理研話し言葉コーパスの概要とその利用法の一例

高橋祐介、岩下志乃、伊藤紀子、岩爪道昭、杉本徹、*小林一郎、菅野道夫
理化学研究所脳科学総合研究センター
*お茶の水女子大学理学部情報学科

1 はじめに

我々は，日常言語による情報処理を行う「日常言語コンピューティング」[岩爪他 2003]というパラダイムを提案している。このパラダイムは選択体系機能言語学[Halliday et al 1999 他]に依拠し、これに基づくテクスト処理システムとしてセミオティックベースを開発している。その構築のため、我々は2001年から2002年にかけて対話音声データの収集を目的とした実験を4回にわたって行い、それをもとに対話コーパスを構築した。実験の内容は、初心者ユーザがエキスパートの助言に基づき、コンピュータを用いて文書作成を行うものである。本発表では、コーパス利用法の一例として、与えられたタスクの進行状況に関するタグをコーパスに付与する手法を提示し、セミオティックベース開発への応用の方法を紹介する。

2 背景と本稿の構成

「日常言語コンピューティング」プロジェクトにおいては、ヒトの知性における言語の果たす役割を重要視し、従来の数値や形式的記号論理に基づく情報処理から、人々の日常の言語使用を用いた情報処理へのパラダイム・シフトが提案される。日常の言語使用において、コンテクストの果たす役割は非常に重要である。したがって、本プロジェクトではテクストにはコンテクストの特質が具現されているとする選択体系機能言語学を採用する。我々はこの言語理論に従ったテクスト処理システムである「セミオティックベース」を構築し、それを利用したシステムを開発・試作してきた[杉本他 2003, Ito et al 2004]。

本プロジェクトにおいて、対象ドメインとなるのは日常的なコンピュータの使用である。セミオティックベース構築のためにはこのドメインに対する言語資料が必要となる。そこで、我々は日常的なコンピュータの使用にかかわる被験者実験を行い、対話音声データを収集した。実験は、ワープロ使用時に関するユーザとインストラクターとの質疑応答、およびコンピュータを使ったユーザの文書作成の支援である。我々は、これらの実験で収集された対話データを基に、書き起こしコーパスと形態素解析情報つきデータベースを作成した。

本稿の構成は次の通りである。3 節ではコーパス収集のために行った実験の概要について述べる。4 節では、構築した書き起こしコーパスと形態素解析情報つきデータベースの概要と仕様を示す。5 節では、コーパスの使用例として、書き起こしコーパスに付与されたステージ情報を用いてステージベースを構築する手法について述べる。

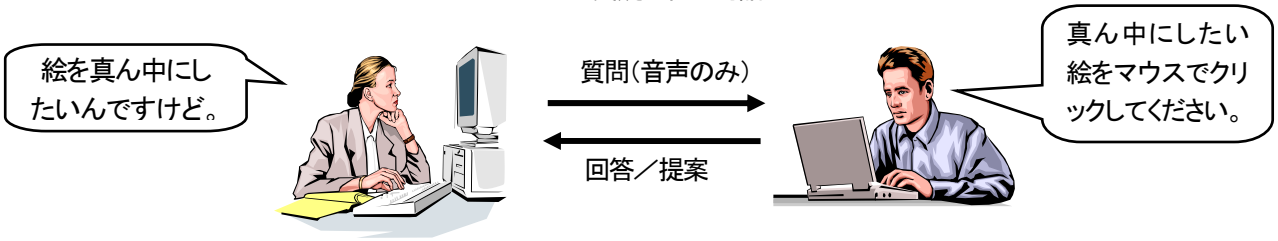
3 実験の概要

被験者実験は、2001 年 6 月から 2002 年 12 月にわたって、4回行われた。すでに述べたとおり、実験は、ワープロを使った文章作成における問題解決をドメインとする。

実験の内容は、回ごとに少しずつ異なるが、概ね質疑応答型と依頼応対型とに分けることができる。質疑応答型においては、ユーザが自ら文書を作成し、操作方法に関してインストラクターに質問するというものである。一方、依頼応対型では、ユーザは直接文書を作成せずに文書の内容に関する支持を秘書に与えて文書を作成してもらうというものである。

例えば、質疑応答型の実験としては、次のようなものがある。ユーザとインストラクターとが一組になってワープロソフトによる文書作成を行う。このとき、ワープロソフトの習熟度という点において、ユーザは基本的に初心者、インストラクターは上級者である。ユーザはワープロソフトを用いてあら

図 1 質疑応答型対話



かじめ与えられた見本と同様の文書を作成する。その際、ユーザはワープロ操作に関する不明点をインストラクターに質問する。一方、インストラクターはユーザの質問を受けて、問題となっているタスクを推定し、解決方法をユーザに教示する。なお、ユーザとインストラクターは音声のみで交信し、画面は共有しない。つまり、インストラクターは実際にユーザが行っている内容に関して直接を直接見ることができないため、質問内容から類推することになる。なお、実験時間は一回あたり二時間で、文書が未完成でも作業は打ち切られる。

一方、依頼応対型の実験としては、次のようなものがある。被験者はユーザ役、秘書役、プロファイリング役、アプリケーション役、アプリケーション操作補佐役に分かれる。ユーザ役はワープロソフトであるテーマに基づいた文書を作るための希望を秘書役に伝える。ここでは双方とも対面せず、マイクとヘッドフォンによる音声のみを使って交信する。別室にはアプリケーション操作役(以後、操作役とアプリケーション操作補佐役(以下補佐役)がおり、秘書役はチャットソフトを利用して非対面で補佐役にユーザの希望を補完して伝える。補佐役は直接対面して音声でユーザの希望を操作役に伝え、操作役はワープロソフトを実際に操作する。このとき操作画面は全員が共有してみることができるようになっている。よって、ユーザは操作された画面を見てさらに希望を秘書に伝えていく。また、プロファイル役はユーザの発話と画面を見ながらユーザの好みや個人情報など、ユーザについて気づいた点があれば記入していく。実験時間は一回当たり2時間30分のものと1時間のものとがあり、いずれも時間が来たら文書が未完成でも作業は打ち切られる。

なお、4回の実験におけるセッション総数は80、実験時間の合計は144.5時間である。

4 理研書き起こしコーパスの内容と仕様

以上のような被験者実験で得られた言語データをもとに、書き起こしコーパスと対話コーパスを作成した。以下、それぞれの仕様に関して述べる。

図2に示した書き起こしコーパスは、依頼応対型のコーパスのうち、ユーザ役と秘書役との対話から抜粋したものである。左から発話番号、タイムスタンプ、発話者(L は秘書役、R はユーザ役)、発話の順に記載されてある。400～500 ミリセカンド程度の発話の空きがあるところで区切って1単位とする。また、一人の発話に関して、200～400 ミリセカンド程度の発話の空きがあるところで読点を入れ、文が終わったと判断できるところで句点を挿入する。

用字法に関しては、一般用語に関しては漢字かな混じり文で一般的に使用されている用字法を用いている。固有名詞に関しては、文字の特定できるものはその文字を利用するが、特定不可能なものに関しては、カタカナを使用している。また、数字は原則として漢数字だが、必要に応じて算用数字を利用している部分もある。英字は原則として使用していない。擬態語はひらがなを用い、擬音語はカタカナを用いている。書き起こし時に聞き取り困難、文字化不可能などの箇所もあったが、それらに関しては、@で表記した。なお、発話のなまけ、言い誤り、言いよどみ、意味不明な言葉の断片に関しては、原則として処理していない。

このほかにも、文字には現れない被験者のアクションや、コンピュータ特有の用語に関してもタグを振ってある。例えば、図2において、作成している文章にかかわる表現に関しては＜ ＞のタグを付与し、ユーザのアクションに関しては[A]のタグを付与し、具体的なアクションを記述する。

次に、形態素解析情報付きデータベースの仕様について述べる。原則として、基本的な仕様そのものは国立国語研究所の『話し言葉コーパス』[前川2001 他]における短単位データベースに準拠してある。つまり、上記の書き起こしコーパスに形態素解析を加え、さらにKWIC情報を付加したものと考えて差し支えない。したがって、形態素ID、転記ファイル、転記情報、前文脈、KWICキー、後文脈、代表形、代表表記、発音形、品詞、活用情報が記載される。なお、用字法は書き起こしコーパスに依存する。

なお、コーパスのサイズは、書き起こしコーパスで約21500KBになる。

5 使用例:ステージタグ付与

本コーパスにステージ情報を付与することによって、コンテクストベースのコンテンツであるステージベース構築する上での資料としたので、その概要を示す。すでに

図2 書き起こしコーパスの例

446	01081.815-01082.837	L	はいその次お願いします
447	01082.628-01083.790	R	はいえ＜場所＞
448	01083.932-01084.196	L	はい
449	01084.338-01085.368	R	＜場所＞でえっと[A 手振り]
450	01086.283-01089.927	R	点点縦@、＜和光市民文化センター＞
451	01087.110-01087.392	L	はい

図 3 ステージ付与

L:	縦サイズで、	用紙サイズ設定	
	じゃアプリケーションのほうはWordを使うということでいいでしょうか	用紙サイズ設定	
R:	はい	用紙サイズ設定	
L:	はいあのタイトルはどうしましょうか	タイトル設定	タイトル決定
R:	タイトルはえ＜クリスマス＞	タイトル設定	タイトル決定

[Iwashita et al 2003]において、本コーパスにおける質疑応答型対話データを基にしたヘルプ出力個人化の提案がなされているが、ここで提示するのは、依頼応対型対話データ利用の一例である。

選択体系機能言語学において、コンテクストを具現するテクストの性質は、具現する言語の意味的・語彙文法的特徴だけではなく、具現するテクストの展開にも現れる[1]。このテクストの展開の仕方はジャンル構造と呼ばれ、ジャンル構造はテクストの展開の段階（ステージ）の遷移として表現される。

セミオティックベースにおいてステージに関する情報はコンテクストベース内にあるステージベースに格納してある。ステージベースには、ステージそのものの情報と、ステージ間の遷移情報とが記述されている。これらの情報を用いて、ステージベースは所与のコンテクストにおけるインタラクションの展開の方法を記述している。

ところで、コーパスに収録した実験に代表されるインタラクションにおいては、二つのコンテクスト情報を分けて考えなければならない。一つは、文書作成に関するユーザと秘書との間の質疑応答あるいは依頼応対に関するコンテクストで、これはコーパスに記述された発話に具現される。一方、このインタラクションの結果完成した招待状などの定型文書も独立した一つのテクストである。これは、差出人のユーザから受取人に対するイベントへの招待というコンテクストが定型文書に具現されたものと考えることができる。それぞれのコンテクストをプライマリコンテクスト、セカンダリコンテクストと区別する[2]。

したがって、我々が想定しなければならないジャンル構造もプライマリジャンル構造とセカンダリジャンル構造の二種類が存在する。プライマリジャンル構造は、ユーザ役と秘書役との間の対話の展開を記述したものであり、セカンダリジャンル構造は、出来上がった文書の構造を記述したものになる。それぞれのジャンル構造は複数のステージによって構成される。よって、それぞれを表現するステージをプライマリステージ、セカンダリステージと呼ぶ。

なお、コンテクストベース構築にあたって、プライマリコンテクストの中にセカンダリコンテクストが埋め込まれているという視点をとる。したがってステージベースにおいても、セカンダリステージは一部のプライマリステージに埋め込まれるという構造をとることになる。

5.2. ステージ情報の付与

共同で招待状を作成するタスクにおけるステージ情報を取得するため、依頼応対型の実験における対話コーパスを基にして、実際にステージ情報を取得した。ここでは、2001 年第 2 回の実験における 9 セッションのユーザ役対秘書役の対話を利用した。まず、書き起こしコーパスと完成した招待状を参考にプライマリステージとセカンダリステージを設定した。

プライマリステージは、対話の進行に応じて、挨拶、タスク同定、ワープロ起動、文書編集、文書確認、保存、印刷の 7 つに設定した。セカンダリステージは、文書編集の下位区分として存在し、実験の結果完成した文書と対話データとを基に、用紙サイズ設定、余白設定、背景装飾、タイトル設定、タイトル調整、本文入力、本文調整、画像挿入、画像調整の9つを設定した。次にコーパスに対してステージを付与した。書き起こしコーパスにおける文の区切りは必ずしも文法的な基準ではないため、まず区切りを節単位に成型し、次にコーパスに対してプライマリステージとセカンダリステージを人手で図 3 のように付与した。図 3 には、左から発話者（L は秘書役、R はユーザ役）、発話（節単位に成型済み）、プライマリステージ、セカンダリステージが記されている。

その結果、プライマリステージに関しては、図4のような遷移が確認できた。つまり、基本的には「挨拶→ワープロ起動・タスク同定→文書編集→状態確認→保存・印刷→作業完了」の段階を経てインタラクションが進行していくということである。ただし、ワープロ起動とタスク同定との段階、および保存と印刷との段階に関してはそれぞれ順不同であり、

[1] 選択体系機能言語学において、コンテクストは活動領域、役割関係、伝達様式の三項の組み合わせとして表現される。また、コンテクストベースの概略については、[高橋他 2002, 岩爪他 2003]を参照。

[2] [山口 2000]は、選択体系機能言語学におけるコンテクストに関して、一次的コンテクスト（状況）と二次的コンテクスト（使用域）とを設定する。ただし、これは一つのコンテクストに対する視点の違いを表すもので、それぞれのコンテクストに対応するテクストは単一のものである。一方、本稿で言うプライマリコンテクストとセカンダリコンテクストは、異なる二つのテクスト、つまりそれぞれプライマリテクスト（コーパス）、セカンダリテクスト（定型文書）に対応する。

図4　プライマリステージの遷移

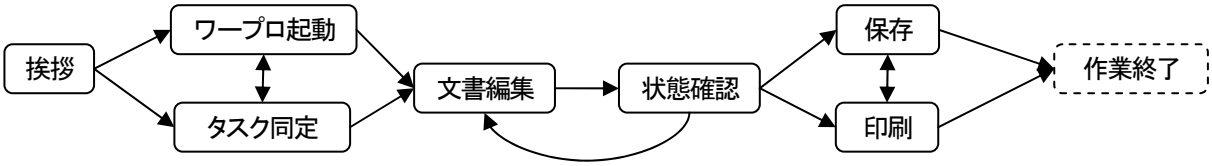
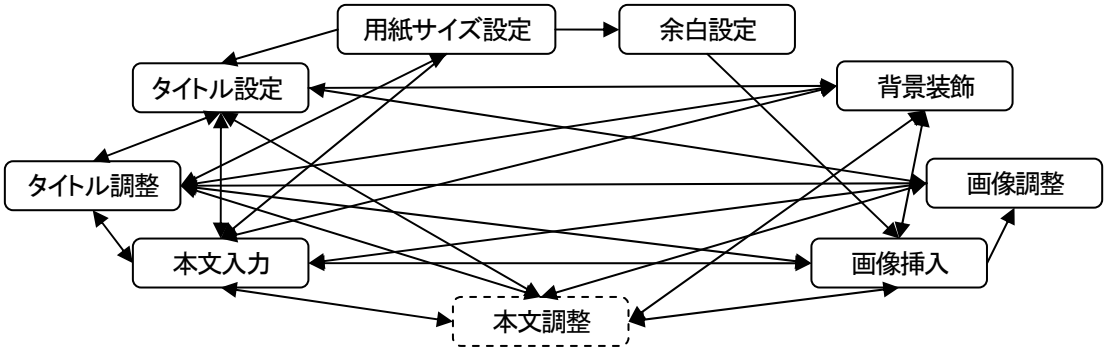


図5　セカンダリステージの遷移



文書編集と状態確認の段階に関しては繰り返しも見られる。また、作業終了の段階に関しては、すべてのテクストで確認できるというわけではない一方、セカンダリステージの遷移に関しては図 5 のように確認できた。これを見てわかるように、遷移に関してはかなり自由度が高い。これは、招待状の作成に関しての手順には自由度が高いことによるものと思われる。

6　まとめ

本稿では、理研話し言葉コーパスの概要を示し、その利用法の一例としてステージ情報の付与について説明した。

なお、本コーパスは 2004 年中にモニター公開、2005 年には完全公開を予定している。当チーム web ページ (http://www.brain.riken.go.jp/labs/lbis/) にコーパス公開に関する情報を掲載する予定である。

付記

本コーパスの構築にあたっては、株式会社アルカディア様のご支援を賜りました。データ整理にあたっては、中村文子さん、佐久間夕美さん、原梨恵子さんのご協力を賜りました。また、国立国語研究所の前川喜久雄先生には、「話し言葉コーパス」に関する情報を賜りました。記して感謝申し上げます。

参考文献

[岩爪他 2003]岩爪道昭、小林一郎、杉本徹、岩下志乃、高橋祐介、伊藤紀子、菅野道夫 2003, 日常言語コンピューティング(第 2 報)―日常言語に基づく計算機資源の管理・実行環境を目指して― 人工知能学会論文誌 18-1, pp. 45-56

[Halliday et al 1999] M.A,K Halliday and Christian M.I.M. Matthiessen 1999, Construing Experience through Meaning: A Language Based Approach to Cognition, Cassell

[杉本他 2003] 杉本徹, 岩下志乃, 岩爪道昭, 小林一郎, 高橋祐介, 伊藤紀子, 菅野道夫 2003, セミオティックベースを使った日常アプリケーションシステム. 2003 年度人工知能学会全国大会, 2G2-09

[Ito et al 2004]Noriko Ito, Toru Sugimoto, Michio Sugeno 2004 A systemic-functional approach to Japanese text understanding. Proceedings of CICLing -2004

[Iwashita et al 2003] Shino Iwashita, Ichiro Kobayashi, Noriko Ito, Toru Sugimoto and Michio Sugeno 2003, Personalization of Help System Output in the Framework of Everyday Language Computing. Proceedings of KES'03, pp.439-445

[前川 2001] 前川喜久雄 2001, スピーチのデータベース, 日本語学 20-5, pp.12-27

[高橋他 2002] 高橋祐介、伊藤紀子、藤城浩子、菅野道夫 2002, セミオティックベースにおけるコンテクスト層の検討, 2002 年度人工知能学会全国大会, 3B1-02

[山口 2000]山口登 2000, 選択体系機能理論の構図―コンテクスト・システム・テクスト, 小泉保(編), 言語理論における機能主義 ―誌上討論会―, くろしお出版, pp. 3-47