際にどのように映像が撮影されているかを抽出し,イメージ化のための知識とすることで,出現しうる様々な状況に適切な構図を与えられるのではないかと考えた.

4.2.2. 構図情報ファイルの作成

ショット解析結果を基に，構図情報ファイルの作成を行う．本研究では構図を決定する情報として，動作タイプ(解析時に無数に出現する動作を,構図決定に利用しやすいように分類したもの)と動作の客体の有無という二つの情報に注目した.これは撮影対象の大半が人物であり，その人物の動作によって構図（撮影方法）に変化が生まれるからである．また，ショット解析の結果，被写体の動作に客体がいる場合といない場合とで，構図に違いが生じている傾向があったためである．例として，図 3 は，映画「ゴーストバスターズ」の中で 2 人の人物が会話しているショットである.条件から動作タイプが SPEAK で，客体があるショットを選択すると，図のようなショットが候補として選ばれる．これは後述の求心と遠心に起因するものと思われる.このように，二つの情報からある程度構図を絞ることができると考えた.

図 3 動作タイプと客体の有無による構図の選択

また,映画は物語の時間の経過や現在の状況がより視聴者に伝わり易いようショット間の関係性やショットの構成が検討されている．そこで，直前に撮影されたショットの構図情報を入力条件に含めることで，前後のショットの関連性が強まり，より典型的なショットの連結ができると考えた.以上から，本研究では，現在のショット（文）における動作タイプと客体の有無，加えて直前のショットの構図情報の組み合わせを入力条件として,ショットの構図情報を決定する.

4.2.3. アニメーションテーブルの生成

セグメントテーブルの情報を入力に,if-then ルールにより構図情報ファイルに格納された構図情報を参照し各ショットの構図を出力する．このとき，入力条件として直前のショットの構図情報が必要になる，直前の情報を参照できないステージユニットの初めの文においては，エスタブリッシングショット(establishing shot)を使用する.これは，物語の時間あるいは空間情報などが変化したとき，場所や人物などの状況説明のためにシーンの初めに多く挿入されるロングショットのことである.

全てのショットの構図情報を決定した後，セグメントテーブルの情報に構図情報を付加し,アニメーションテーブルと呼ぶ CSV ファイルとして出力する．以下の表 3 に，実際のアニメーションテーブルにおける構図情報格納部分を示す.

表 3 アニメーションテーブル（構図情報部分）

ショットサイズ	ショット内の人数	コンポジション	ショットサイズ	ショット内の人数	コンポジション
ls	2	la	cs	2	fr
ls	1	la	fs	1	fr
ls	1	la	fs	1	fr
ls	1	la	cs	1	fr
null	null	null	cs	1	fr
null	null	null	ms	1	fr
null	null	null	cs	1	fr
null	null	null	cs	1	fr
ls	2	la	fs	1	fr

4.3. 配置情報

本研究における配置情報とは，TVML 空間上での人物の配置，カメラの配置，物の配置を指す．TVML 上で構図情報通りのアニメーションを生成するために，空間のどこに人物を置き，どこにカメラを設置し，またどのカメラで撮影するかという情報の決定が必要となる．

配置に関する情報は，登場人物やイベントといった情報とは異なり，テキストの情報と TVML のコマンドを一対一で対応させることが困難であるため，ある条件に対してどのカメラでどのように撮影する，というルールを構築し，if-then で対応させる必要がある．本研究では，アニメーションテーブルに格納された構図情報をはじめとする各情報を決定要因として配置を決定する．次節で配置のための撮影ルールについて詳述する．

4.4. 配置決定ルールの作成

映画文法によれば,コンポジションは，被写体の数によってシングルショット，ツーショット，グループショットに分けられる．ここでいう被写体は，人物の数に限らず，そのショットの動作に関わる能動や受動の主体の数である．動作の両端の能動と受動の両方が画面内に収まっているものを求心(afferent)ショットといい，どちらか一端が画面の外にあるものを遠心(efferent)ショットという．求心と遠心における被写体の数は，あくまでその動作にとっての能動主体と受動主体を単位とするものである．すなわち，ツーショット，あるいはグループショットの場合でも，あるグループがユニットとしてまとまり，それぞれの動作が同じ方向性を持つならば，実質的にシングルショットと同様の扱いになる．

このような求心と遠心の考えをもとにして,(構図情報,動作タイプ,動作主の人数)を配置決定の入力条件とし,これらの情報の組み合わせごとの配置を, 実際に TVML 上での出力を確認しながら配置情報ファイルとしてまとめた.そしてアニメーションへの変換時にテーブルの情報を入力に if-then ルールによって配置の決定を行う．

5. TVML 言語への変換

5.1. TVML とは

TVML(TV program Makin Language)は，NHK 放送技術研究所が 1996 年に提案した, テレビ番組を記述するインタプリタ型言語である．この言語で記述されたスクリプトを TVML Player と呼ばれるソフトウェアで再生すると，リアルタイム CG や音声合成などにより作られたアニメーションを見ることができる．

TVML によるアニメーションには仮想空間が存在し，CG によるスタジオセットやキャスト，小道具，カメラがあり，これらを自由に配置することで自在なアニメーションを作ることができる.また,インタプリタ型であるため,スクリプトで記述された１つのコマンドが本研究でのアニメーションにおける 1 つのイベントに対応し,時間軸に沿って，アニメーションが進んでいく．

これらの仕様は背景セットの変化によるステージの変化や，人物の入退場によるシーン変化等のセグメンテーションの表現を可能にし.また 1 イベント１ショットのカメラの逐次的な操作でも構図を自由に表現できるなど,その自由度から物語理解のシミュレーションを目指す本研究に非常に適しているといえる．

5.2. TVML ファイルの生成

アニメーションテーブルの情報を入力に，システムによって TVML ファイルへ変換し出力していく．本研究では，１文１イベントを１ショットとみなし，システムがテーブルの情報を一度全て読み込んだ後，一文ずつ情報を参照し，逐次変換して TVML ファイルとして書き出していく. 変換の中身は以下の通りである．

5.3. 変換プログラム

初めに，アニメーションテーブル中の空間に関する情報から TVML の空間に背景と照明を設定する.実際の処理としては,空間情報に値があった場合，空間の変化があったと判断し，ステージ内の空間情報を参照して背景セットの設定を行う．

背景と照明の設定が終了すると,次に空間内（ショット内）に登場する人物や物のモデルの呼び出しを行う．現在の文のショット内人数を参照し，呼び出す人物の数を決定する．この際ショット内人数が実際の登場人数より少なかった場合は，ショットに表示する人物を選択する処理を行う．ショット内人数が 1 人であれば，動作の主体を呼び出す．2 人以上であった場合は，主人公の有無，動作主，被動作主をチェックし，これらを優先して呼び出す等のように配置ルールに基づき呼び出すモデルを決定する．このとき，物情報に値が入っていれば，対応する小道具モデルを呼び出す．

人物や物のモデルの呼び出し，設定が行われた後，アニメーションテーブルの配置決定の入力条件を基に, if-thenルールで配置情報ファイルを参照し,キャラクター，撮影カメラの配置を決定していく．条件分岐は以下のようになる．

1. ショットサイズ(cs, ms, fs, ls)，ショット内人数(1~6)，コンポジション(fr, re, sh,l a)を判定する．
2. 動作の人数(1~4)を判定する．
3. 動作タイプ(1~8)を判定する．

これらの手順により，配置はほぼ決定できる．しかし 2 人の人物が会話を行っている場合等に，続けて同じ分岐になってしまうことがある．この場合には，異なる人物が全く同じ配置で発話するような違和感のある映像になる.そこで対照となる配置を用意しておくことで,図 4 のように左右で振り分けを行う．

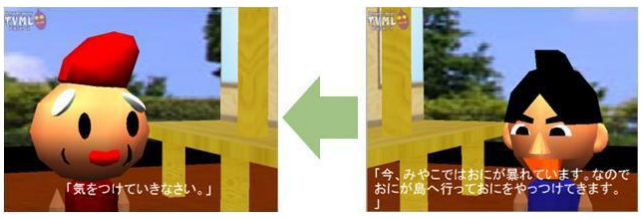

図 4 同方向を向くことを避ける処理

配置の設定後,最後に実際に人物に実行させる動作や，字幕の表示等の設定を行う．

まず原文を表示する．次に，動作情報から人物に実行させる動作を決定し，キャラクターに実行させる．会話文であれば，キャラクターに会話させる．また，ショットごとの時間を調整するために，イベント終了後の待機時間を設定する．

以上の変換を，アニメーションテーブルの最後の行まで繰り返し行い，呼び出されたスクリプトを TVML ファイル（拡張子.tvml）に変換し，出力する．

6. 実験方法

6.1. 材料

子どもを対象とした童話から,「桃太郎」と「鶴の恩返し」を入力の物語テキストとして使用した．それぞれのテキストは,自然言語処理による意味解析を容易にするため 1 文 1 述語からなるように単文化を行っている．

6.2. 手続き

まず，材料である物語テキストを一文ごとに解析し，意味解析を利用してそれぞれの文についての 5 つの次元の情報を抽出する．また,これらの情報にシーン分割等の微分的，積分的計算処理を施す．ここまでの全ての情報は，CSV 形式のセグメントテーブルというファイルに格納する．次に，上記セグメントテーブルの情報を基に構図情報ファイルを参照し，アニメーションの構図を決定，構図情報を追加したものをアニメーションテーブルとして出力する.そして，構図情報に基づいた映像化のために,TVML 空間上での人物やカメラの配置を決定した上で，アニメーションテーブルの情報を TVML ファイルに変換する．

以上の処理によって最終的に出力された TVML ファイルを TVML Player に読み込ませることで，アニメーションを生成する．物語テキストからアニメーションを自動生成する一連の流れをもって,物語理解についてのメンタルシミュレーションとする．

7. 結果と考察

二つの物語テキストに対し,システムによるシミュレーションを行ったところ,どちらの物語においても物語テキストからの状況モデルを構築する5つの次元の情報の抽出や,それに基づくセグメンテーション,さらには視覚的イメージ化のための構図の決定に成功した.「桃太郎」では,ほぼ想定通りの違和感のないアニメーションの生成にも成功した.

図 5 桃太郎アニメーション出力

「鶴の恩返し」においては,現状ではTVMLファイルへの変換部が「桃太郎」を想定して作られているため,モデルが存在しない空間や人物については既存の物で代用し,表示出来ないイベントもいくつか出現する結果となった.しかし,移動や発話等の共通であり,かつ物語の多くの部分を占める行為については,同様に違和感のないアニメーションの生成に成功した.完全とは言えないものの,物語テキストからのアニメーション生成に成功したことにより,メンタルシミュレーションに必要な情報,ひいては人間の物語理解に必要な情報の指針を確認することが出来たのではないかと考える.

ただし，これは非常に限定的な状況下での成功であり，現状では自然言語処理の困難さ,アニメーションに必要な 3D モデルの不足といったことがボトルネックとなっている.テキストからの情報導出の改良,汎用 3D モデルの導入等により，様々な物語，または物語以外のテキストにも対応できるようにした上で,さらなるシミュレーションを行う必要があるだろう.

8. 結論

本研究では,物語テキストからアニメーションを自動生成するシステムを構築することを通して人間が物語テキストから状況モデルを構築する過程について検討した.そして構築したシステムによりメンタルシミュレーションを行った.状況モデルに関する 5 次元の情報については, Zwaan の主張とは少し異なる情報を抽出したが,意図や因果といった情報は元々映像として表現することが困難なためか,行為/状態で代用しても問題なく出力できたと言える.また,次元の変化によるセグメンテーションに関しても,システムで問題なく処理できた.しかし,現状背景セットの変更等による場面転換の表現やシーン単位で登場人物を考えることによるショット構図や配置調整にしか活用できていないなど,分割によって得られる情報を上手く利用しきれていない面があり,導出する情報や映像での表現の仕方について再検討する必要があるだろう.そして,映画の解析結果を用いた視覚的イメージ化については,構図決定による映像への没入感が,視点が固定されたアニメーションよりも高く,イメージ化の知識として有効であったと考える.しかし今回，入力条件に対して最も使用回数の多かった値のみを採用したため．使用した構図は実際の映画で使用されたショットのうち50%に満たない情報である.この値の有効性の検証と，サンプル数の増加が必要であると考える.

本研究の結果から,アニメーション生成に必要な情報の指針が得られた.このことは同時に,メンタルシミュレーションにおける心的表現や，それに伴う計算，推論に必要な情報について指摘することも可能にしたといえる.

参考文献

[1] Van Dijk, T. A., and Kintsch, W,(1988) “Strategies of Discourse Comprehension ”,New York: Academic Press.

[2] Zwaan R.A and Radvansky G.A,(1998) “Situation Models in Language Comprehension and Memory”, Psychological Bulletin,vol.123,No.2,pp162-185.

[3] Zacks,J.M., & Kurby,C.A. ,(2008) “Segmentation in the perception and memory of events:”, Trends in Cognitive Sciences, 12(2), pp72-79.

[4] Zwaan R.A,(2004) “The Immersed Experiencer：Toward an Embodied Theory of Language Comprehension”, the Psychology of Learning and Motivation,Vol.44,pp35-62.

[5] D. Arijon, 岩本憲児, 出口文人（訳）(1980) 映画の文法.