様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 １ 日現在

機関番号：３４３１０
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ００６７０

研究課題名（和文）英語話し言葉コーパスを用いた談話論理構造パターンの認知言語学的分析

研究課題名（英文）A Cognitive Linguistic Analysis of Discourse Logic Patterns Using Corpora of Spoken English

研究代表者

長谷部　陽一郎（Hasebe, Yoichiro）

同志社大学・グローバル・コミュニケーション学部・教授

研究者番号：９０３５３１３５

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：本研究では、第1に、英語における談話論理展開の構造を記述するには、階層的な意味構造を構築・理解するメカニズムだけでなく、談話の時間軸に沿った概念構造の更新プロセスを適切に扱える枠組が必要であることを明らかにし、新たなモデルとして「談話の積層構造モデル」を開発・提案した。第2に、英語プレゼンテーションのデータを用いた話し言葉コーパス・システムTED Corpus Search Engine (TCSE)を開発した。本システムでは、単なる話し言葉の断片的なデータでなく、談話の「展開」を精査するのに適したデータを多数保持しており、談話構造モデルの開発と検証に役立てられる。

研究成果の学術的意義や社会的意義
認知言語学の領域においては、ことばの意味の構造を記述するために様々な枠組みや理論的概念が提案されてきた。談話の構造についても多くの考察が行われているが、実データを用いた具体的な研究と、体系的なモデル化とを両立させた研究は多くない。本研究は、理論的考察、データ処理、検索システムの開発を並行的に進め、研究コミュニティの中でさらなる発展が期待できる分析の枠組みを構築することに成功した。公開済のシステムTCSEは、テキストだけでなく音声や映像と共に談話の展開を精査することができるため、研究者だけでなく、英語教育に携わる教師や、英語による「語り」の手法を学びたい学習者にとっても有用なツールとなっている。

研究成果の概要（英文）：In this study, two key accomplishments were achieved. First, it was clarified that a framework capable of effectively handling not only the mechanism for constructing and comprehending hierarchical semantic structures but also the mechanism for updating conceptual structures along the temporal axis of discourse is essential for describing the architecture of logical development in English discourse. As a result, a new model called the "Stack-Based Structure Model of Discourse" was proposed. Second, the TED Corpus Search Engine (TCSE), a spoken language corpus system utilizing English presentation data, was developed to construct and validate the model. This system houses a vast amount of data suitable for examining the "unfolding" of discourse rather than just fragmented spoken language utterance data. It can be employed for the validation of discourse structure models.

研究分野：認知言語学

キーワード：談話構造　論理展開　話し言葉コーパス　認知言語学　構文　英語教育　言語の線条性　モナド

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

認知言語学の領域では、Ronald Langacker（1987, 1991, 2008）の認知文法（Cognitive Grammar）理論を中心として、言語表現の概念構造を記述する試みが長らく行われてきた。Adele Goldberg（1995, 2006, 2019）に代表される構文文法（Construction Grammar）においても、認知文法とはやや異なった視点から、多くの考察が行われてきた。これらの理論では典型的には語・句・文といったレベルの言語表現が扱われる。談話レベルの研究も行われてはいるが、次の 2 点において、研究の進展はやや滞っていると言わざるを得ない状況であった。

第 1 に、談話レベルの研究では、それ以下の言語単位に関する研究とは異なった理論的道具立てが必要になることである。談話を考えるにあたっては、語・句・文のレベルでの議論では必ずしも大きく関わってこない（もしくは相対的な重要度が低い）問題が重要となる。例えば、談話の中で言及された言語表現やそれらによって想起される概念要素が、経過する時間の中で、談話の参与者にどのように保持され、そして続く談話の中でどのように活性化されるかといった問題である。

第 2 に、使用するデータに関してである。理論言語学の研究においては、研究者自身による例文を用いることが一般的であり、前述の Langacker や Goldberg による研究も例外ではない。近年、認知言語学における用法基盤モデル（usage-based model）の考え方が一般的となり、次第にコーパスから得られた実際の発話が例として用いられることが増えてきたが、実際のところ、語・句・文のレベルにおける考察においては、作例であっても、コーパスからの例であっても、それほど大きな違いはない。しかし、談話レベルの問題について詳細な考察を行うためには、ある程度の量の発話が積み重ねられてできた実際の事例データを確保することが不可欠である。

本研究は以上のような状況のもとに着想された。

２．研究の目的

伝統的に、理論言語学の目的は、「言語とは何か」という究極的な問いに関し、それを構成する様々な側面のいずれかに焦点を当てて、新たな知見を得ることと考えられてきた。本研究における基本的な姿勢もこの伝統的な見方と大きく異なるわけではない。自然言語による「談話」というものが話者の中でどのように概念化されているか、そしてそれが実際の場面の中でどのように運用されているかという問題について考えることが主たる目的である。これにより、認知言語学の枠組みでこれまで論じられてきた理論的概念が言語現象の規模に関わらず妥当であることが示され、また、談話レベルにおける現象を扱うにあたって必要な新たな考え方や道具立てが明らかになると期待できる。

しかしながら、これまで一般的であった理論言語学の研究手法をそのまま談話の研究に用いることは難しい。これまでの多くの研究では、それぞれの目的に応じて、新たに作成した例文や、コーパスから採取した例文をデータとして用いてきたが、談話の分析においては作例に大きく頼ることはできない。一方で、既存のコーパスから分析や検証に適した談話事例を効率的に採取することも困難である。そこで本研究は、談話に関する理論的な探究に加えて、それを可能にするための談話コーパスのシステム自体を開発することを目的として設定した。

このように、談話に関する理論的研究を進めながら、同時に談話コーパスのシステムを開発し公開することには、副次的な利点が存在する。第 1 に、理論的分析に用いた事例データに誰もがアクセス可能であるため、本研究の一環として行われた分析を再検証し、さらなる議論の俎上に載せることが容易である。第 2 に、システムは談話構造のパターンをモデル化する本研究の主たる目的以外の研究用途に利用することも可能であることから、談話事象に対する多角的な分析を推進するきっかけともなり得る。第 3 に、談話に特化したコーパス・システムが比較的少ない現状において、本研究で開発したシステムは、言語研究のみならず、言語の教育や学習にも役立てられる可能性がある。

３．研究の方法

ひと言で談話といっても、実際には様々な種類のものが存在する。モノローグ的なものとして構成される談話もあれば、ダイアログ的な談話もある。また、書き言葉であっても、話し言葉であっても、言語学的にはある種の談話としてみなすことができる。本研究では、このように多岐にわたる対象の中から、英語の話し言葉による談話に含まれる論理構造に焦点を当てて、認知言語学（とりわけ認知文法と構文文法）の観点から分析と記述を試みた。

談話事例の採取を効率的に行うため、TED（https://ted.com）が公開している英語プレゼンテーションのトランスクリプトを言語コーパスとして検索することができる TED Corpus Search Engine（TCSE）を用いた（Hasebe 2015; 長谷部 2018）。TCSE は本研究の代表者が開発・公開しているシステムであり、これに新たな機能とデータを加える作業を行った。

これらの作業を経て、本研究では談話レベルでの現象を記述・分析するためのモデルを提案するに至ったが、その過程では、これまでの認知言語学研究におけるモデル化の手法を精査し、これに倣うことに努めた。認知言語学における様々な理論的概念は、必ずしも自然言語だけに特化していない一般的認知能力や外部世界における普遍的構造（例：メタファー的思考の背後にある「写像」の構造）の存在に依拠している。そこで、談話というプロセスの概念構造をモデル化するにあたっても、そのような能力や構造に目を向け、語・句・文のレベルの現象に対する既存のモデルとの整合性と連続性を担保する形でのモデル化を目指した。

４．研究成果

(1) 談話の論理展開パターン

自然言語による発話は、一般的に、談話の中で何らかの機能や効果を発揮することが期待される。中には、例えばある種の音声形式を用いて驚きや不満を表出するときのように、情意的側面が強調されるものもあるが、多くの場合、状況や文脈と論理的な関連性を持ち、そのような関連性があることを前提として、話者と聞き手は、連続的に発せられる発話から、いわば談話空間（discourse space）と呼ぶべき概念構造を脳内に構築する。またそうした概念構造は、話者と聞き手との間で大きく相違するものとはならず、大枠において並行的もしくは鏡像的なものとなることが想定される。

長谷部（2021）では、英語の談話結合子（discourse connectives）に着目して、そのような論理的構造の中から、とりわけ重要度ないしは使用頻度が高いとみられる下記の(i-a)～(i-c)のパターンを取り上げて分析・モデル化を行うと共に、TCSE から得た多数の事例を用いた検証を実施した。

(i) a. Connectives of causal relationship
b. Connectives of adversative relationship
c. Connectives for elaboration

(i-a)はある種の因果関係によって先行文脈（前件）と新たな発話（後件）の概念構造が関連づけられるタイプであり、聞き手にこれを促す働きを担う結合子の代表例として so, therefore, because, since が挙げられる。次の(i-b)は逆接的または対比的な関係によって前件と後件を関連づけるパターンであり、although, in contrast, instead などの結合子例である。また(i-c)は先行文脈内の要素に対する詳述的内容を導入する働きを担う結合子のタイプであり、in fact, in other words, that is (to say)などが例である。

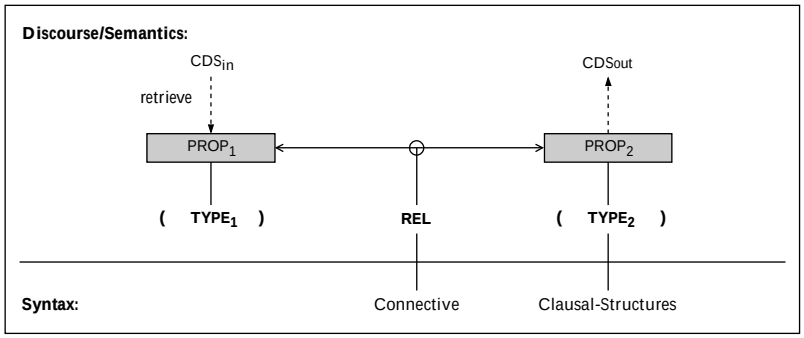

図１：談話結合子の構造

長谷部(2021)では、こうした談話結合子による論理展開パターンが一種の構文(construction)として捉えられることを指摘すると共に、必要な理論的整備を行った。具体的には、Ronald Langacker が認知文法の枠組みの中で導入した現在時発話空間(current discourse space, CDS)の概念を用いて、談話結合子構文を一種の関数として定義した。これは、入力値として与えられた CDS から第１の命題的内容（PROP1）を抽出し、後続する第２の命題的内容（PROP2）と結びつけた内容によって CDS を更新した結果を出力とする関数である（図１を参照）。

さらに、Sweetser（1991）が示した３つの意味領域―内容領域（content domain）、認識様態領域（epistemic domain）、発話行為領域（speech-act domain）―が、談話結合子の概念構造にも深く関わっていることを示した。ただし、Sweetser（1991）はこれらの意味領域が多分に相互排他的な関係にあり、文脈に応じて単一の意味領域が選択されるかのように論じているのに対し、本研究では、３つの意味領域の区分が必ずしも厳密ではなく、また、個々の事例において複数の意味領域が、前景化の程度は異なるものの同時並行的に関わっており、談話の論理構造を支える重要な基盤となっていることを指摘した。

(2) 談話の線条性とモナド的構造

図 1 に示したような線条的な認知処理は「内容物を包みから取り出し、それに変更を加えて再度包み直す」という wrap/unwrap のメタファーを用いて表示することもできる。このように、線形的な処理の連続的な適用を wrap/unwrap のメタファーを用いて記述することは、ソフトウェア工学における関数型プログラミングのパラダイムで論じられるモナド(monad)の概念と一致する。

モナドとは数学の圏論（category theory）における概念に由来しており、関数型プログラミングの文脈においては、次の 3 つの条件を満たす、ある種のデータ型として捉えられている。

(ii) a. データを環境に閉じ込めてモナドを生成する手続き（return）が存在する。
b. モナドの環境からデータを取り出して処理を行い、再び環境に閉じ込める手続き(bind)が存在する。
c. モナドの処理過程で入れ子になった環境を平坦化する手続き（join）が存在する。

このような観点から、長谷部（2024 予定）では、本研究の談話論理構造モデルを「談話の積層構造モデル（stack-based structure model of discourse)」と名付け、図 2a および図 2b のような 2 次元および 3 次元の構造体として記述する手法を提案している。それは、談話を各時間点の発話により導入される概念要素が、先行文脈の中で構築された構造を有機的に発展させていくプロセスとして記述する枠組みである。またそれは、(1)で示した考察を、「記憶」に関連する基本的認知メカニズムと、数学的背景とソフトウェア工学における有用性という「必ずしも自然言語だけに特化していない」構造に一致／依拠した枠組みでもある。このように、言語における概念構造の記述にある種のメタファー（ないしはアナロジー）が関わってくることは、本研究にとって重要な点である。Hasebe（2024 予定）はこの点に関する探究の一環として執筆した Littlemore（2019）の書評論文である。

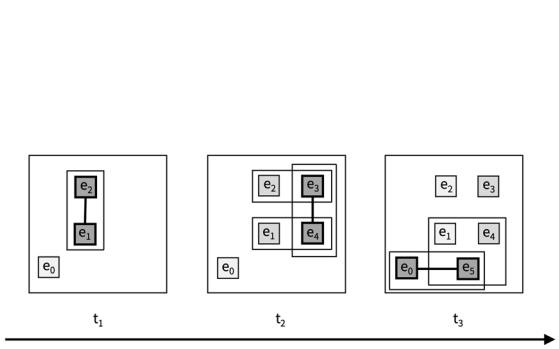

図 2(a)：積層構造モデルの例（2D）

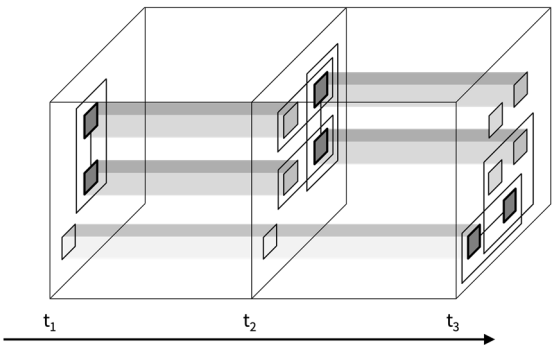

図 2(b)：積層構造モデルの例（3D）

(3) TED Corpus Search Engine (TCSE)

前述の理論的考察を進めるにあたっては、英語における話し言葉による談話の形態の 1 つとして 10 分から 20 分程度のプレゼンテーションのトランスクリプトを主たる分析の対象とした。このような形態のデータは、話し言葉による談話の例として必ずしも「自然」であるとは言い難い。しかし、話し手（講演者）から聞き手（聴衆）に向けて行われるトークの談話は、先行する文脈をほとんど持たない開始時点から、ある目的を達成して迎える終了時点まで、1 つ 1 つの発話が、先行する発話に何らかの形で関連付けられながら局所的な「談話ユニット」を構成し、またそれらがさらに大きなユニットを構成するような形で積み重ねられていく。それは、自然言語による論理展開の構造として 1 つの典型と呼べるだろう。そこで、そのようなデータを本研究だけでなく、広く言語（英語）の研究、教育、学習に役立てられるよう、TCSE に新たな機能とデータを追加した。

本研究の過程で追加または改善した機能と利用可能なデータの詳細を(iii)に示す。またシステムに格納しているデータの基本情報を表 1 に示す。

(iii) a. テキストに含まれる談話結合子のハイライト機能
b. TF-IDF に基づいたトピック表示機能
c. 主要構文のリストから事例を検索する機能
d. 高性能な解析エンジン（spaCy）による統語解析結果の向上
e. 名詞チャンクの自動検出による高度な構文検索機能

表1：TCSE の基本データ（2023 年 5 月）

検索可能トーク数（英語）	4,938
検索可能要素（語）の総数（token）	9,906,358
検索可能要素（語）の総数（type）	95,844
利用可能対訳言語数	34
日本語で検索可能なトーク数	4,102

なお、TCSE を活用した言語学分野及び言語教育分野の研究はすでに数多く行われており、その一部を TCSE のウェブサイトで示している（https://yohasebe.com/tcse/biblio）。

(4) 新たな展開と今後の展望

本研究の実施期間後半には、大規模言語モデル（large language model, LLM）の存在が自然言語処理や人工知能に関連するコミュニティで話題に上るようになり、2022 年 11 月には OpenAI による ChatGPT の公開により一般的にも広く知られるようになった。ここまでに示したように、本研究では認知言語学の概念や枠組みに基づいた理論的研究、実際の言語データを用いた検証、分析の再現性と拡張性を念頭においた言語データのコーパスシステム化を並行的に行ってきたが、人間による自然言語の入力を期待通りに解析し、それに応じた自然言語の応答をきわめて自然なテキストで返すことができる GPT をはじめとする LLM の技術は、理論言語学や言語教育の領域においても今後、非常に大きく広範な影響を及ぼすことは間違いないと思われる。

そのような中、(2)の「談話の線条性とモナド的構造」に示した本研究の成果は単に認知言語学における理論的な貢献となるだけでなく、理論言語学研究の成果を自然言語処理技術の向上に役立てたり、あるいは、自然言語処理技術の成果として生み出された高性能な LLM を理論言語研究に活用するための「接点」となり得ると考えられる。そこで本研究では、OpenAI のテキスト補完 API を用いて、「モナド的構造としての談話」という観点から、LLM に基づく人工知能エージェントとの会話が可能なプログラムを設計・開発した（長谷部 2023）。本プログラムは実験的段階ではあるが、すでにオープンソース・ソフトウェアとして利用可能であり、本研究で行った（1）認知言語学の概念や枠組みに基づいた談話に関する理論的研究、（2）実際の言語データを用いた検証、（3）分析の再現性と拡張性を念頭においた言語データのシステム化を今後さらに推進していくための重要な足掛かりになると期待できる。

＜引用文献＞

Goldberg Adele. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg Adele. 2006. *Constructions at Work: The Nature of Generalization in Language. Oxford*: Oxford University Press.

Sweetser, Eve.1991. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.

Hasebe, Yoichiro. 2015.“Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks”*Procedia* 198, 174-182.

長谷部陽一郎. 2018.「TED Corpus Search Engine：TED Talks を研究と教育に活用するためのプラットフォーム」『英語コーパス研究』25, 159-172.

Hasebe, Yoichiro. 2021.“An Integrated Approach to Discourse Connectives as Grammatical Constructions”PhD dissertation, Kyoto University.

長谷部陽一郎. 2023.「Monadic Chat：テキスト補完 API で文脈を保持するためのフレームワーク」『言語処理学会大 29 回年次大会発表論文集』3138-3143.

Hasebe, Yoichiro 2024（予定）.“Metaphors in the Mind: Sources of Variation in Embodied Metaphor”*English Linguistics* 40.

長谷部陽一郎. 2024（予定）.「談話の積層構造モデル―言語の線条性と概念構造の展開に関する試論―」『認知言語学論考』18.

Langacker Ronald W. 1987. *Foundations of Cognitive Grammar,* Vol.1, *Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker Ronald W. 1990. *Foundations of Cognitive Grammar,* Vol.2, *Descriptive Application*. Stanford: Stanford University Press.

Langacker Ronald W. 2008.

Littlemore, Jeannette. 2019. *Metaphors in the Mind: Sources of Variation in Embodied Metaphor*.Cambridge: Cambridge University Press.

＜ソフトウェア＞

TED Corpus Search Engine (TCSE)　https://yohasebe.com/tcse
Monadic Chat　https://github.com/yohasebe/monadic-chat

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 長谷部陽一郎	４．巻 10
２．論文標題 英語as叙述構文の認知的考察	５．発行年 2021年
３．雑誌名 コミュニカーレ	６．最初と最後の頁 1-20
掲載論文のDOI（デジタルオブジェクト識別子） 10.14988/00028121	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 長谷部陽一郎	４．巻 －
２．論文標題 Monadic Chat:テキスト補完APIで文脈を保持するためのフレームワーク	５．発行年 2023年
３．雑誌名 言語処理学会第29回年次大会発表論文集	６．最初と最後の頁 3138-3143
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yoichiro Hasebe	４．巻 40
２．論文標題 Metaphors in the Mind: Sources of Variation in Embodied Metaphor	５．発行年 2024年
３．雑誌名 English Linguistics	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 長谷部陽一郎	４．巻 18
２．論文標題 談話の積層構造モデル―言語の線条性と概念構造の展開に関する試論―	５．発行年 2024年
３．雑誌名 認知言語学論考	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　1件）

1．発表者名 長谷部陽一郎
2．発表標題 データ構造としての「構文」―構成性、線条性、記号性の観点から―
3．学会等名 同志社ことばの会 2021年度 年次大会 特別研究発表（招待講演）
4．発表年 2022年

1．発表者名 長谷部陽一郎
2．発表標題 語の意味的粒度とコロケーションに関する試論
3．学会等名 英語コーパス学会2018年度春季研究会
4．発表年 2018年

1．発表者名 長谷部陽一郎
2．発表標題 談話結合子による論理展開のパターンについて
3．学会等名 同志社ことばの会 2022年度 年次大会
4．発表年 2023年

1．発表者名 長谷部陽一郎
2．発表標題 Monadic Chat:テキスト補完APIで文脈を保持するためのフレームワーク
3．学会等名 言語処理学会第29回年次大会（国際学会）
4．発表年 2023年

〔図書〕　計1件

１．著者名 石川 慎一郎・長谷部 陽一郎・住吉 誠	４．発行年 2020年
２．出版社 開拓社	５．総ページ数 288
３．書名 コーパス研究の展望	

〔産業財産権〕

〔その他〕

(1) TED Corpus Search Engine (TCSE)
URL: https://yohasebe.com/tcse

(2) Monadic Chat
URL: https://github.com/yohasebe/monadic-chat

(3) 博士学位論文
著者名：Yoichiro Hasebe
論文名: An Integrated Approach to Discourse Connectives as Grammatical Constructions
学位授与大学：京都大学
学位授与年月日: 2021年1月25日
DOI: https://doi.org/10.14989/doctor.k22900

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関