

効用値からみた文法伝搬： 屈折による曖昧性除去と適用コストのバランスから

的場 隆一 中村 誠 東条 敏
北陸先端科学技術大学院大学
情報科学研究科

概要 自然言語の文法は無限に長い文を生成する能力を持っている．これには単語間の依存関係を表す統語上の制約だけではなく、構文や意味の曖昧性を解消するための表層上の制約も含まれる．この制約として一致 (agreement) や格変化 (case marking) が存在する．これは性、数に関する素性や、格要素などにより単語を屈折させることで、単語間の依存関係を示すための制約である．一般的には語順に関する制約が緩い言語ほど、この機能に関する制約は厳しい傾向にあるといわれている．本研究の目的は、言語理解における屈折の効用と語順に関する制約の関係からどのような文法が伝搬されやすいのかを調査することである．本研究では、用いられている文法の複雑さに関して相対する 2 種類のコーパスを適用し、屈折の効果と語順の関係について調査し、更に文の伝達性を定量的に測る効用値関数を導入した．また、利得行列を作成し話者・聞き手の各方針に対する分布の変遷を示した．これにより、屈折の重要性を示すことができた．また、第 2 言語獲得を模倣した実験では、聞き手は格変化による屈折を無視した文法を用いることが有利な方策となり、初期段階のピジンにみられる文法素性の脱落現象を観測することができた．

1 はじめに

自然言語の文法は無限に長い文を生成する能力を持っている．これには単語間の依存関係を表す統語上の制約だけではなく、構文や意味の曖昧性を解消するための表層上の制約も含まれる．この制約として一致 (agreement) や格変化 (case marking) が存在する．これは名詞が持つ性、数に関する素性や、修飾する動詞にかかる格要素などにより単語を屈折させることで、単語間の依存関係を示すための制約である [3]．一般的には語順に関する制約が緩い言語ほど、この機能に関する制約は厳しい傾向にあるといわれている [1]．この屈折により文の曖昧性が解消できる一方で、母語話者ではない者にとってはこの機能は学習の障害になっていると考えられる．実際、複数の言語が接触する特殊な環境で発生する共通語であるピジンの初期段階ではこの機能が失われる傾向にある．また、この段階のピジンでは語順は定まっておらず、母語の語順に関する制約が強く残っている [6]．本研究では、言語理解における屈折の効用と語順に関する制約の関係からどのような文法が伝搬されやすいのかを調査することを目的とする．

Jäger[2] は、能格言語と呼ばれる、他動詞に修飾する代名詞が特殊な格変化をする言語に着目し、言語理解に対する格変化の効用をモデル化した．ここでは、言語使用者の負担と伝達性の関係を効用関数を用いて定量化し、進化ゲーム理論による言語伝搬の解析を行っている．本研究においても、この効用関数の概念を取り入れ、格、人称、数に関する屈折機能を考慮した新たな効用関数を定義する．また、人間の持つ言語に対する知識、能力だけではなく、それにより出力された言語を扱う上で構文の重要性を取り去ることは避けるべきであると考えた．そこで、計算機への実装を

考慮した結果、HPSG(Head-driven Phrase Structure Grammar)[5] による構文解析を人間が行う言語処理と仮定し、効用関数に適用するデータに英語コーパスを解析することで得られる情報を用いた．HPSG は全ての品詞、カテゴリー、部分木などを素性構造で表現しており、本研究で注目している語の屈折を 1 つの素性として扱うことが可能である．

本研究では、話者は相手に理解される言葉を話すことを重視し、また、聞き手は話者の言葉の理解に労力をかけたくないという観点に立ち、話者と聞き手がどのような方策を取れば互いに高い効用が得られるかを調査する．このため、理解にかかる労力と文の曖昧性に基づく効用関数の定義を行い、語の屈折と語順の関係からどのような文法が伝搬されやすいのかを考察する．具体的には、語の屈折を無視した文法とそうでない文法で英語文と語順を入れかえた英語文を解析したときの効用値を求め、初期段階のピジンにみられる現象と比較する．本研究結果は、文法素性を操作することによる話者、聞き手の文法使用の方策の変化を見えるという点で第 2 言語獲得や言語進化の分野に貢献すると考えられる．

2 効用値測定モデル

本研究では、話者は英語文か英語文の語順を入れかえた文を話し、聞き手は屈折を考慮した文法か考慮しない文法で聞く状況を考えている．このため、語順を入れかえた文を解析することができ、更に屈折に関する素性を無視することができるパーサを作成した．この解析結果から文法ごとの効用値を計算し、レプリケータ・ダイナミクスにより話者と聞き手の人口動態から、語順と屈折にかかわる素性の単一化にかかるコストのバランスを考察している．本節では、実験に用

いた文法とパーサの説明, 解析結果を評価するための効用関数の定義, およびレプリケータ・ダイナミクスの説明を行う。

2.1 語順入れかえ文

実験に適用した文は, コーパス中の文の語順を以下の手順で人工的に入れかえたものである。語順の入れかえ方はランダムではなく次の2種類のルールで入れかえる。

- ルール 1
SVO 語順を SOV 語順に入れかえるルール
- ルール 2
SVO 語順を OSV 語順に入れかえるルール

ルール 1 はスペイン語やイタリア語などのラテン語から派生した言語に見られる。また, ドイツ語やオランダ語のような SOV-V2 語順を取るドイツ語派生の言語もこの語順を取る。ルール 2 は長距離依存と呼ばれ, 英語の関係詞節や疑問形などで取られる語順である [3]。

これら2つのルールにしたがい, 主語と目的語を取る動詞 “love” による文 “[Subject] love [Object]” は “[Subject] [Object] love” や “[Object] [Subject] love” の語順に入れかわる。また, これらの文を受理できるよう英文法用 HPSG の一部のスキーマを改良した文法を導入した。HPSG は ID-スキーマ (ID-schema) とプリンシプル (principle) で文法が構成されており, これらは文脈自由文法の書き換え規則にあたるものである。

実験では, 通常の英語文の集合 (*Normal*) と語順を入れかえた文の集合 (*Scrambled*) の2種類を適用した。

2.2 語順入れかえ文対応パーサ

ここでは実験で使用したパーサの特徴について説明をする。大きな特徴は2つある。1つは任意の素性の単一化を無視してパースすることができることであり, もう1つの特徴は2.1節で説明した手順で語順を入れかえた文をパースすることができる機能を持つことである。パースの出力は述語項構造 (Predicate-Argument Structure: *PAS*) である。本研究においては, この述語項構造を出力することを文の理解と定義する。この述語項構造が1文に対して複数個出力されるとき, その文が曖昧であるとする。1つ目の特徴により, “she love me” や “her love me” といった数や格に誤りを含む文でもその素性を無視することによりパース可能となる。

例として, 素性の単一化を考慮したときと無視したときに現れる解析結果の違いを示す。対象とする文は, “she loves me” を前節で説明した手順で入れかえた “me she loves” とする。まず, 文法の素性単一化

を考慮した場合, “me” は対格, “she” は主格とわかるため文が表す意味は一意に定まり曖昧性はない。一方, 一致素性の単一化を無視した場合, “I love her” または “she loves me” の2つの解釈ができるようになり曖昧性が発生する。

実験では, 2つの文法を適用した。1つは人称, 数, 格を考慮した通常の英文法 (Person, Number, Case: *PNC*) であり, もう1つは格についての単一化を無視した文法 (Person, Number: *PN*) である。

本研究の目的の1つは, ピジン化に見られるような文法を扱う労力と円滑な会話の相互関係を観測することである。この, 語順を入れかえた文を入力とすることや, 素性を無視して解析を行うことでピジンを模倣していると考えられることができる。

2.3 効用値

我々は, 1文のもつ曖昧性と屈折にかかわる素性の単一化にかかるコストから文の効用値を計算する関数をモデルに導入した。この効用値は Jäger が提案した効用に基いている [2]。1文の持つ効用値 U は,

$$U = \frac{N}{|PAS|} - w \cdot C(unif) \quad (1)$$

このとき, $|PAS|$, N , $C(unif)$, w はそれぞれ *PAS* の数, 文長, 語の屈折にかかわる素性に関する単一化の回数, $C(unif)$ のウエイトである。

PAS の数は曖昧性を表しており, 式 (1) の第1項目は利得を示している。この値が大きければ曖昧性が小さいことを意味する。第2項目は語の屈折にかかわる素性を単一化することによってかかるコストを表している。ウエイト w は正の係数で, 第2項目の優先度を操作するパラメータである。

2.4 レプリケータ・ダイナミクス

本研究での話者の方策は, 英語文を話す *Normal* と, 語順入れかえ文を話す *Scrambled* である。一方, 聞き手の方策は, 人称, 数, 格を考慮した文法で解析する *PNC* と, 格を無視した文法で解析する *PN* である。我々は, 話者, 聞き手がそれぞれの方策を取ったときにどの解に安定するかを式 (2) のレプリケータ・ダイナミクス [2] により計算した。

$$\begin{aligned} \frac{\Delta x_i}{\Delta t} &= x_i((Ay)_i - \langle x \cdot Ay \rangle) \\ \frac{\Delta y_i}{\Delta t} &= y_i((Bx)_i - \langle y \cdot Bx \rangle) \end{aligned} \quad (2)$$

このとき, x_i , y_i はそれぞれ方策 i における話者, 聞き手の人口である。 A と B はそれぞれ話者, 聞き手の報酬を表す利得行列である。2人ゲームにおける利得行列は以下のように定義できる。プレイヤー1と2の方策集合を $S_1 = \{1, \dots, m\}$, $S_2 = \{1, \dots, n\}$ とする。この

とき、プレイヤー1の方策を行、プレイヤー2の方策を列として、プレイヤー1が方策 i を、プレイヤー2が方策 j を取ったときのプレイヤー1の利得を a_{ij} 、プレイヤー2の利得を b_{ij} とする。これらの利得で構成される $m \times n$ 行列AとBをそれぞれのプレイヤーの利得行列と呼ぶ。本研究でのプレイヤーは話者と聞き手であり、利得行列は2行2列となる。また、各要素は、取りうる4種類の方策の組(*Normal*, *PNC*)、(*Normal*, *PN*)、(*Scrambled*, *PNC*)、(*Scrambled*, *PN*)でコーパスを解析して得られた効用値の平均値である。

3 実験：一致素性と語順の効果

3.1 実験手順

本節では実験手順を説明する。我々のパーサはルールベースであり、文法の不備からコーパス内の全ての文をパースすることができない。よって、最初にパースできる文のみをコーパスから抽出し、文の集合 *Normal* をつくる。次に、抽出した文の語順を、2.1節で示したルール1,2により入れかえた文の集合 *Scrambled* をつくる。これらの集合を2.2節で説明したパーサを用い、文法 *PNC* で解析を行う。また、格による屈折に関する素性の単一化を無視した文法 *PN* でも解析を行う。これらの解析結果から効用値を計算し、話者と聞き手それぞれの方策の組における利得行列を作成する。本実験での話者の効用は方策 *Normal* か *Scrambled* で発話し、それを聞き手に理解されることである。一方、聞き手にとっての効用は、方策 *PNC* か *PN* により構文解析を行う際、いかにコストをかけずに文を理解するかである。最後に、この利得表をもとにレプリケータ・ダイナミクスの位相図を作成する。

3.2 実験設定

本節では行った実験の設定について説明する。実験に使用したコーパスは Penn Tree Bank に保存されている Wall Street Journal (以下 WSJ) コーパスと Air Ticket Information Service (以下 ATIS) コーパスである。これらのコーパスの特徴は、文の複雑さに関して相反しており、WSJ コーパスは文法的に複雑で長い文語文が多く含まれているのに対し、ATIS コーパスは比較的短く文法的にも複雑ではない口語文が多く含まれている。最初の抽出作業により WSJ コーパスから 5,856 文が得られた。また、語順入れかえ作業により 4,390 文を作成した。同様に、ATIS コーパスからは合計で 210 文得ることができた。これらの文に対しランダムに 10,000 文を復元抽出によって選択し効用値を計算した。1 文に対し複数の解析結果が得られた場合、*PAS* の数は複数となり、その文は曖昧性を含む。効用値のウェイトは $w = 0$ 、 $w = 0.1$ とした。

4 結果と考察

ウェイト $w = 0.1$ での WSJ と ATIS の実験結果をそれぞれ表1, 2に示す。表の各値は“話者の効用値 \ 聞き手の効用値”となっている。話者の効用値は式(1)のうち、コストを考慮しないものであるため、 $w = 0$ の場合、聞き手の効用値は話者のものと同値となる。

表 1: 実験結果：WSJ

	<i>Scrambled</i>	<i>Normal</i>
<i>PNC</i>	1.065\0.797	1.567\1.303
<i>PN</i>	0.989\0.753	1.557\1.301

表 2: 実験結果：ATIS

	<i>Scrambled</i>	<i>Normal</i>
<i>PNC</i>	1.232\1.037	2.067\1.912
<i>PN</i>	1.051\0.934	2.059\1.951

表1, 2は、聞き手が *PNC*、*PN* のどちらの方策を選んでも、話者は *Scrambled* よりも *Normal* の言語を話した方が高い効用を得ることを表している。これより、たとえ屈折に関する素性を考慮しても語順が入れかわることにより曖昧性が大きくなり効用値が低下することが示された。つまり、語順を入れかえても文の長さは変化しないため式(1)の利得を表す項の $|PAS|$ が大きくなることがわかる。これは語順の制約が厳しい英語では、語順を入れかえた文は通常の英文よりも動詞の主格、対格を特定しにくいからである。その結果、効用値は *Normal* > *Scrambled* となったと考えられる。

ウェイト $w = 0$ は、聞き手が屈折にかかわる素性を考慮するためのコストに無頓着な状態である。つまり、聞き手は屈折にかかわる素性を通常から母語として使っており、これの使用になんら負担を感じない状態にあると考えることができる。この状態では、ATIS コーパス、WSJ コーパス共に話者と聞き手の方策の組 (*Normal*, *PNC*) がナッシュ均衡点であることがわかる。ナッシュ均衡点とは、あるゲームを行ったときに全てのプレイヤーに対して戦略 S が他のプレイヤーの戦略の組に対する最適応答である点のことである[4]。この結果は、例えば話者・聞き手の分布が (*Scrambled*, *PN*) に偏っていようが (*Normal*, *PNC*) の方へ変遷することを示している。すなわち、コストがかからない状況下では屈折にかかわる素性を適用した文法の方が高い効用値を得ることができることを示している。また、WSJ コーパス、ATIS コーパスともにナッシュ均衡は (*Normal*, *PNC*) であることから、4種類の方策で比較すると英文法はこの中では一番優れていることがわかる。更に、本稿には揭示してはいないが、語順入れかえ文を入力としたとき、屈折にかかわる素性全ての単一化を無視した文法が一番効用値が低かった。これより、人称、数、格といった屈折にかかわる素性は文を構成する各単語が指し示す対象を特

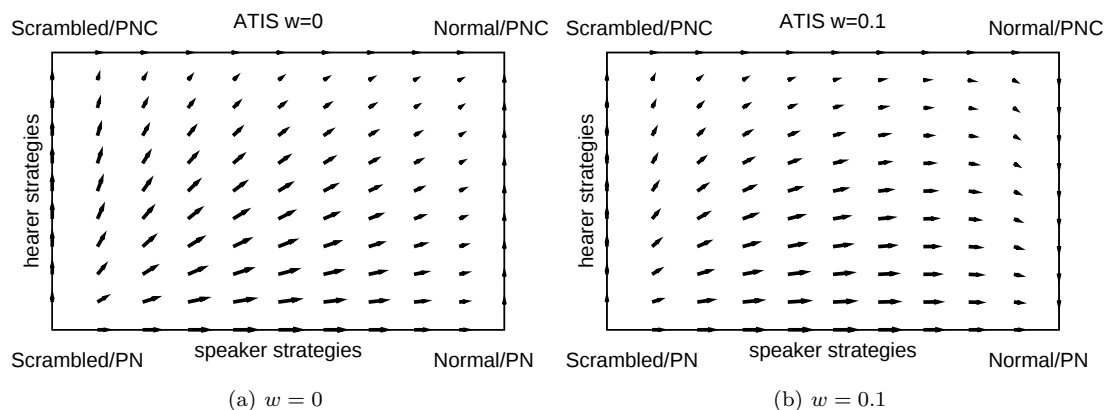


図 1: 各方策に基づく人口遷移: ATIS

定することに役立っていると考えられ、*PAS* の数を減らすことに貢献していると思われる。

一方、ウエイト $w = 0.1$ は、聞き手が屈折にかかわる素性の単一化にコストを払わなければならない状態である。つまり、第 2 言語学習者のように聞き手がこのような素性を扱うことに不馴れな状態にあると考えることができる。WSJ コーパスでは、 $w = 0$ の場合と同様に、方策の組 (*Normal*, *PNC*) がナッシュ均衡点であった。これは、WSJ コーパスでは、格についての単一化を無視することによって削減されるコスト以上に格が無いことによって式 (1) の利得の項が小さくなってしまふことを意味する。この理由として、このコーパスに含まれる文は格に関する一致を考慮しなければならない文が全体の 10.9% とあまり含まれておらず、格についての単一化を無視しても劇的なコストの削減にはならないという理由でこのような結果になったと考えられる。ATIS コーパスの結果を図 1 に示す。この図は、話者・聞き手それぞれの利得行列を用いたレプリケータ・ダイナミクスの位相図である。式 (2) の適用により、 Δt ごとに変遷した各方策組における話者と聞き手の人口を示している。横軸は各方策における話者の分布を示しており、縦軸は各方策における聞き手の分布を示している。図 1(a) より、ウエイト $w = 0$ では収束点は (*Normal*, *PNC*) であるが、ウエイト $w = 0.1$ では図 1(b) より収束点が (*Normal*, *PN*) に移動したことが確認できる。これは、ATIS コーパスは口語コーパスであるため含まれている文に使用されている人称代名詞の数が全体の約 3 割となっており、格を考慮するコストが影響するためと考えられる。また、口語文中心の単純な文で構成されている ATIS コーパスにおいて“*PN*”が有利な方策となったことは初期段階のピジンにみられる文法素性の脱落現象と一致する。

最後に、2 つのコーパスの効用値を比較すると ATIS コーパスの方が値が大きいくことがわかる。これは、WSJ コーパスに含まれる文の方が 1 文あたりの単語

数が多く、文法構造が複雑であるため、文が曖昧性を多く含んでいたことにより効用値が小さくなったものと考えられる。長い文の *PAS* の方がより複雑な構造をもっているにもかかわらず、今回は効用値を求める関数 (式 (1)) に *PAS* の複雑さを考慮する項を定義しなかった。したがって、文長に関しては正確に効用関数に反映されていない。

5 おわりに

本研究では、用いられている文法の複雑さに関して相対する 2 種類のコーパスを適用し、屈折の効果と語順の関係について調査し、更に文の伝達性を定量的に測る効用値関数を導入した。また、利得行列を作成し話者・聞き手の各方策に対する分布の変遷を示した。

語順を入れかえた状態では屈折に関する素性があった方が効用値が高くなったことから屈折の重要性を示すことができた。また、第 2 言語獲得を模倣した (コストにかかるウエイトを増大させた) 実験では、聞き手は格変化による屈折を無視した文法を用いることが有利な方策となり、初期段階のピジンにみられる文法素性の脱落現象を観測することができた。

参考文献

- [1] R. Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, 2002.
- [2] G. Jäger. Evolutionary Game Theory and Typology: a Case Study. In *Proceedings of the 14th Amsterdam Colloquium*, pp. 108–117, 2003.
- [3] 亀井孝, 河野六郎, 千野栄一 (編). 言語学大事典. 三省堂, 1996.
- [4] 岡田章 (編). ゲーム理論. 有斐閣, 1996.
- [5] I. Sag, T. Wasow, and E. Bender. *Syntactic Theory - A Formal Introduction. Second Edition*. CSLI PUBLICATIONS, 2003.
- [6] L. Todd, 田中幸子 (訳). ピジン・クレオール入門. 大修館書店, 1986.