特定領域研究「日本語コーパス」平成18年度公開ワークショップ（研究成果報告会）予稿集

特定領域研究「日本語コーパス」

平成18年度公開ワークショップ（研究成果報告会）予稿集

平成 19 年 3 月 17 日、18 日

文部科学省科学研究費特定領域研究
「代表性を有する大規模日本語書き言葉コーパスの構築：
21世紀の日本語研究の基盤整備」
総括班
JC-G-06-01

特定領域研究「日本語コーパス」

平成18年度公開ワークショップ（研究成果報告会）予稿集

2007年3月17日（土）／3月18日（日）

Program [プログラム]

3月17日（土）第一日目

13:00　■開　会

13:00～13:50　■特定領域研究の紹介

「特定領域研究『日本語コーパス』― 目標、進捗状況、そして夢 ―」 前川 喜久雄（国立国語研究所）

13:50～15:00　■デモンストレーション

国立国語研究所 KOTONOHA コーパス等のデモンストレーション

■招待講演

15:00～15:50　「語彙調査からコーパスへ」 宮島 達夫（国立国語研究所名誉所員）

15:50～16:40　「大規模テキスト処理の時代」 長尾 眞（情報通信研究機構理事長）

16:40～16:50　休　憩

16:50～17:50　■パネルディスカッション「コーパスが拓く可能性」

田野村 忠温（大阪外国語大学）／砂川 有里子（筑波大学）／田中 牧郎（国立国語研究所）

荻野 綱男（日本大学）／奥村 学（東京工業大学）

17:50　■閉　会

3月18日（日）第二日目

9:00　■開　会

9:00～11:15　■計画班研究進捗状況報告（15分×8班、途中休憩15分）

●データ班　山崎 誠　●ツール班　松本 裕治　●電子化辞書班　伝 康晴　●日本語学班　田野村 忠温

休　憩

●日本語教育班　砂川 有里子　●言語政策班　田中 牧郎　●辞書編集班　荻野 綱男

●言語処理班　奥村 学

11:15～13:15　■昼食及びデモ・ポスターセッション（順不同）

「『現代日本語書き言葉均衡コーパス』におけるサンプリングの概要」

丸山 岳彦、柏野 和佳子、山崎 誠、佐野 大樹、秋元 祐哉、稲益 佐知子、吉田谷 幸宏

「『現代日本語書き言葉均衡コーパス』における著作権処理について」

森本 祥子、前川 喜久雄、小沼 悦、新井田 貴之、松下 愛、吉田谷 幸宏、神野 博子、大石 有香

「『現代日本語書き言葉均衡コーパス』における電子化フォーマットの概要」

山口 昌也、高田 智和、北村 雅則、間淵 洋子、西部 みちる

「『現代日本語書き言葉均衡コーパス』における短単位の概要」

小椋 秀樹、小木曽 智信、小磯 花絵、冨士池 優美、相馬 さつき、渡部 涼子、服部 龍太郎

「日本語コーパスでのSketch Engine実装の試み」

投野 由紀夫

「確率的単語分割ツールとその利用」

浅原 正幸

「タグ付きコーパス検索ツールの開発」

谷口 雄作、新保 仁

「『日本語コーパス』用Yahoo！知恵袋データについて」

岡本 真、木戸 冬子、佐古 智正

13:15～16:10　■計画班研究発表（20分×8班、途中休憩15分）　※発表順

「『現代日本語書き言葉均衡コーパス』の基本設計について」

山崎 誠

「セグメントとリンクに基づくコーパス・アノテーション・ツールの設計と実装」

徳永 健伸、乾 健太郎、野口 正樹、三好 健太、飯田 龍、小町 守

「単独ラベラによる大規模アクセントラベリングとそれを用いた統計的アクセント結合処理の実装」

峯松 信明、黒岩 龍

「因子分析を用いた程度副詞と述語等の共起関係の研究試論」

服部 匡

休　憩

「日本語教育における語彙シラバスの作成について」

山内 博之

「国語教育と語彙指導」

鈴木 一史

「共起関係およびコロケーションに関する研究の流れ

― 計量言語学分野、自然言語処理分野および辞書データなどを中心に ―」

荻野 綱男、荻野 孝野

「語彙概念構造辞書の構築による意味役割分析」

竹内 孔一

16:10～16:15　休　憩

16:15～16:40　■全体討議

16:40　■閉　会

Contents [目次]

特定領域研究の紹介

3月17日（土）第一日目　13:00〜13:50

特定領域研究『日本語コーパス』― 目標、進捗状況、そして夢 ―

▶前川 喜久雄（国立国語研究所）

特定領域研究「日本語コーパス」－目標,進捗状況,そして夢－

前川喜久雄（領域代表者：国立国語研究所研究開発部門）†

Priority-Area "*Japanese Corpus*" Project: Goals, Progress, and Dreams

Kikuo Maekawa (Dept. Lang. Res., National Institute for Japanese Language)

１．本領域全体の目標

文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（略称「日本語コーパス」)は平成 18 年 7 月に採択が内定し、同年 9 月から活動を開始した。研究期間は平成 22 年度までの 5 年間である。

本領域にはふたつの目標がある。ひとつは現代日本語のコーパス言語学的研究の基盤を整備するために現代日本語書き言葉の大規模な均衡コーパス(balanced corpus)を構築することである。このコーパスには必要な著作権処理を施して誰でも利用できるものとする。もうひとつの目標は構築途上のコーパスを様々な領域で活用してコーパス日本語学の可能性を探り、同時に構築中のコーパスを評価することである。活用と評価の試みは、狭義の言語学だけでなく、国語教育・日本語教育・辞書編纂・自然言語処理などの領域で実施する。

大規模なコーパスとその活用法が整備されれば日本語の言語学的分析が進展することは当然である。しかし大規模コーパスの影響はそれだけにとどまるものではない。日本語に関する様々な知的活動が面目を一新する可能性がある。表 1 は筆者がこれまで各方面に対して本領域（および後述する KOTONOHA 計画）で構築するコーパスの価値を説明する際に利用してきた表である。特定領域研究の最終ヒアリングのプレゼンテーションでもこの表を利用した。研究費獲得のために作成する効能書は大風呂敷になりがちであるが、本領域の場合、ほとんど掛け値なしにこれだけの効能を期待できると考えている。実際、次節で紹介する本領域の計画研究班はこれらの領域の大部分をカバーしたものになっている。本稿の後半では、これらの用途のいくつかについて、私の考えを述べることにするが、その前に本領域の内部構造について説明しておこう。

表１．書き言葉均衡コーパスに想定される用途

日本語学	主観を排した言語分析 現代日本語の実態に即した文法・語彙の分析
日本語教育	基本語彙、基本構文、共起関係
国語教育	教育用基本語彙の選定
辞書編纂	用例収集、共起関係
心理学・認知科学	**実験における言語刺激の統制**
自然言語処理	**統計的学習データ、アルゴリズム評価用データ**
音声合成・認識	**言語モデルの学習**
国語政策	**常用漢字の見直し、正書法の提案**
文化資源	**未来の文化財としての価値**

† kikuo@kokken.go.jp

２．本領域の構成と計画班の目標

図１に領域全体の構成を示した。本領域は総括班と計画研究班８班から構成されている。特定領域研究では複数の研究班から構成されるグループを研究項目と呼ぶが、本領域の研究項目は「A01 コーパスの構築」と「B01 コーパスの評価」の２項目であり、前者には３班、後者には５班が属している。これにくわえて平成 19 年度からは項目 B01 に関する小規模な研究が５件程度、公募によって発足する予定である。以下に各班の目標を簡単に紹介しておく。

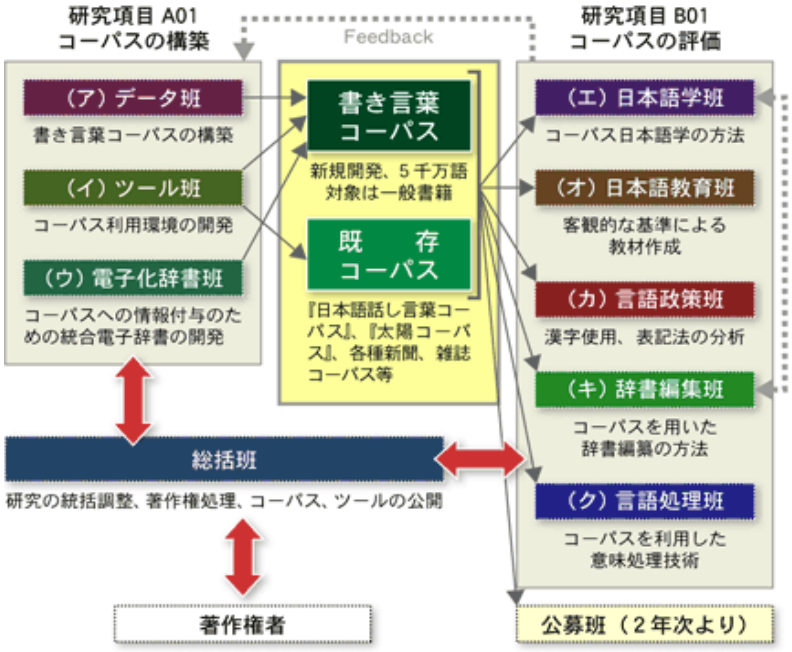


図１．本特定領域研究の構成

２.１　データ班

データ班（班長：山崎誠、国語研）は本領域全体の要をなす計画班であり、「現代日本語書き言葉均衡コーパス」のうち書籍部分、約 5000 万語を構築することが目標である。実際には、コーパスの設計、サンプリング、著作権処理、電子化（文字および文書構造の XML 表現）、形態論情報の５サブグループに分かれて作業を実施している。

２.２　ツール班

ツール班（班長：松本裕治、奈良先端大）の目標は、様々な基礎・応用分野において書き言葉コーパスを有効に利用するために必要とされる研究用情報を付与する（タグ付けする）ために必要とされる自動解析システムおよびタグ付け支援ツールの構築である。タグの仕様を定め、コーパスのサブセットに対して実際にタグ付けを実施することもおこなう。

２.３　電子化辞書班

電子化辞書班（班長：傳康晴、千葉大）の目標は、形態素解析システム用電子化辞書 UniDic を整備・拡充・改良し、本領域がめざす大規模書き言葉コーパスの構築を支援するととも

に、日本語学・日本語教育学・自然言語処理・音声情報処理など、多様な目的に適した統合的な電子化辞書およびその利用システムを提供することにある。UniDic の整備作業はデータ班の形態論グループと密接に協力して実施されている。

２.４　日本語学班

日本語学班（班長：田野村忠温、大阪外大）は、具体的な事例研究を通して書き言葉コーパスの価値を明らかにし、日本語の新しい研究領域・手法を開発するとともに、学界に対してコーパスを用いた日本語研究の啓蒙・普及を図ることを目標とする。

２.５　日本語教育班

日本語教育班（班長：砂川有里子、筑波大）の目標は、データ班と協力して日本語教材コーパスを構築し、日本語教材で用いられている日本語ならびに学習項目の内容や配列について実態を把握することにある。また書き言葉コーパスを利用した日本語教材作成方法および日本語教育のためのコーパス検索ツールについても検討する。

２.６　言語政策班

言語政策班（班長：田中牧郎、国語研）の目標は、書き言葉コーパスを利用して国語施策と国語教育に役立てることのできる語彙表と漢字表を作成し、それらを活用する方法を開発することである。

２.７　辞書編集班

辞書編集班（班長：荻野綱男、日本大）は、全体としてコーパスを用いた辞書編纂方法の研究を目標としている。個別的にはコーパスを利用したコロケーション辞書の概念設計と試作、統語論的観点による辞書作成、コーパスによる実態調査をふまえた語義分析と辞書での記述方法の研究などのテーマに取り組んでいる。

２.８　言語処理班

言語処理班（班長：奥村学、東京工業大）の目標は、書き言葉コーパスを利用して意味解析にかかわる自然言語処理研究を発展させることにある。

２.９　総括班

総括班（班長：前川、国語研）は、計画研究班間の連絡調整、成果の広報および外部評価に関わる業務をうけもつ。またデータ班と協力してコーパスを公開するために必要な著作権処理業務を実施する。

３.『現代日本語書き言葉均衡コーパス』

図１の中央には本領域で構築するコーパスが「書き言葉コーパス」として描かれており、その規模は 5000 万語と想定されている。また、この図には描かれていないが、特定領域の計画書には本領域で構築するのは書籍のコーパスであると明記されている。書籍コーパス

は、それ単独では現代日本語書き言葉全体の均衡コーパスとはなっていないことに注意が必要である。

国立国語研究所研究開発部門言語資源グループでは、本領域と同じ 2006～2010 年度の期間に、雑誌、新聞、その他の書き言葉を対象とするコーパスを構築する。それらと本領域開発の書籍コーパスをあわせた全体が書き言葉全体の均衡コーパスとして機能することになる。この全体を『現代日本語書き言葉均衡コーパス』と呼ぶ。英語名称は *Balanced Corpus of Contemporary Written Japanese* であり BCCWJ と略する。

図２に BCCWJ の概念図を示す。ここからわかるように、BCCWJ は 3 種類のサブコーパスから構成されている。BCCWJ の設計については、明日の研究発表セッションで山崎氏が発表することになっているが、ここでもその特徴を簡単に紹介しておこう。

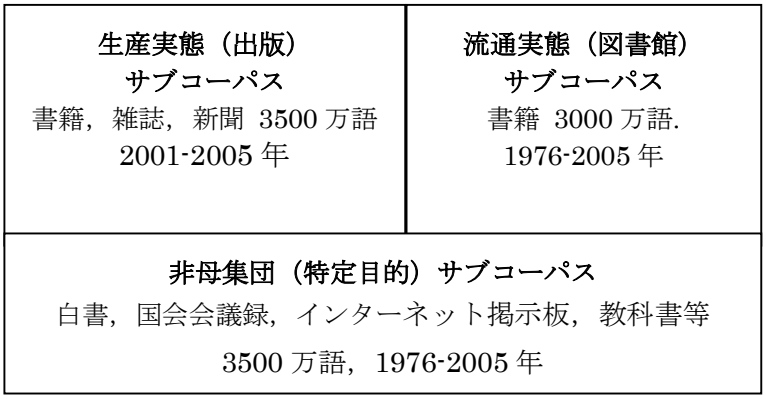


図２.『現代日本語書き言葉均衡コーパス』を構成するサブコーパス

３.１　生産実態サブコーパス

BCCWJ は「生産実態」「流通実態」「非母集団」の三つのサブコーパスから構成されている。生産実態サブコーパスは 2001 年から 2005 年のあいだに出版された書籍、雑誌、新聞の文字の総体を母集団としたコーパスである（そのような母集団をどのように規定するかについては明日のポスターセッションでの丸山らの発表参照)。このコーパスでは、文字数さえ同一であれば、大ベストセラーもゾッキ本も同じ一冊として扱われる。つまり、本サブコーパスは日本語テキストを「異なり」(type)の観点から把握しようとするコーパスなのである。

３.２　流通実態サブコーパス

しかし、コーパスユーザーのなかには、生産よりも受容の実態に興味をもつ人も少なくないだろう。その場合、100 万部のベストセラーに含まれるテキストは 100 冊しか売れなかった本のテキストの 1 万倍の確率でコーパスに採録されるべきであろう。しかし現実には書籍、雑誌の実売部数を正確かつ悉皆的に把握することはほぼ不可能である。

そこで我々は公立図書館に収蔵されている書籍を母集団とするコーパスを構築することを考えた。一定数以上の図書館に共通して収蔵されている書籍は、ある程度まで社会に流通し、受け入れられた書籍であるとみなすことができるだろうと考えたのである。これが図２の「流通実態」サブコーパスである。

現在東京都下の公共図書館に収蔵されている全書籍のうち ISBN が付与されているものは異なりで約 100 万冊である。そのうち例えば 10 以上の自治体（区や市）の図書館に収蔵されているという基準をたてると、これを満たす書籍は異なりで約 48 万冊である。流通実態サブコーパスの母集団となるのは、このようにして規定された書籍の集合（に含まれるすべての文字）になる予定である。

流通実態サブコーパスは、誰の興味もひかなかった本および公序良俗に反するなど種々の理由で公共図書館にふさわしくないと判断された本が除外される点と、30 年程度の時間の幅をもった書籍が対象となっている点で、生産実態サブコーパスの書籍部分とは本質的に異なるコーパスになる。

３.３　非母集団サブコーパス

図２下部は「非母集団」サブコーパスである。生産、流通コーパスの対象とはなりにくいが特定領域の計画班ないし国語研の研究のために必要とされるデータ（教科書や白書の類)、影響力が大きかったと考えられる書籍（ベストセラー、教科書)、典型的な書き言葉との比較対象のために重要と判断されるもの（WEB 上のテキスト、国会会議録）などが含まれる。そのなかには母集団からの無作為抽出でサンプルを得るものもあるが（例えば白書や国会会議録)、母集団を確定できないもの（ウェブのテキストなど）や母集団は確定できるが無作為抽出をおこなわないもの（教科書類）もあるので「非母集団」と呼んでいる。

３.４　サンプル長について

このように BCCWJ では、少なくとも生産実態および流通実態サブコーパスにおいては、明確に規定された母集団からサンプルを無作為抽出することによって母集団の特性を偏りなく反映したコーパスを構築しようとしている。言語研究のために無作為抽出法を利用することは、諸外国は知らず、こと我が国の言語研究では珍しい試みではない。国立国語研究所による語彙調査では新聞、雑誌、教科書、テレビ放送などの母集団からの無作為抽出法が 1950 年代から実施されてきている。しかし、BCCWJ の生産実態サブコーパスにおけるように、新聞、雑誌、書籍という複数のジャンルをまたがって構成される母集団に対して、層化無作為抽出法を適用するのは日本でも（そして当然世界でも）初めての試みであろう。

ただし、国立国語研究所による従来の統計的語彙調査における無作為標本抽出と BCCWJ における標本抽出との間には重要な相違点もある。それは、抽出するサンプルの長さである。これまでの語彙調査のサンプルが、数十字程度の非常に短いサンプル長を利用しているのに対して、BCCWJ のサンプル長は 1000 文字とはるかに長くなっている。これはコーパスとしての利用を考えれば当然のことである。また BCCWJ では長さを 1000 字に固定したサンプルの他に、文書の構造を反映した可変長サンプルも同時に採取することにしている（山崎 2007、丸山ほか 2007)。

４．KOTONOHA 計画

図３は国語研のコーパス整備計画 KOTONOHA の概念図である。KOTONOHA は明治から現代にいたる近現代日本語の全体像を把握するためのスーパーコーパスであり、多数の要素コーパスから構成されている。BCCWJ もそのひとつであるが、図中の「太陽」と「CSJ

（日本語話し言葉コーパス）」は国語研が既に公開を済ませたコーパスであり、それぞれ近代語書き言葉と現代語話し言葉を対象としたものである（国語研 2005, 2006, 前川 2004）。

BCCWJ は KOTONOHA の最重要構成要素であるが、KOTONOHA の整備は BCCWJ の完成後も継続される。図にはその開発の候補となるコーパスも示されている。「太陽」と BCCWJ を繋ぐ書き言葉コーパス、CSJ で十分にカバーできなかった対話や雑談を対象とした現代語話し言葉コーパスのふたつである。また、BCCWJ は 2011 年の公開後も新しい日本語に対応するため、数年おきに生産実態サブコーパスを拡張してゆく予定である。

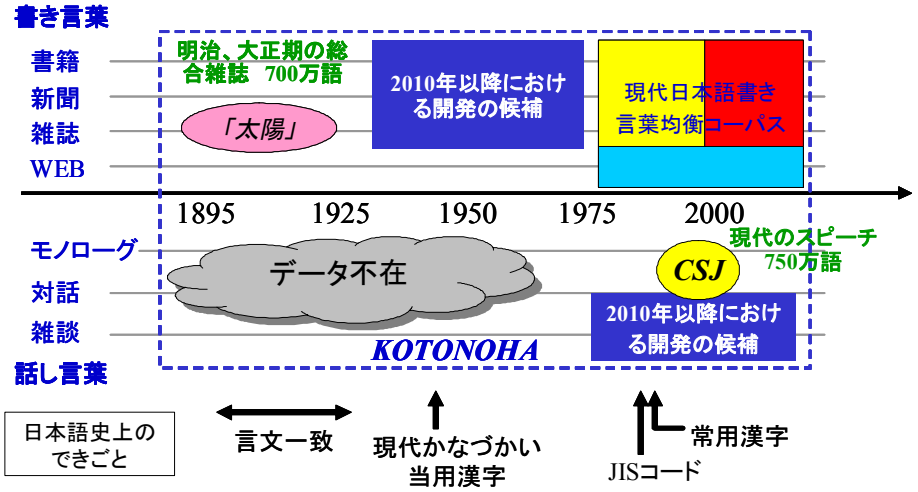


図３．国立国語研究所のコーパス整備計画 KOTONOHA

５．BCCWJ 構築の現状

BCCWJ は 2006 年（平成 18 年）4 月から本格的な構築を開始した。当初は国立国語研究所の運営費交付金に依拠して構築をすすめたが、同年８月以降は特定領域研究の予算も利用できるようになった。ここで両予算の切り分けについて一言しておこう。３節で述べたように、特定領域研究のデータ班が構築するのは、BCCWJ のうち書籍に関係する部分である。具体的には生産実態サブコーパスの一部と流通実態コーパス全体が書籍に関係する。この他に非母集団サブコーパスに過去 30 年間のベストセラーを対象としたコーパスを加える計画もあるので、その部分にも特定領域の経費を利用する可能性がある。この点を指摘したうえで、以下では BCCWJ 全体についてまとめることにする。BCCWJ の構築作業には、以下の４ステップを踏んで実施される。

５.１　サンプリング

サンプリングと総称される作業には、１）母集団の確定、２）母集団からのサンプル無作為抽出、３）抽出されたサンプルに該当する書籍、雑誌等の現物ないしコピーの入手、その他の作業が含まれている。２）と３）が分離しているのを不思議に思う方があるかもしれないが、これは今日では、サンプルを無作為抽出する作業が電子化された出版データ（国会図書館が作成する J-BISC の元データなど）を用いてコンピュータ上で仮想的に実施されるからである（丸山ほか 2007）。

今年度は最初に白書データ（非母集団サブコーパスの一部約 500 万語）のサンプリング

を終了した。次に生産実態サブコーパス全体の母集団を確定し、そのうち書籍部分（サブコーパスの約 76%に相当）についてサンプルの無作為抽出を終えた。さらにそのうち 2500 サンプル分について当該箇所のコピーを作成した。

５.２　著作権処理

著作権処理は BCCWJ 構築において最も困難が大きいと予想していた作業であるが、この１年の経験で予想以上に困難な作業であることがわかってきた。この問題については６節で触れることにする。本年はまとめて大量に著作権をクリアできそうな案件から処理を進めた。

A) 国会会議録（30 年分）
B) インターネット掲示板（ヤフー知恵袋、1 年半分）
C) 政府刊行白書（30 年分）

の三点については交渉がほぼ終了した。いずれも膨大な量のデータであるが、それぞれから 500 万語ずつを無作為抽出して、非母集団サブコーパスに格納する予定である。これらのデータについては、できるだけ早い時期にインターネット上でデモンストレーション用に試験公開する予定である。

他に、生産実態サブコーパスに含まれる新聞データについて、大手新聞社３社（読売、毎日、産経）からデータを提供していただけることになった。同じく書籍データには 13000 件程度のサンプルが含まれる予定であるが、このうち 2500 サンプルについて今年度中に著作権処理作業を実施する予定である。具体的な処理方法については森本ほか(2007)参照。

５.３　電子化

電子化とはサンプルとして抽出されたテキストを機械可読形式に整える作業である（山口ほか 2007）。日本語の文字集合としては JISX0213:2004（いわゆる JIS 第 3,4 水準）を、それを表現する文字コードとしては Unicode（UTF16）を利用する。

BCCWJ 構築作業で電子化という場合、いわゆるアノテーションは含まれないが、文字および文書構造に関する基本的な情報に関するタグ付けは電子化作業の一部として実施している。文字に関するタグには、JISX0213:2004 の表外字であることを示す missingCharacter、ruby（フリガナ）、superScript（上付文字）などの特殊な活字組を処理するためのタグ、誤字・誤植を示す correction などがある。文書構造に関するタグには、sample（サンプル全体）, article（記事全体）, cluster（タイトルに対応するまとまり）, paragraph（段落）, sentence（文）などがあり文章を階層構造に沿って表現する。現在までに政府刊行白書データ（1500 サンプル、500 万語）のタグ付けが終了した。

５.４　形態論情報付与

日本語テキストを語に分割してその品詞情報を付与するのが形態論情報付与作業（形態素解析）である。日本語は分かち書きの習慣が存在しないために、語の定義自体が議論の対象となる言語であるが、BCCWJ では CSJ がそうであったように、「短単位」と「長単位」という二種類の単位に則った二種類の形態論情報を提供する予定である。1 億語のテキストに形態論情報を付与する作業は、当然ながら自動化される必要がある。この目的のために、電子化辞書班とデータ班は協力して電子化辞書 unidic の拡張と整備を続けており、年度当

初に約 46000 語であった短単位見出し語数を現在 10 万語以上まで拡張した(小椋ほか 2007)。拡張された unidic と形態素解析ソフト「茶筌」をあわせ用いた場合の解析精度を白書のデータ約 500 万語分で評価してみたところ約 95%であった。これは単語境界の設定、代表形・代表表記・品詞の付与がすべて成功した場合の精度である。

６．個人情報保護との関係

BCCWJ の構築作業を開始してほぼ 1 年が経過する現在、我々が直面している最大の困難は予想どおり著作権処理の問題である。我々は当初、著作権無償利用に依頼をどの程度許諾していただけるかに不安を抱いていたが、実際に最大の障壁となっているのは、むしろ個人情報保護法である。2003 年に成立し 2005 年に施行された個人情報保護法によって、個人情報取扱事業者は個人情報の厳密な管理を要請されており、出版社の大部分はこの義務を負っている。サンプルの著作権者から利用許諾をいただくためには、まず先方の連絡先を知る必要があるが、個人情報保護法はそのような情報を簡単には提供させないための法律なのである。

この問題については現在出版各社と交渉をすすめている最中である。幸い BCCWJ および KOTONOHA 計画の価値は各社とも躊躇なく認めてくださるので、今後とも鋭意交渉をすすめてゆくつもりである。しかし正直なところ思いもかけぬ伏兵に出会った気がしている。個人情報保護法によって著作権者との連絡がつけられない状況には、どうにも納得しがたいものがある。この法律が著作権者のありうべき利益を阻害していることにはならないのだろうか。

７．コーパスが拓く可能性

以上、本特定領域の目標を述べコーパス構築の現状を報告した。コーパスの応用面に関する研究の進捗状況については、私が拙い紹介をするよりも、明日の午前中に各班長による研究進捗状況報告が予定されているので、そちらをお聞きねがいたい。

以下では BCCWJ あるいはさらに理想的な均衡コーパスが完成されたときに、どのような研究上の可能性が拓けてくるかについて、自由に私見をのべさせていただく。一部、夢に属する話題にも触れることになるので、そのつもりでお聞きいただきたい。

７.１　Corpus-based と corpus-driven

コーパス言語学の世界では corpus-based investigation と corpus-driven investigation を分けて考えようという主張がある(Tognini-Bonelli, 2001)。前者は、従来から言語研究において検討されてきた諸問題をコーパスを利用して解決しようとする研究である。一方後者は、コーパスそのもののなかから従来の言語研究では認識されてこなかった現象を発見し、それを解決しようとする研究である。前者にとってコーパスは研究ツールであるが、後者にとってのコーパスは研究対象そのものである。この主張に従えば、corpus-driven な言語研究は従来の言語研究と或る意味で隔絶したものになる必然性がある。その姿はどのようなものになるのだろうか。この問題を考える手がかりは文法性(grammaticality)という概念に見出せるように思える。

７.２　文法性判断

文法研究では文の文法性判断をおこなうが、その判定が研究者によって異なることがある。文の適格性の判断に幅がありうるという事実は言語の本質を考察するうえで非常に重要な問題である。例えば以下の文の文法性判断を要求されたとき、これを非文と判断する人は少なくないだろう。

(1) 昨晩、あるいは昨夜おそく、このあたりは雨が降ったです

しかし、これは実際に用いられた日本語である。しかも 40 年以上にわたって 60 刷を重ねてきたロングセラーに見つかる用例である[1]。翻訳だから日本語がおかしいのだ... というのはこの場合理屈にならない。翻訳者は立派な日本語母語話者だからである。

手元にある種々のテキストデータを検索してみると話し言葉らしい用例がぽつぽつとみつかる。(2)~(4)は「文芸春秋」の座談会、(5)は国会会議録、(6)は CSJ 中の用例である。もちろん Google 等の検索でも類例を発見できる。

(2) まさに正岡子規だったですよ

(3) それだもんで参っちゃったですよ

(4) ああ、これは本腰を入れなきゃいかんと思ったですね

(5) 政府は一体具体的に何をやったですか

(6) 初めて海外に行ったですよ

これらの用例を読んで、それが用いられたであろう文脈を想像してみる。そうすると私などは(1)を非文と断定しにくく感じられてくる。適当な合理化の口実が与えられれば、むしろ適格文にすら思えてくる。本例の場合「ああ、話し言葉ならたしかにこう言うこともあるな」と思えてくるのである。

もうひとつ例を挙げておこう。(7)は作家今東光が書いた随筆の一節である[2]。

(7) 僕たちは警察に信頼して好いと思う

私はこの例については誤植の可能性が捨てきれないと考えてきたが、先日研究所の同僚の井上優さんにきいたところ、ある種の動詞の補語における「を」と「に」の転換は稀ではないとのことであった。その場で井上さんがあげた動詞の例は「協賛する」であったが、実際「青空文庫」中に(8)を見出すことができる[3]。

(8) 日本の法律は内閣または各省が立案して、議員はこれを協賛するという立
て前となっていた

---

1 バルドゥイン・グロルラー著、阿部主計訳「奇妙な跡」、江戸川乱歩編「世界短編傑作集２」創元推理文庫、1961 年初版。

2 今東光「赤線消ゆ・東光辻説法」半藤一利編『「文芸春秋」にみる昭和史（三）』文芸春秋, 1988（原文の『文芸春秋』への掲載は 1948 年）.

3 中井正一「国立国会図書館について」。広辞苑第 5 版には「明治憲法の下で、帝国議会が、法律案および予算案を有効に成立させるために統治権者である天皇に対し必要な意思表示をすること」とある。現在の「協賛する」とは語彙的意味の外延が若干異なっているかもしれない。

「～を協賛する」は「青空文庫」中にこの一例だけのようだが、「～に信頼する」の例はもっと簡単に見つかる。

(9) 生活を維持するに足る詩的天才に信頼したために胃袋の一語を忘れた[4]
(10) 安心して、僕に信頼したらよかろう[5]
(11) あまりに現在の脆弱な文明的設備に信頼し過ぎているような気がする[6]
(12)まつは、善良で私に信頼し、同時に無智だ[7]

これらは明治生まれの文筆家の日本語である。その時期の日本人にとっては「～に信頼する」が適格文であったことが窺われる。この場合もやはり、一度(9)以下の例を体験してしまうと私はもう現代語としても(7)を非文とする気がなくなってしまう。自分自身が「～に信頼する」と書くことはないかもしれないが（ただし絶対にとはいいきれない)、(7)を適格文として受容することにはこだわりがなくなってしまうのである[8]。

７.３　文と非文の境界

従来の言語研究、特に生成文法理論では文と非文の境界は明確に（二値的に）定まるものと考えてきた。適格文の集合を白で、非文の集合を黒で表現すれば、図４の左パネルの状態である。しかし、文法性判断にゆれが存在する状態が稀な例外でないとすれば、文と非文の境界はむしろ連続的な変化としてとらえるべきだろう。色に例えるならば右パネルのようなグラデーションである。

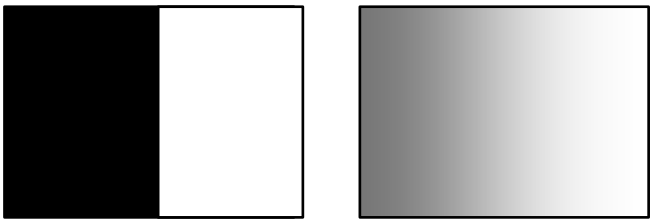


図４．文法性判断の離散性と連続性

Corpus-driven な言語学がめざすべき目標のなかには、このようなグラデーション（すなわち文法性の程度をあらわす連続量）の計算法と、グラデーションが何に起因するかを説明しうる言語理論の構築が含まれていてしかるべきだろう。

第一の目標については、自然言語処理や音声認識で利用されている統計的な言語モデル（N グラムなど）が、単語列の生起確率を与えるという形で、現在でも或る程度まで情報を提供してくれる。非文ないしは非文に近い文の生起確率はコーパスから計算できないというのは単純すぎる考え方であり、巨大なコーパスとクラス化された言語情報を用いれば推定が可能になる。Pereira(2000)は初期の生成文法の有名な例文（Colorless green idea sleeps furiously と Furiously sleep ideas green colorless）の生起確率を推定している。

---

4 芥川龍之介「河童」
5 夏目漱石「二百十日」
6 寺田寅彦「石油ラムプ」
7 宮本百合子「文字のある紙片」
8 井上(2001)はここで例とした類の現象をとりあげた一般向け解説である。

第二の目標については、文法性判断のゆれに関与する可能性がある要因を悉皆的に探りださねばならない。そのなかには、上でもみたように、言語のレジスターの差、言語の変化（時間変化）などが含まれる。これらはいずれも言語共同体の多様性を生み出す要因として従来から指摘されてきたものであり、社会言語学の領域で多くの研究が積み重ねられてきている(例えば Labov, 1972 参照)。

それらの研究を継承して、より広範な説明力を有する理論を構築するためには、多くのレジスターにまたがる言語資料を大量に分析する必要があるのだが、ここで、Google 等の検索データはそのような分析目的のためには不適当であることを指摘しておきたい。インターネット検索では、テキストのレジスターやジャンル、あるいは書き手の年齢や性別などの情報、すなわちサンプルの社会的属性を知ることが非常に困難だからである。上述第二の目標を達成するためにはきちんと設計された大規模な均衡コーパスが絶対的に必要である。

８．超巨大コーパスの夢（まとめにかえて）

言語の文法性判断は、上でみたように、言語資料との接触経験に強く影響されることがある。そのため、文法性判断に関する個人差を完全に解明するためには個々人がそれまでの人生においてどのような言語資料に接触してきたかについての知識が必要だと考えられる。しかし、そのような知識は、得ることが可能だろうか。

このような可能性を考えることは二十年前ならば笑うべき妄想であった。現在でも夢と呼ぶべきだろう。しかし個人が一生涯に接する程度の言語資料をコーパス化することに、もはや技術上の問題は存在しない。書き言葉は当然のこと、話し言葉であっても、ただ記録（録音）するだけであれば、個人が一生涯に発する程度の音声は、圧縮すればテラバイト級のハードディスクに収めることができる[9]。音声認識技術の発展によってはそれを実用上十分な精度で自動認識することもできるようになるだろう。

書き言葉に関していえば、実は個人の言語接触歴をすべて記録する必要もない。十分に大きな均衡コーパスが存在すれば、個人の言語接触歴をシミュレートすることができると考えられるからである。「～を協賛する」、「～に信頼する」などの書き言葉中心の表現であれば、年齢、性別、学歴、専門、趣味、職業、読書傾向などの社会的属性から、対象となる個人が過去に当該言語表現に接触した確率の期待値を計算できる可能性がある。

そのような計算を可能にするコーパスはどの程度の規模になるだろうか。私は試みに 2005 年 1 年間に自分が読んだすべての和書の記録をとってみた。その結果は約 2600 万文字、短単位になおせば 1530 万語程度になった。この調査は単行本だけを対象としたもので、新聞、雑誌、WEB 文書、マンガ、論文等を除外している。それらを適当に按配すれば 1 年で 2000 万語以上の書き言葉に接触していると思われる。仮にこのような接触状態を過去 30 年間継続してきたと仮定すれば、私がこれまでに接触した言語資料の総体は少なくとも 6 億語を超えることになる。BCCWJ 程度の規模のコーパスでは、私程度の読書量の人間の経験をカバーすることすらできないことがわかる[10]。少なくとも数十億語、望ましくは百億語規

---

9 東京工業大学の古井貞熙教授のご指摘による。

10 ただし読書傾向の差が語彙の差に与える影響程度のことは BCCWJ でも検討できそうだ。そのために必要となる書き言葉との接触量についての社会調査も特定領域研究で実施する予定である。

模の均衡コーパスが必要になりそうだ。

実際、世界を見渡してみてもコーパスが巨大化する趨勢が明らかである。Gigaword corpus という言葉が現実味をもって語られるようになってきた(Huang, 2006)。著作権の問題さえ解決できれば、将来の均衡コーパスは百億語規模にまで到達するのではなかろうか。ブラウンコーパスの四半世紀後に構築された BNC は前者の百倍のサイズを達成しているのである。情報技術の進歩によってコーパスの構築コストは今後とも低下してゆくだろう。2030 年代半ばに百億語の均衡コーパスが実現されても私は驚かないつもりである。

文　献

井上優(2001).「問 13」『新「ことば」シリーズ 14 言葉に関する問答集』国立国語研究所, pp.36-37.

小椋秀樹、小木曽智信、小磯花絵、冨士池優美、相馬さつき、渡部涼子、服部龍太郎(2007).「『現代日本語書き言葉均衡コーパス』における短単位の概要」（本予稿集）.

国立国語研究所(2005).『太陽コーパス』(国語研資料集 15), 博文館新社.

国立国語研究所(2006).『日本語話し言葉コーパスの構築』(国語研報告書 124), 国立国語研究所.

前川喜久雄(2004).「『日本語話し言葉コーパス』の概要」 日本語科学, 15:1, pp.111-133.

丸山岳彦、柏野和佳子、山崎誠、佐野大樹、秋元祐哉、稲益佐知子、吉田谷幸宏(2007).「『現代日本語書き言葉均衡コーパス』におけるサンプリングの概要」（本予稿集）.

森本祥子、前川喜久雄、小沼悦、新井田貴之、松下愛、吉田谷幸宏、大石有香、神野博子(2007).「『現代日本語書き言葉均衡コーパス』における著作権処理について」（本予稿集）.

山口昌也、高田智和、北村雅則、間淵洋子、西部みちる(2007)。「『現代日本語書き言葉均衡コーパス』における電子化フォーマットの概要」（本予稿集）.

山崎誠(2007).「『現代日本語書き言葉均衡コーパス』の基本設計について」（本予稿集）.

C-R. Huang (2006). “Automatic acquisition of linguistic knowledge: From Sinica corpus to gigaword corpus”, *Language Corpora: Their compilation and Application.* (Proceedings of the 13th NIJL International Symposium), pp.41-48.

W. Labov. *Sociolinguistic Patterns.* Philadelphia, Univ. Pennsylvania Press, 1972.

F. Pereira (2000). “Formal grammar and information theory: Together again?” *Philosophical Transactions of the Royal Society*, 358(1769): pp.1239-1253.

E. Tognini-Bonelli (2001). *Corpus linguistics at work (Studies in Corpus Linguistics: 6)*, Amsterdam/ Atlanta, GA: John Benjamins.

関連 URL

特定領域研究「日本語コーパス」http://www.tokuteicorpus.jp/

KOTONOHA 計画 http://www2.kokken.go.jp/kotonoha/

筆者個人 http://www2.kokken.go.jp/~kikuo/public/KMHP1.html

招待講演

3月17日（土）第一日目　15:00〜15:50

語彙調査からコーパスへ

▶宮島 達夫（国立国語研究所名誉所員）

語彙調査からコーパスへ

宮島達夫（評価委員：国立国語研究所名誉所員）†

From Word Survey to Language Corpus

Tatsuo Miyajima (Researcher Emeritus, National Institute for Japanese Language)

わたしは国語研究所で３０年以上はたらいており，その大部分は語彙調査の仕事だった。今研究所がとりくんでいるコーパスの作製は，語彙調査の発展という面があり，そのような観点からお話をしたいと思う。

１．国語研究所と現代語研究：統計調査と記述

国立国語研究所が創設されたのは，戦後まもない１９４８年１２月である。最終的には，言語政策に役立てるのが目的だったが，現代語の研究を確立したことは，研究所の最大の功績といってよいものである。今からみると奇妙にみえるかもしれないが，現代語の研究は国語研究所の成立によってはじまったのである。それ以前の「国語学」の対象としていたのは国語史であって，それも奈良・平安からせいぜい鎌倉・室町どまりだった。創立当初の研究をふりかえると，やることすべてが新しい，という熱気が感じられる。新しかったのは，現代語という対象だけではない。若い研究者たちは，つぎつぎに新しい研究方法を身につけていった。人文系の研究では個人研究が中心だが，国語研究所では創立のときから個人研究ではなく共同研究を建て前としてきた。このことが，（当時としては）大規模な語彙調査や全国にわたる方言調査を可能にした。まだめずらしかった録音機で，いちはやくナマの音声を録音して研究したり，理系にしかなかった電子計算機を文系でまっさきに導入したりなど，機器の使用にも積極的だった。社会言語学的な実態調査は世界的にみても早いものに属する。そのような新方法のひとつとして，統計の活用がある。

語彙・文法を対象とした研究について，おもなものをあげると，下の表のようになる。調査の名まえは略称をつかった。くわしい名称は参考文献を参照。

研究所創立後まもなく，国立国語研究所資料集２『語彙調査　ー現代新聞用語の一例ー』（１９５２）が出た。これは朝日新聞１ヶ月の統計をとったもので，いわば語彙調査の習作，といった感じのものである。また，その前年，国立国語研究所報告３『現代語の助詞・助動詞　ー用法と実例ー』（１９５１）が刊行された。戦前にも口語文法の本はいくつも書かれが，それは実態の調査にもとづいたものではなかった。本格的な調査のうえにたった文法書はこれがはじめてである。

こうして，一方では統計的手法をつかって言語の全体像を巨視的にながめる行き方，もう一方では微視的に言語事実を記述する行き方が，国語研究所の研究のなかに生まれた。

† miya_tt@nifty.ne.jp

	全体的統計	統計的分析	個別の記述	実例の提示
現代語の助詞・助動詞			◎	◎
朝日新聞調査	◎			
婦人雑誌調査	◎	○	○	
談話語の実態		◎		
総合雑誌調査	◎	○		
話しことばの文型			◎	○
『郵便報知』(明治)調査	◎	○		
(コンピューター導入)				
雑誌九十種調査	◎	○	○	
新聞調査(電算機)	◎			
(『分類語彙表』初版)				
動詞・形容詞・アスペクト			◎	○
教科書調査(電算機)	◎	○		
『中央公論』経年調査	◎	○		
テレビ調査(電算機)	◎	○		
国定読本索引	◎			
話し言葉コーパス			○	◎
『太陽』コーパス			○	◎

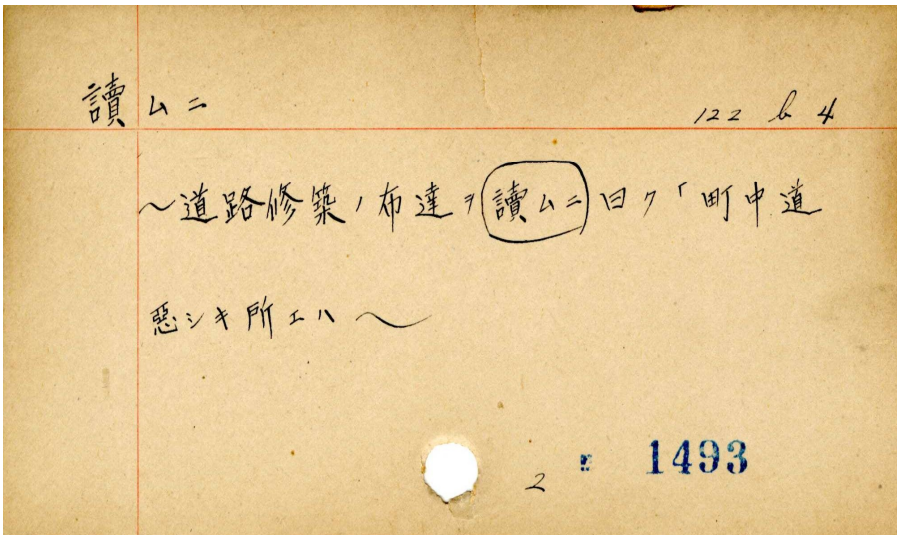
２．カードによる用例採集

語彙調査にあたっては，単語をカードに書きとって５０音順にならべ，集計する，という方法がとられた。しかし，この方法は手間がかかる。それで，婦人雑誌調査の途中から，あたらしい方法がとりいれられた。それは，前もって調査すべき箇所をカードに印刷しておき，採集すべき単語なり漢字なりに○をつける，というやり方である。その複製カードも，最初は手書きのガリ版だったのが，総合雑誌・雑誌九十種の調査では，邦文タイプライターによる謄写印刷になった。それでもカード作成にあたっては厳密に校正する必要があったが，動詞・形容詞の記述や『中央公論』の経年調査では，原文をコピーしてカードをつくるようになり，校正のわずらわしさもなくなった。

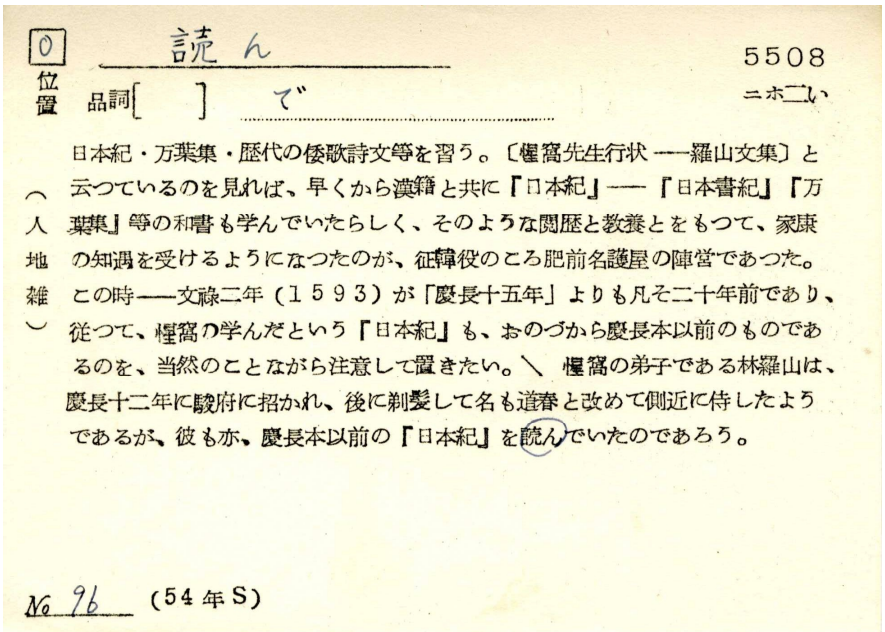
この＜発明＞は語彙調査の能率をあげただけではない。カードには，かなり長い文脈がはいるから，用法の分析に役立てることができる。また，たくさん作っておけば，あとになって思いついたテーマをしらべることもできる。総合雑誌や雑誌九十種の調査に使ったカードは動詞・形容詞の記述に役立ったし，動詞・形容詞の記述のために印刷したカードはアスペクトの研究に利用された。いわば，これはコーパスへの一歩手前の，コンピューター導入以前としては最後の段階である。ある意味ではコンピューターなしのコーパスといってもいいものである。国語研究所の初期に『語彙調査』（頻度の調査）と『現代語の助詞・助動詞』（用法の記述）とに分裂していた研究が，雑誌九十種調査で統合されたのは，

カードのおかげである。ヨーロッパやアメリカでも用例を採集するのにカードが使われたことは，たしかだが，タイプライターでの印字以外に，コピーによって複製するところまでいったかどうか，知りたいところである。

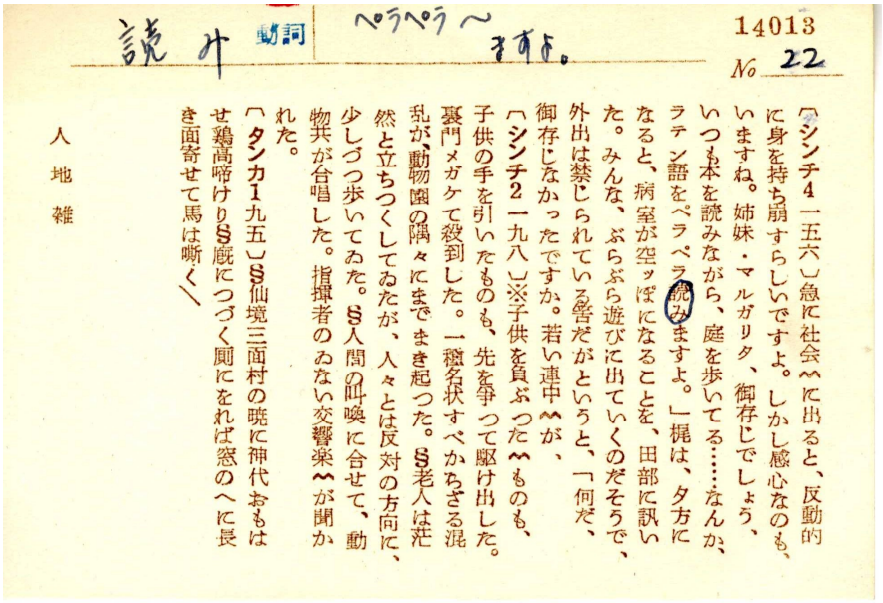
カードの例（手書き：郵便報知）



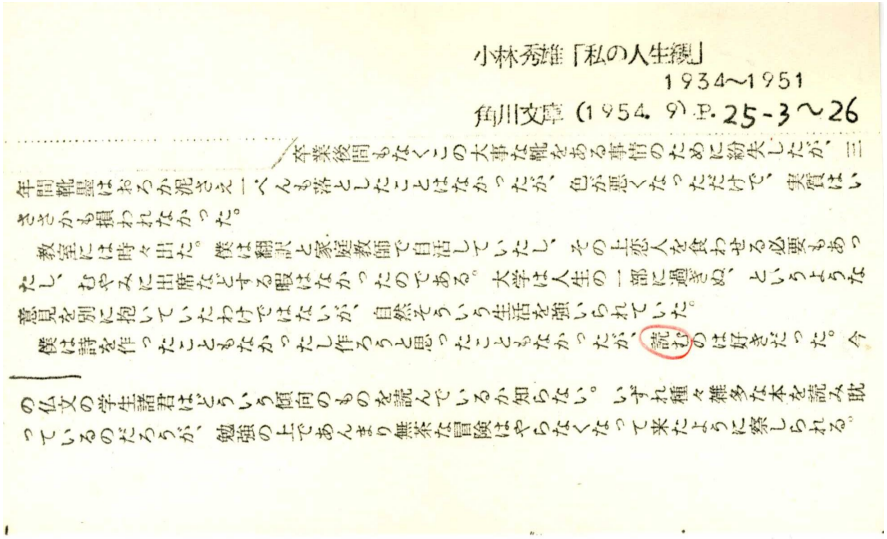
カードの例（邦文タイプ：総合雑誌）



カードの例（邦文タイプ：雑誌九十種）



カードの例（コピー印刷：動詞記述）



雑誌九十種の調査データは最近カードからコンピューターに入力された。また，これを受けついだ雑誌七十誌の調査は，最初からカードを使わずにコンピューター利用によってなされた。これらは，著作権の関係で公開しておらず，国語研究所内部の利用にかぎられているので，コーパスとしては不完全である。

なお，実例の調査にカードを利用したのは，国語研究所だけではない。研究所の外でも，古典語の研究には古くから使われていただろう。とくに意識的に実例主義をとり実行して成果をあげたのは，奥田靖雄氏を中心とする言語学研究会である。ただし，そこで作られた膨大なカードのおおくは，文庫本の文学作品を切り抜いてはりつけたもので，カードを

印刷・複製して利用する方向には，あまりいかなかった。

３．雑誌の語彙調査と基本語彙

日本語の語彙調査は，基本語彙・基本漢字をあきらかにする，という目的をもっていた。雑誌九十種の調査結果を記述した国研報告 21 は「調査結果は実態の記述をまずもって目標とするが，それにとどまらず，基本語彙の選定その他の国語国字問題を考える際の参考資料としても役立つ事を念願した。」(p.1)とのべている。おなじ趣旨のことは，それ以前の婦人雑誌・総合雑誌の調査報告書にもみえる。対象としてまず婦人雑誌をとりあげたのは，女性用語の研究のためではなく，婦人雑誌が生活に密着した記事をのせていて，生活基本語彙をしるのにふさわしい，と考えられたからだった。しかし，最大の範囲をカバーした雑誌九十種の調査でも，日本語の基本語彙をあきらかにする，という目標にはへだたりがある。

140 位「身頃」	343 位「増資」
513 位「ぬいしろ」	356 位「当社」
718 位「えりぐり」	479 位「配当」
785 位「ダーツ」	753 位「投資」

など，裁縫や経済関係の用語が上位にならんでいる一方，５１５８位に
「うたがう，応援，おもちゃ，金持ち，勘，看板，煙，交替，故障，さっさと，姿勢，市民，乗客，新年，スキー，スピード，センター，竹，ためる，知恵，散る，テープ，眠い，のんびり，話しかける，ハンカチ，昼間，文明，ボーナス，翻訳，満員，見本」
など，もっと基本的ではないかと思われるものが，ずっと下にある。(同じ順位の単語がたくさんならんでいるのは，そのような順位づけの方針をとったためである。）１９９４年の雑誌７０誌の調査結果も同様である。ここでは，１０００位以内に

417 位「素材」	434 位「ポイント」
561 位「設定」	648 位「ＣＤ」
671 位「性能」	828 位「ソフト」
747 位「本体」	858 位「アルバム」

など，特定の専門分野に多い単語があるのに対し，５０００位に
「梅，ガソリン，固まる，規則，給料，許可，故障，知らせ，杉，すし，ソース，そっと，近ごろ，爪，つらぬく，どっち，番地，ひげ，孫，間近，まね，都，夢中，役所」
などがある。

国語研究所が日本語教育のための基本語彙を選定したとき，直接参考にしたのは語彙調査の結果ではなく，いわば語彙調査にともなう副産物のようなかっこうで作られてきた『分類語彙表』だった。(ただし，旧版『分類語彙表』には雑誌九十種の上位語に印がつけてあるから，間接的には参考になったところもあるはずである。)

国語研究所の語彙調査は，雑誌九十種からあとは，日本語全体の基本語彙をあきらかにし，国語問題に寄与する，という大きな目標をたてていない。つぎの新聞の調査は，語彙

調査の結果をだすよりもコンピューター利用の成果をだすためのものだった。中学・高校教科書やテレビ用語の語彙調査は，たしかに全国民の言語生活に大きな影響をもつが，むしろ日本語の重要な１つの側面，位相をあきらかにする，という位置づけがふさわしいようにおもう。こうして，いろいろ問題はあるものの，雑誌九十種調査は，依然として日本語を代表する統計的調査としての位置をしめている。

４．雑誌の語彙調査と英語のコーパス

国語研究所による雑誌九十種の調査は，１９６２年の発表当時，語彙の統計調査として世界最高の水準にあっただけでなく，現在もそうだといっていいものである。英語の統計とくらべて評価してみよう。最初の本格的なコーパスは，アメリカ英語の Brown コーパス(1967)，ついでこれにならったイギリス英語の LOB[Lancaster-Oslo-Bergen]コーパス(1982)である。

４．１．見出し語立て

最初発表された形では Brown コーパスも LOB コーパスも, 見出し語立てをしていない，ナマの語形の統計であって，take,takes,took,taken,taking が別語とされる一方，助動詞の can と名詞の can とは区別がなかった。その後，品詞の区別はついたが，土手の bank と銀行の bank のように人間の目で判別しないと区別がつかないものは，そのままになっている。単語ごとに分かち書きされる表音文字の世界では，ナマの語形の統計は，コンピューターを使えば，なんの苦労もなくできる。しかし，それだけなら数分ですむところ，見出し立てをしようとすれば，数年かかるかもしれない。一方，漢字かなまじり文では，単語ごとに切るのにたいへんな労力がかかる。英語では「The National Institute for Japanese Language」が６つの単語からできていることは，だれでも分かるし，コンピューターの自動認識にも問題がない。しかし，「国立国語研究所」をどう切るか，という問題には，何とおりかの答えがある。おなじだけの労力を，その先同語別語の判別をして見出しを立てることにつぎこむとしても，いわば五十歩百歩である。だから，コンピューターを利用するようになっても，日本の語彙調査は人間の手と目で同語別語の判別をして見出し語をたてるのを原則とした。例外は，コンピューターを使った最初の大規模語彙調査である新聞用語の調査で，できるだけコンピューターにやらせるという方針のため，「一月」には和語の「ひとつき」と漢語の「いちがつ」がふくまれ，「いった」「言った」「行った」「言う」「行く」は，それぞれが別語とされた。こんな語彙表をだしても，あまり意味はない。それ以後の国語研究所の語彙調査は，教科書もテレビ用語も，機械と人間の共同作業で見出し立てまでした結果をだすのを原則とし，したがってあまり膨大な量を処理することはできなかった。

４．２．統計

雑誌九十種の調査のレベルがたかいというのは，見出し語が立ててあるだけでなく，標本抽出に厳密な無作為抽出がとられ，統計的な管理がしっかりしているからである。母集団は１９５６年度の雑誌九十種合計 226,358 ページ。そこから８分の１ページを単位に，

227 分の１，7,983 箇所を抽出し，１箇所あたり，ほぼ 55 語になる。また，度数のたかい語に使用率の推定精度をつけたことも，類のないことである。Brown コーパス・LOB コーパスは，やはり無作為抽出をしているが，ある段階で主観をまじえているようで，厳密には，母集団がはっきりしない。

４．３．代表性

しかし，雑誌九十種の代表するものは，あくまで母集団としての１９５６年度の雑誌九十種各号の総体，２２６，３５８ページであって，それ以上ではない。ここでは標本と母集団との関係がはっきりしているから，たとえば「～に」と「～へ」の量的比較を標本についてだけでなく母集団についても推定することができる。しかし，われわれがほんとうに知りたいのは，ある年度の雑誌九十種についてではなく，日本語についてである。学問的な態度をつらぬくかぎり，われわれは，一歩母集団をはなれれば「日本語の書きことば」全体についてはもちろん，つぎの年の雑誌九十種についてもなにもいえない，ということを，みとめなければならない。これは雑誌について全数調査をしたとしても同じである。一方，英語のコーパスの製作者は「(これらのコーパスの）真の代表性は，文章の重要なカテゴリー・下位カテゴリーをふくめるよう計画的に心がけ，盲目的な統計的選択にまかせなかったことから生ずる。」とのべている。統計をぎせいにしても，かれらがまもろうとしたのは，英語をよりよく＜代表する＞コーパスをつくることだった。それが成功したことは，Brown コーパスでアメリカ英語を，LOB コーパスでイギリス英語を代表させ，それらを比較するという研究がされていることからも分かる。tea がイギリスに，coffee がアメリカに多い，などというのは調べるまでもないが，女性をあらわす she,girl,woman がイギリスに，男性をあらわす he,boy,man がアメリカに多いというのは，調べてみてはじめて分かった結果だし，標本抽出のゆれでないとすれば，これがなにを意味するのかは，興味がもたれる点である。日本の調査は，雑誌九十種でも教科書でも「これが日本語だ」といえるほどの代表性はない。

今や「大規模」とはよべないかもしれないが，国語研究所が実施した程度の各種語彙調査が引きつづいてなされている点では，日本語は英語とならぶ特別な言語であるようだ。リーチというイギリスの学者は，最近も，ドイツ語について１００年以上前のケーディングの調査をあげているほどである。欧米で語彙調査がすくないのは，単なる度数だけでは利用価値がすくないからではないだろうか。見出し語立てをしない調査ならすぐにできるが，結果が役にたたない，見出し語立てをするには，たいへんな労力を必要とし，それに見合うだけの価値がない，ということだろう。日本語と同様に，単なる度数だけで基本語彙をきめるのは，むずかしいはずである。英語圏では，語彙調査にそそいできた努力はコーパスにむけられた。そして，重要視されるのは，統計的な精密さよりも言語全体を代表していることである。British National Corpus(BNC)は延べ１億語で，大部分は書きことばだが，そのほかに１割の話しことばがはいっている。話しことばが１割で書きことばが９割というのは，もちろん，現実の言語活動の比率とは関係ない。むしろ，逆の話しことば９割，書きことば１割というほうが，実際に近いだろう。では，なぜ量的にはすくない

書きことばを重視するのか。その理由は，言語的に似ている会話ばかり大量に集めてもしかたがない，それよりも，言語的変種のすべての範囲をふくむものを入れたい，ということのようである。

５．雑誌の語彙調査と日本語のコーパス

しかし，日本語の書きことば全体についての統計がないので，便宜上語彙調査の結果を利用するよりしかたがまい。たとえば，日本語の語種分布については，今でも雑誌九十種調査の結果がよくひかれる。以下に，あたらしい雑誌７０誌の結果とあわせて，延べ語数の比率をあげる。

	和語	漢語	外来語	混種語
雑誌九十種(1956)	53.9	41.3	2.9	1.9
雑誌７０誌(1994)本文	41.5	45.9	10.7	2.0

これでみると，最近は漢語の量が和語をおいこしたこと，外来語が激増していることがわかる。しかし，これは，あくまで雑誌のものだから，新聞や単行本までふくめるとどうなるか，というのが知りたいところである。

国語研究所では，最近『話し言葉コーパス』『「太陽」コーパス』という２つのコーパスをつくった。これらは，ひとつひとつの音や単語をしらべるのには，ひじょうに有効だが，巨視的に日本語がどうなっていたかをみるわけにはいかない。「太陽」コーパスと雑誌の調査結果をみることにしよう。

	寝台	ベッド	食卓	テーブル	汽車	電車	列車	幹線
１８９５年(太陽)	4	-	7	2	120	50	26	5
１９０１年(太陽)	17	-	19	7	116	15	143	3
１９０９年(太陽)	16	1	4	24	135	173	38	4
１９１７年(太陽)	16	-	20	5	86	66	27	1
１９２５年(太陽)	23	21	21	22	93	51	47	6
....								
１９５６年(九十種)	9	25	9	24	26	33	33	2
１９９４年(７０誌)	1	18	18	42	4	29	17	12

年によって，かなり変動があるが，「寝台→ベッド」「食卓→テーブル」という大勢にあることや，「汽車」が激減したことは分かる。「幹線」がふえたのは，もちろん「新幹線」のせいで，古い例は一般用語としての「幹線」である。「太陽」コーパスでは，必要があれば，文脈も出せる。「右鐵道の幹線に於ても亦電氣發動機の採用を見るに至るや否やの一事あるのみ」（１８９５年９号「全國鐵道概覽」）。このように，いちいちの単語について近代１００年の動きは分かるが，語種全体の動きは分からない。その点では，上にあげたような語彙調査の結果にはかなわない。

今めざしているコーパスでは，国語研究所の２つの伝統，語彙調査があきらかにした巨視的な観点と，『現代語の助詞・助動詞』から用例カード・「太陽」コーパスにうけつがれ

た微視的な記述とが総合されることを希望する。

６．コーパス利用とコーパス言語学

終わりに，コーパスの効用ということに関して，わたしの考えをのべておこう。

コーパスは道具である。それは言語観・研究法のちがいにかかわらず，あらゆる言語研究者にとって役にたつはずのものである。ただし，話し手の直観を基準にすればいい，とする生成文法の立場とは両立しにくいかもしれない。チョムスキーは，コーパスにたいして，はっきり否定的な意見をのべている。ただし，かれの発言は１９６２年，まだ大規模なコーパスが現実的なものになっていなかった時期のものである。その後，コーパスおよびそれにもとづく研究が飛躍的に発展した現在にあっては，その効用を頭から否定することはできないだろう。生成文法家にとっても，コーパスによって得た例文を話し手の直観で吟味して使えばいいのだから，利用価値がないことはないはずだ。げんに，「生成文法を学ぶ人のために」という副題のついた『言語研究入門』という本のなかにも，「コーパス言語学」という章がある。たしかに，コーパスを絶対視するのはまちがいである。「ゆく秋の大和の国の薬師寺の塔の上なるひとひらの雲」（佐々木信綱）という有名な歌には「の」が５つかさなっている。夏目漱石の「吾輩は猫である」には「何でも天璋院様の御祐筆の妹の御嫁に行った先きの御っかさんの甥の娘なんだって」という，６つの「の」がかさなった修飾語があり，実際の文では，これ以上に「の」がかさなる文をみつけるのは，むずかしいだろう。しかし，文法の規則としては，「の」のかさなりに限界がない，という一般化をしなければならない。と同時に，このような重なりは，現実にいくつがどのくらいあるのか，という頻度の記述も必要である。

特定のコーパスという一定量の範囲で調査する，という必要はない。ある学者は「わたしが言いたいことは，ただ１つ，コーパスはできるだけ大きいほうがよい，そして，どんどん大きくならなければいけない，ということである」といっている。１億語のコーパスを利用したら，そこでやめずに，例文をふやしたらいいのである。そのさい，分野・文体のバランスがくずれないように注意する必要はあるが，それはコーパスをつかわない例文採集でもおなじである。また，コーパス利用を計量的なものにかぎる必要もない。「～らしい」と「～ようだ」のちがいをしらべるのに，計量的にしらべるのは，均衡コーパスの長所をいかすことだが，使用例数をかぞえなくても，大量の用例をしらべることでコーパスは活用されている。単にコーパスを利用して言語現象をしらべた，という研究を「コーパス言語学」とよぶ必要はない。コーパスの第１の価値は，膨大な用例の量にある。これからの研究には，当然それを利用すべきだが，それは「コーパス言語学」でも「用例言語学」でもない。大量の例文をしらべることは，まさに言語学の王道，限定語なしのザ・言語学だ，というのが，わたしの立場である。

(2007.3.17)

（参考：国立国語研究所報告書類）

No.3　現代語の助詞・助動詞　―用法と実例―〔1951〕
資料集 2 語彙調査―現代新聞用語の一例―〔1952〕
No.4　婦人雑誌の用語　―現代語の語彙調査―　〔1953〕
資料集 6 分類語彙表〔1964〕
No.8　談話語の実態〔1955〕
No.12 総合雑誌の用語　―現代語の語彙調査―前編〔1957〕
No.13 総合雑誌の用語　―現代語の語彙調査―後編〔1958〕
No.15 明治初期の新聞の用語〔1959〕
No.18 話しことばの文型　―対話資料による研究―〔1960〕
No.21 現代雑誌九十種の用語用字　第 1 分冊総記・語彙表〔1962〕
No.22 現代雑誌九十種の用語用字　第 2 分冊漢字表〔1963〕
No.23 話しことばの文型　―独話資料による研究―〔1963〕
No.25 現代雑誌九十種の用語用字　第 3 分冊分析〔1964〕
No.37 電子計算機による新聞の語彙調査 I　〔1970〕
No.38 電子計算機による新聞の語彙調査 II 〔1971〕
No.42 電子計算機による新聞の語彙調査 III〔1972〕
No.43 動詞の意味・用法の記述的研究〔1972〕
No.44 形容詞の意味・用法の記述的研究〔1972〕
No.48 電子計算機による新聞の語彙調査 IV 〔1973〕
No.56 現代新聞の漢字〔1973〕
No.76 高校教科書の語彙調査 I [1983〕
No.78 日本語教育のための基本語彙調査〔1984〕
No.81 高校教科書の語彙調査 II [1984〕
No.82 現代日本語動詞のアスペクト [1985〕
No.87 中学校教科書の語彙調査 [1986〕
No.89 雑誌用語の変遷 [1987〕
No.99 高校・中学校教科書の語彙調査 分析編 [1983〕
No.112 テレビ放送の語彙調査 I [1995〕
No.114 テレビ放送の語彙調査 II [1997〕
No.115 テレビ放送の語彙調査 III [1999〕
資料集 14 分類語彙表―増補改訂版―〔2003〕
日本語話し言葉コーパス〔2004〕
No.121 現代雑誌の語彙調査〔2005〕
No.122 雑誌『太陽』による確立期現代語の研究（「太陽コーパス」）〔2005〕
No.125 現代雑誌の表記〔2006〕

招待講演

3月17日（土）第一日目　15:50〜16:40

大規模テキスト処理の時代

▶長尾 眞（情報通信研究機構理事長）

大規模テキスト処理の時代

長尾　真（情報通信研究機構）

The Age of Large Scale Text Processing

Makoto Nagao (National Institute of Information and Communications Technology)

１．テキストデータの収集

過去の新聞のデータだけでなく、グーグルがやり始めているように、膨大な数の本の全文をコンピュータに入れられる時代となってきた。さらに Web 上に存在するテキストを機械的に収集することによって数億、数十億文のテキストを集めることができる。しかし例文翻訳や統計翻訳などに必要な対訳テキストデータ（パラレルコーパス）を大量に入手することは未だに困難である。

２．スーパーコンピュータの利用

グリッド構造のスーパーコンピュータを利用すれば何億文というテキストデータが解析できる時代となってきた。形態素解析や構文解析（格構造解析）の精度が非常に高くなったことによって、自動学習のメカニズムを導入して精度を徐々に上げてゆけるという良いサイクルに入ってきた。

３．用例翻訳の方式

文を句単位に分割し、それぞれの句の対訳句を用例辞書から取り出し、目的言語の文法に従って組み上げることによって翻訳文を作り出す方式である。そこで用例辞書（対訳句辞書）を用意することが必要となり、膨大なパラレルコーパスから用例を取り出すことになる。この用例辞書の大きさがどれほどになるかは現時点では不明であるが、同義語、類義語を一つにすることにしても、少なくとも数十万用例の辞書が必要となるだろう。

４．統計翻訳の方式

一つの文（または語列）を他言語のどの文（または語列）に対応させるのが確率論的に最も良いかを膨大なパラレルコーパスの統計的解析によって推定する方法である。言語の知識を全く使わない方法であるが、無限の対訳テキストがあれば良い結果が得られるとしている。英語・中国語間の翻訳にも既にかなり良い成果を得ているが、実用的な見地からは句という概念を持ち込んで統計計算の有効性を高める方向に動いている。

５．用語集・オントロジー・格構造辞書の構築

機械翻訳だけでなく、情報検索などにも必要な用語集・オントロジー・格構造辞書の充実が大切である。そのためにも膨大なテキスト・コーパスの解析を行い、単に語を抽出す

るだけでなく、語と語の関係を明らかにしてゆくことが必要である。

６．辞書について

１９９０年に岩波書店から情報科学辞典を出したが、そこでは各項目（専門用語）について、(１）内包的説明（定義)、(２）外延的説明（定義)、(３）関係的説明（対立・対比、上位・下位、部分・全体、性質・属性・機能、目的、因果・継承、歴史等）などの記述を与えた。また類似の語との差異がどこにあるかを示すことによって、その語の意味の範囲を明確にする努力もした。一般用語の辞書についても、こういった説明をするとともに、できるだけ豊富な用例を与えること、さらに“このような用法は誤りである”といったことも付け加えた、内容豊かな辞書があって欲しいものである。

７．文法とは？

このようにコンピュータによってあらゆる微妙な表現が集められ利用される時代になってくると、言語の文法はコンピュータにとってどのような意味を持つか、コンピュータが使えるだけ厳密に作れるか、といった問題が出てくるだろう。

今日の文法は言語のあらゆる微妙な表現にきっちりと対応できない。文法はメタの世界での概念的なもので、現実の場に直接対応しないものであるといえばそれまでであるが、コンピュータが安心して使える基本的な文法規則だけでもはっきりと整備して示されると、それ以外のところについては用例辞書などによって処理することができ、文法と辞書の役割分担が明らかになるのだが？

計画班研究進捗状況報告

3月18日（日）第二日目　9:00〜11:15

●データ班	山崎 誠
●ツール班	松本 裕治
●電子化辞書班	伝 康晴
●日本語学班	田野村 忠温
●日本語教育班	砂川 有里子
●言語政策班	田中 牧郎
●辞書編集班	荻野 綱男
●言語処理班	奥村 学

平成 18 年度研究進捗状況報告:データ班
（代表性を有する現代日本語書籍コーパスの構築）

山崎 誠 （班 長：国立国語研究所研究開発部門） †
小椋 秀樹（分担者：国立国語研究所研究開発部門）
柏野和佳子（分担者：国立国語研究所研究開発部門）
高田 智和（分担者：国立国語研究所研究開発部門）
間淵 洋子（分担者：国立国語研究所研究開発部門）
丸山 岳彦（分担者：国立国語研究所研究開発部門）
森本 祥子（分担者：国立国語研究所情報資料部門）
山口 昌也（分担者：国立国語研究所研究開発部門）
大和 淳 （分担者：横浜国立大学大学院国際社会科学研究科）

Progress Report of the Year 2006: ‘Data handling’ Group

Makoto Yamazaki (Dept. Lang. Res., National Institute for Japanese Language) †
Hideki Ogura (Dept. Lang. Res., National Institute for Japanese Language
Wakako Kashino (Dept. Lang. Res., National Institute for Japanese Language)
Tomokazu Takada (Dept. Lang. Res., National Institute for Japanese Language)
Yoko Mabuchi (Dept. Lang. Res., National Institute for Japanese Language)
Takehiko Maruyama (Dept. Lang. Res., National Institute for Japanese Language)
Sachiko Morimoto (Dept. Lang. Res., National Institute for Japanese Language)
Masaya Yamaguchi (Dept. Lang. Res., National Institute for Japanese Language)
Atsushi Yamato (Int. Graduate School of Social Sciences, YNU)

１．データ班の目標

データ班の目標は、タイトルにも示すとおり、「代表性を有する現代日本語書籍コーパスの構築」である。この研究課題は、国立国語研究所の研究プロジェクト「大規模汎用日本語データベースの構築とその活用に関する調査研究」と連携して『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese 、以下 BCCWJ と略す）を構築することにある。書籍に関する部分だけを取り出すと全体像が分からなくなるため、適宜、BCCWJ の他の部分についても触れながら今年度の進捗状況を記述する。

２．全体設計の変更

特定領域研究が発足する前から BCCWJ の設計はスタートしていたが、山崎(2006)、山崎他(2006)で紹介したコーパスの設計にその後大きな変更を加えたので、ここで改めて全体設計を紹介する。

これまでは、BCCWJ は「生産実態サブコーパス」（出版物を母集団とするサブコーパス）と「流通実態サブコーパス」（図書館所蔵の書籍を母集団とするサブコーパス）の２つから構成され、生産実態サブコーパスは、サンプルの長さを 1,000 字の固定長とし、収録語数は 1,000 万語、一方、流通実態サブコーパスは、サンプルの長さを１万字を上限とする可変長

† yamazaki@kokken.go.jp

として、収録語数は１億語という設計であった。
この設計を次の図１のように変更した。

<table>
<tr><td>生産実態サブコーパス
約 3,500 万語
書籍、雑誌、新聞
固定長＋可変長</td><td>流通実態サブコーパス
約 3,000 万語
書籍
固定長＋可変長</td></tr>
<tr><td colspan="2">非母集団サブコーパス
約 3,500 万語
白書、法律、国会会議録、検定教科書
ベストセラー、Web 掲示板等
可変長（一部固定長＋可変長）</td></tr>
</table>

図１　BCCWJ の基本構成

大きな変更点は、「非母集団サブコーパス」を追加したことである。ここには、現代日本語の実態を把握する上で欠かせない資料で、生産実態、流通実態の２つのサブコーパスでは充分に量が集まらないもの、特に、国立国語研究所の研究活動において必要なものなどを収録する。主に、単一のジャンルや媒体を対象とし、特定の研究目的での利用を想定している。ここに収める資料は、必ずしも母集団を設定してサンプリングをする必要はない。

また、以前の設計では、サンプルの長さに関して、生産実態が固定長、流通実態が可変長と分けていたが、サンプルの長さと母集団とを相関させる必要はないと判断し、生産実態サブコーパス、流通実態サブコーパスともに、固定長と可変長のサンプルを取得することにした。２つのサンプルは別々のサンプリングで取得するのではなく、１回のサンプリングにおける同一の「サンプル抽出基準点」をもとに同時に取得する。従って、固定長サンプルと可変長サンプルとは重なる部分を持つことになる。

以上が大きな変更点である。以下、サンプリング、著作権処理、電子化、形態論情報付与の各作業工程に沿って進捗状況を報告する。

３．サンプリング

生産実態サブコーパスにおいては、書籍、雑誌、新聞の母集団を文字数により推計し、その比率に基づいて、各媒体の構成比を決定した。結果は表１に示すとおりである。なお、文字数計測調査及び推計の詳細は、丸山他(2007)を参照されたい。

表１　生産実態サブコーパスにおける各媒体の構成比率

	総文字数	構成比率	固定長 合計語数	必要サンプル数	可変長 合計語数
書籍	485.40億字	74.14%	741.4万語	12,604サンプル	2891.5万語
雑誌	105.16億字	16.06%	160.6万語	2,730サンプル	481.8万語
新聞	64.16億字	9.80%	98.0万語	1,666サンプル	98.0万語
合計	654.72億字	100.00%	1,000万語	17,000サンプル	3471.3万語

各媒体の母集団は、以下のように決定した。書籍については、国立国会図書館の所蔵目録である J-BISC に収録されたものから 2001 年～2005 年に発行されたものを抜き出し、書き言葉コーパスに収録する適切性条件でふるいを掛け、317,117 冊を母集団とした。

雑誌の母集団は、2001 年～2005 年の間に社団法人日本雑誌協会に加盟していた出版社の発行する雑誌 1,259 タイトル、総数 55,779 冊と決定し、リストを作成した。

新聞の母集団は、2001 年～2005 年に発行された、全国紙 5 紙（朝日、毎日、読売、日経、産経)、ブロック紙 3 紙（北海道、中日、西日本)、地方紙 8 紙（河北新報、新潟日報、京都新聞、神戸新聞、中国新聞、高知新聞、愛媛新聞、琉球新聞）と決定し、リストを作成した。

今年度は、上記の中から書籍 2,500 サンプルのサンプリングを行った。そのほかのサブコーパスの進捗状況は以下の表を参照されたい。この中には、言語政策班で活用する検定教科書、日本語教育班で活用する日本語教科書のデータも入っている。なお、Yahoo! 知恵袋のデータについては、岡本他(2007)を参照のこと。

表２ サンプリングの進捗状況

	生産SC			流通SC	非母集団SC					
メディア	**書籍**	**雑誌**	**新聞**	**書籍**	**白書**	**国会会議録**	**Yahoo!**	**法令**	**検定教科書**	**日本語教科書**
必要サンプル数	12,604	2,730	1,666	12,604	1,500	?	39,969	?	?	?
母集団定義	済	済	済	済	済	済	済	済	不要	不要
リスト作成	済	済	済	済	済	未	未	未	済	済
総文字数	済	済	済	未	不要	不要	不要	不要	不要	不要
台帳作成	済	未	未	未	済	未	未	未	不要	不要
サンプリング	2500	未	未	未	済	未	未	未	不要	不要
データ入力	2500	未	未	未	済	未	未	未	着手	着手
今年度実績	2,500	0	0	0	完了	未定	完了予定	未定	121冊	中級30冊

４．著作権処理

作家団体（日本文藝家協会、日本推理作家協会、日本児童文芸家協会、日本児童文学者協会、日本ペンクラブ）にお願いしていた事前許諾の準備が整い、年度内に各作家団体の会員に発送する予定である。事前許諾とは、その人の作品がサンプルとして当たった場合に、コーパスへの収録を前もって承諾してもらう依頼である。

また、昨年度から行ってきた新聞社 4 社（朝日、読売、毎日、産経）との交渉もほぼまとまり、3 社と覚書を交わすことができた。

著作権処理の手順等の詳細については、森本他(2007)を参照されたい。

５．電子化

(1)主要なタグセットの設計をほぼ完了した。タグの種類は、付与される対象により、サンプルに関するタグ、文字・表記に関するタグ、文書構造に関するタグの３つに分けられる。

・サンプルに関するタグは、サンプル全体に関する情報で、サンプルの ID（書誌情報データベースへの参照が可能)、固定長・可変長の別、サンプル抽出基準点などが記述される。

・文字・表記に関するタグは、ルビ、誤植の訂正、外字、囲み文字、上付き・下付き文字などを表現する。

・文書構造に関するタグは、『太陽コーパス』の仕様を引き継ぎ、それを拡充したものである。文書の階層構造、図表、引用、注記など論理的な役割が明確な文書要素を表現する。

実際のタグの一覧やマークアップの例は、山口他(2007)を参照されたい。なお、以上の電子化の仕様については、今年度中に一般公開を予定している。

(2)白書サンプルの XML 化

白書 1500 サンプル 500 万語の入力が完了し、入力精度が 99.95%以上を確認したのち、11 月 10 日にプレーンテキスト版を全文検索ソフト『ひまわり』に同梱の形で特定領域内で公開した。年度内には Web 上での試験公開（全文検索）を開始する予定である。

６．形態論情報付与

(1)解析用辞書データの整備拡充（電子化辞書班と連携）

電子化辞書班と共同で構築を進めている UniDic に格納する辞書データ（短単位データベース）の整備拡充を行った。整備拡充に用いた資料は、『岩波国語辞典』第 6 版、ＣＳＪ、ipadic、cannadic、郵便番号データなどである。2 月 9 日現在、語彙素の数は 102,919、書字形の数は 132,581 に達している。なお、以上の情報を格納した解析用辞書 unidic ver.1.3.0 は、関連ツールとともに近く一般に公開する予定である。

(2)言語単位の設計について（ツール班と連携）

文節認定基準の見直し、及び、短単位と ipadic との対応付けにおける問題点（品詞と単位の長さ）を検討する検討会を 4 回開催した。

(3)学習用データの作成（ツール班と連携）

形態素解析システムの学習用データとするため、BCCWJ に収められたサンプルから、新聞、雑誌、書籍、白書、その他のそれぞれから 20 万語合計 100 万語を選び、人手修正を加えて短単位・長単位とも精度 99%以上のデータを作成する。このデータには、ツール班で文節区切り、係り受け情報などを付与する。

今年度は、白書の 20 万語を非母集団サブコーパスに収められた白書データのうち、2001～2005 年分から選定した。

７．成果物・イベント

今年度は、以下の２冊の報告書を刊行する予定である。

(1)『現代日本語書き言葉均衡コーパス』短単位規程集 Version 1.2（執筆：小椋秀樹）

(2)『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法―現代日本語書き言葉の文字数調査―（執筆：丸山岳彦、秋元祐哉）

また、2006 年 11 月 7 日に「現代日本語書き言葉均衡コーパス仕様説明会」を開催した。

参考文献

岡本真、木戸冬子、佐古智正(2007)「日本語コーパス」用 Yahoo!知恵袋データについて，本予稿集収録

小椋秀樹、小木曽智信、小磯花絵、他(2007)『現代日本語書き言葉均衡コーパス』における短単位の概要，本予稿集収録

丸山岳彦、柏野和佳子、山崎誠、他(2007)『現代日本語書き言葉均衡コーパス』におけるサンプリングの概要，本予稿集収録

森本 祥子、前川 喜久雄、小沼 悦、他(2007)『現代日本語書き言葉均衡コーパス』における著作権処理について，本予稿集収録

山口昌也、高田智和、北村雅則、他(2007)『現代日本語書き言葉均衡コーパス』における電子化フォーマットの概要，本予稿集収録

山崎誠(2006)「代表性を有する現代日本語書き言葉コーパスの設計」，第 13 回国立国語研究所国際シンポジウム：言語コーパスの構築と活用，pp.63-70

山崎誠、 前川喜久雄、 田中牧郎、他(2006) 「代表性を有する現代日本語書き言葉コーパスの設計」，言語処理学会第 12 回年次大会発表論文集，pp.440-443

平成 18 年度研究進捗状況報告：ツール班
（書き言葉コーパスの自動アノテーションの研究）

松本裕治（班長：奈良先端科学技術大学院大学情報科学研究科）†
徳永健伸（分担者：東京工業大学大学院情報理工学研究科）
乾健太郎（分担者：奈良先端科学技術大学院大学情報科学研究科）
橋田浩一（分担者：産業技術総合研究所情報技術研究部門）
橋本泰一（分担者：東京工業大学大学院情報理工学研究科）
浅原正幸（分担者：奈良先端科学技術大学院大学情報科学研究科）
野口正樹（協力者：東京工業大学大学院情報理工学研究科）
飯田龍　（協力者：奈良先端科学技術大学院大学情報科学研究科）
小町守　（協力者：奈良先端科学技術大学院大学情報科学研究科）
谷口雄作（協力者：奈良先端科学技術大学院大学情報科学研究科）

Progress Report of the Year 2006: Tools and annotation group

Yuji Matsumoto (Nara Institute of Science and Technology)
Takenobu Tokunaga (Tokyo Institute of Technology)
Kentaro Inui (Nara Institute of Science and Technology)
Koiti Hasida (Advanced Industrial Science and Technology)
Taiichi Hashimoto (Tokyo Institute of Technology)
Masayuki Asahara (Nara Institute of Science and Technology)
Masaki Noguchi (Tokyo Institute of Technology)
Ryu Iida (Nara Institute of Science and Technology)
Mamoru Komachi (Nara Institute of Science and Technology)
Yusaku Taniguchi (Nara Institute of Science and Technology)

１．ツール班の目標

本特定領域研究で構築されるコーパスに対し，形態素，統語，文脈情報など様々なレベルのタグ付けを対象とし，タグの詳細度の検討，および，実際にコーパスへタグ付けを効率よく行うための自動タグ付けツールの構築，高いタグ付け精度を保証し・管理するためのタグ付け支援環境の構築を目標とする．

タグ付けの詳細度については，単語分かち書き，品詞付与，係り受けや句構造解析等の構文解析，用言だけでなく事象を表す体言に対する項構造解析，照応解析等の指示対象の特定や共参照の解析，意味・談話構造解析など，さまざまな言語情報についてのタグ設計およびタグ付け基準の設定を行う．異なるレベルの情報間の整合性を保証しつつタグ付け作業を可能にするための統合的なタグ付け方式の設計を行う．後者については，設計されたタグ方式に従ってコーパスを作成しつつ，タグ付きコーパスからの機械学習に基づいてタグ付けの自動化を行う言語解析システムの構築を行なう．また，様々なレベルのタグ付け作業に共通に利用できる汎用的なタグ付きコーパス管理システムの設計と開発を行う．さらに，タグ付きコーパス公開のための Web ベースのブラウザやコーパス検索ツールの構築を行なうことを班の目標とする．

† matsu@is.naist.jp

２．ツール班の研究・開発項目

ツール班が担当している研究開発項目を，次の４つに分類して要点をまとめる．

２．１　タグ付けの仕様

本年度は，形態素解析の分かち書き基準と語彙について，コーパス班および辞書班とのミーティング行い，短単位に基づく分割基準の詳細化を行なった．文節分かち書き，係り受け解析，句構造解析などの解析基準の仕様設計や，照応解析，共参照解析，文書構造解析などの文書解析についても仕様の設計を順次行なう．また，タグ付きコーパスの表現法についても検討を行い，関係データベース，ＸＭＬ，ＲＤＦなどによるデータの記述とその相互変換について検討する．

２．２　自動タグ付けツール

ツール班に所属するメンバーは，これまで，形態素解析システム「茶筌」，日本語係り受け解析システム「南瓜」，統語解析システム，日本語統語解析システム MSLR などを構築してきた．また，照応解析や述語項構造解析，文章構造のアノテーションスキーマなどの研究も行なっている．これら様々なレベルのタグ付きデータを構築するとともに，自動解析システムの開発，および，解析支援システムを並行して開発する．

２．３　タグ付け作業支援ツール

これまで，ツール班のメンバーは，形態素解析・係り受け解析済みのコーパスを格納し，検索・表示・修正を行なうコーパス管理ツール「茶器」(Matsumoto, Asahara, et al 2006)，句構造解析済みコーパスの検索・表示・修正を行なう「eBonsai」(野口，市川，橋本，徳永 2006)，文や文間の意味関係をラベルとする有向グラフとしてコンテンツの作成支援するセマンティックオーサリングツール(Hasida 2003)などを開発している．ツール班では，これらを発展・利用するのと並行して，これらの機能（の一部）を拡張し，統合したより汎用性の高いツールの開発を目指す．

コーパスに対する様々なタグ付けを支援するためのシステムが種々構築されているが，上記の我々のシステムを含めて，多くは特定のタグ付けだけを対象にしたものであった．汎用のタグ付け仕様とツール化を目指した研究の一つに AGTK(Annotation Graph Toolkit) (Cotton and Bird 2002)がある．AGTK では，コーパス内の任意の部分文字列をセグメントとして定義し，これをノードとするグラフとしてコーパスへのアノテーションを行なう．我々の班が対象とするアノテーションについても，コーパス中の特定の範囲（セグメント）へのタグ付け，および，セグメント間の関係のタグ付け，という形に整理できるので，AGTK の考え方をさらに整理することにより，汎用のタグ付け仕様の設定と汎用のタグ付け支援システムの構築を目指す．

２．４　タグ付きコーパス検索・公開用ツール

本特定領域研究で構築されるコーパスは適宜公開される予定であるが，そのサンプルを簡単に検索・ブラウズできる環境を構築し，ネットワーク上で自由にアクセスできる環境を実現することが好ましい．上記のタグ付け支援ツール以外に，外部公開用のコーパス検索・表示ツールを作成する．

３．本年度の活動

３．１　班の活動状況

本年度のツール班独自の活動としては，本特定領域研究が開始された８月後半より５回の班会議を主に東京地区で開催した．それ以外に，全体会議や総括班会議への参加，コーパスへの形態素解析情報付与のための短単位基準の詳細化と辞書構築へ向けた打ち合わせを，コーパス班，辞書班と合同で数回行なった．また，班内でも共通のツール化へ向けてサブメンバーによる会合が数回持たれた．特に，汎用性の高いタグ付けツールを目指して開発が行なわれている aTagrin（後述）に関しては，遠隔会議による打ち合わせを適宜行なった．本年度，ツール班全員が出席して開催した班会議は次の通りである．

・第１回ツール班会議

日時：2006 年 8 月 31 日（木），場所：NAIST 東京事務所（東京都田町）

議題：コーパス解析とアノテーションに関するこれまでのメンバーの実績報告と班の今後の活動計画について話し合った．

・第２回ツール班会議

日時：2006 年 10 月 13 日（金），場所：産業技術総合研究所（秋葉原ダイビル）

議題：ISO/TC37 におけるコーパスアノテーションの標準化に関する動向調査報告，および，今後の共通化ツール構築の方針について議論した．

・第３回ツール班会議

日時：2006 年 11 月 17 日（金），場所：NAIST 東京事務所（東京都田町）

議題：AGTK(Annotation Graph Toolkit) (Cotton and Bird 2002)の論文紹介の後，この方法の利点欠点について議論し，共通化ツールの設計方針を議論し，基本的な仕様の確認を行なった．

・第４回ツール班会議

日時：2007 年 1 月 19 日（金），場所：NAIST 東京事務所（東京都田町）

議題：汎用性を目指した共通化タグ付けツールである aTagrin の現状の報告．ChaKi Web API の設計に関する報告，および，公開コーパス用のブラウザツールとして開発している「茶杓」の現状紹介を行なった．

・第２回ツール班会議

日時：2007 年 2 月 21 日（水），場所：産業技術総合研究所（秋葉原ダイビル）

議題：今年度の活動のまとめとして各メンバーからの報告と，来年度計画について話し合った．

３．２　本年度の研究活動の現状と成果

各メンバー（グループ）の本年度の主な活動の現状と成果についてまとめる．現状のコーパス作成支援ツールとしては，形態素解析・係り受け解析コーパスの検索・作成支援システム「茶器」，および，句構造解析コーパスの検索・作成支援システム「eBonsai」がある．図１は，茶器の係り受け構造検索のスナップショットを，図２は，eBonsai の統語的

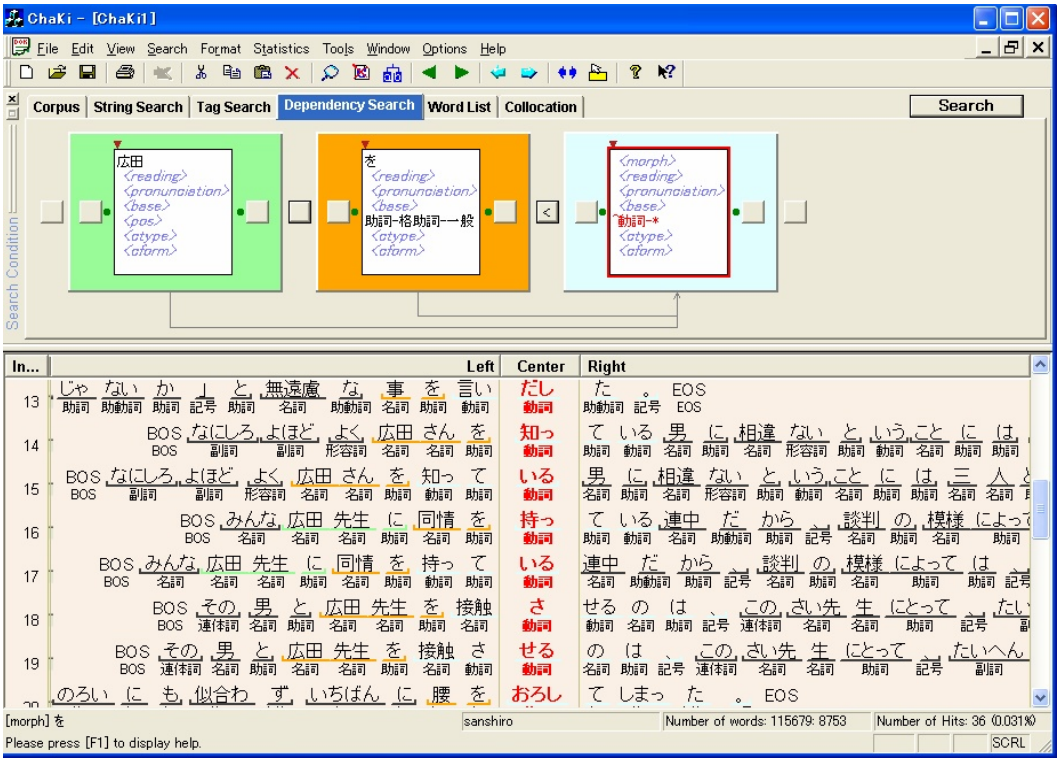


図１．茶器の係り受け解析検索のスナップショット

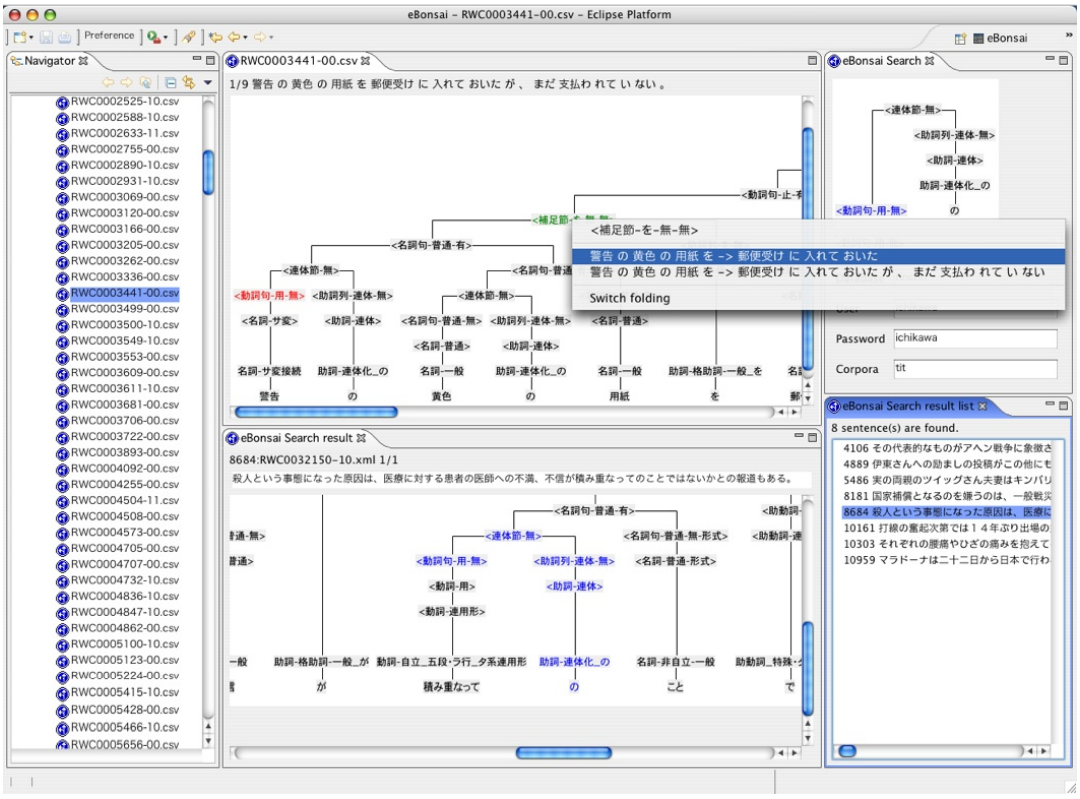


図２．eBonsai の句構造木の編集作業のスナップショット

曖昧性解消作業のスナップショットを示す.

本年度の主な活動と成果は，次の通りである．松本，浅原らは，これまで開発してきたコーパス検索ツール茶器を拡張し，複合語の取り扱いと複合語の構成要素を表示する機能の追加を行なった．また，今後他のシステムとのデータ共有の利便性を考え，形態素解析・係り受け解析済みコーパスのためのデータベーススキーマの整理を行なった．さらに，新しいドメインのテキストの解析精度を高めるために未知語処理を考慮した頑健な形態素解析法の設計を行なった．浅原は，様々な分かち書き基準に対応するため，確率的単語分割ツールの設計と実装を行なった.

乾らは，機械学習に基づく照応解析システムの洗練，および，述語項構造解析システムの設計と実装を行なった（飯田，小町，乾，松本 2007b）．また，これらのシステムの学習・評価データとして，NAIST テキストコーパス（飯田，小町，乾，松本 2007a）を構築し，一般公開した．図３に，述語項構造解析システム SynCha のデモ画面を示す．入力された文章に含まれるすべての述語（用言およびサ変名詞）に対して，その項となっている名詞の一覧が表として示されている.

徳永, 乾らは, セグメントとリンクに基づく汎用タグ付けツール aTagrin の設計を行い，プロトタイプシステムを作成した（野口，三好，徳永，飯田，小町，乾 2007）．コーパスには様々な種類のアノテーションが行なわれる可能性がある．しかし，そのほとんどは，文書の特定の部分（セグメント）に対するラベル付け，あるいは，セグメント間の関係を

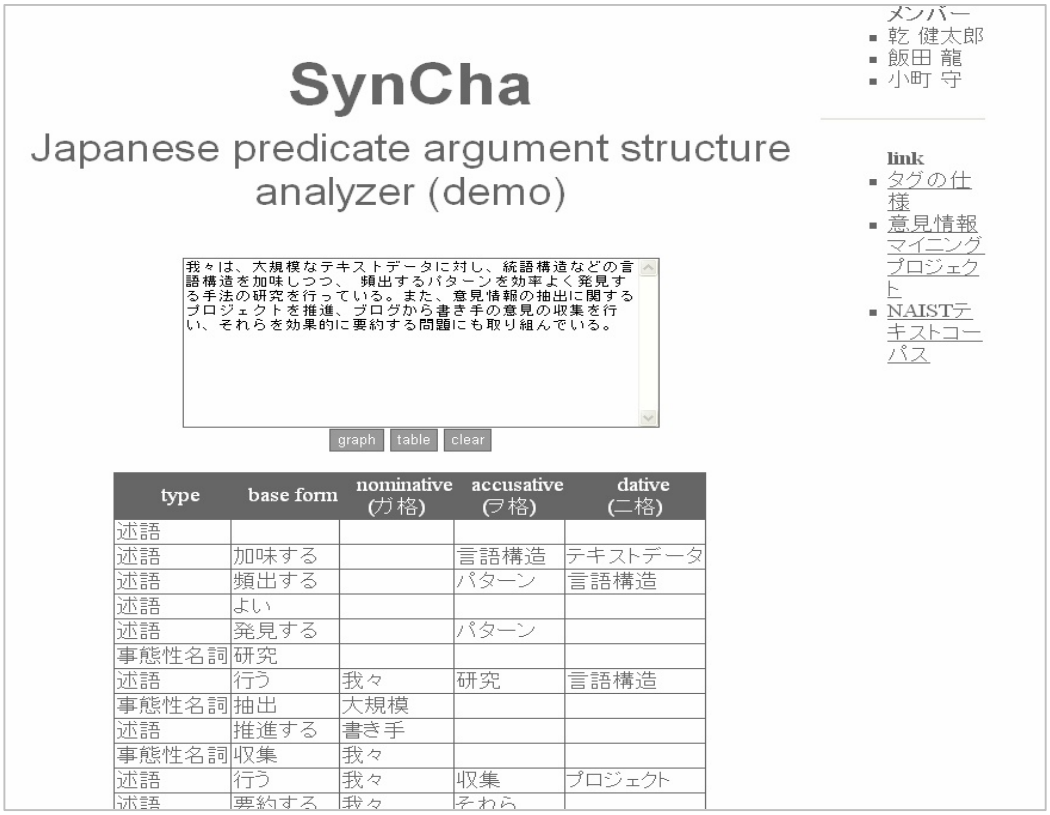


図３．項構造解析システム SynCha のデモ画面

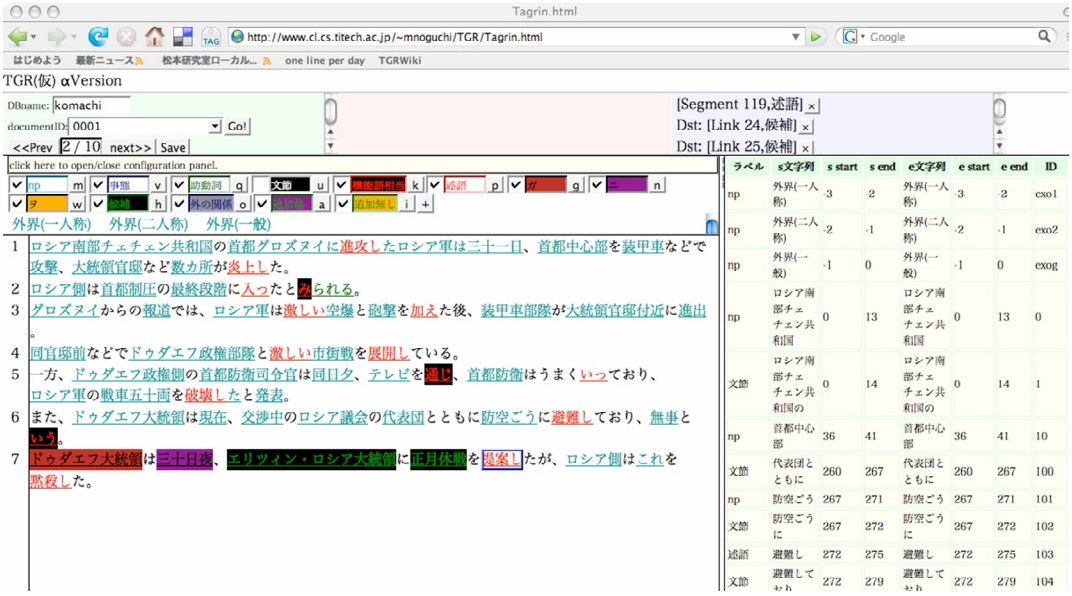


図４．汎用タグ付けツール aTagrin の動作画面

表すリンクのいずれかに分類することができる．aTagrin では，セグメントとリンクのデータフォーマットを定義し，特に，リンクについては，ソースセグメント，ターゲットセグメント，リンクの名称以外に，リンクに方向性があるか（有向／無向の区別），および，関係が遷移的(transitive)かどうかを記述することができるようにし，リンクの意味が明確になるようにしている．例えば，リンクが無向かつ遷移的であるならば，それらに結び付けられたセグメントは集合あるいは同値関係を表すと考えられる．また，それらの要素の間には一つでもリンクが張られていればよく，集合内のすべてのセグメント間に関係を記述する必要がなくなる．図４に，aTagrin の動作画面を示す．文章中の任意の範囲（セグメント）をマウスで指定してラベル付けを行なったり，セグメント間の関係を指定したりすることができる．タグ付け結果は，文書中に色分けによって示されるのと同時に，表形式の一覧によっても示される．

橋本らは，Web ベースのコーパス検索ツールの構築と，コーパス内の単語の共起分析のためのグラフ表示モジュールの作成を行なった．ある特定の語を指定して，その語と共起関係にある語の一覧を棒グラフによって表示できる．また，2 種類の共起グラフを示すことによって，それらの差異を視覚的に示すことができる．このツールは他のツールのプラグインツールとして用いることを考えている．

橋田は，これまで構築してきたセマンティクエディタを，拡張 RDF による言語的構造の一般的表現と可視化が行なえるように拡張し，照応や共参照を含む談話構造の表示・編集ツールを構築している．本年度は特に，ノードとリンクのタイプ(談話関係等)の選択を効率化するプラグインの開発，および，談話グラフからテキストをインタラクティブに生成するツールの開発を行なった．図５に，談話構造のグラフ表示の例を示す．通常のＲＤＦとはことなり，グラフのノード内に他のノードを埋め込んだ記述が可能となっている．今後は，談話構造タグ付きテキストと談話グラフの相互変換をサポートする予定である．

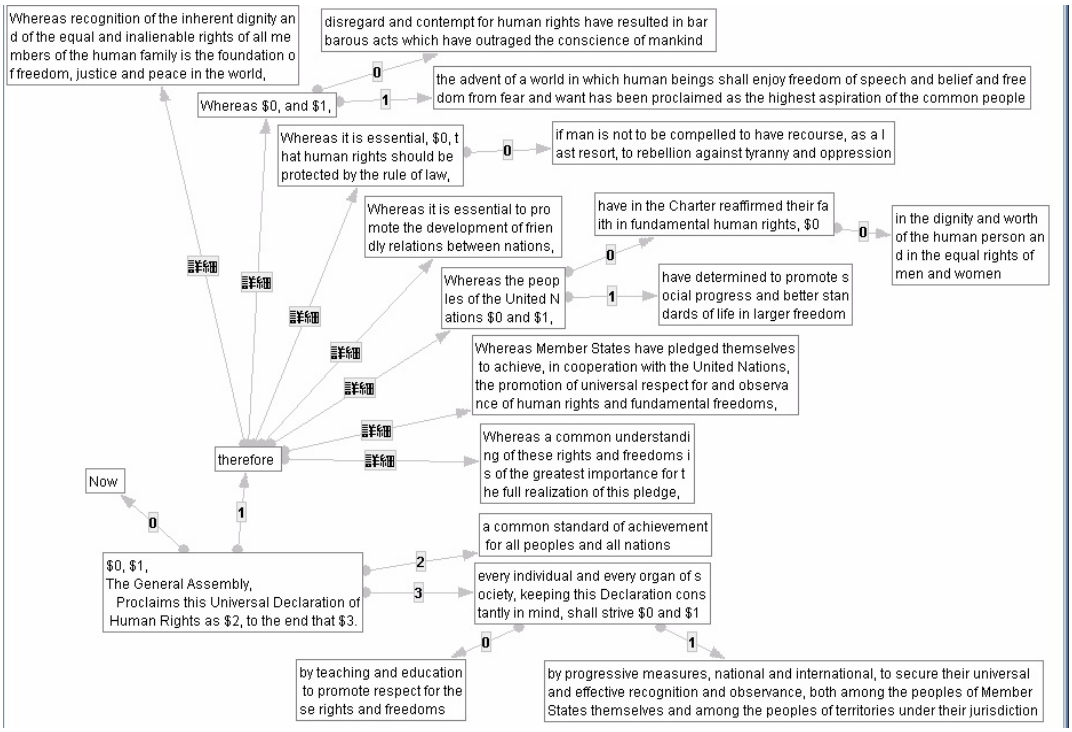


図５．照応・共参照を含む談話構造

その他，公開用のコーパスを Web 経由で検索し，結果を KWIC 形式で表示するシステム「茶杓」（谷口，新保，浅原，松本 2007）を試作した．

４．おわりに

本年度は，個別のタグ付けについて，それぞれのグループで独自のツールの開発を進めながら，共通のツールについての議論を進行させた．異なるレベルのタグ付けの間の関係やデータフォーマットの共通化の検討には，共通のコーパスを対象にし，実際にタグ付けを進めながら検討することがより効率的である．来年度は，一般公開を予定しているいくつかの分野のサブコーパスを班全体の共通データとして，そのデータへの異なるレベルのタグ付けを行い，データフォーマットの整理を行なっていきたい．また，自動タグ付けツールやタグ付け支援システムを実用レベルまで効率化し，コーパス班の作業の支援を行いたい．

文献

Scott Cotton, Steven Bird (2002) “An Integrated Framework for Treebanks and Multilayer Annotations,” Proceedings of the Third International Conference on Language Resources and Evaluation, pp. 1670-1677. (http://agtk.sourceforge.net/)

飯田龍，小町守，乾健太郎，松本裕治 (2007a),「NAIST テキストコーパス: 述語項構造と共参照関係のアノテーション」，情報処理学会第 177 回自然言語処理研究会, 2007-NL-177, pp.71-78.

飯田龍，小町守，乾健太郎，松本裕治 (2007b),「日本語書き言葉を対象とした述語項構造と

共参照関係のアノテーション: NAIST テキストコーパス開発の経験から」, 言語処理学会第 13 回年次大会論文集.
小町守, 飯田龍, 乾健太郎, 松本裕治 (2007),「事態性名詞の項構造解析における共起尺度と構文パターンの有効性の分析」, 言語処理学会第 13 回年次大会論文集.
Koiti Hasida (2003) "Distributed Semantic Authoring as Foundation of Semantic Society," in Notes on From Semantic Web to Semantic World Workshop conjoint with JSAI2003.
Yuji Matsumoto, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Ohtani, Toshio Morita (2006), "An Annotated Corpus Management Tool: ChaKi," Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp.1418-1421.
野口正樹, 市川宙, 橋本泰一, 徳永健伸 (2006) 「構文木付きコーパス作成支援統合環境 eBonsai の新しいインターフェース」 言語処理学会第 12 回年次大会. pp.751-754.
野口正樹, 三好健太, 徳永健伸, 飯田龍, 小町守, 乾健太郎 (2007) 「セグメントとリンクに基づくアノテーションツールの設計と実装」, 言語処理学会第 13 回年次大会論文集.
谷口雄作, 新保仁, 浅原正幸, 松本裕治 (2007) 「タグ付きコーパス検索ツール『茶杓』」, 言語処理学会第 13 回年次大会論文集.

平成 18 年度進捗状況報告：電子化辞書班
（多様な目的に適した形態素解析システム用電子化辞書の開発）

伝康晴（電子化辞書班班長：千葉大学文学部）†
山田篤（分担者：京都高度技術研究所）
峯松信明（分担者：東京大学大学院新領域創成科学研究科）
内元清貴（分担者：情報通信研究機構）
小木曽智信（分担者：国立国語研究所）

The development of a multi-purpose electronic dictionary for morphological analyzers

Yasuharu Den (Faculty of Letters, Chiba University)
Atsushi Yamada (ASTEM)
Nobuaki Minematsu (Graduate School of Frontier Sciences, The University of Tokyo)
Kiyotaka Uchimoto (National Institute of Information and Communications Technology)
Toshinobu Ogiso (National Institute for Japanese Language)

1. はじめに

本計画班の目的は、従来開発を進めてきた形態素解析システム用電子化辞書 UniDic (伝ほか 2002)を整備・拡充・改良することにより、（１）本研究領域が目指す大規模書き言葉コーパスの構築を支援するとともに、（２）日本語学・日本語教育学における語彙・文法調査研究、自然言語処理における構文・意味解析研究、音声情報処理におけるテキスト音声合成研究など、多様な目的に適した統合的な電子化辞書およびその利用システムを提供することにある(伝ほか 2007)。この目的を達成するために、本年度は以下のことを行なった。

（１） 本研究領域で用いる短単位辞書を関係データベースとして実装し、10 万語を超える辞書情報の登録を行なった。
（２） 辞書データベースと学習コーパスから形態素解析システム用辞書を生成するプログラムを作成し、形態素解析システム ChaSen で運用・評価を行なった。
（３） 語の複合に伴う音変化・アクセント変化に関するデータを作成し、調査・モデル化を行なった。
（４） 中・長単位の自動構成に関するデータ・プロトタイプシステムを作成した。

2. 短単位辞書データベースの構築

2.1. 辞書データベースの設計

UniDic では、語形の変異や表記の揺れに対応するために、語彙素・語形・書字形・発音形からなる階層的な見出し設計を採用しており（図 1 左）、各階層に対してさまざまな属性を記述している（表 1）（カッコ付きのものは計画中）。それぞれの階層を関係データベースのテーブルとして実現し、ID を用いた参照関係によって階層関係を表現した。さらに、活用による語尾変化や語の複合による語頭・語末音の変化（連濁や促音化）を派生するために変化の型を語形テーブルの属性として、それぞれの変化型の取りうる変化形を変化形テ

† den@cogsci.L.chiba-u.ac.jp

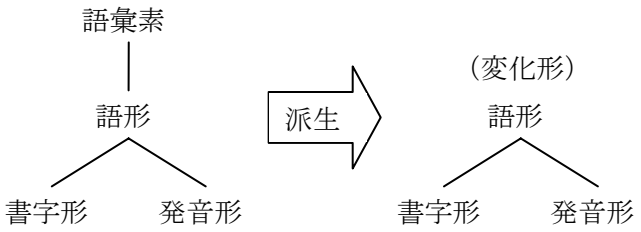


図 1　階層的見出し

表 1　各階層に記述される属性

階層	見出し属性	その他の属性
語彙素	語彙素読み・語彙素表記 語彙素細分類・類	（語種）・（意味分類）
語形	語形基本形・品詞・活用型	語頭変化型・語末変化型・簡略活用型 語頭変化結合型・語末変化結合型
書字形	書字形基本形	活用型書字形分類・語頭変化型書字形分類 語末変化型書字形分類・仮名書字形
発音形	発音形基本形	活用型発音形分類 アクセント型・アクセント結合型

ーブルとして記述し、テーブルの結合操作によってすべての変化形を派生するという方式を取った（図 1 右）（本年度は語尾変化のみ実装）。

2.2.　辞書データベースシステムの開発

前項の設計を持つ辞書データベースシステムをデータ班と共同で構築した。本システムは、語彙素・語形・書字形・発音形の見出し表と、活用表などの変化表を国立国語研究所内のサーバ上に格納したもので、テーブルの結合操作によって、コーパス中に出現する可能な出現形を全展開した語彙表を派生できる。さらに、UniDic 品詞体系で記述されたコーパスをデータベースに格納し、語彙表と対照しながらコーパスを修正していくことを可能にした（図 3 参照）。これらのシステムは Microsoft SQL Server 2005 を用いて構築した。

また、このシステムを利用するためのクライアントとして、辞書データベースに見出しを登録し修正するためのフォームと、辞書と関連づけながらコーパスを修正するためのフォームを開発した。前者は、階層的見出しを木構造表示し、階層構造をそのままの形で編集可能にしたものである。各種の検索機能のほか、語彙素・語形などの枝ごとに移動・コピー・削除を行なう機能を持つ（図 2）。後者は、UniDic 品詞体系で記述されたコーパスを辞書と関連づけながら修正してゆくためのツールである。誤解析されたテキストを分割・結合し直して、語彙表を参照しつつ正しい解析結果を入力する機能を持っている。

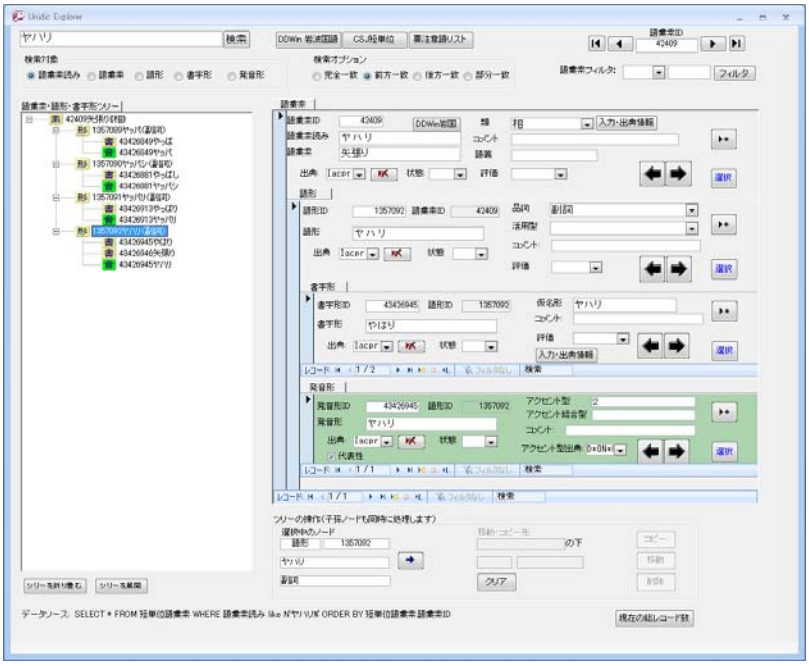


図 2　辞書登録フォーム

これらのクライアント用アプリケーションは Microsoft Access のフォームとして作成した。いずれも文字入力やボタンの押下などのフォーム上の操作に対応して SQL 文を発行し、サーバ上で処理を行ない、結果を受け取って表示する。

2.3.　辞書情報の登録

前項のシステムを用いて、データ班と共同で、日々10 名近くの作業者が辞書とコーパスの修正を行なっている。辞書の規模は、語彙素で約 108,000、書字形で約 138,000 となっている（2007 年 2 月 22 日現在)。

さらに、電子化辞書班では、これらの辞書項目に対して、『大辞林』『NHK 日本語発音アクセント辞典』『新明解日本語アクセント辞典』および『日本語話し言葉コーパス』の音声データをもとにして、約 10 万件の発音形のアクセント型の付与・修正を行なった。

3.　形態素解析システムでの運用

3.1.　形態素解析システム用辞書生成プログラムの開発

構築した電子化辞書を形態素解析システムで利用するために、形態素解析システム用辞書生成プログラムを開発した。本プログラムは、コーパス中に出現する可能な出現形を全展開した語彙表と学習コーパス・学習パラメタからコーパス中の語の分布を統計的に学習し、自動形態素解析に利用可能な辞書を生成するものである（図 3)。形態素解析システムとして ChaSen を、統計モデルには拡張 HMM (浅原・松本 2002)を採用した。学習コーパスには、書き言葉の『RWCP テキストコーパス』と話し言葉の『日本語話し言葉コーパス』をおもに用いた。作成した形態素解析システム用辞書は本年度末に公開する予定である。

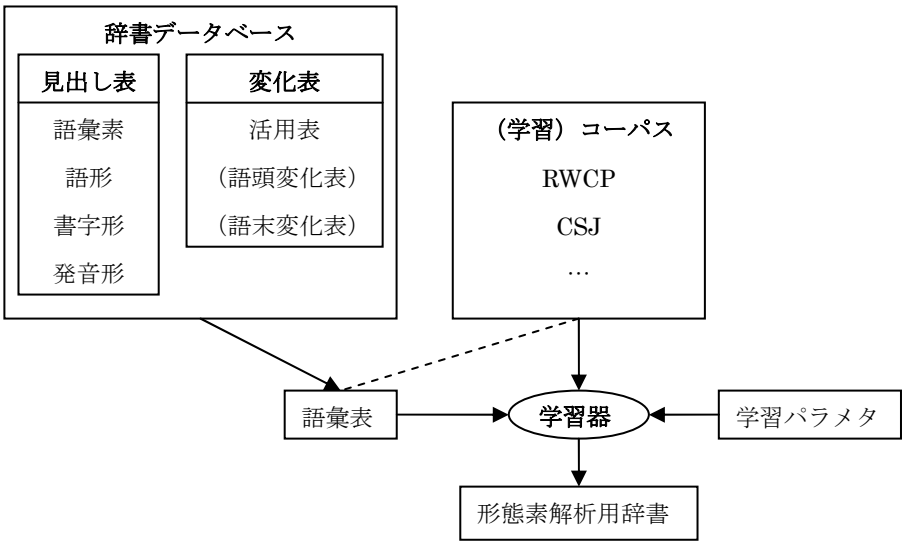


図 3　形態素解析システム用辞書生成過程

3.2.　形態素解析用 GUI の開発

自動形態素解析を手軽に行なうためのツールとして Windows 用の GUI「茶まめ」を作成した（図 4)。コンピュータに不慣れな文系研究者でも形態素解析を研究に生かすことができるようにすることを目的としており、基本的にコピー&ペーストとボタンの押下といった簡単な操作だけで任意のテキストの解析を行なうことができる。形態素解析システム用辞書の Windows 版インストーラとセットにして配布する予定である。

図 4　形態素解析用 GUI「茶まめ」

3.3. 形態素解析前処理ツールの開発

世の中で用いられている電子化文書中には様々な形で数字を含む文字列が出現する。形態素解析において、これらを適切に処理できるようにするための前処理ツール (numtrans) を作成した。UniDic が採用している短単位では、位取りされた数字は一・十・百・千の桁で分割される。そこで、たとえば「４，５００」を「４」「,」「５」「０」「０」の並びと解析するのではなく、「四千」「五百」と解析するためには、形態素解析の前段階で、文字列に対する処理として「４，５００」を「四千五百」に変換する必要がある。このような処理は自動で 100%の精度を出すことは難しく、最終的には人手によるチェックが必須であるが、コーパス作成の際の人手修正の負担軽減を狙っている。

処理手順は、まず入力テキストから数字列の部分を切り出し、その前後を見て変換モードを決定し、最後に適切な変換を行なう。数字列の切り出しの際には、算数字・漢数字だけでなく、デリミタとして用いられる記号類も含めて、変換候補として切り出す。変換モードは、数字列に前接・後接する文字列、および、当該数字列に含まれる記号の情報をもとにヒューリスティクスを用いて判断する。

変換モードとして、通常の位取り表記モード以外に、日付モードや時間モード、電話番号モード（変換しない）等がある。日付モードの場合、たとえば「２００３．１０．１６」を「二千三年十月十六日」に変換する。時間モードの場合、たとえば「３：４５：２３」を「三時四十五分二十三秒」に変換する。

現在、本ツールは、XSLT 1.0 を用いて実装しており、任意の XSLT 処理系で動作可能である。変換処理を施した部分には XML のタグを挿入し、元の表記が復元可能なようにしている。本稿の形態素解析システム用辞書を用いて、数字列を含む電子化文書を形態素解析する際の前処理ツールとして、本ツールも公開する予定である。

3.4. 形態素解析結果の評価

作成した形態素解析システム用辞書を評価した。学習コーパスを用いた評価結果を表 2 に示す。コーパス全体を学習と評価に用いたものを上段、コーパスの 9 割を学習に残りの 1 割を評価に用いたものを下段に示す。評価は、単位境界が正解、品詞・活用型・活用形まで正解、語彙素まで正解の 3 段階で示し、数値はいずれも F 値（再現率と適合率の調和平均）である。従来の形態素解析システムの評価で使われる品詞認定までては十分に高い精度といえるが、本研究で目標とする語彙素認定で 98%以上という数値には達していない。

さらに、データ班にこの辞書を提供し、『現代語書き言葉均衡コーパス』の白書データを解析して､数詞を除く自立語1万単位について精度を見積もったところ､語彙素認定で95.8%程度であった。学習コーパスと比べて低いのは、ジャンルの違いなどによると思われる。

表 2　形態素解析システム用辞書の評価（F 値）

	単位認定	品詞認定	語彙素認定
学習データ	99.5%	98.2%	97.8%
テストデータ	99.3%	97.5%	97.1%

誤解析の原因の追究のために、ChaSen による解析途上に出現する途中解も含めた可能な全解をコスト値とともにラティス状に表示する GUI を作成した。作業員 1 名がこのツールを用いて誤解析の主要な原因の調査を行なっている。

4. 音変化・アクセント変化の処理

4.1. 音変化の処理

形態素解析の結果、各語に付与された発音形に対し、語の複合による語頭・語末音の変化（連濁や促音化）を処理する必要が生じる場合がある。たとえば、書字形「一」を持つ数詞に対する発音形は「イチ」「イッ」「ヒト」「ヒ」「ヒー」がある。これらのうちから適切なものを文脈に応じて選択することは、テキスト音声合成などへの応用では重要な課題であるが、現状の形態素解析技術で実現するのは難しい。

このような音変化を扱うための形態素解析後処理ツールを開発した。現在、対象としているのは、おもに数詞と助数詞である。このような音変化は一定の語の境界をまたいでは起こらない。そのため、音変化処理の前にチャンキング処理を行ない、中単位の範囲内で音変化の処理を行なうことを想定している。

語末の音変化は、後続する語との関係で決まる。UniDic では、後続語が与える当該語への影響を語末変化結合型として記述している。このため、当該語の語彙素読み (lForm)・語彙素表記 (lemma)・発音形 (pron)・語末変化型 (fType)・語末変化形 (fForm) と後続語の語末変化結合型 (fConType) を引数とする以下のような関数を定義し、その値により最尤の発音形を選択するような枠組みが考えられる。

$$F(\mathrm{lForm}, \mathrm{lemma}, \mathrm{pron}, \mathrm{fType}, \mathrm{fForm}, \mathrm{fConType})$$

語頭の音変化についても、当該語の語頭変化型・語頭変化形と前接語の語頭変化結合型を用いて同様の定式化ができる。

このような音変化を処理するツール ChaOne を開発した。現状は上記の関数部分は規則ベースのテーブルで処理している。たとえば、後続する助数詞の語末変化結合型を用いた数詞の発音形の選択について、もう少し詳しく見てみると、助数詞「本」の語末変化結合型は"B1S6SjShS"と定義している。このうち、数詞「一」との接続に関する部分を取り出すと"1S"となる。規則テーブルから F (イチ, 一, イッ, チ促, 促音形, 1S) = 1.0 が得られ、発音形として「イッ」が選ばれる。本ツールも辞書の公開と同時に公開する予定である。

現状では上記のように規則ベースの処理を行なっているが、普通名詞の連濁等のより一般的な音変化に対応するために、統計ベースの処理に移行することを検討している。このために、条件付確率場 (CRF) を用いた予備実験を行なった。今後、学習に用いる前接語・後続語の素性の検討を予定している。

4.2. アクセント変化の処理

UniDic はテキスト音声合成への応用も想定しており、アクセント型・アクセント結合型の属性を保有している。アクセント型は各語を単独で発声した場合のものであり、音声合

成に使うためには、語の複合に伴うアクセント変化などを考慮して、文発声時のアクセントを導く必要がある。このようなアクセント変化（アクセント結合）は、従来、規則によって記述・実装されてきた(匂坂・佐藤 1983; Minematsu et al. 2003)。UniDic のアクセント結合型属性もこの方式を踏襲している。しかし、現状の規則ではあらゆるアクセント結合現象を網羅しているとは言い難く、誤った出力がしばしば見られた。そこで、処理精度の改善を目指すとともに、今後における電子化辞書の在り方を模索するため、統計的手法によるアクセント結合処理の検討を行なった。

統計的手法によって学習・推定を行なうには、正解ラベルの付与されたテキストコーパスが必要となるが、それに適したコーパスは存在しなかった。『日本語話し言葉コーパス』はアクセントラベルを持っているが、話し言葉であるがゆえの乱れや、話者が多数に渡ることによるアクセントの揺れを含み、必ずしも適切ではなかった。そこで、単一のラベラによるアクセントコーパスを独自に作成することとした。ラベラには、文に対してアクセント句境界・アクセント核の位置を付与させ、さらに、文に出現する全ての自立語に対して単独発声時のアクセント核位置を付与させた。ラベラは東京生まれ・育ちであり、合唱部に所属する比較的音感の鋭い大学生 6 名に対して、日本語アクセントに対する教育を施した上で、試験により選抜した者である。アクセント句・核は明確に（物理的に）定義できるものでないため、アクセント句境界は「ピッチの句頭上昇が見られる箇所」、アクセント核は「ピッチが急激に下降する箇所の直前のモーラ（句内の出現回数は制限しない）」と定義し、直感的なラベリングをしやすいようにした。使用した文は、新聞記事読み上げ音声コーパス（JNAS）で使用されているもの（毎日新聞記事から抽出した 16,178 文および ATR 音素バランス文 503 文）である。現段階では、このうちおよそ四分の一についてのラベリングが完了している。

このようにして作成したアクセントコーパスを使用して、統計的手法によるアクセント結合処理を行なった。アクセントコーパスの完成分を、学習用 3,581 文（25,692 アクセント句）、評価用 527 文（3,533 アクセント句）として分け、統計モデルとしては条件付確率場（CRF）を用いた。CRF は、観測データに対する出力ラベルを学習するに際し、ある箇所での観測データと出力ラベルの組に対する特徴（観測素性）や、隣り合う出力ラベルの組に対する特徴（遷移素性）を考え、これらの素性の重要度を変化させることによって学習をするモデルである。観測素性に工夫を加えることによってさまざまな改善を行なったが、アクセント結合規則から得られる知見を観測素性に取り入れたものでは、全アクセント句のうち 93.6%（自立語と付属語の 2 語で構成される単純なアクセント句では 98.3%）で適切な結果が得られた。規則に基づく従来手法では解決できなかった多くの問題への対処ができ、処理精度の大幅な改善に成功した。その一方で、複合名詞などでは係り受けの情報や、意味情報に基づくアクセント変化が観測されるものも少なくなく、統計／規則による処理を越える枠組みを模索する必要がある。

詳細については、本予稿集の(峯松・黒岩, 2007)をご覧いただきたい。

5. 中・長単位構成システムの開発

5.1. 長単位解析システムの開発

短単位から長単位を自動構成するシステムのプロトタイプを作成した。システムには統計的チャンキングモデルを採用した(Uchimoto et al., 2004)。短・長単位情報が付与された『日本語話し言葉コーパス（CSJ）』を用いた実験では、入力となる短単位の情報が適切な場合、約 99%の精度で長単位を解析できることが分かっている(Uchimoto & Isahara, 2007)。

ここでは、さらに、話し言葉コーパスを用いて学習した統計モデルを書き言葉コーパスに適用することを行なった。統計モデルの学習に用いる話し言葉コーパスとして、CSJ のうち人手修正が施された 396 講演（931,282 短単位、773,677 長単位）を用いた。書き言葉コーパスとしては、『京都大学テキストコーパス（Version3.0）』を用いた。京都大学テキストコーパスは、1995 年の毎日新聞記事から抽出した約 4 万文に形態素・文節係り受け構造を付与したものである。このコーパスの語の定義は本研究の短単位とは異なるため、上記学習コーパスを用いて学習した、単語ベースの統計モデルに基づく短単位解析システムにより、再解析した。このうち、1 月 1 日分の記事 1,129 文（29,154 短単位）を用いた実験では、長単位解析の前に、短単位を人手修正し、学習コーパスには現れない句読点や括弧の記号を削除することにより、概ね適切な長単位解析結果が得られることを確認した。詳細な精度評価と本研究領域の書き言葉コーパスへの長単位情報付与は今後の課題である。

5.2. 短単位間の係り受け構造の付与と中単位の認定

中単位は語の内部構造に従った単位であり、長単位を超えない範囲で、直接的な係り受け関係を持つ、隣接する短単位同士を結合したものとして定義できる。今年度は、この定義に従って中単位情報を付与したデータを作成した。

まず、CSJ のうち 40 講演に短単位間の係り受け構造を付与した。短単位間の係り受け構造の付与基準は、文節間の係り受け構造の付与基準(内元ほか, 2004)を参考に新たに作成した。CSJ には文節間の係り受け構造はすでに付与されているため、今回は文節内の短単位間の係り受け構造に限定した。各講演の係り受け構造は、一次作業者 2 名と一次作業者とは別の二次作業者 1 名の計 3 名により付与した。一次作業者は各講演にそれぞれ係り受け構造を付与し、二次作業者は一次作業者の付与結果のずれをチェックし最終判断を行なった。

係り受け構造の付与には、図 5 のような支援ツールを作成して用いた。右下のウィンドウには 1 行に 1 短単位が表示され、短単位間を線で結ぶ形で係り受け関係が表現されている。上部にはそれぞれ、短・中・長単位が表示され、左下のウィンドウには中単位に特有の語が表示されている。ここで表示されている中単位は、短単位間の係り受け構造をもとに、長単位を超えない範囲で、直接的な係り受け関係を持つ、隣接する短単位同士を結合することにより、自動的に認定されたものである。この定義の妥当性については、今後さらに検討する必要がある。

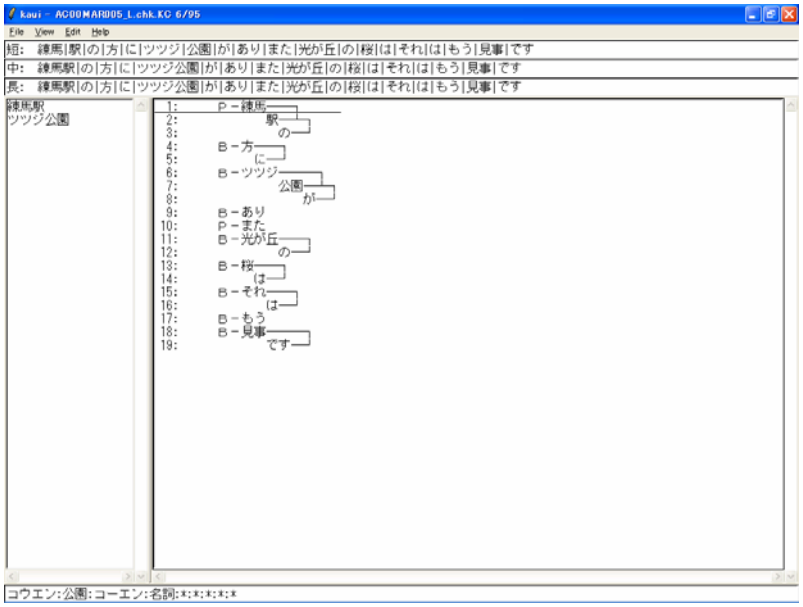


図 5　短単位間の係り受け構造の付与を支援するツール

6.　おわりに

計画研究開始当初に設定した本年度の目標であった

- 短単位辞書データベースの開発と運用
- 短単位辞書の拡充・整備（語彙素数 7 万以上）
- 形態素解析システムでの運用（解析精度 97%以上）
- 音変化・アクセント変化処理の改良
- 中・長単位構成プロトタイプシステムの開発

は概ね達成された。次年度以降に積み残した課題として、形態素解析システムの高精度化、とくに語彙素認定の改善が挙げられる。現在、データ班と共同で、誤解析事例を蓄積しながら、誤解析の原因の究明を進めている。語彙素認定については、そもそも現状の拡張 HMM モデルでは対処が難しい。今年度、音変化・アクセント変化処理で採用した条件付確率場（CRF）を語彙素認定処理にも採用することを検討している。

文献

浅原正幸・松本裕治 (2002). 「形態素解析のための拡張統計モデル」 情報処理学会論文誌, 43, 685-695.

伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治 (2002). 「話し言葉研究に適した電子化辞書の設計」 第 2 回「話し言葉の科学と工学」ワークショップ講演予稿集, pp. 39-46.

伝康晴・山田篤・峯松信明・内元清貴・小磯花絵・小木曽智信 (2007). 「多様な目的に適した形態素解析システム用電子化辞書の開発」特定領域「日本語コーパス」平成 18 年度全体会議予稿集, pp. 21-26. (http://www.tokuteicorpus.jp/result/pdf/2006_017.pdfよりダウンロード可能)

Minematsu, N., Kita, R., & Hirose, K. (2003). Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion. *IEICE Transactions on*

Information and Systems, *E86-D*, pp.550-557.
峯松信明・黒岩龍 (2007). 「単独ラベラによる大規模アクセントラベリングとそれを用いた統計的アクセント結合処理の実装」 (本予稿集).
匂坂芳典・佐藤大和 (1983). 「日本語単語連鎖のアクセント規則」 電子通信学会論文誌, *J66-D*, pp.847.856.
内元清貴・丸山岳彦・高梨克也・井佐原均 (2004). 『『日本語話し言葉コーパス』における係り受け構造付与』http://www2.kokken.go.jp/~csj/public/members_only/manuals/dependency_2004MAR30.pdf
Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. *In Proceedings of IJCAI*, pp. 1731-1737.
Uchimoto, K., Takaoka, K., Nobata, C, Yamada, A., Sekine, S., & Isahara, H. (2004). Morphological analysis of the Corpus of Spontaneous Japanese. *IEEE Transactions on Speech and Audio Processing*, *12*, pp. 382-390.

平成 18 年度研究進捗状況報告：日本語学班

コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発

田野村忠温（班長：大阪外国語大学外国語学部）
服部匡（分担者：同志社女子大学学芸学部）
杉本武（分担者：筑波大学大学院人文社会科学研究科）
石井正彦（分担者：大阪大学大学院文学研究科）

Progress Report of the Year 2006: 'Japanese Linguistics' Group

Tadaharu Tanomura(Osaka University of Foreign Studies)
Tadasu Hattori(Doshisha Women's College of Liberal Arts)
Takeshi Sugimoto(University of Tsukuba)
Masahiko Ishii(Osaka University)

１．日本語学班の研究目的

日本語の研究において、語句の用例の収集手段としてコーパス（電子媒体の言語資料）を用いることが昨今一般化してきた。これは研究者の内省や少数の用例に依存せざるを得なかった従来の日本語研究法に大きな改善をもたらすものであり、日本語研究史という大きな観点から見てもまさに画期的な変革を意味すると言ってよい。

しかし、コーパスの意義は、単に用例収集の簡便な代用手段たり得るにとどまらない。資料の規模がきわめて大きく、しかも、大量のデータを柔軟かつ効率的に処理できるという特徴を有するこの新しい形態の言語資料の出現により、従来の方法では望むこともできなかった様々なタイプの日本語研究への道が拓かれようとしている。

日本語学班は、具体的な事例研究の試行を通して日本語研究におけるコーパス利用の価値を明らかにし、日本語の新しい研究領域・手法を開発するとともに、それにより学界に対してコーパスを用いた日本語研究の啓蒙・普及を図ることを目的とする。と言っても、最終的な目的は日本語研究そのものの発展・質的向上への寄与にある。コーパス利用法の開拓と普及はそのための強力な手段として理解されるべきものである。

また、本特定領域研究の中心的な成果となる、我が国初でありかつ将来日本語研究の標準的資料として広範に利用されるであろう大規模な書き言葉コーパスの構築の進行に伴い、それを日本語研究に適用し、その過程で得られた知見をコーパスの構築にフィードバックすることをも目的とする。

２．本年度の研究内容

各班員が事例研究の試行を通して日本語研究におけるコーパス利用の可能性を模索するとともに、研究会などの場においてそうした研究の成果や日本語研究におけるコーパス利用の諸問題について討議した。

本特定領域研究で構築中のコーパスはまだ使用することができないため、文学作品や新

聞記事などの既存の電子媒体の日本語資料を用いて研究を進めた。また、特定領域内部で利用可能となった数種類の日本語データも試用し、日本語研究における資料としての性格などについても検討を行った。

本年度の各班員の具体的な研究内容は以下の通りである。

Ａ　コピュラ諸形式の分布の分析（田野村）

日本語研究の隆盛により、主要な文法の問題のあれこれに関しては、多数の研究者が集中的に取り組み微に入り細をうがつ議論が重ねられている。しかし、そうした状況の一方で、非常に基礎的でありながら顧みられず、表面的な言語事実の観察さえ手付かずの状態にとどまっている文法の問題もある。その１つに「ＡはＢだ（である、です）」という文型に代表されるコピュラ述語文（いわゆる名詞述語文）の問題がある。

コピュラ述語文をめぐる問題は多岐にわたるが、まず考察を要する基本的な問題として、①コピュラの形式にはどのようなものがあるのか、②それらの諸形式はどのような条件に基づいて生起するのか、の２つがある。

コピュラの形式に関しては、「だ」「です」などのいわゆる指定（断定）の助動詞とその活用形がコピュラの代表例であるが、そのほかにも、例えば「でいらっしゃる」「でおいでになる」のように一般には指定の助動詞とはされないコピュラもある。また、逆に、「なら」のように指定の助動詞の一活用形とされるもののコピュラと考えるべきではないものもある。「これは偽物φに違いない」のように無形のコピュラφがその位置に潜在していると理解すべき場合もある。詳しく述べる余裕はないが、ほかにも検討を要することがらがいくつもある。

次にコピュラの諸形式の分布の問題もなかなか複雑である。従来の研究では、例えば“コピュラ述語に「のだ」が付くときはコピュラが「な」に変わる”とか“「ようだ」の直前ではコピュラが「の」に置き換えられる”といった具合に、コピュラに後続する特定の要素の観点から――他の後続要素の場合のことには無関心に、しかも、気紛れかつ断片的に――コピュラの異形態の交替が指摘されるにとどまっていた。しかし、これをコピュラの観点から見直せば、然るべき観察・分析が必要であることが明らかになる。いくつかの例を挙げれば、

(1)　太郎が学生｛である／　だ／×な／×の／×φ｝から～
(2) 私は太郎が学生｛である／　だ／×な／×の／？φ｝と思う
(3)　太郎が学生｛である／×だ／×な／×の／×φ｝ものの～
(4)　太郎が学生｛である／×だ／？な／　の／×φ｝はずだ
(5)　太郎が学生｛である／×だ／×な／×の／　φ｝なら～

に見るように、生起可能なコピュラ形式の範囲はその後続要素に応じてさまざまな異なりを見せるのである。中には、「だけ」のように、用法によって可能なコピュラ形式が異なるものもある。

(6) a. 太郎が学生｛である／×だ／？な／×の／×φ｝だけだ
　　b. 太郎が学生｛である／×だ／？な／×の／　φ｝だけに～

さて、以上は主に自分の内省に基づく考察の結果である。いくつかの例文で「？」を付けたところは、誤りとは言わないが、不自然であるか、もしくは、くだけた物言いにのみ適した言い方であるように感じる。しかし、内省に頼ってコピュラ形式の選択の適否を判断すること自体困難であり、個人差や日本語資料の種類などによる違いの存在も予想され

る。

コピュラ形式の分布は上述のようにコピュラに後続する要素に依存するだけなく、先行する要素にも依存しており、状況は非常に複雑である。本年度は、コピュラ形式の分布の様相をいくつかの種類のコーパスを使って調査するとともに、日本語研究におけるコーパスの種類や規模の問題について一般的な見地から若干の考察を行った。コピュラ形式の分布の問題は複雑で不明瞭な点も多いことから、その全貌を正確に見極めることは容易ではない。来年度も引き続きこの問題に取り組むとともに、同時に、コピュラ述語文の構文・意味を始めとする新たな問題の調査・分析を通してコーパス利用の新たな可能性を探る予定である。

B　副詞と述語の共起関係の分析等（服部）

①多変量解析の手法を用いた、程度副詞－述語の共起関係の分析

語の共起特性に関する新分析手法の探究の一環として、程度副詞と述語の間の共起関係を、大量データ（新聞記事）に因子分析の手法を適用することによって分析し、それに基づいてそれぞれの程度副詞の特徴づけを行うことを試みた。

程度副詞の構文的・意味的特徴による分類・体系化に関しては、主に内省判断に基づいた提案がいくつかあるが互いに齟齬する点もあり、実用例に照らして妥当性が疑わしい点もある。また、そもそも程度副詞を離散的なカテゴリーへ分類したり体系という観点から捉えたりすることが適切であるのかどうかも自明ではないように思われる。

一切の先入観なく実際のデータのみからそれぞれの程度副詞の、述語との共起における特性を抽出するため、新聞記事の大量データから程度副詞－述語を切り出し、その組み合わせに対して因子分析の方法を適用することにより、各々の程度副詞の共起特性を明らかにすることを試みた。対象とするデータは新聞 13 年分、約 16 億字である。

この目的のため、原データの抽出と解析手法の検討を行った。文末まで 10 字以内の述語に対象を限定して、一定以上の回数出現する述語のすべてと程度副詞との共起例のすべてを新聞記事データから切り出した。当該の述語は 555 個、程度副詞は 15 個あり、組み合わせは 8325 になる。そのうち共起例のある組み合わせは 2526 個，共起例のない組み合わせは 5799 個である。

それぞれの組み合わせに共起例があるかないかの二値データに因子分析を適用して 4 因子を抽出し、因子得点と述語の意味を参照して各因子への解釈を与えた。

この結果を、従来の内省（実例視察）に基づく分類・体系化と比較検討すると、渡辺実氏による程度副詞の体系化・分類に概ね一致しているが「ずいぶん」という語に関しては相違が観察された。また、程度副詞を量的なものとそうでないものに分類するいくつかの説（工藤浩氏、森山卓郎氏、など）に対して、大筋で矛盾しない結果が得られた。

これらの検証を通じて方法の有効性が確認でき、今後、従来研究が行われていない言語現象の研究における発見手順として同様の方法を利用しうる可能性があると考えられる。

②　「～先」における意味関係の研究

「○○(二字漢語動名詞)先」の、新聞コーパスでの用例を精査したところ、「受注先（受注者/被受注者)」、「支援先（支援者/被支援者)」、「依頼先（依頼者/被依頼者)」のように二義で用いられるものがあることが分かった。このような事実は従来報告されていない。

「○○先」の意味を調査すると、全体としてゴール（目標点・到達点）を表すものと、相手方を表すものとがある（両者重なる部分もある）ことが分かる。また前者は全体とし

て表すものが動名詞「○○」に対する二格の関係にある場合が大部分であるが、ヲ格その他に対応するものも一部ある。ゴールとしても相手方としても解釈可能な「○○先」には二義性を生じるものがあり、多くは定型的な行為にかかわるものである。その他、「○○先」はいわゆるトコロ名詞としての性格を持つことなど、いくつかの意味的な特徴を明らかにした。

さらに「基地移転先」「資源供給先」「就職希望先」「部品調達先」のように「先」の前に複数の漢語要素を持つものについても同様に二義性を生じる例があることが分かり、それらの形式を項構造等の観点から分類した。

このような調査・分析は、大量データの利用によってはじめて可能になったものである。

C　複合格助詞の分析（杉本）

単純形式の格助詞の場合、例えば「～することに(決めた)」「～するのを(見る)」のようにコト節やノ節の形で補文をとることができるが、これは出来事を格成分としてとることができるという述語の特徴によるものであると考えられる。一方、複合格助詞の場合、「によって」「に際して」などのように、その複合格助詞自体の特徴により補文をとることができる。また、このような場合、補文と複合格助詞から成る成分は、格成分と言うより、従属節に近くなる。この点で、複合格助詞と呼ばれるものも、接続助詞に近くなっていく。

本年度は、このような複合格助詞のうち「によって」を取り上げ、用例の分類を行いながら、分析を行った。一つには、「によって」がコト節をとる場合に着目した。「によって」は意味的には手段や原因などの用法を持つが、この用法の違いは、主文の述語、コト節の述語の特徴やテンスなどによって決定されていると考えられる。コト節をとる「によって」としては、手段の用法と原因の用法とがあるため、この二つの用法に関して、コト節の述語の特徴、テンスの現れ方を見た。直観的にも、「その会社は、最新機器を導入することによって業界トップに躍り出た」のようにコト節がル形の場合は手段の解釈になるのに対して、「その会社は、最新機器を導入したことによって業界トップに躍り出た」のようにタ形にすると、原因の解釈がしやすくなる。毎日新聞から収集したデータを検討した結果、(1)手段の用法の場合、コト節がル形になること、(2)原因の用法の場合、コト節がタ形になることが多いが、ル形のものもいくらか見られること、(3)ただし、その場合、主文もル形になることが多いこと、などがわかった。さらに、毎日新聞から、コト節をとる「によって」句が「ながら」「つつ」による A 類の従属節に含まれる用例、コト節をとる「によって」句が擬似分裂文の焦点に現れている用例を採集したところ、少数ではあるが用例が採集され、そのほとんどがコト節がル形をとる、手段の用法のものであることがわかった。ただし、コト節がタ形をとる用例自体が相対的に少なく、この結果が信頼できるものであるのかどうかは検討の余地がある。このような稀な用例の場合、かなり大量のデータを用いない限り(今回使用したデータは毎日新聞 10 年分である)、文法的であるとしても実際には出現しない可能性があり、現実的にそれを検証することが難しい。このような問題をどのように克服するかが課題となる。

また、「によって」には、受動文の動作主の用法も持つが、これと手段、原因の用法とは

類似性を持つ。そこで、毎日新聞のデータをもとに、受動文の「によって」句と動作主の関係、受動文の動作主の「に」句と「によって」句の共起を検討した。このような現象は内省による判断がしにくく、一般的に、内省による判断がしにくい現象は、コーパスを用いる価値が高いものの、コーパスを用いた分析に馴染みにくい現象もある。このような問題に、どのようにコーパスを活用していくか、その方法論の開拓が課題となる。

次に、上のような現象が他言語でも見られるのかを確認するために、韓国語との対照を行った。韓国語の"e uihae"は、日本語の「によって」と似た語構成、用法を持つが、テンスの制限に関して、日本語とは異なることが明らかになった。ただし、これは、韓国語の"e uihae"が理由の用法も併せ持つという用法上の差である可能性がある。

「によって」は、「ために」のような理由、目的を示す形式、「おかげで」「せいで」のような原因を示す形式との関係も含めて検討する必要もある。これらの形式は接続助詞的なものとされるが、このようなものも含めた、原因、理由などを示す格、接続形式を包括的に分析することが、来年度の課題となる。

D　通時コーパスを使った現代語の語彙変化の研究（石井）

新しい「コーパス語彙研究」の可能性として、

A．この特定領域研究でつくられる「代表性を有する大規模な書きことばの均衡コーパス」を用いて可能になる（新しい）研究

B．上記「均衡コーパス」とは違ったタイプのコーパスを用いて可能になる（新しい）研究

の二つを掲げ、計画期間の前半ではBを中心に、(「均衡コーパス」が本格的に利用できる)後半ではAを中心に、各種コーパスの作成やそれを用いた調査を具体的に行うことを計画している。

前半で行うBの研究としては、

(1)通時コーパスを使った、現代語における語彙変化の研究

(2)画像付きコーパスを使った、言語行動（としての単語使用の）研究

後半で行うAの研究としては、

(3)均衡コーパスを使った、単語の社会・文化的意味の研究

(4)均衡コーパスを使った、基本語彙をはじめとする語彙論の主要概念の検証ないし精密化

を、具体的なテーマとする予定である。(4)は、「新しい研究」とはいえないが、大規模な均衡コーパスを使って、先行研究の内容を検証したり、従来の小規模なコーパスや平コーパスでは実現できず「今後の課題」とされているような問題を扱ったりするものである。

本年度は、Bの(1)「通時コーパスを使った、現代語における語彙変化の研究」を主たる課題とし、20 世紀後半の日本語における語彙の変化をとらえ得る「通時コーパス」の一つとして、毎日新聞のコラム「余録」欄の、1950・60・70・80・90・2000 年の各１年分・計６年分を入力したコーパスを試作した。

新聞のコラムを対象とするのは、各年の延べ語数がほぼ等しくなる、文体も期間を通してほぼ同じである、話題が適当に分散していて特定の語彙に集中せず、また、まったく拡散してしまうということも少ない、と考えたからである。

今年度中には入力を終える予定であるが、現在のところ、各年平均で約 26 万字、自立語の延べで約８万語程度、全体では 50 万語弱のコーパスになるものと見込んでいる。この程度の規模で、たとえば、個々の単語の語誌を記述することがどの程度可能かを検討することが、当面の問題となる。

試みに、このコーパスの、すでに入力を終えた半年分を使って「外人／外国人」の使用頻度を調べてみると、表１のようになった（空欄は０、カッコ内は結合用法・外数）。また、表２は、雑誌『中央公論』から各年延べ１万語をサンプリングした国立国語研究所『雑誌用語の変遷』(1987)における「外人／外国人」の用例数である。

（表１）「余録」コーパスにおける「外人／外国人」の用例数

	1950	1960	1970	1980	1991	2000
外人	5(1)		1(2)	2		
外国人	2	2	8	3(1)	6	3

（表２）『中央公論』調査における「外人／外国人」の用例数

	1906	1916	1926	1936	1946	1956	1966	1976
外人	1		1					
外国人		1			2			

新聞をはじめとするマスコミにおいて、「外人」は「外国人」に置き換えられたと考えられるが、表１は、そのことをおおよそ示しているように思える。今後、データ量を各年１年分に増やすことで、より精度の高い調査結果を得ることができるものと考えられる。このほか、調査年の間隔を短くしたり、他の新聞のコラムについても同様にコーパス化するなどして、新聞コラムのコーパスの、通時コーパスとしての可能性と問題点を明らかにしていきたい。

３．研究会・会議の開催

・9/25 第１回研究会を開催（大阪）
　研究発表５件（班外の方１件）および全体討議
　班外・領域外の方を含め 16 名参加
　領域公開サイトに実施報告を掲載
　領域内部サイトにプログラム、配布資料、質疑応答・全体討議の記録を登録
・11/20 非公開の会議を開催（東京）
　今後の活動方針の相談、各種情報・意見交換
　領域公開サイトに実施報告を掲載

・1/14 第２回研究会を開催（大阪）
研究発表５件（班外の方１件）および全体討議
班外・領域外の方を含め 10 名参加
領域公開サイトに実施報告を掲載
領域内部サイトにプログラム、配布資料を登録

４．班間の交流状況（総括班会議への班長の参加を除く）

・9/9-10　全体会議　田野村・杉本・石井が参加
・9/25　日本語学班研究会　班外・領域外より 12 名の方が参加
・11/3　コロケーション研究会　杉本が参加
・11/7　コーパス仕様説明会　田野村・杉本が参加
・12/11　辞書編集班研究会　田野村が参加
・1/14　日本語学班研究会　班外・領域外より６名の方が参加
・2/26　辞書編集班研究会　田野村が参加
・3/17-18 公開ワークショップ　全班員が参加予定

５．その他

・メーリングリストを用いて日常的に情報・意見交換
・コーパスを使った日本語研究の文献リストを作成中
・本年度の研究成果報告書を作成中

文献

石井正彦(2006)「日本語研究における探索的データ解析の有用性」土岐哲先生還暦記念論文集編集委員会編『日本語の教育から研究へ』（くろしお出版）

石井正彦(2007)『現代日本語の複合語形成論』（ひつじ書房）

杉本武(2006)「複合格助詞『にとって』の意味と文法機能」藤田保幸・山崎誠編『複合辞研究の現在』（和泉書院）

杉本武・李昇祐(2007)「複合辞の日韓対照――『によって』と"e uihae"――」『文部科学省科学研究費補助金 基盤研究 B(2)「諸外国語と日本語の対照的記述に関する方法論的研究」研究成果報告書』（研究代表者:青木三郎、筑波大学）

田野村忠温(2006)「コピュラ再考」藤田保幸・山崎誠編『複合辞研究の現在』(和泉書院)

田野村忠温(2007 刊行予定)「コーパス言語学と語彙」斎藤倫明・石井正彦編『これからの語彙論』（ひつじ書房）

服部匡(2006)「『～どころか』、『～どころで（は）ない』とその周辺の諸表現――あわせて、『～ばかりか、～はおろか』等との比較――」藤田保幸・山崎誠編『複合辞研究の現在』（和泉書院）

服部匡(2007)「因子分析を用いた程度副詞と述語等の共起関係分析の試み――新聞コーパスのデータから――」『同志社女子大学総合文化研究所紀要』 第 24 号

付記

現在作成中の今年度の日本語学班研究成果報告書は領域の皆様全員にお送りする予定ですが、ご関心をお持ちいただける領域外の方にもお送りいたします。詳しくは３月下旬ないし４月に班長の Web サイト（http://tanomura.osaka-gaidai.ac.jp/）に掲載予定の案内をご覧ください。

今後、領域内外の皆様からの日本語学班へのご支援、ご協力をお願いいたします。電子資料の特性を生かした、新しいアイディアに基づく日本語研究の試行例をお持ちの方は随時ご連絡いただければ幸いです。

平成 18 年度進捗状況報告：日本語教育班

（代表性を有する書き言葉コーパスを活用した日本語教育研究）

砂川 有里子 （班　長：筑波大学大学院人文社会科学研究科）†
井上 優 （分担者：国立国語研究所日本語教育基盤情報センター）
小林 ミナ （分担者：早稲田大学大学院日本語教育研究科）
滝沢 直宏 （分担者：名古屋大学大学院国際開発研究科）
投野 由紀夫 （分担者：明海大学外国語学部）
山内 博之 （分担者：実践女子大学文学部）

Progress Report of the Year 2006: 'Japanese Language Education' Group

Yuriko Sunakawa (University of Tsukuba)
Masaru Inoue (National Institute for Japanese Language)
Mina Kobayashi (Waseda University)
Naohiro Takizawa (Nagoya University)
Yukio Tono (Meikai University)
Hiroyuki Yamauchi (Jissen Women's University)

１．日本語教育班の研究目的

日本語教育班の研究目的は,「現代日本語書き言葉均衡コーパスを日本語教育に活用する方法の開発」である。従来教師の経験と勘にもとづいて作成されていた日本語教科書，教材，シラバスについて，客観的な言語データにもとづいて考えるべき部分を見極め，コーパスを日本語教育に活用する方法について検討する。具体的には，次の 3 つの課題に取り組む。

・課題１「日本語教材コーパスの作成と分析」
・課題２「書き言葉均衡コーパスを活用した日本語教材作成法の開発」
・課題３「日本語教育のためのコーパス活用ツールの開発」

日本語教育の内容について考える際に重要なのは，日本語母語話者の日本語使用の実態に関する信頼できるデータと，日本語教育に必要な日本語情報を的確に抽出する方法である。他の研究班と連携して，コーパスに含まれる文章や付加情報の内容について検討することは，書き言葉均衡コーパスの応用分野を広げることにつながる。また，教師や学習者が大規模日本語コーパスを活用できるようにすることは，コーパスの有効活用という観点からも重要である。

２．平成 18 年度の研究経過

２．１　班会議の開催

班会議を 6 回開催した。各回の日時，場所，主なテーマと発題者などは以下のとおり。

† sunakawa@sakura.cc.tsukuba.ac.jp

第１回　2006 年 7 月 31 日（月）　早稲田大学西早稲田キャンパス

・特定領域全体に関する概略説明（前川）

・日本語教育班の研究計画に関する検討

第２回　2006 年 8 月 22 日（火）　東京大学駒場キャンパス

・日本語教育班の研究計画に関する検討

第３回　2006 年 10 月 1 日（日）　早稲田大学西早稲田キャンパス

・「日本語教科書コーパス」に関する検討

・研究課題に関する構想発表

a　語彙文法シラバスの作成

b　コーパスにもとづく日本語教育文法の再考

c　コロケーション研究の日本語教育への応用

d　英語教育での経験にもとづく，日本語教育へのコーパスの活用

e　「用例用法辞書」とコーパス

第４回　2006 年 11 月 3 日（日）　国立国語研究所

・班会議

a　「日本語教科書コーパス」の進捗状況（井上）

b　日本語教育のための「語彙シラバス」の作成（山内）

c　中国の「日本語教科書コーパス」の紹介（曹大峰）

・プレ・コロケーション研究会

a　発題：「日本語のコロケーション研究」（滝沢）

b　ディスカッション

第５回　2006 年 12 月 17 日（日）　早稲田大学西早稲田キャンパス

・「日本語教科書コーパス」の進捗状況（井上）

・研究課題に関する発表

a　日本語学習者のためのコロケーション辞書作成のための基礎的研究（滝沢）

b　コーパスを活用したジャンル別文法シラバスの作成（砂川・小林）

c　日本語教育のための語彙シラバスの作成（山内）

d　Sketch Engine の日本語インターフェイス構築（投野）

第６回　2007 年 2 月 11 日（日）　早稲田大学西早稲田キャンパス

・「日本語教科書コーパス」の進捗状況（井上）

・均衡コーパスに付与を希望する情報に関する検討。

２．２　平成 18 年度の進捗状況

平成 18 年度は，「１．日本語教育班の研究目的」で掲げた 3 つの課題について，次のことを行った。

課題１「日本語教材コーパスの作成と分析」

・日本語教材コーパスの位置づけについて検討した。当初は独立のコーパスを想定していたが，他の媒体（特に学校教科書）との比較ができるように，基本的な仕様を言語政策班で作成する「教科書コーパス」に合わせて，非母集団コーパスに組み込む方向で進めることにした。

・現状においてまず実態把握が必要なのは中級以上の教材であると考え，中級向けの総合教科書と読解教材のコーパスの作成に着手した。

課題２「書き言葉均衡コーパスを活用した日本語教材作成法の開発」

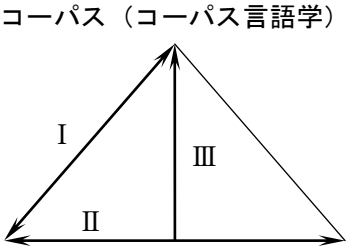
・他研究班との関係を視野に入れながら，以下の方向でコーパスを日本語教育に活用する可能性について検討した。

・日本語学習者のためのコロケーション・リストの作成
・ジャンル別文型シラバスの構築
・語彙シラバスの構築

課題３「日本語教育のためのコーパス活用ツールの開発」

・日本語でSketch Engineのインターフェイスを実現するための検討を開始した。

他研究班との連携は，日本語教育班にとって重要な課題である。日本語教育とコーパス（コーパス言語学）ならびに他の研究班との関係は，概略次のように整理できる。



Ⅰ：a　コーパス（コーパス言語学）の日本語教育への応用
　　b　日本語教育からコーパス（コーパス言語学）への提案
Ⅱ：他研究班との知見の共有（共通利益の追求）
Ⅲ：他研究班との連携によるコーパス（コーパス言語学）への提案

平成18年度は，このうち「Ⅱ　他研究班との知見の共有（共通利益の追求)」の可能性について検討した。

言語政策班とは,「教科書コーパス」と「語彙リスト，語彙シラバス」という２つの観点を共有している。日本語教育において，学校での教科教育を含む年少者対象の日本語教育の問題がクローズアップされるようになったこともふまえ，日本語教育班が作成する「日本語教科書コーパス」は，言語政策班が作成する「教科書コーパス」と基本的に同じ仕様とし，相互の比較ができるようにする。「教科書コーパス」を文型シラバスという観点から分析することも今後の課題となる。

「語彙リスト，語彙シラバス」については,「教育基本語彙」「語彙の難易度付与」という点で，日本語教育班と言語政策班の研究は重なる部分が大きい。コーパスから取得できるどのような情報に着目して語彙リストを作るか，また，そのためにコーパスにどのような情報を付与するか，さらに，何をもとにして語彙の難易度を決定するかといった点について，より具体的な検討を言語政策班，データ班と共同で進めたいと考えている。

辞書編集班とは,「コロケーション」という観点を共有している。辞書編集班で行われるのは日本語母語話者向けの辞書を念頭においたコロケーション研究なのに対し，日本語教育班で行うのは，日本語学習者向け日本語辞書の構築を念頭に置いた「非母語話者向けコロケーション・リスト」の作成である。この研究は，辞書編集班および日本語学班の「複合辞」研究とも問題意識を共有している。第 4 回の班会議の後に班の枠をこえた「プレ・コミュニケーション研究会」を開催したが，今後，他研究班と共同の研究会を立ち上げるなどして，コロケーションに関する知見の相互共有を行いたい。

研究分担者の井上が国立国語研究所で担当している研究プロジェクト「日本語学習のための用例用法辞書のモデル作成」も，学習者用日本語辞書に関連する研究である。この研究では，日本語学習者に対して説明が必要な「意味上（使用上）の単位」を母語話者の判断に基づいて抽出し（例 :「ここに来て…」がこのまとまりで「この段階に至って」という意味を表す)，その単位の用法に関する文法的・語用論的・社会言語学的情報を記述する。この研究は，日本語教育班で行うコーパスを利用した非母語話者向けコロケーション・リストの作成と相互補完的な関係にある。

ツール班とは,「日本語教育に役立つ検索ツールの作成」について共同で考える必要がある。その際に最も重要な問題は,「日本語教育で 1 つの表現項目として扱われる単位と，コーパスにおける解析単位とが必ずしも一致しない」ことである。現在，日本語で Sketch Engine のインターフェイスを実現するための検討を始めているが，日本語教育班としては,「どのような検索が日本語教育のために有効か」ということを，できるだけ具体的な形で見極め，提案を行うことが必要である。

３．平成 19 年度以降の計画

３．１　平成 19 年度

平成 19 年度は，各課題について次のことを行う。

課題１「日本語教材コーパスの作成と分析」

- 平成 19 年度末までに中級向け総合教科書と読解教材のコーパスを作成する。

課題２「書き言葉均衡コーパスを活用した日本語教材作成法の開発」

- 日本語学習者向けのコロケーション・リストに求められる条件について検討し，リストの試作を行う。
- 既存のコーパス等を利用し，文章のジャンルによる文型の出現の様相について予備的調査を行う。
- 語彙の難易度を測るための予備的調査を実施する。

課題３「日本語教育のためのコーパス活用ツールの開発」

- 日本語教育のために有効な検索について問題の整理を行う。
- 日本語コーパス・データを Sketch Engine 実装で検索することを試行する。

３．２　平成 20 年度以降

各研究課題について具体的な展開を図る。

平成 18 年度進捗状況報告：言語政策班

（言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用）

田中 牧郎（班　長：国立国語研究所研究開発部門）†
相澤 正夫（分担者：国立国語研究所研究開発部門）
棚橋 尚子（分担者：奈良教育大学教育学部）
野村 敏夫（分担者：桜美林大学文学部）
鈴木 一史（協力者：東京大学教育学部付属中等教育学校）

Progress Report of the Year 2006: 'Language Policy' Group

TANAKA Makiro　（National Institute for Japanese Language）
AIZAWA Masao　（National Institute for Japanese Language）
TANAHASHI Hisako（Nara University of Education）
NOMURA Toshio　（Obirin University）
SUZUKI Kazufumi　（Tokyo University Secondary Education School）

１．言語政策班の研究計画

国語施策と国語教育に役立てることのできる語彙表と漢字表を，代表性を有するコーパスに基づいて作成し，それらを活用する方法を開発することを目的とし，次の五つの課題を設定している。括弧内は主たる担当者である。

ａ）難解用語の抽出と言い換え（田中牧郎）
ｂ）常用漢字表・人名用漢字等の在り方に関する調査研究（相澤正夫）
ｃ）文章作成における語彙選択指導（野村敏夫）
ｄ）概念体系構築のための語彙指導（鈴木一史）
ｅ）国語力向上のための漢字指導（棚橋尚子）

これらの課題に取り組むための共通の基盤として，次のデータベースを構築する。

Ⅰ）「現代日本語書き言葉均衡コーパス」に基づいた語彙表・漢字表
Ⅱ）「教科書コーパス」と，これに基づいた語彙表・漢字表

「現代日本語書き言葉均衡コーパス」は，本領域のデータ班及び国立国語研究所で構築されるコーパスを指し，「教科書コーパス」は，言語政策班で構築するものである。

これらのコーパスと語彙表・漢字表を活用して，国語施策・国語教育の分野で新しい施策や指導法を開発するのに役立つ基礎資料を作成し，提供することを目指す。

２．平成 18 年度の研究経過

２．１　班会議の開催

班会議を５回開催した。各回の日時，場所，主なテーマと発題者などは次の通りである。

第１回　平成 18 年 8 月 17 日　国立国語研究所
- 言語政策班の研究計画（１）

第２回　平成 18 年 9 月 10 日　国立国語研究所
- 言語政策班の研究計画（２）

† mtanaka@kokken.go.jp

第３回　平成 18 年 10 月 9 日　国立国語研究所
- 教科書コーパスの設計（田中牧郎）
- 語彙表・漢字表の設計（田中牧郎）
- 類縁語想起に関する学生の実態（棚橋尚子）

第４回　平成 18 年 11 月 23 日　桜美林大学新宿キャンパス
- 教科書コーパス作成の進捗状況について（田中牧郎・河内昭浩）
- パイロットコーパス・白書コーパスを用いた語彙表の作成と分析（近藤明日子）
- 語彙・文法の成績と他領域との関係について（鈴木一史）
- 学生の類縁語想起の実態と語彙資料作成への課題（棚橋尚子）
- 教え合いに基づく作文作成支援システムの構築（山口昌也）

第５回　平成 18 年 12 月 25 日　桜美林大学新宿キャンパス
- 漢字政策へのコーパスの活用（相澤正夫）
- 大学生の語彙習得支援とコーパス活用の可能性（野村敏夫）

班会議では，各研究課題における視点や方法の探索と，共通基盤としての教科書コーパスや語彙表・漢字表の設計について検討討議した。以下，その内容を具体的に報告する。

２．２　各研究課題における視点・方法の探索

a) 難解用語の抽出と言い換え

国立国語研究所の「外来語言い換え提案」（国立国語研究所 2006a）を行う際に，白書・新聞のコーパスから分かりにくい外来語の候補を抽出した経験を踏まえ，コーパスから難解用語を抽出する明示的な手順を開発したい。コーパスの各サンプルに，ジャンル，媒体，読者などの層情報を与えておき，単語の出現が各層にどのように偏るのかを数値化し，難解用語の範囲を定めていく。当面は事例研究として，医療分野における難解用語の抽出方法を研究する。コーパスから得られる難解用語の範囲と，医療従事者や国民一般の判断による難解用語の範囲とが，どの程度一致し，どの程度乖離するかについても考察したい。

b) 常用漢字表･人名用漢字等の在り方に関する調査研究

文化審議会国語分科会では，現在，常用漢字表の見直しについての審議が進められている。この審議に役立てるために，国立国語研究所は，雑誌の語彙調査のデータ（国立国語研究所 2005a）に基づいた，漢字音訓表（国立国語研究所 2005b），表記一覧（国立国語研究所 2006b）を作成した。この実態調査では，表外字（「狙う」など），表外音訓（「育（はぐく）む」），など常用漢字表の範囲を逸脱した漢字使用が広く行われていることや，交ぜ書き（「ら旋」など）のような，常用漢字表に従うことによる不自然と思われる例の存在が，種々明らかになっている。一方，常用漢字表には，国民一般が社会生活を営む際に読み書きできることが望まれる範囲という性格もあるので，国民の漢字能力についての考察も必要である。コーパスによるデータと，国民に対する調査データとの突き合わせが望まれる。

c) 文章作成における語彙選択指導

語彙選択指導にコーパスを活用するためには，①指導にあたり効果的に活用できる，コーパスに基づいた語彙表や文例集の在り方を追求し，これを作成すること，②語彙表や文例集（電子媒体によるものを含む）を用いて語彙選択指導を行うための方法を開発すること，が有効だと考えられる。平成 18 年度は，上記①・②を念頭に置きつつ，マニュアルを用いた大学生の手紙文作成指導を行い，使用語彙における資料参照の影響等について考察した。同一クラスで自由形式の手紙とマニュアルに則った手紙を書かせたところ，後者で

は，伝統的な語句・表現の採用増加や，不慣れな語句の使用による誤用発生などが確認できた。語句の用例活用等を通じ，“ 使いこなせる語彙 ” とする方策を追求する必要がある。上記①の実現のため，学習者の語彙力の実態把握も当面の課題となっている。

d) 概念体系構築のための語彙指導

学習者の概念を広げるための語彙とその指導のあり方を検討している。基礎的調査の段階で，学習者の語彙・文法事項は他領域との相関が高い。この結果を踏まえた語彙指導として，検索ソフト「ひまわり」を使って，ネット上にある「青空文庫」の検索の授業を行った。学習者は他の言語使用状況を自ら検索することで，自己の言語運用能力の伸張を図ることができると考えている。学習者を対象とした調査として，語彙と読解，語彙と他領域の因果関係の調査を予定している。また，今後の方向性として，言語コーパスを基礎とした語彙指導のあり方，学習者の語彙能力と読解力や概念形成との関係性の検証，またそのために教科書コーパスとの比較による学習語彙の選定を目標とする。

e) 国語力向上のための漢字指導

『小学校学習指導要領解説国語編』（1999）では，学年別漢字配当表に配当された漢字について「書きの方が習得に時間がかかるという実態を考慮し」，「次の学年までに定着を図る」としている。2003 年に実施された総合初等教育研究所の調査（総合初等教育研究所 2005）では，配当学年の 1 学年上の書きの定着状況がよいことが導かれ，この方針が一定の効果をあげていることが証明された。しかしながら，この調査で書きの正答率が低かった漢字について大学生 140 名に解答させたところ，正答率が 50 パーセントを切る漢字が少なからず存在した。このことから理解できることは，日常的に使用しない漢字語彙（例えば，「円い」や「読本」など）は成人後も定着しにくいということであり，今後コーパスを活用した指導法の開発を考えていく余地があるということである。また，「思いの外」を「思いの他」と書く解答者が多く存在し，一般社会で使用する漢字と常用漢字との「ずれ」について検討する必要があることについても看取できた。

２．３　教科書コーパスの設計

文部科学省の検定済み教科書（小中高等学校用）を対象として，次の二種類の教科書コーパスを構築する計画である。

(1) 全教科コーパス

2005 年度使用の全教科の教科書。各教科・各学年において，もっともよく使用されている教科書１種ずつについて，その全文を対象とする。500～600 万語程度の見込み。社会に出るまでに学んでいることが期待される，現段階での国民の実質的なスタンダードを語彙の面から把握する基礎資料として位置付ける。

(2) 国語科コーパス

1976 年～2005 年度使用の国語科の教科書。各学年においてもっともよく使用されている教科書２～３種ずつについて，その全文を対象とする。600～700 万語程度の見込み。言語教育の科目としての国語科を特に重視し，歴史的な視点でもデータを分析できるようにする。

タグの仕様や形態論情報の付与方法などは，「現代日本語書き言葉均衡コーパス」とほぼ同様であるが，全文コーパス，教科書という媒体の特性を引き出しやすいように，一部を

変更する。なお，教科書コーパスの一部は，「現代日本語書き言葉均衡コーパス」に組み込む予定であり，相互の対応を取りやすい設計にする。現在，「全教科コーパス」の作成作業に着手したところである。

２．４　語彙表・漢字表の設計

言語政策班の基盤データの一つとして，「現代日本語書き言葉均衡コーパス」と「教科書コーパス」の汎用語彙表を作成する。この汎用語彙表は，各コーパスに形態論情報を付与したものをＳＱＬデータベースとし，語形・表記・品詞・語種など語彙情報と，ジャンル・媒体・著者・読者などサンプル情報の双方から，検索や抽出ができるものとする。汎用語彙表のなかに，漢字情報も格納し，漢字表としても活用できるように工夫する。

国語施策，国語教育の具体的な実践面で必要となる種々の語彙表を，汎用語彙表をもとに，目的に応じて作成していく。作成する語彙表の一例をあげると，課題ａ）について「国民が見聞きする機会のある難解な医療用語リスト」，課題ｂ）について「高校卒業時までに学ぶ必要のある漢字リスト」などである。

現在，データ班で作成が進んでいる白書のデータをもとに，語彙表・漢字表の作成を試行しながら，設計を進めているところである。

３．平成 19 年度以降の計画

3．1　平成 19 年度

平成 19 年度は，共通基盤としての教科書コーパス作成と汎用語彙表の作成に，資源を集中的に投入し早期の整備を進め，平成 19 年度末までには利用可能にする。各研究課題で探索してきた視点に基づいて，コーパスデータを生かした具体的な分析に順次着手する。

また，これまでの検討で浮上してきた，コーパスから得られるデータと対照する，国民や児童生徒の語彙や漢字の能力についてのデータの取得方法について，集中的に議論する。

3．2　平成 20 年度以降

各研究課題について具体的な展開を図る。国民や児童生徒の語彙や漢字の能力について調査を行った場合，その調査データとコーパスによるデータを突き合わせた研究も進める。

文　献

国立国語研究所(2005a)『現代雑誌の語彙調査―1994 年発行 70 誌―』国立国語研究所報告 121,pp.1-721

国立国語研究所(2005b)『「現代雑誌の語彙調査」に基づく漢字音訓一覧表』国立国語研究所,pp1-211

国立国語研究所(2006a)『分かりやすく伝える 外来語言い換え手引き』ぎょうせい,pp.1-275

国立国語研究所(2006b)『「現代雑誌の語彙調査」に基づく表記一覧』国立国語研究所,pp1-555

総合初等教育研究所（2005）『教育漢字の読み・書きの習得に関する調査と研究』総合初等教育研究所,pp1-195

文部省（1999）『小学校学習指導要領解説　国語編』東洋館,pp1-189

平成 18 年度研究進捗状況報告：辞書編集班
（コーパスを利用した国語辞典編集法の研究）

荻野綱男（班　長：日本大学文理学部）
近藤泰弘（分担者：青山学院大学文学部）
矢澤真人（分担者：筑波大学大学院人文社会科学研究科）
丸山直子（分担者：東京女子大学文理学部）

Progress Report of the Year 2006: 'Dictionary compilation' Group

Tsunao Ogino (College of Hum. and Sci., Nihon University)
Yasuhiro Kondo (College of Literature, Aoyama Gakuin University)
Makoto Yazawa (Graduate School of Hum. and Soc. Sci., Tsukuba University)
Naoko Maruyama (College of Arts and Sci., Tokyo Woman's Christian University)

１．辞書編集班の研究目的

辞書編集班は、全体として、コーパスを用いた辞書編集の方法を研究する。

既存のコーパスを利用してこのような課題を追求する一方、特定領域研究の進行とともに利用に供される新しいコーパスによる用例分析を行い、新開発コーパスの特性を明らかにするとともに、それが辞書編集にどのように役立つかを明らかにする。

国語辞書のどういう部分の記述にコーパスを活かすかという点では、主な分担課題を四つ設定して、それぞれを分担者が研究することにする。

２．辞書編集班の全体としての活動

辞書編集班では、分担者ごとに四つの小グループに分かれて、それぞれの研究を進めるとともに、班全体として研究進捗状況を確認しあい、今後の研究方向を考えるための会議を開催してきた。

班会議は、2006.8.29 2006.12.11 2007.2.26 の３回開催した。

また、年度末には、辞書編集班の初年度の研究成果をとりまとめた報告書を印刷する予定である。

３．今年度の研究進捗状況

３．１　コロケーション辞書の概念設計と試作（荻野綱男・荻野孝野）

第１グループでは、「コーパスを利用したコロケーション辞書の概念設計と試作」ということで研究を進めている。最終年度までには、大規模コーパスを用いてコロケーション辞書の一部を実際に記述する予定である。今年度は、以下のような３項目を重点的に行った。

[1] コロケーション情報の整理の試み

WWW をコーパスとして利用して、コロケーション情報を人手で抽出し、一部の整理を試みた。

姫野昌子(2004.6)『日本語表現活用辞典』研究社の記述を改善することを当面の目標とする。

整理の過程で、次のような問題点があることがわかった。これらをどう扱うかを考慮しなければ、望ましい辞書記述はむずかしいと思われる。

問題点１：「の」などによる名詞の修飾をどう考えるか

「～をあんじる【案じる】」の場合、「行く末」が多数あらわれるが、「行く末を案じる」だけでは、やや不自然に感じられる。実際の用例では「この国の、恋愛の、教育の」などの修飾語句が「行く末」についており、まったく自然である。ということは、動詞と名詞の

コロケーションの記述の場合、名詞までにとどめると、不十分な記述になることがあるということになる。

問題点２：単に名詞を列挙するだけでなく、シソーラスで表現されるような「カテゴリ」を考えるべきではないか。
「～をうつす【写す】」の場合、「富士山、京都、横浜」などの<場所>、「月、紅葉、雲」などの<自然>、「目次、型紙、板書」などの<物>などがあらわれる。これらは単語を列挙して記述するのでなく、カテゴリで記述する方が（記述の手間が減り、有効性を広げるために）望ましいと考えられる。しかし、カテゴリとして何を設定すればいいかという問題があり、コロケーションとしてカテゴリとともに「単語」も記述する必要があると考えられ、どこまで含めるかという問題が発生する。

問題点３：動詞の意味の分類をどうするか---意味によってコロケーションが違う
「～におちこむ【落ち込む】」の場合、「日本海、落とし穴、淵」のように物理的に落ち込む場合と、「うつ状態、失意のどん底、失恋、悲しく憂鬱な気分」のように精神的に「落ち込んだ気持ちになる」場合とがある。これらをどう区分するかという問題がある。場合によっては、どちらとも判別しがたい（両方で解釈できる）ものがある。

問題点４：複数の名詞との同時共起をどう扱うべきか
「～が～に消える」の場合、「貯金が学費に、国政調査活動費が飲み食いに、資金が経費に」といったように複数名詞との同時共起を記述する必要がある場合がある。しかし、一方では名詞単独の用例もあり、用例からは単独と複数の区別がむずかしい。一見単独のふりをしていて、実は複数が省略されている場合もある。省略の場合、そもそも、用例として現れていないものを加えていいかという研究方法論上の問題もある（用例主義と内省主義の違い）。

問題点５：並立助詞「と」の処理
「～をくぎる【区切る】」の場合、「始まりと終わり、男湯と女湯、日常と非日常」など「と」で結ばれる名詞があらわれることがある。この場合、「と」で結ばれる名詞の範囲は、必ずしも限定的でないし、記述しはじめるときりがない。これが（動詞と）名詞のコロケーションだろうかという疑問もある。

[2] 文献調査（日本語学分野）
主として国語年鑑から 1960-2006 年の日本語学分野の研究文献を調査し、コロケーションや共起関係、結合価などをあつかった文献を抜き出し、どんな研究が行われてきたかを概観した。
約 50 年間でコロケーションに関連する論文は 48 本, 書籍は 9 冊, 辞書は 2 冊があった。
1960 年代までは、見るべき研究はなかった。1970 年代から、文法研究の分野で、いくつかの論文があらわれるが、その数はあまり多くない。仁田義雄や水谷静夫の研究が先駆的なものであり、結合価文法に関する研究が多かった。1980 年代から 1990 年代にかけては、コロケーションの位置づけに関する理論的な研究や、具体的な語を取り上げてコロケーションを記述するような研究が行われるようになった。2000 年代からは実例多数に基づく数量的研究が見られるようになる。

[3] 文献調査（自然言語処理分野）
詳しくは、このシンポジウムで「結合価およびコロケーションに関する文献調査」として発表するので、ここでは省略する。

３．２　日本語複合辞の研究（近藤泰弘・坂野収・多田知子・岡田純子）

[1] 複合辞の定義再考

まず最初に複合辞の定義についての従来の研究を調査した結果、次のようなものが主要な結果として集成された。

（永野賢の定義）
○単なる構成要素のプラス以上の意味をもっていること
○類語（意味の近似した他の助詞や複合助詞）の中にあって、独特の意味やニュアンスを分担していること。
○構成要素の結合が固着していること。

（松木正恵の複合辞の分類）
○[第１種複合辞] 助詞・助動詞のみが二語以上複合したもの
○[第２種複合辞] 実質的意味が稀薄になった形式名詞を中心に複合したもの
○[第３種複合辞] 用言が実質的意味を稀薄にしたもの（形式用言）を中心に複合したもの

（松木正恵の「複合辞化」を計る尺度）
○構成要素の（結びつきの）緊密化の度合い
○形式名詞・形式用言の形式化（＝実質的意味の稀薄化）の度合い
○形式用言の文法範疇（≒ヴォイス・テンス・ムードなどの文法的表現の区別）喪失の度合い

（田野村忠温の分類）
○単純辞・複合辞

（土屋その他の定義）
○複合辞とは、いくつかの後が複後してひとかたまりの形となって非構成的な意味を持ち、辞的な機能を果たす表現である。
○複合辞とは、複数の形態素がひとかたまりとなって、一つの機能語相当として働く表現である。

以上より、本研究での複合辞の仮の定義として次を採用したい。

> 複数の形態素がひとかたまりとなって、非構成的な意味を生じ、ひとつの機能語相当語として働くものである。したがって、体言や用言を中心とする形態では、必ず文法化（機能語化）の現象が見られる。またそれらは次のようなものと関係がある。
> ○[機能語化] 複合辞・イディオム・複合接続詞
> ○[複数の形態素] 複合辞・コロケーション・慣用句・イディオム・文型・複合動詞・複合接続詞
> ○[非構成的な意味] 複合辞・慣用句・イディオム・文型・複合接続詞
> ○[文法化] 複合辞・イディオム・乙類単純辞・複合動詞後項・複合接続詞

また、以上の定義や分布から、複合助詞・助動詞（付属語）だけでなく、複合接続詞（自立語）も研究の対象にするのが妥当であろう。(「・・・うえ（で)」と「そのうえ」との連関を記述しやすい。)「複合辞」として一括するかどうかは今後の課題である。

[2] 複合辞の統計的研究

そして、これらを元に、統計的にコーパスから複合辞候補を抜き出す実験を行った。

まず、コロケーションの研究において、重要な指標として t スコアと MI スコアがある。t スコアは二つの語の間のコロケーションの確からしさを計る指標であり、MI スコアは、二つの語の間のコロケーションの強さを計る指標である。MI スコアは大きさの異なる複数のコーパスを資料として別々に採取されたものを比較することも有効であるが、t スコアは別の大きさのコーパスから取られたものを比較することはできない。MI スコアは低頻度でも興味深いコロケーションを見つけることができるが、低頻度では値が異常に高くなるためその部分での信頼性には問題もある。

本研究では３つの付属語形態素が連続するようなものについて、その t スコアと MI スコアを計算した。その結果、t スコア順は、単純頻度順に似るが文法的複合辞らしいものを抽出する指標としては、MI スコアよりも好ましいことが明らかになった。一方、MI スコアは文末の特徴的表現を抽出できることがわかった。

これ以上のグラム数（形態素数）の場合にはいろいろな技法が開発されているようであるので、次年度以降、それらによって複合辞の統計的な性質を利用したコーパスからの自動抽出を試みたい。

３．３　コーパスによる語義分析と辞書での記述方法（矢澤真人・橋本修）

[1] 本年の活動

コーパスを用いた国語辞典構築を最終目標として，今年度はその前段階として，国語辞典のブランチ（見出し語のもとに立てられた①や②などの分類）の立て方について検討を加えた。具体的には，以下の研究活動を行った。

[2] 国語辞典の障碍に関する研究

国語辞典を利用する際に障碍となる点について，一つ一つ検討を加え，求めるのがどのブランチの意味であるかを選択する「ブランチ選択の障碍」が電子型国語辞典の最大の障碍になることを示し，それを解消させるためにブランチの立項を工夫するとともに，コロケーション情報を利用した連語引きや，素性分析をふまえたブランチ立てが有効であることを論じた。

[3] 形容詞意味記述の解釈の揺れ

既存の国語辞典で立てられたブランチがどの程度適切であるかを図るため，以下のように実態調査を行っている（継続中）。

[3-1] 調査方法

大学生に，新聞から抽出した多義形容詞の実例が明鏡国語辞典・岩波国語辞典のブランチのどれに当てはまるかを記入させる。対象とする形容詞は２０語（明鏡・岩波ともに同じ見出し語で３ブランチとなり，同音異表記の形容詞が見出し語にたっていないもの），１語につき４人が同一の４００例に対して回答したものを集計して，どの程度一致するか，どのブランチで解釈の揺れが生じるかを調べる。

[3-2] 調査の進展状況

明鏡国語辞典について，形容詞１４語（２人回答１１語／３人回答３語）について集計済み。明鏡国語辞典の残り（上記１４形容詞のうち１１語の２人分の回答，３語の１人分の回答，および６形容詞の４人分の回答）の集計と，岩波国語辞典第５版についての調査・集計は，現在継続中。

[3-3] 調査の中間報告

「さびしい」（二者平均 53%；三者一致 40%）と「濃い」（二者 46%）をのぞき，その他の１２形容詞は，ほぼ 80%前後の一致率であった。個人による判定の差は見られたが，特定個人による影響はあまり見られなかった。

また，ブランチ数の多寡による差もほとんどなく，ブランチ数９の「浅い」の二者一致率が 80%，ブランチ数６の「近い」の二者一致率 94%，ブランチ数３の「さびしい」の二者平均一致率が 50%と，ブランチの多さが判定の揺れの原因ではないことが知られた。

[3-3-1] 「さびしい」

寂しい	A	B	C	平均
①	23%	15%	21%	20%
②	32%	40%	36%	36%
③	45%	14%	32%	30%
×	0%	31%	10%	14%
合計	100%	100%	100%	100%

淡い	D	E	F	平均
①	62%	61%	61%	61%
②	13%	12%	16%	14%
③	19%	27%	23%	23%
×	5%	0%	0%	2%
合計	100%	100%	100%	100%

（×は該当ブランチなしとの回答）

「淡い」は調査者３人がほぼ同じような選択をしているのに対し，「さびしい」は個人の認定差がかなり大きく，特に A は③と×，B は②と×の判定が特異であった。

ABC の三者が全く異なった判定をしたものも見られたが，ほとんどが，①と③と×の判定の組み合わせであった。属性形容詞（①）と感情形容詞（③）のどちらと判定するかの揺れと見られる。

1) **「日本開催なのに，町中でW杯の文字がほとんどなくて寂しい」と知恵を絞った。**（A③ B× C①）
2) **いま，テレビの時代劇枠が少なくなっているのは寂しい限り。**（A③ B× C①）

さびしい（明鏡語釈）
①人やものが少なくて，にぎわいを感じさせないさま。
②寄り添うものがあってほしいのに，それがなくて孤独な気持ちである。
③あるべきものがなくて，物足りない気持ちである。

[3-3-2] 「濃い」

「濃い」の二者間の対応

	H①	②	③	④	⑤	⑥	⑦	×	
G①	69	0	4	0	0	0	0	0	73
②	4	27	8	0	0	1	0	0	40
③	0	0	21	0	0	0	0	0	21
④	0	89	3	56	0	9	1	0	158
⑤	0	1	2	0	0	0	0	0	3
⑥	1	8	2	0	0	8	0	0	19
⑦	0	0	1	0	0	0	1	0	2
×	4	5	64	1	0	3	6	0	83
	78	130	105	57	0	21	8	0	399

「濃い」の判定の揺れは，G④：H②，G×：H③に集中している。
④「そのような傾向や可能性が高い。濃厚だ。強い。「どちらかと言えば洋犬の血が―」
②「成分の濃度が濃い。特に味の刺激が強い。濃厚だ。「塩分が―」「酸味が―」

３．４　辞書記述のためのコーパス利用（丸山直子・星野和子）

[１] 辞書班第４グループの研究目的

コーパスを辞書記述に役立てる方法や、辞書記述に役に立つコーパスの性質の検討を行う。

語義記述（語釈）、例文掲載、その他（類義語・反義語の記述ほか）辞書記述に必要な項目を洗い出し、実際にコーパスをもとに辞書記述を行うことで、どの項目にコーパスを利用することができるかを検討する。さらに、辞書記述に役に立つコーパスとはどういうものかについても検討する。

中でも、特に、辞書において重要な役割を果たす例文を充実させるために、動詞及び名詞について、格情報をコーパスから抽出する方法について検討する。

[2] 今年度の活動のまとめ

[2-1] 辞書に必要な情報の調査として文献調査および辞書編纂者への聞き取りを行った。辞書編纂者から聞き取った結果は、報告書にまとめた。語の選定法、釈義のあり方（意味とは指し方であること、メタ釈義、意味区分等）、用例の挙げ方、注記について、編集態度、改修の仕方等、多岐にわたる内容となった。今後、これらの調査結果をもとに、辞書記述に必要な項目は何かを検討する。

[2-2] コーパスの収集

既存のコーパスからどのような情報が抽出できるかの調査を行うため、新聞ＣＤの購入、ネット上で獲得できるデータの開拓を行った。また、05 年度に採取した雑談の話しことばコーパスの整備を行った。

[2-3] ト格要素をとる体言の用例収集と分析を行った。名詞について、取り得る格情報を既存のコーパスより抽出し、辞書記述に生かすことが目的である。これまで動詞については格情報が重視されてきたし、名詞が述語になる場合も、ある程度、取り得る格成分について記述されてきたが、述語でない名詞が文の格成分に影響を与えることについては、体系的な記述がない。そのあたりを辞書の例文の記述に反映させるにはどうすればよいか検討するために、コーパス情報を使用する。

[2-4] 動詞の格情報とその記述法を、役割を表すニ格を中心に考察した。動詞の格情報について明示的に記述している国語辞典（『日本語新辞典』（小学館）と『新明解国語辞典』（三省堂））の格情報を、『日本語新辞典』が基本語としている動詞 576 語を対象に比較し、従来の格の研究（ＩＰＡＬ、『日本語における表層格と深層格の対応関係』（国立国語研究所報告 113）、『基本動詞用法辞典』（大修館書店）、『現代言語理論と格』（石綿敏雄，ひつじ書房）等）と照らし合わせた。既存のコーパスを利用して、用例を収集。今後、類語間で統一のとれた記述法を模索する予定である。

平成 18 年度進捗状況報告：言語処理班 （代表性のあるコーパスを利用した日本語意味解析）

奥村 学 (班長: 東京工業大学)[1]
白井 清昭 (分担者: 北陸先端科学技術大学院大学)
竹内 孔一 (分担者: 岡山大学)
中村 誠 (分担者: 北陸先端科学技術大学院大学)
杉山一成 (協力者: 東京工業大学)

Progress Report of the Year 2006: 'Natural Language Processing' Group

Manabu Okumura (Tokyo Institute of Technology)
Kiyoaki Shirai (Japan Advanced Institute of Science and Technology)
Koichi Takeuchi (Okayama University)
Makoto Nakamura (Japan Advanced Institute of Science and Technology)
Kazunari Sugiyama (Tokyo Institute of Technology)

1 研究目的

日本語を対象にした言語処理研究では，形態素解析，構文解析について研究が進み，高精度なツールの開発も行われてきており，それらのツールが日本語学，日本語教育など他の研究分野でも広く利用されるようになってきている．その一方で，意味解析については依然研究が遅れており，一般に利用可能なツールの開発レベルにまで解析精度が到達していない．また，代表性のあるコーパスを用いた言語処理研究は，これまでそのようなコーパスが存在しなかったため，日本語に関してはまったく行われてこなかったと言って良い．そこで本研究課題では，研究項目 A で構築する代表性のあるコーパスを用いた実証研究を行う．具体的には，以下の 3 つを柱とした日本語意味解析手法の開発を行う．

1. 機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築
   タグ付コーパスから学習した多義性解消システムによりタグ付コーパス作成コストの軽減を図るとともに，作成されたコーパスを用いて bootstrap 的に多義性解消システムの性能向上を図る．

2. 単語の新語義，新用法の自動発見手法の開発
   時を経るにしたがって単語の意味は変化し，新しい意味が生まれることが知られている．今回構築されるような，時間幅を伴うコーパスで顕著に見られるこの言語現象を自動的に発見する手法を開発する．1) で開発する多義性解消手法で特定できない語義は新語義と考えられるため，2) は 1) のシステムの自然な拡張と言える．

3. 語彙概念構造に基づく動詞の意味構造の自動構築法の開発と語彙概念付与システムの開発
   語彙概念構造は動詞の振る舞いに関する分析から動詞の意味をそれが取る名詞同士の意味関係で記述する言語学に基づく意味構造である．文の意味構造は，1) で特定される単語の語義と 3) で抽出される意味構造の統合により得ることができる．

本研究課題で開発する手法は，領域内の少なくとも，コーパス日本語学研究，日本語教育研究，辞書編纂研究に寄与することを想定している．たとえば，代表性のあるコーパスにおける単語の語義の頻度分布の情報は，日本語教育において，どの単語を教育上用いるべきか，どの語義 (意味) につい

[1]oku@pi.titech.ac.jp

て教えるべきか等，客観的な教材作成における単語，意味の選択において重要な役割を果たすと言える．また，単語の語義ごとの代表的な用例集が半自動作成できれば，教材作成の有用な基礎データとなる．さらに，単語の新しい意味が自動的に発見できれば，辞書編纂作業を強力にサポートすることができる．

これまでの言語処理研究は，新聞コーパスを主に対象としてきた．その結果，種々のツールも新聞に依存したものになりがちだった．本研究課題では，代表性のある，様々なジャンルのコーパスを元にすることで，特定のジャンルに依存しない汎用的な手法の開発を実現できる可能性がある．さらに，今回構築されるコーパスの時間幅 (10〜20 年程度) の中で，時間とともに生じる単語の新しい語義を自動的に発見する手法を開発するが，このような試みは過去にはほとんど行われていない．

日本語の語義タグ付コーパスには，EDR コーパス (20 万文)，RWC コーパス (3000 記事) があるが，いずれも代表性のあるコーパスを元にしていない．海外では，代表性のあるコーパスの上にタグ付けを行うことで，代表性のある語義タグ付コーパスの構築が進んでおり，日本語においての構築は急務であると考える．また，語彙概念構造に基づく辞書構築は，文系の言語学の知識と理系の情報処理の知識の両方が必要なことから，日本語では立ち遅れている．英語では，延べ約 1 万語の動詞に対して開発されており，さらに深層の格を付与したコーパス (Propbank) まで構築されている．このため本研究課題では，日本語において大規模な辞書を自動構築する手法の開発を目指す．

2　機械学習手法を用いた単語の語義同定

東京工業大学の研究グループでは，「機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築」を目的とし，今年度は，機械学習手法を用いて単語の語義同定を行う手法の検討を行うとともに，ベースラインとなるシステムを開発した．

本研究の問題設定は，白井 (白井, 2003) の報告にあるのと同じであり，テキスト中の単語の各出現について，辞書中のその単語の語義の区分に基づき，語義を同定するものである．我々も，元となる辞書として，岩波国語辞典を用いることにした．

白井 (白井, 2003) の報告にもある通り，近年の語義曖昧性解消研究では，訓練コーパスを用いて分類器を学習し，その分類器により語義を同定する (ある単語の出現がその単語の語義のうちどの語義の出現であるか分類する) 手法が採用されることが多く，また，より良い性能を得られている．そこで，我々も，機械学習に基づく手法を採用することとし，今年度は，比較的良い性能が得られると報告がある，以下の 3 つの手法を用いることで，ベースラインシステムを開発した．

1. Support Vector Machines(SVM),
2. Naive Bayes(NB),
3. Maximum Entropy(ME)

NB を用いた語義曖昧性解消手法については，3.1.1 節で詳しく述べる．SVM, ME については，それぞれ (Kudo, 2000; Ratnaparkhi, 1996) を参照して頂きたい．なお，SVM は本来 2 値分類器である．語義が 3 つ以上の単語では，多値分類問題を扱うことになるが，その場合，‘one versus rest’ 法を用いて SVM を拡張することで実現している．

2.1　実験

訓練用コーパスとしては，「SENSEVAL-2 日本語辞書タスク」で訓練用コーパスとして配布されたのと同じデータ (RWC コーパス) を用いた．毎日新聞の 1994 年の 3,000 記事に人手で語義タグ付けが行なわれたコーパスである．品詞が名詞，動詞，形容詞のいずれかであり，岩波国語辞典に見出しがある，多義の単語，総数で 148,558 語に語義タグが付与されている．

評価用データも，「SENSEVAL-2 日本語辞書タスク」で用いられた 100 単語 (名詞，動詞がそれぞれ 50 単語) をそのまま用いており，実験は，10 分割交差検定により行なった．評価尺度としては，精度を用いるが，fine-grained scoring(語義の完全一致)(白井, 2003) を評価基準としている．

表 1: 3 つの機械学習手法による精度

学習法	精度
SVM	0.761
Naive Bayes	**0.794**
Maximum Entropy	0.773

表 2: 語義数ごとの平均精度 (SVM)

語義数	精度
1 (4 単語)	1.000
2 (40 単語)	0.826
3 (18 単語)	0.802
4 (21 単語)	0.706
5 以上 (17 単語)	0.575

表 3: 語義数ごとの平均精度 (Naive Bayes)

語義数	精度
1 (4 単語)	1.000
2 (40 単語)	0.840
3 (18 単語)	0.830
4 (21 単語)	0.748
5 以上 (17 単語)	0.656

2.2 用いる素性

解析に用いる素性としては, 以下に上げるものを用いた. 今年度は, ベースラインシステムの開発であることから, 先行研究で用いられている代表的なものを用いている.

- 形態素素性
  対象単語の周辺を Chasen で形態素解析し, その結果を素性として利用する.
  - 対象単語および, 前後 2 語までの単語の表記,
  - 対象単語および, 前後 2 語までの単語の品詞, 品詞細分類
- 構文素性
  対象単語の周辺を Cabocha で係り受け解析し, その結果を素性として利用する.
  - 対象単語が名詞の場合，その名詞が係る動詞,
  - 対象単語が動詞の場合，その動詞のヲ格の格要素

2.3 実験結果

3 つの機械学習手法を用いた多義性解消ベースラインシステムの実験結果を表 1 に示す. また，3 つの手法による, 語義数ごとの平均精度を表 2〜4 に示す.

2.4 今後の予定

機械学習手法を用いたベースラインシステムが完成したので, 今後はこのシステムをベースにした研究を進めていく予定である.

表 4: 語義数ごとの平均精度 (Maximum Entropy)

語義数	精度
1 (4 単語)	1.000
2 (40 単語)	0.858
3 (18 単語)	0.803
4 (21 単語)	0.731
5 以上 (17 単語)	0.543

1. 有効と考えられる素性を追加することによる解析精度の向上

2. 代表性のあるコーパスを用いた手法への拡張

代表性のあるコーパス中には, 複数のジャンルのテキストが混在していることになる．したがって, コーパスは, いくつかのジャンルごとのサブコーパスに分割できることになる．そのため, コーパス全体で学習したモデルと, サブコーパスで学習したモデルを併用し, 対象とするテキストのジャンルごとにモデルを使い分ける多義性解消手法が可能となる．今後は, このように, 代表性のあるコーパスを利用することの特徴を活かした多義性解消手法を開発していく予定である．

3 未定義語義の判別

北陸先端科学技術大学院大学の研究グループでは未定義語義の判別に関する研究を行った．未定義語義とは，ここではあらかじめ辞書などに定義されていない単語の意味を指す．テキスト中の単語に対し，その単語の意味が定義された語義のいずれかであるか，あるいはそれ以外の未定義語義であるかを判別する手法の開発に取り組んだ．従来の語義曖昧性解消 (Word Sense Disambiguation;WSD) があらかじめ定義された語義の中から該当する語義を選択するのに対し，本研究はその単語が辞書などで定義されていない意味で使われているか否かを含めて語義の判定を行う点が異なる．未定義語義の判定は，様々な自然言語処理アプリケーションに有用であるだけでなく，辞書編纂作業のサポートへの応用も期待できる．

3.1 提案モデル

まず，本研究における問題設定について述べる．本研究の目的は，ある文に出現する対象単語の語義を決定することにある．単語の語義は，辞書などで定義されている n 個の語義 $s_1, \cdots, s_n$ (以下, 既定義語義と呼ぶ)，ならびに未定義語義 s_{n+1} のいずれかとする．そして，これら $n+1$ 個の語義の中から，与えられた単語の意味として適切なものを 1 つ選択する．本研究では以下の 2 つの手法を提案する．

3.1.1 モデル 1

WSD のためのモデルとして Naive Bayes モデルを用いる．すなわち，式 (1) の確率モデルをコーパスから学習する．

$$P(s_k) \prod_{f_i \in c_j} P(f_i|s_k) \tag{1}$$

c_j は対象語 (語義を決めたい単語) を含む文を，f_i は c_j に含まれる素性を表わす．素性とは文中から得られる WSD にとって有効と思われる情報である．ここでは，対象語の前後に現われる単語，その品詞，文脈内の自立語など，既存の語義曖昧性解消でよく用いられる素性を用いた．$P(s_k)$ は語義 s_k の出現確率，$P(f_i|s_k)$ はある文が語義 s_k を持つ単語を含むとき，その文中に素性 f_i が生起する確率である．語義の曖昧性を解消する際には，未定義語義を含む全ての語義 s_k について式 (1) を推定し，それが最も大きい語義を正しい語義として選択する．

式 (1) 中のパラメタ $P(s_k)$, $P(f_i|s_k)$ は教師なし学習のひとつである EM アルゴリズムによって学習する．ここでは，基本的には Manning らによって提案されたアルゴリズム (Manning and Schütze, 1999) を用いる．彼らの手法では，まずパラメタの初期値をランダムに決定する．次に，平文テキストを学習データとし，与えられたパラメタから $P(s_k|c_j)$ (ある文脈 c_j 中の対象単語の語義が s_k である確率) を推定する E-step と，推定された $P(s_k|c_j)$ からパラメタを再推定する M-step を交互に繰り返し，最終的なパラメタ $P(s_k)$, $P(f_i|s_k)$ を得る．

EM アルゴリズムは，初期値の与え方によって推定されるパラメタが大きく異なるという問題点がある．このため，初期パラメタをランダムに設定するのではなく，語義タグ付きコーパスから得られる統計情報をもとに，より信頼性の高い初期パラメタを設定することを試みた．

まず，語義の生起確率 $P(s_k)$ の初期パラメタを式 (2), (3) のように設定する．

$$P^0(s_k) = (1-I_u)\frac{O(s_k)}{\sum_{k=1}^{n} O(s_k)} \quad (1 \leq k \leq n) \tag{2}$$

$$P^0(s_{n+1}) = I_u \tag{3}$$

I_u は未定義語義 s_{n+1} に与えるべき初期パラメタで，定数とする．本研究では $I_u = 0.1$ とした．一方，既定義語義 s_k については語義タグ付きコーパスから最尤推定する．式 (2) 中の $O(s_k)$ はコーパスにおける s_k の出現回数である．なお，式 (2) で $(1-I_u)$ という項をかけているのは $\sum_{k=1}^{n+1} P(s_k) = 1$ という制約を満たすためである．

一方，$P(f_i|s_k)$ の初期パラメタは以下のように求める．まず，既定義語義 s_k については，$P(f_i|s_k)$ の大きい上位 n 個の素性集合 F_{s_k} と，それ以外の素性集合 $\overline{F_{s_k}}$ に分ける．ここで $P(f_i|s_k)$ は語義タグ付きコーパスから最尤推定で求める．そして，F_{s_k} 中の素性に対する初期パラメタ $P^0(f_i|s_k)$ の値を $\overline{F_{s_k}}$ 中の素性に対する初期パラメタの r 倍とする．また，F_{s_k} 中，$\overline{F_{s_k}}$ 中の各素性については，$P^0(f_i|s_k)$ の値は全て等しく設定する．$\sum_{f_i} P(f_i|s_k) = 1$ という制約から，具体的な設定式は (4) のようになる．

$$\begin{cases} P^0(f_i|s_k) = \dfrac{r}{|F| + (r-1)n} & f_i \in F_{s_k} \\ P^0(f_i|s_k) = \dfrac{1}{|F| + (r-1)n} & f_i \in \overline{F_{s_k}} \end{cases} \tag{4}$$

一方，未定義語義 s_{n+1} については，$P(f_i|s_k)$ の大きい上位 n' 個の素性集合 F' と，それ以外の素性集合 $\overline{F'}$ に分ける．先ほどと異なるのは，s_k 毎に上位の $P(f_i|s_k)$ を選択するのではなく，全ての f_i,s_k の組について $P(f_i|s_k)$ を比較し，その上位 n' 個の素性を異なりで選別するということである．そして，F' 中の素性に対する初期パラメタ $P^0(f_i|s_k)$ の値を $\overline{F'}$ 中の素性に対する初期パラメタの $1/r'$ 倍とする．F' 中の素性に対する初期パラメタを低く設定したのは，F' は既定義語義とよく共起する素性の集合であり，未定義語義とはあまり共起しないと考えたためである．具体的な設定式を (5) に示す．

$$\begin{cases} P^0(f_i|s_{n+1}) = \dfrac{1}{|F| + (r'-1)n'} & f_i \in F' \\ P^0(f_i|s_{n+1}) = \dfrac{r'}{|F| + (r'-1)n'} & f_i \in \overline{F'} \end{cases} \tag{5}$$

n,r,n',r' は語義タグ付きコーパスから得られる統計情報をどれだけ初期パラメタの設定に反映させるかを調整する役割を持つことに注意していただきたい．n または n' が 0 のとき，あるいは r または r' が 1 のときは，語義タグ付きコーパスを使わずに初期パラメタを全て一様分布に設定することに相当する．一方，n,n' や r,r' の値を増やせば増やすほど，語義タグ付きコーパスから得られる統計情報を信頼し，初期パラメタの設定に反映させることになる．

3.1.2 モデル 2

以下の 2 段階で語義の判別を行う．

Step 1. 既定義語義か未定義語義かの判定

対象単語の語義が，既定義語義 ($s_1 \sim s_n$ のいずれか) か未定義語義 (s_{n+1}) のどちらであるかの二値判定を行う．この二値判定を行うモデルは，モデル 1 と同様に EM アルゴリズムによって学習する．すなわち，初期パラメタを式 (2)〜(5) のように設定し，E-step と M-step を繰り返して確率モデルの最終的なパラメタ $P(s_k)$ と $P(f_i|s_k)$ を推定する．モデル 1 と異なるのは，語義の数が既定義語義と未定義語義の 2 つである点のみである．

表 5: 実験結果

	R_u	P_u	F_u	A_a
モデル 1	0.16	0.82	0.26	0.50
モデル 2	0.46	0.33	0.38	0.49
最尤推定 NB	–	–	–	0.53

Step 2. 既定義語義の判別

Step 1. で既定義語義と判定された場合，その語義が s_1 から s_n のいずれであるかを判定する．これは通常の語義曖昧性解消の問題設定と全く同一である．したがって，既存の教師あり学習に基づく語義曖昧性解消をそのまま適用することができ，比較的高い精度で語義を判別できることが期待される．本研究では，既定義語義の判別に語義タグ付きコーパスから最尤推定した Naive Bayes モデルを用いる．

3.2 実験

提案手法の評価実験を行った．語義を決定する対象単語として以下の 10 語を用いた．

> 気持ち (3)，教える (2)，決める (5)，情報 (3)，朝 (2)，世紀 (2)，電話 (2)，非 (3)，
> 与える (3)，条件 (2)

括弧内の数値は EDR 概念辞書で定義されている語義の数である．これらの対象単語は，予備調査によって未定義語義で使われる可能性の高い語を調べて選んだ．

EM アルゴリズムによる学習に用いる正解なしコーパスとして毎日新聞の 12 年分のコーパスを，既定義語義が付与されたコーパスとして EDR コーパスを利用した．テストデータとして，学習に用いなかった毎日新聞の記事の中から対象単語毎に 100 語選択し，正解語義を人手で付与した．のべ 1000 語のうち未定義語義を割り当てたのは 258 語である．未定義語義の例としては,「電話」という単語が持つ (電話番号) という意味や,「朝」という単語が持つ (北朝鮮) という意味などがある．

実験結果を表 5 に示す．R_u，P_u，F_u は未定義語義判別の再現率，精度 (適合率)，F 値である．一方，A_a は既定義語義，未定義語義全てに対する語義曖昧性解消の正解率である.「モデル 1」「モデル 2」はそれぞれ 3.1.1，3.1.2 節で述べた語義判別手法を表わす．一方,「最尤推定 NB」は語義タグ付きコーパスから最尤推定によって学習された Naive Bayes モデルを用いる手法を表わす．この手法は常に既定義語義のいずれかを選択し，未定義語義は出力しない．

表 5 から，未定義語義判別の F 値 (F_u) は，既定義語義と未定義語義を同時に判別するモデル 1 よりも，まず未定義語義を判別し，次に既定義語義を判別するモデル 2 の方が高いことがわかった．とはいえ，両者の F 値は決っして高くない．また，全体の正解率 A_a は，未定義語義の存在を考慮しない手法 (最尤推定 NB) が最も高かった．このことは，提案手法による未定義語義の判別が必ずしも有効ではなく，改善の余地があることを示唆している．

今後，未定義語義の検出の正解率を向上させるために，語義の判別に用いた素性の有効性を検証することを検討している．すなわち，既定義語義とよく共起する素性と未定義語義とよく共起する素性を比べ，それらに差異があるかどうかを調べる．両者に差異がある素性を発見し，それらを確率モデルに反映させれば，より正確に未定義語義を検出できるようになると考えられる．

4 LCS 辞書の評価

岡山大学の研究グループでは日本語コーパスから語彙概念構造辞書 (Lexical Conceptual Structure, 以下 LCS) を半自動で構築する研究を行っている．LCS とは動詞の意味を統語構造との対応をとりつつ動詞間の含意関係を整理して記述する枠組みであり，言語学で主に研究され現在も発展中の理論的枠組みである．これに対して記述できる範囲を限定した形で電子化した辞書 (約 1200 語の動詞を

対象) が竹内 (竹内, 2004) によって公開されており，この辞書を元にしてコーパスによる自動的な拡張をおこなうことを考えている．本年度はまず作成しようとする LCS 辞書の体系がどの程度言語処理で有効であるかについて述語と項との関係を抽象化してとらえる意味役割付与システムを作成し，その精度をとおして LCS 辞書の有効性を評価したので報告する．評価の枠組みを固めておくことで，これから自動的に LCS 辞書を拡張した場合に比較による辞書の良さが評価できる．

4.1 LCS 辞書評価の枠組み

まず LCS 辞書とはどのような情報を持っているのか具体的に説明した後意味役割との対応関係について述べて評価の枠組みについて説明する．

LCS は動詞の意味を項の間の関係で記述するとともに構造的に記述することで動詞間の関係を記述するものである．例えば「あげる/渡す」という動詞の LCS は [x CAUSE [BECOME [y BE AT z]]] という式で表される．これは「x が y を z にある状態にするという変化結果を引き起こす」ことを示しているとともに「y が z にあるように変化する」と「y が z にある」が含意されていることを LCS 内の部分構造が示している．

一方，意味役割は文の述語と項との関係を抽象化したラベルであり，これを文に対して付与することができると情報の集約を行うことができる．例えば以下の例文の場合，

- 彼が　彼女に　プレゼントを　あげる/渡す

「彼」が Agent,「彼女」が Goal,「プレゼント」が Theme となり，これによって「あげる/渡す」のどちらの表現も Theme(「プレゼント」) が Goal(「彼女」) に移動するように Agent(「彼」) が仕向けたことを示している．移動や変化という概念を中心にこのような抽象的な役割分析を行うことで事実関係の集約が可能となる[2]．

上記で示した LCS と意味役割の記述的対応関係から明らかなように LCS 辞書には動詞がどのような意味役割を持っているのかを記述している．そこで LCS 辞書を元に動詞に対応した項の意味役割を付与するモデルを作成し，その精度で LCS の体系が実際どの程度のうまく辞書化できているかを測定する．

4.2 意味役割付与タスクとモデル化

4.2.1 問題設定

言語処理においてどの程度の意味役割の種類が必須であるかはまだ定まっていない．そこで先行研究とデータ分析から主要な意味役割を表 6 の 11 種類に定めた．この 11 種類の意味役割は名詞に対して決定されるのではなく，文における述語の持つ動作・状態をある典型的な骨格 (フレーム) として考えたときに仮定できるものである．LCS 辞書はこの典型的な振る舞いの型を動詞ごとに書いた言語資源であるので LCS 辞書を用いて表層格表現と意味役割を結びつける表層深層対応規則を作成することができる．図 1 には「発足する」という動詞の LCS 辞書から意味役割フレームが 2 種類対応する可能性を示している．それぞれのフレームは文脈と名詞のタイプによってどちらかが選択される．これらのフレームの違いを文で表現すれば「彼が野球チームを発足する/ 野球チームが発足する」である．つまり LCS はある動詞が選択できる意味役割フレームのセットを与えており，ここから名詞の概念や文脈を利用してどのフレームが文で表現されていたかを推定するのが意味役割タスクである．ただし上記の意味役割セットの中には Time や Location といった個別の動詞に依存しない付加詞も決定する必要がある．これは LCS 辞書とは別に名詞の概念と文脈により推定するモデルを別に用意する必要がある．

4.2.2 意味役割付与モデル

意味役割付与モデルの全体像を図 2 に示す．処理手順としては以下のようになる．

[2]例文ではプレゼントが結果として今どこにあるのかという事実関係.

表 6: 意味役割セット

意味役割	説明	例
Agent	動作を引き起こす主体	彼が (Agent) 茶碗を洗う
Experiencer	ある心理事象を体験する者	彼が (Experiencer) 風邪をひく
Instrument/Cause	ことがらを起こす原因	雨で (Instrument) かさが濡れる
Theme/Object	変化の対象	おもちが (Theme) 焼ける
Source	移動・変化の起点	神戸港から (Source) 出港する
Goal	移動の終点，変化結果	彼に (Goal) ボールを投げる
Location/Place	場所や位置	喫茶店で (Location) 会う
Time	時間を表す役割	6 時に (Time) 待ち合わせる
Scene	状況・場面	国際会議で (Scene) 発表する
Path	経路や経由点	ウィーンを (Path) 旅する
Reason	動作の理由	雨で (Reason) 遠足が中止になる
Opponent	動作の基準となる相手	彼に (Opponent) 反対する

- 述語と項の組み合わせを取得 (イベント単位の取得)
- LCS 辞書からの意味役割フレーム候補の取得
- 名詞の概念体系を利用した意味役割フレームの選択
- 付加詞の意味役割分析

動詞	格の制約	意味役割との対応
発足	人, が ::	Agent, が,Theme, を, Goal, に
発足	物, が ::	Theme, が,Goal, に

図 1: LCS と表層深層の対応規則

以下順に説明する．

述語と項の組み合わせの取得 まず入力文に係り受け解析器 (Cabocha) を適用し動詞，サ変名詞と係り関係にある項を同定する．一つの述語 (動詞) に対して項が複数あり一つの事象単位をあらわすことから，ここではイベント単位と呼ぶことにする．イベント単位はガヲ二格といった格助詞だけを対象にするのではなく，図 3 に示すように「彼女たちの (Agent) 送金」のように接続助詞「の」によるサ変名詞との関係も対象にする．

LCS 辞書からの意味役割セット候補の取得 LCS 辞書そのものは式で表現されているだけで (4.1 節参照) 意味役割ラベルが存在しない．しかし LCS は意味役割と対応関係があることから，LCS 辞書からあらかじめ意味役割ラベルの形式に変換した制約付き表層深層対応規則を作成する．図 1 に例を示す．図中の「意味役割との対応」が表層格に対応した意味役割を示しており「格の制約」がそのフレームを選択する場合の名詞に対する制約である．

名詞の概念体系を利用した意味役割フレーム選択 名詞の概念体系辞書を利用して表層格に対応する名詞の分類に従って意味役割フレームの選択を行う．例えば「研究部門が発足する」が入力された場合，LCS 辞書の対応規則 (図 1 参照) によってガ格の名詞が人か物かのどちらに近いかによって意味役割フレームが選択される．この近さを EDR 概念体系辞書の構造を利用して測定する．この例文では「部門」が人に近いのか物に近いのかで判断する (詳細な計算方法は (下村・竹内, 2006) を参照).

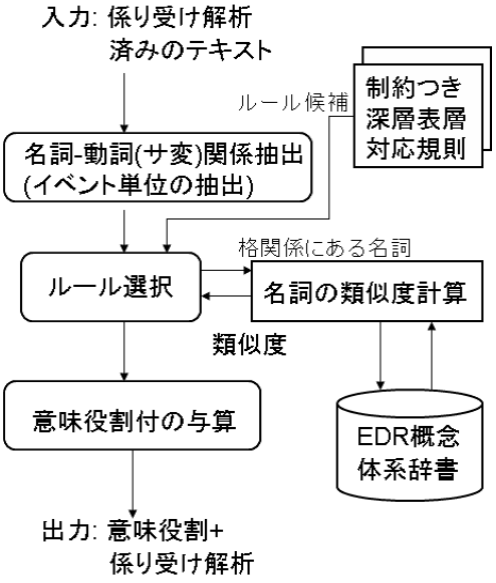


図 2: 意味役割付与システムの全体図

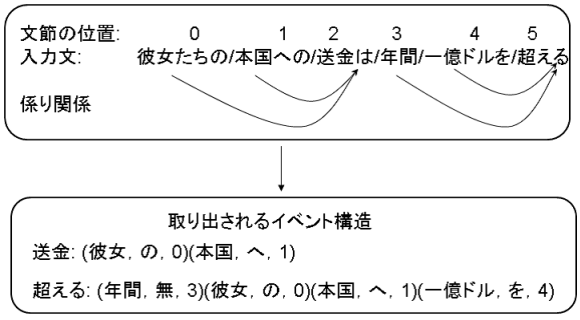


図 3: イベント単位の抽出例

付加詞の意味役割分析 意味役割の中で深層格 Instrument，Reason，Location，Scene，Time および表層格におけるデ格やマデ格は付加詞であるため動詞で決定すべきものではない．ここでは簡易的に各意味役割に対して名詞概念が対応すると仮定して処理を行う．具体的には上記の深層格に対してそれぞれ，Instrument:道具，Reason:現象，Location:場所，Scene:イベント，Time:時間と概念を設定して名詞がどの概念に対応しているか EDR 概念体系辞書を利用して決定する．例えば「飛行機で到着する」という文の場合,「到着する」のデ格は LCS 辞書では記述されていないため「飛行機」を EDR 概念体系で「道具」「現象」「場所」「イベント」のどの概念に近いかを判定する (詳細は (下村・竹内, 2006) 参照)．これにより付加詞による意味役割を付与する．

4.3 実験と考察

実験対象は毎日新聞 95 年版のうち京都コーパスに収録されている 100 文とし，人手により意味役割を付与したタグ付きコーパスを作成した．ただし LCS 辞書が小規模であるため LCS 辞書に載っている動詞のみを実験の対象とする．評価はイベント単位と単語単位で行った．実験結果を表 7 に示す．イベント単位で 7 割程度の精度がでており，他の実験結果と単純な比較はできないが，肥塚ら (肥塚他, 2007) の FrameNet による意味役割推定に比べて高い精度を示している．このことから本提案モデルの有効性は示せたと考えられる．表 8 に実験で誤ったものを原因別にまとめたものを示す．もっとも主要な誤り要因は名詞概念のカテゴリ分けであり，ついで機能語の判別，その次に LCS 辞書の不備が挙げられる．この誤り分析の結果から LCS 辞書はかなり正確に機能していることがわかる．よって意味役割付与タスクにおいて限定的にではあるが既存の LCS 辞書が有効であることを示すことができた．

4.4 まとめ

約 1200 語の LCS 辞書と EDR 電子化辞書による概念体系を利用して意味役割付与システムを作成し，その精度から LCS 辞書の評価をこころみた．意味役割付与の精度は高く，誤りの多くは名詞のカテゴリー分析に関するものであった．LCS 辞書に起因する誤りは相対的に少なく LCS 辞書は精度良く構築されていると判断できる．今後，既存の LCS 辞書を利用した半自動構築について手法を整理し，実行していく予定である．

表 7: 名詞単位, イベント単位での結果

	正解/全体	精度
名詞単位	219/268	81.71 %
イベント単位	117/165	70.90 %

表 8: 誤り解析

原因	誤り数/誤り全体	割合
名詞のカテゴリ分け	30/49	61.2 %
機能語	9/49	18.3 %
LCS の格の不備	4/49	8.1 %
定義	2/49	4.08 %
その他	4/49	8.16 %

文献

肥塚真輔、岡本紘幸、斎藤博昭、小原京子 (2007).「日本語フレームネットに基づく意味役割推定」, 自然言語処理, 14 巻 1 号, pp.43–66.

Taku Kudo (2000). *TinySVM: Support Vector Machines*. `http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/index.html`.

Christopher D. Manning and Hinrich Schütze (1999). *Foundations of Statisitcal Natural Language Processing*: The MIT Press.

Adwait Ratnaparkhi (1996). "A Maximum Entropy Model for Part-of-Speech Tagging," in *Proceedings of EMNLP 1*, pp. 133–142.

下村拓也、竹内孔一 (2006).「名詞の概念体系を利用した規則に基づく意味役割付与システムの構築」, 情報処理学会自然言語処理研究会, *175-NL-2006*, pp.13–20.

白井清昭 (2003).「SENSEVAL-2 日本語辞書タスク」, 自然言語処理, 10 巻 3 号, pp.3–24.

竹内孔一 (2004).「語彙概念構造による動詞辞書の作成」, 第 *10* 回言語処理学会年次大会, pp.576–579.

デモ・ポスターセッション

3月18日（日）第二日目　11:15〜13:15

『現代日本語書き言葉均衡コーパス』におけるサンプリングの概要

▶丸山 岳彦、柏野 和佳子、山崎 誠、佐野 大樹、秋元 祐哉、稲益 佐知子、吉田谷 幸宏

『現代日本語書き言葉均衡コーパス』における著作権処理について

▶森本 祥子、前川 喜久雄、小沼 悦、新井田 貴之、松下 愛、吉田谷 幸宏、神野 博子、大石 有香

『現代日本語書き言葉均衡コーパス』における電子化フォーマットの概要

▶山口 昌也、高田 智和、北村 雅則、間淵 洋子、西部 みちる

『現代日本語書き言葉均衡コーパス』における短単位の概要

▶小椋 秀樹、小木曽 智信、小磯 花絵、冨士池 優美、相馬 さつき、渡部 涼子、服部 龍太郎

日本語コーパスでのSketch Engine実装の試み

▶投野 由紀夫

確率的単語分割ツールとその利用

▶浅原 正幸

タグ付きコーパス検索ツールの開発

▶谷口 雄作、新保 仁

『日本語コーパス』用Yahoo！知恵袋データについて

▶岡本 真、木戸 冬子、佐古 智正

『現代日本語書き言葉均衡コーパス』における サンプリングの概要

丸山 岳彦* （国立国語研究所研究開発部門）
柏野 和佳子 （国立国語研究所研究開発部門）
山崎 誠 （国立国語研究所研究開発部門）
佐野 大樹 （国立国語研究所研究開発部門）
秋元 祐哉 （国立国語研究所研究開発部門）
稲益 佐知子 （国立国語研究所研究開発部門）
吉田谷 幸宏 （国立国語研究所研究開発部門）

Outline of Sampling Method in the Balanced Corpus of Contemporary Written Japanese

Takehiko Maruyama* (Dept. Lang. Res., National Institute for Japanese Language)
Wakako Kashino (Dept. Lang. Res., National Institute for Japanese Language)
Makoto Yamazaki (Dept. Lang. Res., National Institute for Japanese Language)
Motoki Sano (Dept. Lang. Res., National Institute for Japanese Language)
Masaki Akimoto (Dept. Lang. Res., National Institute for Japanese Language)
Sachiko Inamasu (Dept. Lang. Res., National Institute for Japanese Language)
Yukihiro Yoshidaya (Dept. Lang. Res., National Institute for Japanese Language)

1 導入

言語コーパスの設計（コーパスデザイン）を考える上で，サンプリングの概念は極めて重要な役割を果たす。どのような母集団から，どのような種類のサンプルを，どのような割合で，どのような方法により収集するか，といった問題を具体的に検討しなければならない点において，サンプリングの方針を定めることは，コーパス構築の基礎を担う作業であると言える。

『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ と記す）』は,「生産実態サブコーパス」「流通実態サブコーパス」「非母集団サブコーパス」という 3 つのサブコーパスから構成される，1 億語規模のコーパスである。このうち生産実態サブコーパスと流通実態サブコーパスでは，母集団を明確に定義し，層別ランダムサンプリングによってサンプルを抽出する。これを実現するために必要となるのは，書き言葉の生産実態および流通実態に関する基礎調査と，母集団の定義，ランダムネス性を保障するサンプリング手法の開発，サンプリング台帳の作成，そして一定の手続きに基づくサンプリングの実施である。

本稿では，BCCWJ の構築にあたり，現在我々が進めているサンプリングの方法論について報告する。紙幅の都合上，生産実態サブコーパスのサンプリング方法について，書き言葉の生産実態に関する調査，母集団の定義，サブコーパス全体のサンプルの構成比の算出方法を中心に述べる。

* maruyama@kokken.go.jp

2　BCCWJ におけるサンプリングの基本方針

はじめに，BCCWJ におけるサンプリングの基本方針を示す。また，生産実態サブコーパスの設計方針について述べ，サンプリングを実施するためには何が必要になるかを示す[1]。

2.1　サンプリングの基本方針

まず，サンプリングの基本方針について示す。BCCWJ には，1976 年以降の現代日本語の書き言葉をサンプルとして格納する。実際には，1976 年以降に発行された書籍・雑誌・新聞・白書などが，サンプルを取得する主な対象となる。ただし，漫画，写真集などのように非言語表現が主たる内容のものや，名簿，年鑑，データ集などのように非文章表現が主たる内容のものは，対象から外す。

サンプル抽出には，層別ランダムサンプリング（層別無作為抽出法）を採用する。母集団を複数の基準により層別し，各層からランダムにサンプルを抽出する。この際，母集団に含まれる「文字数」を推計し，各層の総文字数の比に応じて構成比を決定する，という方針を採る。文字数によって母集団のサイズを把握し，一定の抽出比でサンプルを取得することにより，代表性を備えたバランストコーパスを実現することを目指す。

2.2　2 種類のサンプル

続いて，サンプリングの基本単位となる，2 種類のサンプルについて確認しておく。BCCWJ では，収録サンプルが備えるべき条件として，以下の方針が立てられている。

- 統計的に厳密な言語調査に耐え得るよう，母集団からの抽出比を重視した設計にする。
- 文体研究・テキスト研究に耐え得るよう，ある程度の文脈を確保した設計にする。

これらの方針に対処するため,「固定長サンプル」「可変長サンプル」という，異なる長さを持つ 2 種類のサンプルを設計した。両者は，母集団からの抽出方法という点において異なる仕様を持つ。

「固定長サンプル」は，母集団に含まれるすべての文字に対して等確率を与えた上で，ある 1 文字をランダムに抽出し，その文字を始点として 1,000 文字の範囲を抽出するサンプルである[2]。すべての文字に対して等確率を与えるために，母集団に含まれる文字の総数をあらかじめ推計しておく必要がある。母集団（＝推計された総文字数）からの抽出比が明確である点で，基本語彙表や漢字表の作成，語彙・文字調査など，統計的な言語研究に向く。また，母集団の層的かつ量的な構造を忠実に反映する点で，代表性を備えたバランストコーパスとしての性格を強く持つ。

一方,「可変長サンプル」は，固定長サンプルと同様，母集団に含まれるすべての文字に対して等確率を与えた上で，ランダムに抽出した 1 文字を含む言語的な構造のまとまり（「章」や「節」など，ただし 1 万字を上限とする）を抽出するサンプルである。文章・談話としてのまとまりを重視したサンプルであるため，テキストの論理構造の把握や文脈の分析，文体の調査などに向く。

可変長サンプルは，BCCWJ を構成する 3 つのサブコーパスのすべてに対して提供される。一方の固定長サンプルは，統計的な調査を行う可能性の高い部分，すなわち，生産実態サブコーパス，流通実態サブコーパス，および，非母集団サブコーパスの一部（白書など）に対して，可変長サンプルと同時に提供される。

---

[1] BCCWJ 全体の仕様と設計方針については，山崎 (2007) を参照。

[2] 実際に 1,000 文字としてカウントするのは，漢字，仮名，数字，アルファベットのみである。句読点や記号類は数えない。

2.3　生産実態サブコーパスのコーパスデザイン

以下では，3 つのサブコーパスのうち，生産実態サブコーパス（以下，生産実態 SC と記す）のコーパスデザイン，およびサンプリングの方針をどのように定めたかについて述べる。

我々はまず，書き言葉の語彙調査を主たる目的として，2001 年から 2005 年の間に国内で発行されたすべての書籍・雑誌・新聞から 1,000 万語分の固定長サンプルを抽出することを計画した。1,000 万語という数値は，統計的な語彙調査のために十分なサンプルサイズを見積もった結果である。

1,000 万語の内訳は，5 年間に発行された書籍・雑誌・新聞に含まれる総文字数を推計し，その比によって求めることにした。例えば,「書籍：雑誌：新聞」の総文字数が「5:3:2」の割合であれば，書籍から 500 万語，雑誌から 300 万語，新聞から 200 万語を取得する，という具合である。これは，書き言葉の生産力を文字の数によって近似的に捉え，サブコーパス全体が書き言葉の生産実態をなるべく忠実に反映する縮図となるように意図したものである。

また，従来行われてきた語彙調査の結果などを参考にして，1 語あたりの平均長を約 1.7 文字と試算した。ここから，1,000 文字を 1 サンプルとする固定長サンプルで 1,000 万語分を達成するためには，17,000 サンプルが必要となるという見積もりを立てた。

さらに，固定長サンプルを取得するために選んだ 1 文字を利用して，可変長サンプルを同時に取得することにした。可変長サンプルの平均長を試算したところ，書籍で平均 3,900 文字，雑誌で平均 3,000 文字，新聞で平均 1,000 文字という結果が得られた。可変長サンプル全体の語数がどれだけのサイズになるかは，固定長サンプル 1,000 万語の構成比が算出されてから決まることになる。

2.4　サンプリングの実施には何が必要か

以上の設計方針に基づき，生産実態 SC でのサンプリングを，概略，次の手順で進めることにした。

まず，2001 年から 2005 年までに国内で発行された書籍・雑誌・新聞の総体を調査し，客観的に把握できる形で母集団を厳密に定義する。次に，書籍・雑誌・新聞それぞれを「層」と見なし，さらにそれらを下位の層に分割した上で，各層に含まれる文字の数を推計する。これにより，生産実態 SC の母集団を文字数という側面から把握し，総文字数の比に応じて各層の構成比を決定する。さらに，層別ランダムサンプリングを行うためのサンプリング台帳を作成する。これは，各層に含まれるすべての文字に対して等確率を与え，その中の 1 文字をランダムに指定するように設計する。指定された 1 文字は「サンプル抽出基準点」となり，この点を基準として，固定長サンプルでは 1,000 文字の範囲が，可変長サンプルではその文字を含む「章」「記事」などの範囲が，それぞれサンプルとして抽出されることになる。あとは，サンプリング台帳で指定されている実際の書籍などを手に取り，一定の手順に従って，印刷紙面からサンプルを抽出する作業を個別に進めていけばよい。

このような手順によってサンプリングを実施するためには，前もって，次の 3 つの点を明らかにしておく必要がある。

- **母集団の定義** — 母集団をどのように定義するか。それをどのように層別するか。
- **サンプル構成比の決定** — 各層に含まれる総文字数をどのように推計し，固定長サンプル 1,000 万語分の構成比を算出するか。
- **サンプリング台帳の作成** — どのような手順により 1 文字をランダムに抽出するか。

以下の各節では，上記の 3 点について順に解説する。

3　母集団の定義 — 5 年間における書き言葉の生産実態

以下では，我々が生産実態 SC における母集団をどのように定義し，どのような基準により複数の層に分割したかについて，書籍・雑誌・新聞の順に述べる。

3.1　生産実態 SC「書籍」の母集団の定義と層別の方法

まず，生産実態 SC「書籍」の母集団の定義と，層別の方法について述べる。

2001 年から 2005 年の間に国内で発行された書籍の実態を把握するために，我々は，国立国会図書館の蔵書目録を利用することにした。国内で出版される出版物は，原則的にすべて国立国会図書館に納本されるため，この蔵書目録を使えば書籍の生産実態を網羅的に捉えることができる。

我々は，国立国会図書館・日本図書館協会からの協力を得て，国立国会図書館の蔵書目録を電子化した市販のデータ「J-BISC（Japan Biblio disc)」をリスト化した。このリストは，国内で発行された書籍の書誌情報（タイトル，著者，発行者，発行年，ページ数，NDC など）から構成される。このリストから，漫画，写真集，人名録のように言語表現・文章表現が主体でないものや，極端にページ数の少ないもの，楽譜，地図，マイクロフィルム，電子出版物などを取り除いたところ，2001 年から 2005 年までに発行された書籍の冊数は合計 317,117 冊，74,911,520 ページという結果を得た。この結果を，生産実態 SC「書籍」の母集団として定義した。

次に，この母集団を層別する基準について示す。ここでは,「日本十進分類法（NDC)」および「発行年」という基準によって層別を行った。「NDC」は日本で用いられている図書館の蔵書分類法であり，書籍の内容により，大きく「0. 総記」「1. 哲学」「2. 歴史」「3. 社会科学」「4. 自然科学」「5. 技術・工学」「6. 産業」「7. 芸術・美術」「8. 言語」「9. 文学」という 10 のカテゴリに分類される[3]。我々は，J-BISC に付与されている NDC（1 桁目）の 10 分類に加え，NDC が付与されていないレコードを「n（null；記録なし)」として，合計 11 の層に分類した。これに「発行年」として 2001 年から 2005 年までの 5 分類を重ね合わせ，書籍全体を合計 55 の層に分割した。

生産実態 SC「書籍」の母集団を NDC で層別した際の冊数とページ数について，表 1 に示す[4]。

表 1: 生産実態 SC「書籍」の母集団

NDC	総冊数	総ページ数
0 （総記）	11,132	2,859,793
1 （哲学）	18,067	4,529,329
2 （歴史）	24,624	6,449,172
3 （社会科学）	62,986	16,059,116
4 （自然科学）	28,745	6,771,958
5 （技術工学）	31,377	6,681,335
6 （産業）	15,332	3,298,313
7 （芸術）	25,387	5,153,531
8 （言語）	5,211	1,196,840
9 （文学）	73,716	18,888,278
n （記録なし）	20,540	3,023,855
合計	317,117	74,911,520

3.2　生産実態 SC「雑誌」の母集団の定義と層別の方法

次に，生産実態 SC「雑誌」の母集団の定義と，層別の方法について述べる。

「雑誌」という概念を，定期刊行物（定期的に刊行される冊子）という観点から捉えると，典型的な「雑誌」として想起される月刊誌や週刊誌だけでなく，学会誌や業界誌などの専門誌や，極めて限られた地域だけで流通しているコミュニティ冊子なども含んでしまうため，その全容を把握すること

[3] 実際の NDC では,「007.637 図形処理ソフトウェア」のように，さらに細かく分類される。
[4] さらに発行年で層別した場合の冊数とページ数を示すこともできるが，紙幅の都合でここでは割愛する。

が困難になる。そこで,「2001 年から 2005 年の間に社団法人日本雑誌協会に加盟していた出版社が発行していた定期刊行物」という条件によって，雑誌を絞り込むことにした。日本雑誌協会の加盟社は国内でも有力な出版社が多く，典型的な「雑誌」の範囲を捉えるのに適切であると判断した。

まず，2001 年から 2005 年の各年における雑誌協会加盟社のリストを作り，対象出版社 102 社を絞り込んだ。さらに,『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）を用いて対象出版社が各年に発行した定期刊行物に関する書誌情報を抽出した。その際，新聞，要覧，コミック，非日本語による定期刊行物などは除外した。その結果，2001 年から 2005 年の間に発行された「雑誌」は，異なりで 1,259 タイトル，合計 55,779 冊，10,414,955 ページという結果を得た。この結果を，生産実態 SC「雑誌」の母集団として定義した。

次に，この母集団を「分野」「発行年」という基準で層別した。「分野」は『雑誌新聞総かたろぐ』で示されている「1. 総合」「2. 教育・学芸」「3. 政治・経済・商業」「4. 産業」「5. 工業」「6. 厚生・医療」という 6 分類である。これに「発行年」として 2001 年から 2005 年までの 5 分類を重ね合わせ，雑誌全体を合計 30 の層に分割した。

生産実態 SC「雑誌」の母集団を「分野」で層別した際の冊数とページ数について，表 2 に示す。

表 2: 生産実態 SC「雑誌」の母集団

分野	総冊数	総ページ数
1. 総合	38,383	7,163,989
2. 教育・学芸	5,456	983,224
3. 政治・経済・商業	3,168	469,282
4. 産業	599	115,172
5. 工業	7,101	1,493,800
6. 厚生・医療	1,072	189,488
合計	55,779	10,414,955

表 3: 生産実態 SC「新聞」の母集団

紙種	総冊数	総ページ数
全国紙	15,950	426,472
ブロック紙	9,570	248,300
地方紙	24,105	523,417
合計	49,625	1,198189

3.3 生産実態 SC「新聞」の母集団の定義と層別の方法

最後に，生産実態 SC「新聞」の母集団の定義と，層別の方法について述べる。

「新聞」という形態には，全国紙，地方紙，スポーツ紙，専門紙，タウン紙など，幅広い様式が存在する。これらをすべて収録対象とすると，研究用途上あるいはコーパス構築の実務上，何らかの障害を生じさせる可能性が高い。そこで,「『全国新聞ガイド』（社団法人日本新聞協会発行）において「全国紙」「ブロック紙」として記載されている日刊新聞」という条件で新聞タイトルを絞り込んだ。さらに，この条件ではカバーできない各地域の有力な地方紙も取り入れ，日本全国で発行されている新聞を含むように調整した。その結果，2001 年から 2005 年の間に発行された「新聞」は，異なりで 16 タイトル，合計 49,625 冊[5]，1,198,189 ページとなった。この結果を，生産実態 SC「新聞」の母集団として定義した。対象となる新聞 16 タイトルを，以下に示す。

全国紙： 朝日新聞，毎日新聞，読売新聞，日本経済新聞，産経新聞
ブロック紙： 北海道新聞，中日新聞，西日本新聞
地方紙： 河北新報，新潟日報，京都新聞，神戸新聞，中国新聞，高知新聞，愛媛新聞，琉球新報

次に，この母集団を「紙種」「発行年」という基準で層別した。「紙種」は全国紙・ブロック紙・地方紙の別，および 16 種の新聞のタイトルとした。これに「発行年」の 5 分類を重ね合わせ，新聞全体を合計 80 の層に分割した。

[5] この場合の 1 冊は，1 つの朝刊または夕刊を指す。

新聞の母集団を「紙種」で層別した際の冊数とページ数の分布について，表 3 に示す。

以上，生産実態 SC の母集団を定義し，層別を行った手順について述べた。

4　サンプル構成比の決定 — 「現代日本語書き言葉の文字数調査」

以下では，生産実態 SC の母集団を構成する各層に含まれる総文字数を推計した調査「現代日本語書き言葉の文字数調査」の概要を述べる[6]。また，調査結果を用いて生産実態 SC の固定長サンプル 1,000 万語分の構成比を求めた結果について示す。

4.1　「現代日本語書き言葉の文字数調査」の概要

この調査は，母集団を構成する各層に含まれる総文字数を推計し，その比を求めることで，1,000 万語分の固定長サンプルにおける各層の構成比を算出することを目的とするものである。

「現代日本語書き言葉の文字数調査」は，概略，以下のような手順によって実施した。まず，母集団を構成する各層を,「判型（紙面のサイズ）」の観点からさらに下位の層に分割した。これにより，例えば「NDC が 9 番台，高さ 21cm の書籍」「分野が 1 番（総合），A4 版の雑誌」などの層ができる。これらの各層に含まれるページの中から，複数のページをランダムに抽出した[7]。次に，抽出した各ページに含まれる文字数を計測し，1 ページあたりに含まれる平均文字数を各層ごとに求めた。これを「キャラクタ密度（character density）」と呼ぶ。このキャラクタ密度を係数として，各層の 5 年間の総ページ数に掛け合わせることにより，各層に含まれる総文字数，および母集団全体の総文字数を推計した。

文字数の計測対象となる範囲は，表紙・広告を除くすべての部分とした。文章内の論理構造の別（本文，図表，脚注，キャプション，ルビ，柱など）や，文字種の別（かな・カナ・漢字・記号・外国語・絵文字など）を問わず，現れた文字要素についてはすべて計測対象とした。

以下では，書籍・雑誌・新聞の順に，文字数調査の手順と結果について示す。

4.2　書籍の文字数調査

まず，書籍の文字数調査について述べる。2003 年に発行された全書籍（65,719 冊，15,544,357 ページ）を対象に，11 種類の「NDC」と 18 種類の「判型」によって行列を作成した。次に，NDC ごとに，全体の 90%までの冊数に含まれる判型の種類を選び出した。この判型に該当する書籍をランダムに抽出し，さらに各冊から 5 ページをランダムに抽出した。このページ内に含まれる文字の数を，人手または OCR によって計測した。最後に，5 ページ中に現れた文字数の平均値を算出し，1 ページあたりに含まれる文字数を示す「キャラクタ密度」とした。実際に計測したのは，227 冊（2003 年に発行された書籍全体の 0.345%），1,135 ページ（同 0.073 ‰）であった。

以上の計測により得られた判型ごとのキャラクタ密度の分布から，回帰直線 y=31.255x+379.5 を得た。判型別のキャラクタ密度の分布と回帰直線を，図 1 に示す。

この回帰直線を用いて，実測しなかった判型の種類についてもキャラクタ密度を計算した。これにより完成した行列を，書籍の「総文字数推計基準表」とした。完成した総文字数推計基準表を，表 4 に示す。行は NDC，列は判型を表す。また，斜体になっている数値は，回帰直線に基づいて算出された値であることを表す。

---

[6] 「現代日本語書き言葉の文字数調査」の詳細については，丸山・秋元 (印刷中) を参照のこと。

[7] 調査に用いたサンプルは，2003 年に発行されたものに限定して行った。

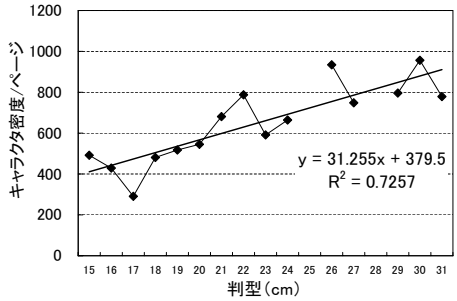


図 1: 判型別のキャラクタ密度と回帰直線（書籍）

表 4: 総文字数推計基準表 （書籍）

	≦15cm	16cm	17cm	18cm	19cm	20cm	21cm	22cm	23cm	24cm	25cm	26cm	27cm	28cm	29cm	30cm	31cm≧	null
0	*410.8*	*442.0*	*473.3*	530.8	557.8	389.8	660.9	502.2	591.2	467.6	*723.3*	1026.4	*785.8*	*817.1*	*848.3*	569.4	*910.8*	*629.9*
1	544.0	418.5	*473.3*	542.2	497.0	529.6	743.9	655.8	*660.8*	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*675.1*
2	516.0	434.0	*473.3*	466.6	471.9	529.8	652.9	747.1	*660.8*	*692.1*	*723.3*	1234.6	851.6	*817.1*	*848.3*	*879.6*	*910.8*	*700.6*
3	413.8	*442.0*	*473.3*	456.0	602.5	603.0	618.6	885.9	*660.8*	*692.1*	*723.3*	1272.6	*785.8*	*817.1*	*848.3*	1674.1	*910.8*	*757.6*
4	*410.8*	*442.0*	290.8	457.0	469.1	631.6	680.8	801.1	*660.8*	*692.1*	*723.3*	1159.7	473.6	*817.1*	*848.3*	580.4	*910.8*	*650.0*
5	430.4	396.0	*473.3*	340.8	409.9	539.0	672.8	815.5	*660.8*	860.8	*723.3*	702.4	544.8	*817.1*	636.2	962.5	*910.8*	*641.0*
6	*410.8*	*442.0*	*473.3*	558.8	513.9	511.6	448.3	1192.7	*660.8*	*692.1*	*723.3*	1254.6	1123.2	*817.1*	*848.3*	441.5	*910.8*	*707.2*
7	503.6	492.8	*473.3*	451.2	573.4	654.5	667.1	638.8	*660.8*	*692.1*	*723.3*	318.5	*785.8*	*817.1*	955.8	1510.8	778.8	*688.1*
8	600.6	345.0	*473.3*	522.2	632.0	561.6	910.4	1052.0	*660.8*	*692.1*	*723.3*	506.6	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*701.3*
9	435.2	487.2	*473.3*	482.8	447.8	501.4	753.3	585.3	*660.8*	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*661.1*
null	*410.8*	*442.0*	*473.3*	*504.5*	*535.8*	*567.0*	*598.3*	*629.5*	*660.8*	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*660.8*

この推計基準表を 5 年間の判型ごとのページ数に掛け合わせ，書籍の 5 年間の総文字数を推計したところ，約 485 億文字という結果になった。NDC ごとの推計結果を表 5 に示す。

表 5: 総文字数の推計結果（書籍）

NDC	総文字数	構成比
0 （総記）	1,636,414,548	3.371%
1 （哲学）	2,597,610,813	5.351%
2 （歴史）	4,301,204,340	8.861%
3 （社会科学）	12,408,321,943	25.563%
4 （自然科学）	5,069,594,034	10.444%
5 （技術工学）	4,615,929,967	9.510%

NDC	総文字数	構成比
6 （産業）	2,196,387,437	4.525%
7 （芸術）	3,258,432,447	6.713%
8 （言語）	888,800,128	1.831%
9 （文学）	9,341,275,486	19.245%
n （記録なし）	2,225,954,208	4.586%
合計	48,539,925,351	100.00%

4.3 雑誌の文字数調査

次に，雑誌の文字数調査について述べる。2003 年に発行された全雑誌（909 タイトル，11,167 冊，2,095,217 ページ）を対象に，6 種類の「分野」と 5 種類の「判型」によって行列を作り，分野ごとに全体の 90%までの冊数に含まれる判型のタイプを求めた。次にこの判型に該当する雑誌の冊をランダムに抽出し，さらに各冊から 5 ページをランダムに抽出した。5 ページ中に現れた文字数を実測し，その平均値を「キャラクタ密度」とした。実際に計測したのは，53 冊（2003 年に発行された雑誌全体の 0.475%），265 ページ（同 0.127 ‰）であった。

実測により得られた値については行列に埋め込んだ。実測しなかった判型については，他の分野で計測したその判型のキャラクタ密度の平均値を割り当てた。この結果を，雑誌の「総文字数推計基準表」とした。完成した総文字数推計基準表を，表 6 に示す。斜体になっている数値は，他の分野で計測した平均値に基づいて算出された値であることを表す。

表 6: 総文字数推計基準表 （雑誌）

分野	A4 系	A5 系	AB 系	B4 系	B5 系
1. 総合	1022.1	1031.2	1413.7	831.1	807.5
2. 教育	765.4	884.8	*1413.7*	*831.1*	878.1
3. 政治	1014.2	*921.4*	*1413.7*	*831.1*	798.2
4. 産業	*1093.2*	*921.4*	*1413.7*	*831.1*	506.6
5. 工業	973.2	*921.4*	*1413.7*	*831.1*	*767.9*
6. 厚生	1273.3	*921.4*	*1413.7*	*831.1*	510.0

表 7: 総文字数の推計結果（雑誌）

分野	総文字数	構成比
1. 総合	7,421,447,806	70.575%
2. 教育	877,875,592	8.348%
3. 政治	456,459,405	4.341%
4. 産業	110,640,958	1.052%
5. 工業	1,468,293,360	13.963%
6. 厚生	180,964,513	1.721%
合計	10,515,681,636	100.00%

この推計基準表を 5 年間の判型ごとのページ数に掛け合わせ，雑誌の 5 年間の総文字数を推計したところ，約 105 億文字という結果になった。分野ごとの推計結果を表 7 に示す。

4.4 新聞の文字数調査

最後に，新聞の文字数調査について述べる。新聞の場合，印刷紙面がほぼ定型であることを考慮し，書籍・雑誌とは異なる方法で推計を行った。まず，全国紙「朝日」「毎日」「読売」「日経」の朝夕刊，各 1 日分（2003 年発行分）をランダムに抽出し，1 冊の各ページに含まれる文字数を人手により計測した。なお，2003 年に発行された新聞は，16 タイトル，9,925 冊，239,638 ページであり，そのうち実際に計測したのは 8 冊（全体の 0.081%），211 ページ（全体の 0.880 ‰）であった。

次に，実際に計測した 1 冊の紙面構成を分析し，1cm^2 あたりに含まれる文字数を面種（いわゆる社会面・政治面など）ごとに算出した。さらに，1 週間分の新聞について面種ごとに面積を計測し，各紙 1 週間分の紙面構成と面積を求めた。面種ごとに求めた 1cm^2 あたりの文字数を，面種ごとの総面積に掛け合わせることにより，新聞 1 週間分に含まれる総文字数を推計した。

以上の手続きにより新聞 1 ページあたりの平均文字数を求め，上記 4 紙のキャラクタ密度を算出した。計測しなかった新聞タイトルについては，朝日・毎日・読売のキャラクタ密度の平均値をキャラクタ密度とした。このキャラクタ密度を 5 年間の総ページ数と掛け合わせ，新聞の 5 年間の総文字数を推計したところ，約 64 億文字という結果になった。新聞タイトルごとの推計結果を表 8 に示す。

表 8: 総文字数の推計結果（新聞）

区分	新聞タイトル	総文字数	構成比
全国紙	朝日新聞	461,946,416	7.200%
	毎日新聞	399,078,164	6.220%
	読売新聞	457,239,476	7.126%
	日本経済新聞	688,108,023	10.725%
	産経新聞	411,250,382	6.410%
ブロック紙	北海道新聞	420,730,356	6.557%
	中日新聞	455,131,442	7.094%
	西日本新聞	420,730,356	6.557%
地方紙	河北新報	368,688,876	5.746%
	新潟日報	351,048,482	5.471%
	京都新聞	357,758,315	5.576%
	神戸新聞	377,170,042	5.879%
	中国新聞	351,048,482	5.471%
	高知新聞	298,127,299	4.647%
	愛媛新聞	246,965,522	3.849%
	琉球新報	351,048,482	5.471%
合計		6,416,070,114	100.00%

4.5 生産実態 SC のサンプル構成比の算出法

「現代日本語書き言葉の文字数調査」により，2001 年から 2005 年に生産された書籍・雑誌・新聞に含まれる総文字数の推計値として，書籍が約 485 億文字，雑誌が約 105 億文字，新聞が約 64 億文字という結果を得た。調査結果全体をまとめて各層の比を算出した結果を，表 9 に示す[8]。

[8] 実際には発行年や新聞タイトル等によっても層別されているが，ここでは割愛する。

表 9: 総文字数の推計結果（全体）

層		総文字数	構成比
書籍	0. 総記	1,636,414,548	2.50%
	1. 哲学	2,597,610,813	3.97%
	2. 歴史	4,301,204,340	6.57%
	3. 社会科学	12,408,321,943	18.95%
	4. 自然科学	5,069,594,034	7.74%
	5. 技術工学	4,615,929,967	7.05%
	6. 産業	2,196,387,437	3.35%
	7. 芸術	3,258,432,447	4.98%
	8. 言語	888,800,128	1.36%
	9. 文学	9,341,275,486	14.27%
	n. 記録なし	2,225,954,208	3.40%
雑誌	1. 総合	7,421,447,806	11.34%
	2. 教育	877,875,592	1.34%
	3. 政治	456,459,405	0.70%
	4. 産業	110,640,958	0.17%
	5. 工業	1,468,293,360	2.24%
	6. 厚生	180,964,513	0.28%
新聞	全国紙	2,417,622,461	3.69%
	ブロック紙	1,296,592,154	1.98%
	地方紙	2,701,855,499	4.13%
合計		65,471,677,100	100%

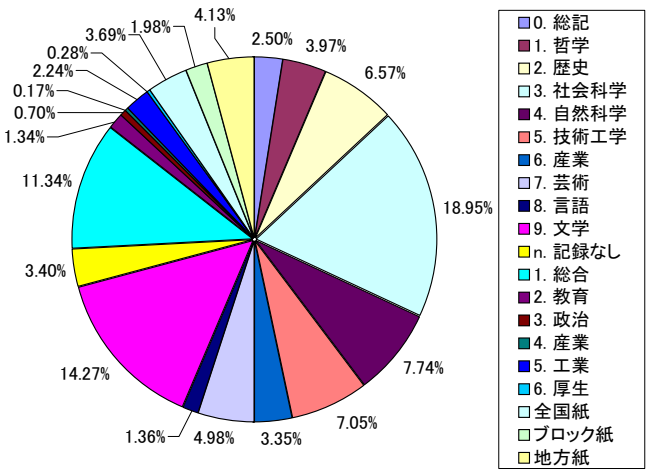


図 2: 固定長サンプル（1,000 万語分）の構成比

表 9 の各層の構成比を，生産実態 SC のうち固定長サンプルで構成される 1,000 万語分の構成比として採用する。その結果得られる全体の構成図を，図 2 に示す。これにより，例えば，2.50%という構成比を持つ書籍の「0. 総記」という層からは，25 万語分のサンプルを抽出すればよいことになる。この場合，1,000 文字の固定長サンプルを 425 サンプルを取得すれば，目標語数が達成される。

この構成比に従って書籍・雑誌・新聞から抽出すべき語数を求め，必要な固定長サンプルの数を算出した。また，そこから得られる可変長サンプルの合計語数を見積もった。結果を表 10 に示す。これで，生産実態 SC の量的なコーパスデザインが完成した。

表 10: 生産実態 SC 全体の構成比

	総文字数	構成比	固定長 合計語数	必要サンプル数	可変長 合計語数
書籍	485.40 億文字	74.14%	741.4 万語	12,604 サンプル	2891.5 万語
雑誌	105.16 億文字	16.06%	160.6 万語	2,730 サンプル	481.8 万語
新聞	64.16 億文字	9.80%	98.0 万語	1,666 サンプル	98.0 万語
合計	654.72 億文字	100.00%	1,000 万語	17,000 サンプル	3471.3 万語

以上,「現代日本語書き言葉の文字数調査」によって母集団に含まれる総文字数を推計し，生産実態 SC の固定長サンプル 1,000 万語分の構成比を定めた過程について述べた。

5 サンプリング台帳の作成

最後に，サンプリング台帳の作成について触れておく。

サンプリングの基本方針としては，母集団中に含まれるすべての文字の中から，ある 1 文字をランダムに抽出する必要がある。我々は，これに近似する方法として，各層に含まれるページをランダムに抽出し，さらに選ばれたページからランダムに 1 文字を抽出する，という手順を採ることにした。

以下では書籍を例として述べる。先述のように，生産実態 SC「書籍」には約 7,500 万ページが含まれており，これらが発行年による 5 分類，NDC による 11 分類の，計 55 の層に層別されている。この情報をデータベース上に展開し，7,500 万の各ページに対して優先順位をランダムに割り振った。さらに，7,500 万の各ページに対して，ページ内の 1 点を指定する座標情報をランダムに指定した。これは，ページに 10×10 の座標枠を割り当て，指定された座標の交点に最も近い文字を抽出するためのものである。交点の直近が白紙や図・写真だった場合に備えて，1 ページあたり 10 通りの交点を，優先順位をつけて指定した。

以上のような手順によって，サンプリング台帳を（論理上は）7,500 万ページ分作成した。1 枚の台帳は，書誌情報，ページ情報，ページ内の座標情報から構成される。その後，55 の各層について，必要となるサンプル数分の台帳を，優先順位の高いものから取得した。

この台帳により，母集団の中から 1 文字をランダムに抽出するという基本方針が，近似的に実現できる。実際の作業では，抽出された 1 文字を「サンプル抽出基準点」として，固定長サンプルでは 1,000 文字を，可変長サンプルでは「章」「記事」などの範囲を，一定の手続きに従って抽出していくことになる。サンプルを抽出する手続きの詳細については，丸山他 (2007) を参照されたい。

6　結語

本稿では，BCCWJ におけるサンプリングの概要として，生産実態 SC のサンプリング方法に関する基本方針について述べた。我々は，書き言葉の生産実態という側面から母集団を定義し，層別をした上で，各層に含まれる総文字数を推計することにより，コーパス全体の構成比を決定するという方針を採用した。これは，母集団の量的な構造をコーパスの量的な構造に適切に反映させるための方針であり，代表性を備えたバランストコーパスとしての BCCWJ が持つ，大きな特徴となっている。

現代では，膨大な量の電子テキストを簡単に入手できるという背景もあり，書き言葉コーパスを作るという作業は比較的単純であるかのように見える。しかしながら，書き言葉の現実的なありさまを精密に捉え，その実態を記述しようとする立場に立つ限り，既存の電子テキストの集積は「書き言葉コーパス」たり得ない。書き言葉の実態を捉えるためのコーパスを構築するためには，綿密な調査に基づく母集団の定義とコーパスデザイン，そしてきめの細かいサンプリングの実践が求められる。

謝辞： J-BISC のリスト化には，国立国会図書館，日本図書館協会より協力を得た。「現代日本語書き言葉の文字数調査」には，東京都立多摩図書館，立川市中央図書館より協力を得た。記して感謝申し上げる。本研究は，文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18〜22 年度，領域代表者：前川喜久雄）による補助を得た。

文献

秋元祐哉・丸山岳彦・吉田谷幸宏・山崎誠・柏野和佳子・稲益佐知子・前川喜久雄 (2007) 書き言葉の総量を捉える —書き言葉はどれだけ生産されるのか—. 『言語処理学会 第 13 回年次大会 発表論文集』. 言語処理学会

丸山岳彦・秋元祐哉 (印刷中) 『『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 —現代日本語書き言葉の文字数調査—』. 特定領域「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-2)

丸山岳彦・柏野和佳子・稲益佐知子・秋元祐哉・吉田谷幸宏・山崎誠 (2007) 書き言葉の構造を捉える —書き言葉の多様な構造とサンプリング手法—. 『言語処理学会 第 13 回年次大会 発表論文集』. 言語処理学会

山崎誠 (2007) 『現代日本語書き言葉均衡コーパス』の基本設計について. 本予稿集.

『現代日本語書き言葉均衡コーパス』における著作権処理について

森本祥子　（国立国語研究所情報資料部門）†
前川喜久雄　（国立国語研究所研究開発部門）
小沼悦　（国立国語研究所研究開発部門）
新井田貴之　（国立国語研究所管理部）
松下愛　（国立国語研究所管理部）
吉田谷幸宏　（国立国語研究所研究開発部門）
神野博子　（国立国語研究所研究開発部門
大石有香　（国立国語研究所研究開発部門）

Copyright Clearing Process in the BCCWJ Project

Sachiko Morimoto　(Dept. Lang. Info. & Res., National Institute for Japanese Language)
Kikuo Maekawa　(Dept. Lang. Res., National Institute for Japanese Language)
Etsu Onuma　(Dept. Lang. Res., National Institute for Japanese Language)
Takayuki Niida　(Dept. Admin. Affairs., National Institute for Japanese Language)
Ai Matsushita　(Dept. Admin. Affairs., National Institute for Japanese Language)
Yukihiro Yoshidaya　(Dept. Lang. Res., National Institute for Japanese Language)
Hiroko Kamino　(Dept. Lang. Res., National Institute for Japanese Language)
Yuka Oishi　(Dept. Lang. Res., National Institute for Japanese Language)

１． なぜ著作権処理をするのか

「現代日本語書き言葉均衡コーパス」の重要な特徴のひとつに，構築したコーパスを公開し，誰もが利用できるようにするという点がある。

日本の著作権法の下では，著作権者に断らずに著作物を使用してよい場合は，非常に限られている。例えば，自分だけが使うために本をコピーする場合などは，そのつど著作権者に断らなくても使用してよいことになっている。しかし，コーパスへの採録はそうした例外にはあてはまらず，取り込む著作物については著作権使用許諾を得ることが不可欠である。

本プロジェクトは，大規模な著作権処理済みデータを採録するコーパスとしては日本で初めてのものである。ここでは，採録対象データすべてについて著作権処理を施し，著作権者から許諾されたデータのみを載せることを基本方針としている。本稿では，その作業について簡単に報告する。

２． 著作権処理の基本方針

本コーパスの構築および構築後の一般公開のためには，著作権法のうち以下の点について使用許諾を得ることが必要となる。

- ○　データベースへの蓄積【著作権法上の複製】
  （データベースの構築・運用に必要なバックアップの複製を含む）
- ○　ネットワーク（インターネット）を介したアクセスに対する提供【著作権法上の自動公衆送信】

† morimoto@kokken.go.jp

○　電子記録媒体での提供【著作権法上の複製・譲渡】
○　研究者等の研究発表活動上のプレゼンテーション【著作権法上の上映】

著作権使用許諾を依頼する際には，この点を明確に伝え，かつ著作権使用料を無料とさせてほしいことを伝えたうえで，著作権者の了解を得ることにしている。

本プロジェクトでは，原則として採録対象サンプルに著作権を有するすべての著作権者に対して著作権使用許諾依頼を行い，使用許諾を得られたサンプルのみ採録することにしている。従って，一つのサンプルに複数の権利者がいる場合には，一人でも拒否すればそのサンプルは採録しない。実際には一つのサンプルに関わる著作権のあり方は多様であり，原則を徹底することは困難な場合も多い。そのため，以下のような方針をたてた。

○　サンプルに引用されている文言は，体裁から明確に引用と判定でき，かつ典拠の明確なもののみ著作権処理をする。

著作者 A が，法で許された範囲内で著作者 B の著作物を自著の中で引用するときは，A は B に著作権使用許諾を得る必要はないが，BCCWJ 構築プロジェクトという第三者が B からの引用を含む A の著作を使用する際は，B の許諾も得る義務が生じる。しかし，A が常に B の著作物についての典拠情報を明確に記載しているとは限らない。また，体裁から明確に引用と判定できない場合もある。われわれは被引用文献の特定について最大限の努力をするが，それでも判明しない場合は，引用部分についての著作権処理を行わない。
これに準ずるが，街頭インタビューでの発言など，著作権者が厳密に特定できない話し言葉の引用も，著作権処理を行わない。

○　著作権使用許諾の判断を，出版社等に依存する場合もある。

サンプル部分を執筆した著作者の権利を最優先し，その著作権者から許諾を得ることを原則とするが，出版社・編者・監修者などが個々の事例を判断し，著作者に代わって許諾を出す場合がある。例えば翻訳本などは，原著者・翻訳者・出版社・日本での出版を取り持つエージェントなど，複雑な権利関係が絡むことがある。このような時は，その特定のサンプルについて関係者間でなされている権利の調整を尊重する。その結果，原著者の許諾まで得る場合もあれば，出版社から一括して許諾を得る場合もある。いずれも，当プロジェクトとしては当該サンプルに責任のある立場での許諾を得たものと理解する。

なお，いずれの場合も，当該部分の著作権者から「使用を許諾していない」との指摘があれば，改めてその著作権処理を行う。

２． 作業の進め方と進捗状況

著作権処理は，サンプリング作業の次の作業として位置付けられている。現在，以下の 3 本の流れに分けてこの作業を進めている。

(1) サンプル箇所に関わる著作権を特定し，個別に処理

サンプル箇所が具体的に特定されたところで，その主たる著者，引用されている文献の著者，共著等の場合に誰がそこを執筆しているか，等，関連著作権をすべて特定し，個々の著作権者に許諾依頼をする。最も基本的な流れである。
なお，この際には，後述するように個人の住所等を調べることが困難であるため，著作権者に連絡をとるには出版社の協力を仰ぐことも多い。

(2) サンプル対象書籍が確定した時点で，許諾依頼

(1)との違いは，ある本の中で実際にどのページがサンプルに該当するか，といった情報が特定される前に，サンプリング対象書籍リストから，「サンプルに関わる著作権が本の著者のみであろうと推定される書籍（単著の小説など）」について，書籍名が判明した時点で著作権者に許諾依頼をするものである。個々のサンプリング作業を待たずに，許諾依頼作業が出来るという点で，作業の効率化が図れる。
なお，この方法で使用許諾を得た書籍であっても，実際のサンプリング時には改めて他に権利者がいないかどうか確認し，必要に応じて著作権処理を行う。

(3) 団体等の協力を得て，サンプル特定前に協力許諾依頼

(2)と同様に作業の効率化を図る方法であるが，ここではサンプル単位ではなく著作権者単位で事前に使用許諾を得られるため，より一層作業の効率化が図れるものと期待している。この流れについては，現時点では，以下の団体からそれぞれ記したような形で協力を得られることになっている。

○ 読売新聞社，毎日新聞社，産経新聞社
各社が著作権を持つ新聞記事等がサンプルに該当した場合は，自動的に採録可とする覚書を締結。

○ （社）日本文藝家協会,（社）推理作家協会,（社）日本児童文芸家協会,（社）日本児童文学者協会，日本ペンクラブの 5 作家団体共同による協力
各団体の会員（原則として散文の作家のみ）の著作物がサンプルに該当した場合に自動的に採録してよいかどうか，あらかじめ上記団体会員に諮る。

○ （株）ヤフー・ジャパン
「Yahoo!知恵袋」のデータの一括提供。

○ 衆議院・参議院
国会会議録のデータの一括使用許諾。

法律上は，著作権使用許諾は口頭で得てもよいことになっているが，本プロジェクトでは，国語研究所所長名で文書による許諾依頼を行い，著作権者へも書面の承諾書の返送をお願いしている。

上記の流れのうち，実際に著作権者に個別に連絡をとる作業に着手しているのは(1)のみであるが，その進捗状況は次の表のとおりである。

表　著作権処理の進捗状況（2006 年 12 月～2007 年 2 月）

サンプル数	1500 件
作家団体経由で一括許諾依頼予定のもの	507 件
個別に著作権処理の必要なもの	993 件
著作権者に連絡済	227 件
許諾	92 件
拒否	3 件
回答待ち	132 件
一部の著作権者に連絡済*	45 件
出版社に依頼予定**	163 件
連絡先調査中	558 件

*　複数の著作権者がいる場合に，その一部にのみ連絡ができているもの

**　個人の連絡先が判明しないため，出版社に転送等を依頼予定のもの

この他に，非母集団サブコーパスに含まれる白書のサンプル 1500 件についても，(1)の手順で著作権使用許諾を依頼しており，これまでに概ね許諾を得ることができている。

３． 問題点

これまでの著作権処理作業を通じて，最大の課題は個人の連絡先を突き止めることであることが強く認識された。これは個人情報保護法の施行による影響である。現在，われわれは，『日本紳士録』『著作権台帳』等の参考図書類を活用したり，インターネットで著者のホームページ等を探して E メールで連絡したりするなど手を尽くしているが，連絡先の特定作業は，なかなかはかどらない。そのため，各書籍の出版社に連絡し，著者への連絡の取り次ぎもお願いしている。出版社によっては 100 件を超える著作権者への連絡を依頼されることとなり，その負担は相当なものである。それにも関わらず，これまで多くの出版社に快く著作権者への連絡の取り次ぎをしていただいており，本プロジェクトの著作権処理は，こうした理解者の好意と協力に拠って成り立っているといっても過言ではない。

本プロジェクトでは，著作権処理も個人情報保護も，いずれも法の趣旨を尊重し，その許される範囲内で努力することを原則としている。しかしその結果，これらの法律が研究を進める上で大きな障壁となることが明らかになってきた。例えば，出版社に著者への手紙の転送を依頼するとして，出版社は著者の許諾なく転送を引き受けてはならない，というのも個人情報保護法の一つの厳格な解釈である。これでは正当な手続きを踏みたいと思っても，そのために権利者に連絡することが叶わないという矛盾が生じ兼ねない。本プロジェクトを通じて著作権処理に関わる法律・制度面での具体的な問題を明らかにすることも，今後の同様な研究活動の展開のために意義あることと考えている。

「現代日本語書き言葉均衡コーパス」における電子化フォーマットの概要

山口昌也† (国立国語研究所研究開発部門)
高田智和 (国立国語研究所研究開発部門)
北村雅則 (国立国語研究所研究開発部門)
間淵洋子 (国立国語研究所研究開発部門)
西部みちる (国立国語研究所研究開発部門)

Outline of Text Encoding Format in the Balanced Corpus of Contemporary Written Japanese

Masaya YAMAGUCHI† (Dept. Lang. Res., National Institute for Japanese Language)
TAKADA Tomokazu (Dept. Lang. Res., National Institute for Japanese Language)
Masanori KITAMURA (Dept. Lang. Res., National Institute for Japanese Language)
MABUCHI, Yoko (Dept. Lang. Res., National Institute for Japanese Language)
NISHIBE Michiru (Dept. Lang. Res., National Institute for Japanese Language)

1 はじめに

本稿では,「現代日本語書き言葉均衡コーパス」 (Balanced Corpus of Contemporary Japanese，以後，“BCCWJ” と表記) における電子化フォーマットの概要について述べる。

本電子化フォーマットは，BCCWJ のサンプリング基準によりサンプリングされた原資料を電子テキストに変換する際の形式を定めるものである。BCCWJ に収録される電子化テキストには，原資料に陽に記述されているテキストのほかに，書誌情報，文書構造情報，文字情報といった，さまざまな情報が XML のタグにより付与される。したがって，本電子化フォーマットが規定するのは，テキストの符号化形式，および，付与情報の記述形式ということになる。

本電子化フォーマットが記述対象として想定するテキスト，および，電子化されたテキストの利用分野は，次に示すとおりである。これらは，BCCWJ と同一である。

- 記述対象として想定するテキストは，現代日本語の書き言葉とし，1976 年以降の (主として) 出版物を対象とする。実際に想定しているのは，書籍，新聞，雑誌，白書，教科書，議事録，Web データなどである。
- 利用分野としては，言語学，国語教育，日本語教育，辞書編集，自然言語処理など幅広い分野での利用を想定する。

この後の節では，次の順序で，電子化フォーマットの概要を説明していくことにする。まず，次節で電子化フォーマットに対する要求分析を行い，その結果に基づいて，設計方針を決定する。次に，3 節で電子化フォーマットの仕様を規定するための XML タグセットを示す。そして，最後に 4 節で本稿のまとめを述べる。

† masaya@kokken.go.jp

2　電子化フォーマットの設計

2.1　電子化フォーマットに対する要求

ここでは，電子化フォーマットの仕様として，何が必要なのかを明確にするために，電子化するテキストの種類，利用方法，コーパスの規模，作成方法という四つの観点から，電子化フォーマットに対する要求分析を行う。

まず，電子化するテキストの種類の観点から要求を考える。BCCWJ の収録対象となる資料としては，書籍，雑誌，新聞，白書，教科書，議事録，Web データ (Yahoo!知恵袋[1]を予定) などが想定されている (山崎他 2006)。したがって，多様な文書構造を持ったテキストを扱う必要がある。例えば，小説のように，文書の階層構造が単純な資料もあれば，白書のように非常に深い階層構造を持った文書もある。さらに，雑誌の中には，図が多用され，レイアウトが複雑で，文書構造が不明確なものもある。このような文書構造上の多様性に加えて，テキストの特性や利用目的を活かすために，利用目的に特化した情報を付与しなければならないものもある。例えば，非母集団 (特定目的) サブコーパスのテキストは，個別の利用目的に対応できるような情報付与が必要になるだろう。以上のことから，次の要求を挙げる。

要求 1 多様な文書形式に対応できるようにすること
要求 2 利用目的に特化した情報付与に対応できるようにすること

次に，想定される利用方法を見てみよう。BCCWJ の利用分野としては，日本語学，日本語教育，国語教育，辞書編纂，自然言語処理などが挙げられている (山崎他 2006)。まず，すべての利用分野に共通して必要なことは，(1) テキストの文字が適切に符号化されていること，(2) 文字，文法，語彙，文体など言語学的な分析に役立つ文書要素に対して，適切にマークアップがなされ，容易に検索できることである。また，辞書編集のための用例収集のように，実際の用例を検索し，それを人間が詳細に分析するといった用途には，用例を理解しやすい形式で表示するための情報が付与されていることが望ましい。さらに，自然言語処理など，工学的な利用を考慮すると，汎用のツールで処理したり，他の言語資源と連係して利用できることが求められる。以上をまとめると，次のようになる。

要求 3 テキストを正確に符号化できること
要求 4 言語学的な分析に役立つ文書要素が適切にマークアップできること
要求 5 計算機処理に適した形式であること
要求 6 利用者が理解しやすい形式で電子化テキストを閲覧できること
要求 7 他の電子化フォーマットとの連係が取りやすいこと

最後に，コーパスの規模と作成方法の観点から考察する。まず，コーパスの規模は 1 億語で，開発期間は 5 年間と予定されている。また，電子化テキストの作成に際しては，Web データや議事録などの一部を除き，紙媒体からの入力を行う。これは，情報付与がまったくされていない状態から電子化することを意味し，(テキストの著者ではない) コーパスの作成者がテキストを解釈した上で，情報付与を行うことになる。したがって，本電子化フォーマットの利用者，つまり，コーパスの作成者とコーパスの利用者の共通理解を得やすいマークアップが必要であると考える。また，人手によるマークアップを行うことが予想されるため，量的にマークアップすることが可能な付与情報かどうかを考慮することも必要である。

要求 8 コーパス作成者，コーパス利用者の共通認識を得やすいマークアップであること
要求 9 人手で構築するのに，実現可能な量の付与情報であること

[1]利用者参加型の質問サイト。http://chiebukuro.yahoo.co.jp/

2.2 設計方針

前節で示した電子化フォーマットに対する要求のうち，電子化するテキスト，利用分野，利用者の多様性を鑑み，次の設計方針を立てた。

- 言語学，国語教育，日本語教育，辞書編集，自然言語処理などの幅広い分野への応用を想定した設計にする
- シンプルで，拡張性を考慮した仕様となるように設計する

これらの設計方針の下で，前節に示した要求に対して，次のように対処する。

- 文書中の論理的な役割が明確であり，かつ，紙面上の物理的な構造が明確な文書要素をマークアップの対象とする。
  - 二つの基準により文書構造が認定されるので，コーパス作成者・利用者の両者にとって共通理解を得やすい情報付与が可能になると考えられる ([**要求 8**])。また，論理的な役割が明確な文書要素がマークアップされるので，言語学的な分析に役立つ文書要素が適切にマークアップされることが期待できる ([**要求 4**])。
  - 論理的な構造ごとに閲覧時の表示形式を工夫し，電子テキストを利用者が理解しやすい形式で表示する (例えば，タイトルとしてマークアップされている場合は，フォントサイズを大きくするなど)。 ([**要求 6**])
- 収録対象の資料に含まれる文字を記述するのに十分な文字規格を採用する。また，ルビ，外字など，文字・表記に関するタグを用意する。 ([**要求 3**])
- 文書記述言語として，XML (eXtensible Markup Language) を用いる。XML は拡張性に優れた文書記述言語であり，多様な文書形式や利用目的に特化した情報付与に対応しやすい ([**要求 1,2**])。また，TEI (Text Encoding Initiative) をはじめとして，多くのコーパスや電子化フォーマットで採用されており,『太陽コーパス』(国立国語研究所 2005) や『日本語話し言葉コーパス』(国立国語研究所 2006) も XML を用いて記述されている。したがって，これらのデータとの整合性も高い。また，XML は，コーパスの記述だけでなく，データ一般の記述に広く用いられており，データ形式の検証，変換，検索などを行う際に，既存のツールを利用できるという利点もある。([**要求 5,7**])
- 量的な観点から，人手でマークアップすることが困難な場合は，自動的，もしくは，半自動的なマークアップを検討する。([**要求 9**])

3 電子化フォーマットの仕様

3.1 概要

本電子化フォーマットの概要は，次のとおりである。

文書記述言語： XML
文字符号化方式： UTF-16
文字集合： JISX0213:2004

BCCWJ の電子化テキストは XML で記述する。電子化フォーマットは，XML の文書型によって規定する。BCCWJ には，一つのサンプルが一つの「記事」に相当する可変長サンプルと，一つのサンプルに 1000 文字を包含する固定長サンプルがある。したがって，2 種類の文書型を定義する。

文字符号化方式は UTF-16 を，文字集合には JISX0213:2004 を採用した。JISX0213:2004 に含まれる文字数は，約 11000 字である。JISX0213:2004 には，現在最も一般的に利用されている JISX0208 の約 6800 字に，第 3，4 水準漢字・非漢字，約 4000 字が追加されている。

JISX0208 ではなく，JISX0213:2004 を採用したのは，(a) 現時点の国内規格では，最も大きな文字集合を持つこと，(b) 印刷字体を考慮した包接基準を持つこと，(c) 他のコーパスとの連係を考慮したこと，などが挙げられる。(a)(b) は，正確な文字の符号化に寄与すると期待される。(c) の例としては，BCCWJ に収録されているものよりも古い時代の資料[2] や，今後発展の見込まれる電子データ[3]がある。

3.2 タグの仕様

本電子化フォーマットでは，56 種類の XML タグを定義した。タグの一覧を表 1 に示す (スペースの関係上，一部のみ)。また，本電子化フォーマットで電子化テキストに変換した例を図 1 に示す。定義した XML のタグは，付与される情報として，付与される情報は，次の三つに大別される。詳細は，この後の節で随時説明する。

- サンプルに関するタグ
- 文字・表記に関するタグ
- 文書構造に関するタグ

3.2.1 サンプルに関するタグ

サンプルに関するタグには，sample と sampling (表 1 参照) がある。sample 要素[4]は，一つのサンプルを表す。sampling タグは，サンプリングポイント (サンプリング時にランダム抽出された文字。詳細は，(丸山 2007 を参照)) を表す。

sample タグには，サンプルに関する情報が属性として記述されている。sampleID 属性値は，サンプル固有の識別番号である。サンプルの書誌情報は，sampleID をキーとして，書誌情報のデータベースを参照する。書誌情報としては，書名，著者，出版社などが提供される予定である。

sample タグの type 属性は，サンプルの種別 (固定長，可変長) を表す。図 1 では，type 属性が “variableLength” となっているので，可変長のサンプルであることがわかる。一方，固定長の場合は，属性値が “fixedLenghth” となる。

3.2.2 文字・表記に関するタグ

文字・表記に関するタグの役割は，二つある。一つは，検索や計算機処理の利便性を高めることである。この役割を持つタグに correction タグがある。このタグは，原文の誤植を訂正した文字であることを表す。次の例は，誤字，脱字，衍字を修正した例である。タグでマークアップした結果が修正結果となるので，誤りを意識せずに，検索したり，計算機処理を行うことができる。修正前の文字は，originalText 属性として保持される。

```
生活基<correction type="erratum" originalText="盟">盤</correction>に
伸びを示し<correction type="omission">て</correction>いる
整備を<correction type="excess" originalText="を" />図るべく
```

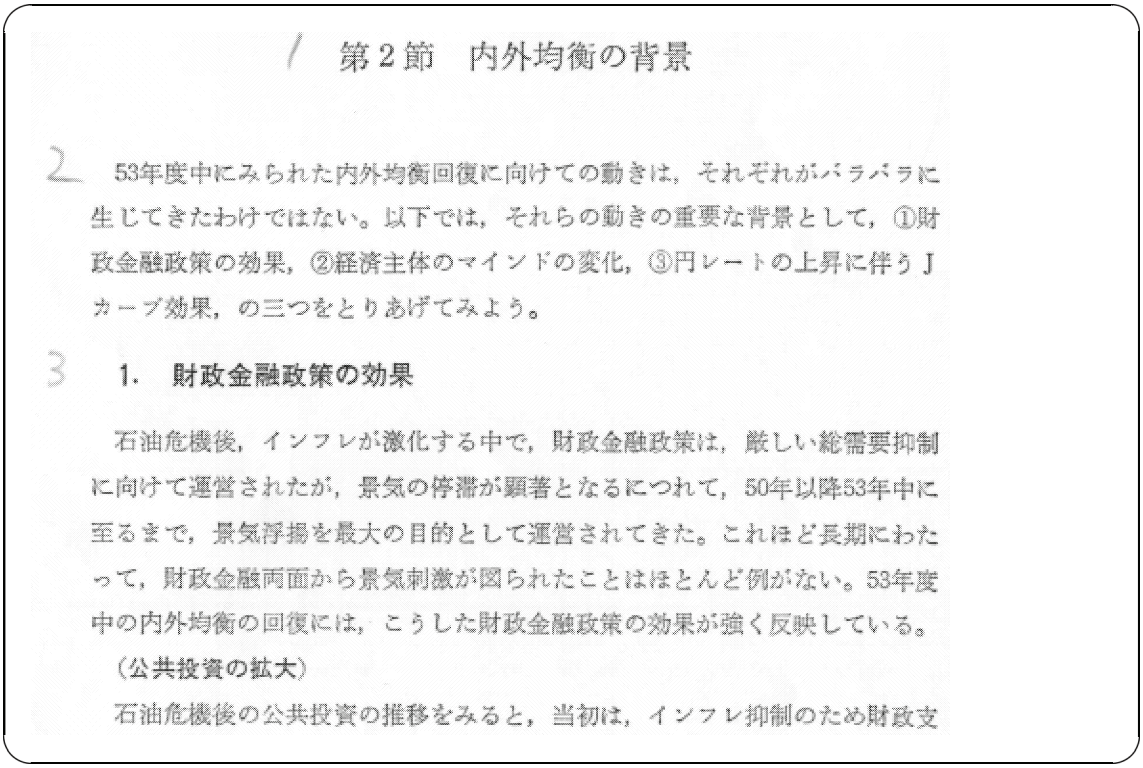
[2]例えば,『太陽コーパス』

[3]PC 用の OS として，現在最も普及している Windows の新バージョンも JISX0213 を採用しているため，JISX0213 で符号化したデータが流通する可能性がある

[4]sample タグでマークアップされている文書要素

表 1: タグ一覧 (一部)

	タグ名	内容
サンプル	sample	サンプリングによって 1 サンプルとされた文書要素
	sampling	サンプリングポイントに関する情報
階層構造 (文書構造)	article	同一著者による，同一テーマのひとまとまりの文書要素
	body	article 要素が対象としている原資料自体
	hierarchy	article 要素の上位階層の構造
	blockEnd	意味のまとまりや形式のまとまりを区切るためのマーカー
	cluster	title 要素が包括する文書要素全体
	titleBlock	title 要素とそれに付随する要素全体
	title	特定範囲の文書要素の内容を代表する記述
	list	箇条書きや名詞句の羅列など，列挙された文書要素の集まり
	paragraph	段落を表す文書要素
	sentence	文に相当する文書要素
図表 (文書構造)	figureBlock	図表・写真・絵などの要素と，それに付随する文書要素をまとめた要素
	figure	図・表・写真・絵など
	figureTitle	図表・写真・絵によって表そうとする事柄を端的に示す「標識」
	caption	図表についての タイトルや説明
	table	表
引用 (文書構造)	citation	当該 article 要素の本文において言及される，他文献からの引用要素
	source	引用文献についての情報 (文献名，著者名，著者情報など)
	speech	発話の引用・書き起こし，心内発話の描写
	speaker	話者を明示的に表した文字列やマーク
	stage	発話の状況，発話者の動作に関する説明・注等
	stageInline	speech 要素内の発話および発話者表示以外の文書要素
	quote	当該 article 要素とは異なる著作物からの引用や，発話・心内発話の引用・描写・書き起こし
注記 (文書構造)	note	注記とその注記の範囲
	noteBodyInline	傍注など行外に付随する形式で現れる注記
その他 (文書構造)	abstract	article 要素，または cluster 要素の概要に相当する文書要素
	authorsData	著作者表示・署名にあたる要素
	contents	目次に対応する文書要素
	profile	著者や登場人物のプロフィールに相当する文書要素
	rejectedBlock	サンプル範囲内において，削除対象となったブロック要素の存在
	verse	詩，和歌，俳句，歌謡などの韻文
文字・表記	ruby	ルビ付き文字
	correction	原文の誤植を訂正した文字
	missingCharacter	JIS X 0213:2004 で規定されている文字以外の文字 (JIS 外字)
	enclosedCharacter	連続や参照などのラベルとして機能している囲み付きの文字
	image	JIS X 0213:2004 が規定する諸記号に含まれていない記号類や絵文字
	cursive	変体仮名
	superScript	数式や化学式などに用いる上付きの文字
	subScript	数式や化学式などに用いる下付きの文字
	fraction	帯分数の中の真分数部分
	delete	抹消線などによって削除された本文要素
	insert	新たに挿入された本文要素
	br	物理改行
	rejectedSpan	サンプル範囲内において，削除対象となったインライン要素の存在



第２節　内外均衡の背景

53年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。以下では，それらの動きの重要な背景として，①財政金融政策の効果，②経済主体のマインドの変化，③円レートの上昇に伴うＪカーブ効果，の三つをとりあげてみよう。

1.　財政金融政策の効果

石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，50年以降53年中に至るまで，景気浮揚を最大の目的として運営されてきた。これほど長期にわたって，財政金融両面から景気刺激が図られたことはほとんど例がない。53年度中の内外均衡の回復には，こうした財政金融政策の効果が強く反映している。

（公共投資の拡大）

石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支

```
<?xml version="1.0" encoding="UTF-16" ?>
<?xml-stylesheet href="sc_check.xsl" type="text/xsl" ?>
<sample sampleID="OW1X_00000" version="20070208" type="variableLength">
<article articleID="OW1X_00000_xxxx" isWholeArticle="false">
<hierarchy>
<titleBlock><title>第１部　内外均衡に向かった昭和５３年度経済</title></titleBlock>
<titleBlock>
  <title>第１章　昭和５３年度の日本経済</title>
  －その推移と特徴－
</titleBlock>
</hierarchy>
<body>
<titleBlock><title>第２節　内外均衡の背景</title></titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。</sentence><sentence>以下では，それらの動きの重要な背景として，...
</paragraph>
<cluster>
<titleBlock><title>１．財政金融政策の効果</title></titleBlock>
<paragraph>
<sentence>　石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，５０年以降５３年中に至るまで，景気浮揚を最大の目的として運営されてきた。</sentence> ...
</paragraph>
<cluster>
<titleBlock><title>（公共投資の拡大）</title></titleBlock>
<paragraph>
<sentence>　石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支出が抑制され，公共事業の伸びは低いものにとどまっていた。</sentence>
```

図 1: 原資料とその電子化テキストの例 (経済白書 昭和 54 年版から引用)

もう一つの役割は，原資料に忠実に電子化テキストを記述することである。この役割を持つタグの例として，ruby，missingCharacter タグの例を次に示す。ruby タグはルビ付き文字を表す。JISX0213:2004 で規定されているいない文字は**=**で代替され，missingCharacter でマークアップされる。missingCharacter タグは，属性として，文字種を表す attribute 属性，Unicode 番号を保持する unicode 属性,『大漢和辞典』の親番号を表す daikanwa 属性，字体記述を行うための description 属性を持つ。

```
<ruby rubyText="ご">語</ruby><ruby rubyText="い">彙</ruby>
<missingChar attribute="HanIdeograph" unicode="U+5AEB"
             daikanwa="M06673" description="女偏に莫">=</missingChar>
```

3.2.3 文書構造に関するタグ

文書構造に関するタグは，論理的な役割が明確な文書要素に対して付与される。表 1 に示したとおり，この種のタグは，(a) 階層構造，(b) 図表，(c) 引用，(d) 注記，(e) その他，に分けられる。本稿では，このうち階層構造に関するタグを中心に説明する。

階層構造に関するタグは，article を最上位の階層として，cluster，paragraph，sentence といった言語的な階層構造を表現する。図 1 から，これらの要素に関係する部分を取り出すと次のようになる。なお，字下げは，下位の階層であることを示す。例えば，図 1 の article 要素直下の階層には，hierarchy と body 要素があることがわかる。

```
article
  hierarchy
  body
     titleBlock          第２節　内外均衡の背景
     cluster
       titleBlock        １．財政金融政策の効果
       cluster
         titleBlock      （公共投資の拡大）
```

article　article 要素は「記事」を想定した要素で,「同一著者による，同一テーマのひとまとまりの文書要素」を表す。なお，BCCWJ では，一つの article 要素に含まれる文字数の上限が 1 万字ということになっているため，必ずしも,「同一著者による，同一テーマのひとまとまりの文書要素」すべてを収録できるとは限らない。例えば，図 1 の白書のサンプルは，1 章 2 節だけしか収録していない。このような場合,「記事」全体を収録できたか否かを表す isWholeArticle 属性は，“false” となる。

hierarchy　hierarchy 要素は，当該の article 要素の上位の階層構造を格納し，当該の article 要素が階層構造上どのような位置づけにあるかを表す。この要素の役割は，収録文字数の制限により,「記事」全体を収録できなかった場合に，当該の article 要素が文書構造中どこに位置づけを明確にすることである。例えば，図 1 の article 要素の場合は，第 2 節からしか取得できていないが，hierarchy 要素に記述されている，部と章のタイトルを参照することにより，この article 要素の位置付けを把握することができる。

cluster　cluster 要素は，章，節といったように，タイトル (titleBlock 要素) を持った，ひとまとまりの文書要素を表す。cluster 要素自体には，章，節といった特定の階層を表すための意味づけを行っていないが，入れ子構造により，階層の上下を表す。例えば，上記の「(公共投資の拡大)」というタイトルを持つ cluster 要素は， 2.1 節に対応する cluster 要

素の子要素なることで，2.1 節の下位構造であることを表現する。なお，cluster には必ず titleBlock が含まれる。この制約を課すことにより，紙面上のデザインなどの物理的な特徴に基づいて，cluster が過度に認定されるのを防ぐことができる。

titleBlock　すでに述べたように，titleBlock 要素は，cluster 要素のタイトルとそれに付随する部分からなる文書要素である。タイトルとその付随部分は，title 要素により，明示的にマークアップされているので，容易にタイトルだけを検索したり，抽出したりすることが可能である。

paragraph, sentence　それぞれ，段落，文に相当する要素である。これらの要素は，テキスト中に大量に含まれるため，人手でタグを付与することは困難である。そこで，paragraph は行頭の空白，sentence は句点などを手がかりに，自動的にタグを付与している。

3.3　他の電子化フォーマットとの関係

テキストを電子的に記述するための形式としては，従来から，TEI や CES (Corpus Encoding Standard) などが提案されている。BCCWJ で新たに電子化フォーマットを策定したのは，次の理由による。まず，TEI は，汎用の電子化フォーマットであるため，仕様が複雑であり，BCCWJ の規模，実施期間を考慮すると，実際に実装するのは困難である。一方，CES は TEI よりもシンプルな仕様であるが，適用範囲として，言語工学やその応用を指向しており，言語学的な分析と工学的な利用の双方を視野に入れた BCCWJ に CES をそのまま適合することは難しい。

それに対して，BCCWJ の電子化フォーマットは，言語学から工学という多様な利用分野を想定しつつ，記述対象のテキストを現代日本語の書き言葉に限定することにより，シンプルで，実際に運用可能なフォーマットを実現するものである。

4　おわりに

本稿では，BCCWJ における電子化フォーマットの仕様について概要を説明した。我々は本仕様に基づいて，これまで，白書のサンプル (1500 サンプル) を電子テキストに変換した。今後，書籍，新聞など白書以外の資料に対して，本電子化フォーマットを適用するとともに，随時仕様を修正していく予定である。なお，本仕様については，Web 上[5] で一般に公開する予定である。詳細な内容については，そちらを御覧いただきたい。

参考文献

Text Encoding Initiative, The XML Version of the TEI Guidelines, http://www.tei-c.org/P4X/index.html

Corpus Encoding Standard, http://www.cs.vassar.edu/CES/

山崎 誠，丸山岳彦，柏野和佳子，他 (2006)「現代書き言葉均衡コーパスの現状」，特定領域「日本語コーパス」平成 18 年度全体会議予稿集，pp.9–16

丸山岳彦，柏野和佳子，山崎誠，他 (2007)「「現代日本語書き言葉均衡コーパス」におけるサンプリングの概要」,「日本語コーパス」平成 18 年度公開ワークショップ予稿集

国立国語研究所 (2005)『太陽コーパス』(国語研究所資料集 15)，博文館新社

国立国語研究所 (2006)『日本語話し言葉コーパスの構築』(国語研究所報告書 124)，国立国語研究所

[5] http://www2.kokken.go.jp/densi/public/wiki/

『現代日本語書き言葉均衡コーパス』における短単位の概要

小椋秀樹　（国立国語研究所研究開発部門）
小木曽智信（国立国語研究所研究開発部門）
小磯花絵　（国立国語研究所研究開発部門）
冨士池優美（国立国語研究所研究開発部門）
相馬さつき（国立国語研究所研究開発部門）
渡部涼子　（国立国語研究所研究開発部門）
服部龍太郎（国立国語研究所研究開発部門）

Outline of short-unit word in the Balanced Corpus of Contemporary Written Japanese

Hideki Ogura　(Dept. Lang. Res., National Institute for Japanese Language)
Toshinobu Ogiso　(Dept. Lang. Res., National Institute for Japanese Language)
Hanae Koiso　(Dept. Lang. Res., National Institute for Japanese Language)
Yumi Fujiike　(Dept. Lang. Res., National Institute for Japanese Language)
Satsuki Souma　(Dept. Lang. Res., National Institute for Japanese Language)
Ryoko Watanabe　(Dept. Lang. Res., National Institute for Japanese Language)
Ryutaro Hattori　(Dept. Lang. Res., National Institute for Japanese Language)

１．はじめに

国立国語研究所が構築を進めている『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese，以下 BCCWJ）には，国語学や国語施策・情報工学をはじめとする幅広い分野での活用を目指して，様々な研究用の付加情報を与える。このうち形態論情報については，言語単位として，コーパスからの用例収集に適した「短単位」と BCCWJ に格納したサンプルの言語的特徴の解明に適した「長単位」の2種類を採用した。この2種類の言語単位に基づいて，更に代表形・品詞等の情報を与える。

本稿では，BCCWJ における言語単位の設計方針，長短2種類の言語単位のうち短単位の認定規定等について述べる。また併せて，形態論情報付与作業の進捗状況，BCCWJ に格納する中央省庁刊行の白書のデータを解析した結果も報告する。

２．BCCWJにおける言語単位の設計

語をどのように定義するかについては研究者によって様々な立場がある。そのため，コーパスの言語単位をどのように規定するかについても様々な立場があり，容易に決めることはできない。

BCCWJ の言語単位の設計に当たっては，まず BCCWJ を日本語研究に利用するために，どのような言語単位が必要か整理した。その上で設計方針を立て，その方針に基づいて単位を設計した。

２．１　言語単位の設計方針

我々は，BCCWJ の言語単位の設計方針として，次の三つを掲げた。

方針１：コーパスに基づく用例収集，各ジャンルの言語的特徴の解明に適した言語単位を設計する。

コーパスの日本語研究への活用としてまず考えられるのは，コーパスから用例を集める

ことである。そのため，BCCWJ を日本語研究で幅広く利用できるようにするには，用例収集に適した言語単位を設計する必要がある。また BCCWJ は，新聞・雑誌・書籍といった複数の媒体を対象としたコーパスであり，内容も政治・経済・自然科学・文芸等と多岐にわたっている。このような BCCWJ の構成から，媒体別・分野別の言語的な特徴を明らかにしていくことが重要な研究テーマになると考えられる。したがって，そのような分析に適した言語単位を設計することが必要になる。

方針２：「日本語話し言葉コーパス」と互換性のある形態論情報を設計する。

国立国語研究所が既に構築したコーパスとして，現代の話し言葉を対象とした「日本語話し言葉コーパス」（Corpus of Spontaneous Japanese，以下 CSJ）がある。BCCWJ と CSJ とは，国立国語研究所が進めているコーパス整備計画において，共に現代語のコーパスとして位置付けられる。BCCWJ の完成後には，これら二つのコーパスを使って，現代語の書き言葉と話し言葉の比較等の研究を行うことが考えられる。そのため，BCCWJ と CSJ とを統一的に扱うことのできる，互換性を持った言語単位を設計する必要がある。

方針３：国立国語研究所の語彙調査における知見を活用する。

国立国語研究所は，1949年の『語彙調査 ―現代新聞用語の一例―』以来，合計10回の語彙調査を実施した。その中で，調査単位（言語単位）の設計や言語事象の処理に関して，様々な知見を蓄積している。そこで，BCCWJ の言語単位の設計や単位認定の際に，これら語彙調査の知見を活用していく。語彙調査の結果は，日本語研究でも様々に活用されており，言語単位の設計等に語彙調査の知見を活用していくことは，BCCWJ を使った日本語研究を進めていくためにも有用であると考えられる。

２．２　BCCWJの言語単位

以上の方針の下，BCCWJ の言語単位について検討した結果，次のような結論を得た。

BCCWJ の言語単位には，方針1で挙げた，用例収集・各ジャンルの言語的特徴の解明という二つの利用目的に応じて，次に示す2種類を採用する。

(1) 用例収集を目的とした短単位

(2) 言語的特徴の解明を目的とした長単位

この短単位・長単位は，いずれも CSJ で採用した言語単位である[1]。また短単位は国立国語研究所が行った現代雑誌九十種調査の β 単位[2]を，長単位はテレビ放送の語彙調査の長い単位[3]を基に設計したものである。このようにして，CSJ との互換性の保持と，国立国語研究所の持つ語彙調査の知見の活用とを図る。

なお，短単位・長単位の認定規定は，CSJ の規定をそのまま用いるのではなく，部分的に改定することがある[4]。このような場合，CSJ の形態論情報の修正も行い，BCCWJ の公開に合わせて，修正した CSJ の短単位・長単位データを公開することを考えている。

３．短単位の概要

短単位は，言語の形態的側面に着目して規定した言語単位である。短単位の認定に当たっては，まず現代語において意味を持つ最小の単位（以下，最小単位）を規定する。その上で，最小単位を文節の範囲内で短単位の認定規定に基づいて結合させる（又は結合させない）ことにより，短単位を認定する。

以下，最小単位の認定規定，短単位の認定規定，短単位の付加情報の概略及びコーパスの言語単位としての短単位の長所について述べる。

３．１　最小単位の認定規定

最小単位は，現代語において意味を持つ最小の単位であり，和語・漢語・外来語・記号・人名・地名の種類ごとに，次のように認定する。

和　語　：　/豊か/な/暮らし/に/つい/て/　　/大/雨/が/降っ/た/の/で/

漢　語 ： /国/語/　　/研/究/所/
外来語 ： /コール/センター/　　/オレンジ/色/
人　名 ： /星野/仙一/　　/ジェフ/・/ウィリアムス/　　/林/威助/
地　名 ： /大阪/府/豊中/市/待兼山町/　　/六甲/山/　　/琵琶/湖/
記　号 ： /図/Ａ/　　/ＪＲ/

上記のように認定した最小単位を短単位認定の必要上，表１のように分類する。

表１．最小単位の分類

分　類		例
一　　般		和　語 ： 豊か　大　雨 … 漢　語 ： 国　語　研　究　所 … 外来語 ： コール　センター　オレンジ …
数		一　二　十　百　千 …
その他	付 属 要 素	接頭的要素　：　相　御　各 … 接尾的要素　：　兼ねる　がたい　的 …
	助詞・助動詞	う　だ　ます　か　から　て　の …
	人名・地名	星野　仙一　大阪　六甲 …
	記　　号	Ａ　Ｂ　ω　イ　ロ　ア　ＪＲ …

上記の分類のうち「付属要素」とは，接頭辞・接尾辞・補助用言のことである。ただし，すべての接頭辞・接尾辞・補助用言を付属要素に分類するわけではない。現代雑誌九十種調査や CSJ に出現したものの中から造語力が高いなど注目されるものを付属要素に分類している。今後，BCCWJ に出現した接頭辞・接尾辞・補助用言からも，造語力が高いものなどを追加していく予定である。

なお，最小単位は短単位認定のために必要な概念として規定するものである。そのため，BCCWJ のサンプルを最小単位に分割することはしない。

３．２　短単位の認定規定

短単位の認定規定は，表１の分類ごとに適用すべき規定が定められている。その規定に基づいて最小単位を結合させる（又は結合させない）ことにより，短単位を認定する。なお，最小単位を結合させる際には，文節境界を超えないという制約を設け，文節と短単位とが階層構造を持つようにしている。

以下，「一般」・「数」・その他に分けて，短単位認定規定の概略を示す。

［１］一般

《原則》

（１）和語・漢語は，2最小単位の1次結合体を1短単位とする。

｜母=親｜　　｜食べ=歩く｜　　｜言=語｜資=源｜　　｜研=究｜所｜　　｜本=箱｜作り｜

（２）外来語は，1最小単位を1短単位とする。

｜コール｜センター｜　　｜オレンジ｜色｜

《例外規定》

（１）省略された外来語の最小単位の扱い

① 省略された外来語の最小単位は，和語・漢語の最小単位と同様に扱う。

｜パソ=コン｜　　｜塩=ビ｜　　｜ピン=ぼけ｜

② 省略された外来語の最小単位と省略されていない外来語の最小単位との1次結合体は1短単位とする。

｜エア=コン｜　　｜マス=コミ｜

（２）1最小単位を1短単位とするもの

① 最小単位が3個以上並列した場合の各最小単位。
|衣|食|住|　　|松|竹|梅|　　|都|道|府|県|
② 類概念を表す部分と名を表す部分とが結合してできた固有名詞のうち，類概念を表す部分と名を表す部分とが共に1最小単位の場合の，それぞれの最小単位。
|さくら|屋|　　|歌舞伎|座|　　|のぞみ|号|

（3）最小単位の3個以上の結合体を1短単位とするもの
① 3個以上の最小単位からなる組織名等の略称。
|日経連|　　|通総研|
② 切る位置が明確でないもの，あるいは切った場合と一まとめにした場合とで意味にずれがあるもの。
大統領		不可解		明後日		殺風景
輸出入		国内外		原水爆		市町村長
大袈裟		大丈夫		二枚目		十八番
ただし二つ以上の漢語の最小単位が並列して，1短単位と結合している場合は，次のように短単位を認定する。
中|小|企業|　　|小|中|学校|　　|都|道|府|県|知事

［2］数

「数」以外の最小単位と結合させない。「数」どうしの結合は，一・十・百・千のとなえを取る桁ごとに1短単位とする。「万」「億」「兆」などの最小単位は，それだけで1短単位とする。小数部分は1最小単位を1短単位とする。

|十|二|月|二十|三|日|　　|七百|五十|二|万|語|　　|五|分|の|二|
|二三十|回|　　|〇|.|四|五|

［3］その他

1最小単位を1短単位とする。

付属要素　：|筒|状|　|扱い|兼ねる|
助詞･助動詞　：|豊か|な|暮らし|に|つい|て|
人名　：|星野|仙一|　　|ジェフ|・|ウィリアムス|　　|林|威助|
地名　：|大阪|府|豊中|市|待兼山町|　|六甲|山|　|琵琶|湖|
記号　：|図|Ａ|　|ＪＲ|

３．３　付加情報

3.2節に示した規定によって認定した短単位には，次に挙げる情報を与える。

代表形　　代表表記　　品詞　　活用型　　活用形

代表形は国語辞典の見出しに，代表表記はその見出しに与えた漢字等の表記に相当するものである。

品詞・活用型・活用形は，CSJ を基に，BCCWJ の短単位解析に用いる解析用辞書 unidic の品詞・活用型・活用形を参考にして細分化を行った[5]。例えば品詞は，CSJ の品詞を基に，次の16種類に分類した。これは学校文法に準ずるものである。

名詞　　代名詞　　形状詞　　連体詞　　副詞　　接続詞　　感動詞　　動詞
形容詞　　助動詞　　助詞　　接頭辞　　接尾辞　　記号　　補助記号　　空白

さらに，これらの品詞を unidic を参考にして用法等の観点から細分化した。例えば名詞は，次の11種類に細分化した。

普通名詞　　サ変可能　　数詞　　助動詞語幹

固有名詞-一般　　固有名詞-人名　　固有名詞-姓　　固有名詞-名
固有名詞-国　　固有名詞-地名　　固有名詞-組織名

３．４　短単位の長所

ここでは，短単位がコーパスの言語単位として，どのような長所を持つのかについて述べておく。短単位の長所としては，次の2点が挙げられる。

長所１：基準が分かりやすく，ゆれが少ない。

これは，短単位の基礎となる最小単位の認定に当たり，個人によってとらえ方に幅のある要素を基準に持ち込んでいないことによる。

なお，基準が分かりやすく，ゆれが少ないという短単位の長所は，作業効率の向上につながるだけでなく，コーパスの使いやすさにもつながる。基準が分かりやすければ，利用者が語を検索する際，どのように検索条件を指定すればよいか迷うことが少なくなる。また，ゆれの少なさ，つまりデータの精度の高さは，分析結果の確かさにもつながる。

長所２：取り出した単位が文脈から離れすぎない。

上で短単位はゆれが少ない単位であると述べたが，実は最もゆれが少ない単位は，短単位ではなく，その基礎となっている最小単位である。それにもかかわらず，最小単位を言語単位として採用しなかったのは，最小単位は文脈から離れすぎるため，日本語の研究に使いにくいからである。

例えば，短単位「気持ち」は「気」と「持ち」の二つの最小単位に分割することができる。もしこのような最小単位でコーパスが解析されていると，動詞「持つ」を検索した際に，「荷物を持つ」などの「持つ」とともに，「気持ち」の「持ち」も検索結果として得られることになる。

しかし，動詞「持つ」の分析を行う際に，「気持ち」の「持ち」まで検索結果に含まれるのは望ましいとは言い難い。それは，実際の文脈の中では，動詞「持つ」として機能していないからである。したがって，コーパスから用例を収集し，分析することを考えた場合，正確に単位認定ができるとしても，最小単位のような単位では問題が多いということになる。

このように考えた場合，短単位は，基準の分かりやすさ・ゆれの少なさという条件を満たしつつ，用例を収集して分析を行うという利用目的にもかなう単位と言える。

４．形態論情報付与作業の進捗状況

2006年度に，国立国語研究所が特定領域研究・電子化辞書班と共同で実施した形態論情報付与作業の進捗状況，BCCWJ に格納する中央省庁刊行の白書のデータ（約500万語）を解析した結果について報告する。

４．１　解析用辞書の整備・拡充

BCCWJ の短単位解析は，自動解析システムを使って行うことにしており，自動解析エンジンには「茶筌」，解析用辞書には unidic を使う。unidic で採用している単位は，短単位とほとんど一致しており，品詞体系も BCCWJ の品詞体系と互換性がある。

自動解析システムを利用することから，解析精度の向上に最も深くかかわる解析用辞書の整備・拡充作業を，2006年度に実施した。この作業では，千葉大学の伝康晴氏が中心になって構築した unidic（見出し語：約46,000語，2006年2月時点）を基に，国語辞典や国立国語研究所の語彙調査等を基に作成されたデータから，unidic にない見出し語を，短単位の認定規定に基づいて分割した上で，順次登録した。その結果，unidic の見出し語は2007年2月末時点で約106,000語になった。

４．２　作業用ツールの作成

一般に，自動解析用の辞書では，語形や表記が異なれば別語として扱うという方針が取

られている。例えば，《オコナウ》という語が「行う」「行なう」「おこなう」といった表記形で出現した場合，自動解析用の辞書では，これら三つの表記形を別語として扱う。また，《トテモ》という語が「トテモ」のほか「トッテモ」という語形で出現した場合も，二つの語形を別語として扱う。

これに対して unidic では，語形や表記が異なっても同じ語であれば，一つの見出し語にまとめるという方針を取っている。つまり，「行う」「行なう」「おこなう」という各表記形は《オコナウ》という語として，「トテモ」と「トッテモ」という各語形は《トテモ》という語としてまとめられる（表２参照）。なお，表２に示したように unidic では，見出し語《オコナウ》に当たるものを「語彙素」，出現した語形（活用語は終止形）に当たるものを「語形」，出現した表記形（活用語は終止形）に当たるものを「書字形」と呼ぶ[6]。

表２．unidic の階層構造の例

語彙素	語形	書字形
オコナウ【行う】	オコナウ	行う
		行なう
		おこなう
トテモ【とても】	トテモ	とても
	トッテモ	とっても

このような unidic の階層構造を保ちながら，辞書データを効率的に拡張・修正・管理できるようにするため，電子化辞書班と共同でマイクロソフト SQL Server 2005 を用いてデータベース（短単位データベース）を構築し，unidic の辞書データを登録した。見出し語の新規追加や見出し語の整理等を行う場合は，解析用辞書にではなく，この短単位データベースに対して追加・修正の作業を行っている。

この追加・修正作業の際には，表２のような階層構造を保ちながら作業する必要がある。そのために，作業者が階層構造を視覚的に把握して，見出し語の新規登録や既登録語の修正を行うことができるような辞書管理ツール（Unidic Explorer）を作成し，利用している。

自動解析で高い解析精度を実現するには，辞書の整備とともに，コーパス修正も重要な作業となる。この修正作業の際には，単位や品詞等の情報が解析用辞書と食い違うことのないよう，コーパスを修正する必要がある。そこで，短単位データベースと連携させてコーパス修正を行うためのツールを開発した。このツールでは，コーパス修正を行う際には，必ず短単位データベースを参照し，データベースから品詞等の情報を取得するようになっている。もしコーパスに付与すべき情報が短単位データベースに登録されていない場合は，データベースにその情報を新規登録した後，コーパス修正を行う。

４．３　白書データの解析精度

BCCWJ に格納する中央省庁刊行の白書のデータ（約500万語）を，2006年10月時点の unidic（見出し語・約10万語）を使って解析した。その解析結果から，数詞を除く自立語1万語を無作為抽出し，短単位境界の認定，代表形・品詞・活用型・活用形の情報付与が正しく行われているかを調べた。結果は，誤解析が536例で，精度は94.64％であった。この536例の誤解析を短単位境界に関する誤りなどの5種類に分けて，その数を表３に示した。この表から，誤解析の約半数が短単位境界に関する誤りであることが分かる。

表３．白書データの誤解析の分類

短単位境界	代表形	品詞	活用型	活用形	合計
253	163	112	0	8	536

以下，短単位境界の誤りについて，要因は何か，具体的にどのような事例があるのか見ていく。

４．３．１　未登録語（語彙素）に起因する誤解析

短単位境界の誤りのうち，unidic に正解の短単位（語彙素）が登録されていなかったことに起因するものは115例ある。このうち，未登録の漢語が構成要素に分割される誤り，特に短単位の規定から漢語の大半を占めている2字漢語が1字ずつに分割される誤りが54例あり，ほぼ半数を占めている。例えば，「各年」「同法」を「|各|年|」「|同|法|」と誤解析した例がある。

漢語に関しては，上記のほかに，複数（特に2語）の漢語を1語に縮約した語を誤解析したものがある。複数の漢語を1語に縮約した語というのは，「教員」と「職員」とを1語に縮約した「教職員」，「中企業」と「小企業」とを1語に縮約した「中小企業」などの語である。3.2節「短単位の認定規定」の[1]一般の例外規定(3)②にあるとおり，前者は分割せずに全体を1短単位とし，後者は「|中|小|企業|」のように分割する。これらの語が，解析用辞書に登録されていない場合，「教職員」は「|教|職員|」のように正解よりも短く分割される。また「中小企業」は「|中小|企業|」と2短単位に分割される。

これらの縮約語は，元々 unidic に余り登録されていない上に，白書には，例えば「給貸与」「政省令」「凍雪害」など，分野に特有の（あるいは臨時的な）縮約語が多く見られる。そのため，これら縮約語が白書に出現した語全体に占める割合は余り高くないにもかかわらず，未登録語に起因する誤解析例の少なからぬ割合を占める結果となっている。

外来語に関しては，その一部を既登録の語として認定することによって，正解より短く分割するという誤りが36例見られた。例えば，経済用語「スタグフレーション」を「|ス|タグ|フレーション|」，外国人名「アーメド」を「|アー|メド|」と誤解析した例がある。前者はその一部を既登録の「タグ」（tag）としたことによる誤り，後者はその一部を既登録の「メド」（目処の片仮名表記）としたことによる誤りである。

４．３．２　未登録の表記形（書字形）に起因する誤解析

4.2節で述べたように，unidic では，表記が異なっても同じ語であれば，一つの見出し語にまとめるという方針を取り，語を階層化した形で登録している。短単位境界の誤解析の中には，正解の短単位は登録されているが，白書に出現した表記形（書字形）が登録されていなかったことに起因するものが38例あった。

例えば,《モノヅクリ》という語は「物作り」「モノづくり」「モノ作り」という表記が unidic に登録されていた。しかし，平仮名表記の「ものづくり」が登録されていなかったため，白書に出現したこの平仮名表記例が「|もの|づくり|」と誤解析された。

外来語に関しては表記のゆれに起因する誤解析が見られる。例えば，「ウェイト」「プレイ」のようなエ列の長音を長音符号を用いずに表記するものと，「ユーザ」「レーザ」のような語末のア列の長音符号を省略しているものとが誤解析された。これらは，unidic には「ウエート」「プレー」「ユーザー」「レーザー」のように長音符号を用いた表記形のみが登録されていた。そのため，一部を既登録の語と認定して「|ウェイ|ト|」「|ユー|ザ|」のように正解より短く分割されていた。

以上のほか，未登録の表記形に起因する誤解析には，公用文の表記の基準にかかわるものがある。特に，交ぜ書き漢語や送り仮名が省略された語の誤解析が多く見られた。

公用文では，常用漢字表にない漢字は使わないという漢字使用の基準がある。そのため，「破綻」のような常用漢字以外の漢字を含む漢語は，「破たん」と交ぜ書きするか「破綻（たん）」とルビ付きで書かれる。このうち，交ぜ書きの表記形が誤解析された。今回の調査では，「研さん」（研鑽），「くん蒸」（燻蒸），「ヨウ素」（沃素），「防あつ」（防遏）などの交ぜ書き表記が出現したが，これらはいずれも漢字部分と仮名部分とに分割されるか，「|防|

あ|つ|」のように1字ずつに分割されていた。

また，公用文では，活用のない語のうち読み誤るおそれのないとされる特定の語について送り仮名を省略するという基準がある。送り仮名が省略された語としては，「打合せ」「立入（検査）」「引下げ」などがあった。これらは「|打|合せ|」「|立|入|」「|引|下げ|」のように2単位に分割されていた。

以上のように，短単位境界に関する誤解析の主な原因は，正解の短単位（語彙素）や白書に出現した表記形（書字形）が unidic に登録されていないことであった。今後，精度向上のために，未登録の語や表記を unidic に登録していく。ただし異表記を網羅的に登録することは，逆に精度を下げる可能性もある。各表記の頻度や出現分野の偏り等を見て，登録するか否かを判断する必要がある。そのためにも，今回のような分野を限定した解析結果の分析を継続的に行い，各分野の語や表記の出現傾向を明らかにしていきたい。

５．終わりに

以上，本稿では，BCCWJ の言語単位の設計方針，BCCWJ で採用した長短2種類の言語単位のうち短単位の認定規定等について説明した。また，2006年度における形態論情報付与作業の進捗状況，BCCWJ に格納する中央省庁刊行の白書のデータを解析した結果も報告した。

BCCWJ の構築計画では短単位解析の精度の目標を98％以上としている。この目標値を達成するため，来年度以降も引き続き電子化辞書班と連携しながら，辞書の整備・拡充を図るとともに，学習用コーパスの整備，それに基づく学習等を進めていく予定である。

注

1 CSJ の短単位・長単位については，国立国語研究所（2006:133-186）を参照。
2 β単位については，国立国語研究所（1962:6-14）を参照。
3 長い単位については，国立国語研究所（1995:49-63）を参照。
4 既に外来語の短単位の認定規定と付属要素とを改定した。外来語の単位認定については，CSJ では「一般」の外来語の最小単位も和語・漢語と同様に2個の1次結合を1短単位とするのを原則としていたが，BCCWJ では1最小単位を1短単位とするのを原則とした。付属要素については，CSJ で付属要素としていた動詞連用形に接続する「掛かる」や形容詞性接尾辞「がましい」等を付属要素としないことにした。
5 unidic の品詞・活用型・活用形については，伝康晴（2006:26）を参照。
6 unidic における語の階層的な定義については，伝康晴（2006:25）を参照。

参 考 文 献

国立国語研究所（1962）『国立国語研究所報告21 現代雑誌九十種の用語用字(1)』秀英出版.
国立国語研究所（1995）『国立国語研究所報告112 テレビ放送の語彙調査 I 』秀英出版.
国立国語研究所（2006）『国立国語研究所報告124 日本語話し言葉コーパスの構築法』.
伝康晴（2006）「多様な目的に適した形態素解析システム用電子化辞書の開発」『特定領域「日本語コーパス」平成18年度全体会議予稿集』, pp. 21-26.（http://www.tokuteicorpus.jp/result/pdf/2006_017.pdf よりダウンロード可能）

付記 本研究は，文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21世紀の日本語研究の基盤整備」（平成18～22年度，領域代表者：前川喜久雄）による補助を得た。

日本語コーパスでの Sketch Engine 実装の試み

投野由紀夫（日本語教育班分担者：明海大学大学院応用言語学研究科）†

Using the Sketch Engine for Japanese Corpora

Yukio Tono (Graduate School of Applied Linguistics, Meikai University)

１．本研究の目的

日本語教育班ではいくつかの柱を立てて現在構築中の日本語書き言葉均衡コーパスの教育利用の研究を進めている。その１つが「教育利用に適したコーパス検索インタフェースの提案」がある。コーパスが綿密に設計・構築されても，それらの情報を抽出する方法が複雑では，検索用途や利用が一部の高度な検索技術を身につけたユーザーに限定されてしまう。逆に単なる文字列検索のみではコーパスとしてのさまざまな特性を十分に活かしきれない。このようなジレンマをある程度解消するには，検索ソフトのユーザー・インタフェースの工夫改善が不可欠で，それにはあらかじめどのような検索結果をユーザー側が要求としているかに関しての予測検討が重要になってくる。

このような問題点に関して，英語コーパス言語学の分野でもさまざまな試みがなされている。コーパス構築環境から検索までの統合環境を提供するもの（CLaRK System），British National Corpus（１億語のイギリス英語均衡コーパス）をもとにしたさまざまなニーズに応える多様な web interface の試み（BNCWeb; VIEWS, Phrases in English など），複数コーパスの自動タグ付与・頻度比較インタフェース（Wmatrix），複数コーパスの統一 web インタフェース（小学館コーパス・ネットワーク）などがある。

そのような中で近年特に注目されている検索システムの１つが Sketch Engine である。これは Adam Kilgarriff らが中心で開発しているコーパス検索インタフェースであるが，最近はこの Sketch Engine が多言語対応で数多くの言語での実装が試みられている。日本語教育班の分担者として日本語コーパス・データを Sketch Engine 実装で検索してみることがどの程度可能か，その際にはどのようなテキスト加工上の技術的な問題点があるか，また教育的な視野からの形態素の単位や文法関係の提示方法なども検討してみたい。今回のポスター発表ではその最も初期段階の研究内容として，Sketch Engine そのものの概要を調査し，一部分の日本語データを Sketch Engine 上で動作させた結果を報告する。

２．Sketch Engine の概要

Sketch Engine の発想は，開発者の Adam Kilgarriff が英語学習辞書編纂に役立つコーパス資源の研究を従来から進めており，ロングマン，マクミランなどの英英辞典プロジェクトのコーパス分析を担当したところから端を発している。従来の辞典編纂においては，COBUILD プロジェクトなどに代表されるように，レキシコグラファー本人が見出し語執筆項目をコーパスで検索し，生のコンコーダンスを読みながら使用傾向の分析を行っていた。これには例文を読み込む膨大な時間が必要であり，人間が単位時間内に分析できる例文の数は限られていることから単語の動きを一般化するのは困難であった。COBUILD ではさらにコロケーション統計として当該単語の前後５語などの共起語の頻度やMI 値などを一覧に

† yukio.tono@gmail.com

してそれを辞典編集者が読むことが出来たが，それでも単語列の位置情報のみであったために，そこからの情報の集約には一定のスキルが必要で，かつそれでも取り出せる情報には限界があった。

Adam Kilgarriff はこれらの辞典編纂に資するコーパス情報の提示方法として，当該単語の文法関係とコロケーションを 1 ページで一覧できる Word Sketch を考案して，マクミラン英語辞典（Rundell 2002）に活用，その内容が最初に紹介されたのは Rundell & Kilgarriff (2002) であった。この中で Kilgarriff は従来の辞典編纂におけるコーパス情報の欠点として(1)に示す 5 点を挙げている：

(1) a. コロケーション統計（MI 値など）で極度に低頻度項目を強調してしまう欠点
b. 見出し語（lemma）ではなく，活用形・表層形（wordform）である点
c. 左右の文脈何語を統計量抽出の基準にするかが曖昧である点
d. 抽出結果に言語学的に重要でない項目，いわゆるノイズが多い点
e. 同一リストに異なる文法関係の項目が混在して一覧される点

これらのうち，(1a) についてはさまざまな統計値が提案され，MI 値の欠点を補うようなものが既に出てきている。また (1b) に関しては見出し語化（lemmatization）の手順を踏むことによって，より適切なリストの抽出がいまや可能となっている。

Sketch Engine は (1c) ～ (1e) の問題を解決するために考案されたインタフェースと言ってもよい。１つの単語に関してその文法関係を定義し，その文法関係ごとにコロケーション統計を抽出するので，(1c) の左右何語というような条件に縛られない。また文法関係で特定の位置における名詞とか形容詞とか品詞単位の抽出をするので (1d) のようなノイズ

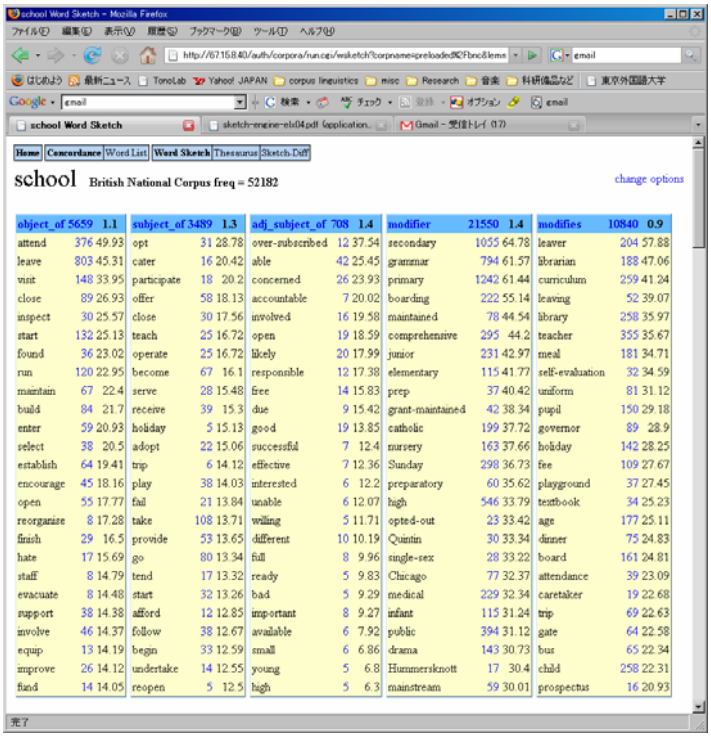


図１：Word Sketch（school）

の混入が少ない。さらに，文法関係を明示的に定義してコロケーションを抽出するために，(1e) にあるような複数の文法関係の項目が混在するというような不具合がない。

図 1 に Sketch Engine で school を検索した結果を示す。school の場合には (1) *object_of*（動詞＋ school）, (2) *subject_of* （school＋動詞），(3) *adj_subject_of* （school ＋ be + 形容詞），(4) *modifier*（形容詞・名詞＋ school），(5) *modifies* （school +名詞）といった文法関係ごとにその位置に来る単語が統計値順に整理されている。その他にも school の後に来る前置詞のパターンが１０種類くらい整理されており，これらを一覧することで school の使い方の概要が把握できるようになっている。

３．Sketch Engine の実装の手順

Sketch Engine にコーパスを実装するためにはいくつかのステップが必要である。英語の場合を例に取ると以下のような段階がある：

(2) a. 見出し語化（lemmatization）
b. 品詞タグ付与（POS tagging）
c. 入力フォーマット（Input formatting）
d. 文法関係の定義（grammatical relations）

日本語の実装の際には，(2a) から (2c) までは形態素解析ツール「茶筅」の出力形式を利用すればよい。最も困難なのは，(2d) の文法関係の記述で，これに関しては大まかに 2 通りの方法がある：

(3) a. 文法関係を注釈付けした構文解析済みのデータを用いる
b. 品詞タグ付与のレベルで，正規表現などを用いて検索式を用意する

(3a) に関しては，日本語の場合には「南瓜」などが形態素解析器として利用できるが，Sketch Engine にどの程度適しているかはまだ検討していない。現在のところ (3b) の品詞タグ付与をしたデータをもとに文法関係を抽出する検索式を IMS Corpus Workbench の CQS (Corpus Query Syntax) をもとに作成する方向で検討をしている。

３.１．日本語データ実装試験

今回は最も初歩的な段階として，日本語テキストのサンプル（夏目漱石『坊ちゃん』）を茶筅により形態素解析し，その出力を 1 行 1 単語（表層形－品詞－見出し語）形式でユニコード（UTF-8）に変換し，Sketch Engine の Corpus Builder の機能を使って web 上のサーバーにアップロードした。図 2 はコンコーダンサー部分が動作している様子である。

ただし，この段階では Word Sketch までは動作していない。英語の文法関係の定義ファイルしかないので，これを日本語に書き換える必要がある。この作業のためには定義ファイルの内容を具体的に理解する必要がある。

図 2：日本語実装の例（コンコーダンス）

３.２．文法関係定義ファイル

Sketch Engine の文法関係定義ファイルは以下のような書式になっている：

(3) 1: [キーワードの文法関係を表す CQS 式]　　2: [抽出したい collocate]

例えば，他動詞の目的語に関しては (4) のような定義式を持つ：

(4) 1:"VB.?" [tag="DT|PRP¥$"]{0,1} "JJ.?"{0,3} "NN.*"{0,2} 2:"NN.*"

日本語を実装する場合にはこの定義ファイルの書き換えが今後の大きな作業となるが，手始めに全面的な改変ではなく，動詞の目的語を抽出するような小さな部分の定義から初めて徐々に動詞と連動する格助詞をノードにしたコロケーションの抽出に迫ってみたい。

４．まとめ

ポスターでは実際に部分的に文法関係の定義式を読み込ませて動作するかどうかを検証した結果を紹介する予定であるが，全面的な実装にはいくつかの障壁がある：

(5) a. 文法関係を定義する部分で，日本語文法の専門家との協議が必要
b. 日本語教育の観点からどのような文法関係を取り出すことが意義があるか？
c. 語順が一定しない日本語の場合，正規表現で抽出することの限界がある
d. 茶筅の形態素解析の単位が細かすぎる問題

これらの点を日本語教育班で検討しながら，２年目の後半にはフル実装を目指したい。

文献

Kilgarriff, A. and Rundell, M. (2002). Lexical Profiling Software and its lexicographic applications: a case study. *Proceedings of EURALEX 2002*. Copenhagen: 807-818.

Rundell, M. (ed.) (2002). *Macmillan English Dictionary for Advanced Learners*. UK: Macmillan.

確率的単語分割ツールとその利用

浅原正幸（ツール班分担者：奈良先端科学技術大学院大学 情報科学研究科）†

A Stochastic Word Segmentation Tool – “Bar++”

Masayuki Asahara (Grad. School of Information Science, Nara Institute of Science and Technology)

１．はじめに

本稿では、今年度作成した確率的単語分割ツール Bar++ について紹介する。確率的単語分割とは 2004 年に森らにより提案されたコーパスに対する単語境界情報付与手法である(Mori 2004)。図１のように各文字境界に、単語境界になるかならないかの確率値を付与する。このような単純なデータ表現手法を用いることにより、任意の部分文字列の見なし頻度をコンパクトに保持することができる。この枠組を用いることにより多種多様な応用を実現することができることを紹介する。

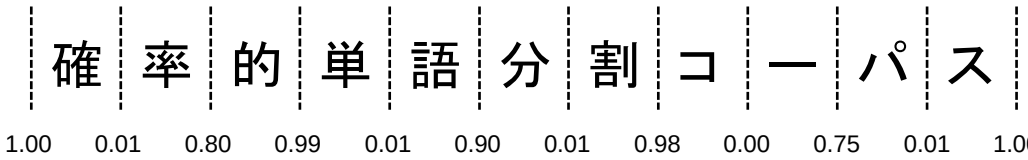


図１：確率的単語分割コーパスの例

２節では確率的単語分割に関する過去の研究について述べる。３節では今回作成したツール Bar++ の構成について示す。４節では確率的単語分割を利用してどのようにコーパス研究に役立てるか、いくつかの応用について検証する。最後にまとめと今後の計画について述べる。

２．確率的単語分割と関連研究

２．１　確率的単語分割

確率的単語分割 (Mori 2004)は N 文字からなる生コーパス（文字列） $x_1 \cdots x_N$ とその隣接する各２文字 $\langle x_i, x_{i+1} \rangle$ の間に単語境界が存在する確率 P_i の２つによって定義される。コーパスの開始位置と終了位置には単語境界が存在するとみなせるため、$P_0 = P_n = 1$ とする。確率変数 X_i を

$$X_i = \begin{cases} 1 & \langle x_i, x_{i+1} \rangle \text{の間に単語境界が存在する場合} \\ 0 & \langle x_i, x_{i+1} \rangle \text{の間に単語境界が存在しない場合} \end{cases}$$

とする(ここで $P(X_i = 1) = P_i, P(X_i = 0) = 1 - P_i$)。各確率変数 $X_0, X_1, \ldots, X_{N-1}, X_N$ は独立であることを仮定する。確率値 P_i を k bits で表現した場合には、これらの情報を保持するために必要な領域は元の文字列に加えて $k(N+1)$ bits である。

２．２　確率的単語分割モデルの推定手法

(Mori 2004)は単語 2-gram の生成モデルに基づく自動単語分割器により確率的単語分割

† masayu-a@is.naist.jp

モデルを構成した。あらかじめ単語単位に分割されたコーパスを用いて自動単語分割システムの境界推定精度αを計算する。次に適用分野のコーパスを自動単語分割し、その出力において単語境界であると判定された点では$P_i = \alpha$とし、単語境界でないと判定された点では$P_i = 1 - \alpha$とする。この推定手法では文字境界に付与される確率値はαもしくは$1-\alpha$のいずれかである。

(森 2006b) は、最大エントロピーモデルを適用し、文字境界毎に単語境界であるか否かの確率値を推定する手法を提案した。文字列$x = x_1 \cdots x_N$が与えられているとして、各文字x_iにその前で単語境界があるか否かのラベル$y_i \in \{B, I\}$を付与する条件付確率値を次のような式を用いて推定する(ラベル B はその文字の前に単語境界がある、ラベル I はその文字の前に単語境界がないことを表す) :

$$P(y_i \mid x) = \exp(\sum_k \lambda_k f_k(x, y_{i-1}, y_i)) \Big/ Z(x)$$

但し
$$Z(x) = \sum_{y_i \in \{B,I\}} \exp(\sum_k \lambda_k f_k(x, y_i, y_{i+1}))$$

ここで、f_kはその文字位置における素性関数で文字や字種などの出現を表す二値関数、λ_kは素性関数に対応する重みである。生成モデルと異なり、確率値推定時に与えられる素性が独立である必要がなく、比較的自由に単語境界の特徴をモデルに反映させることができる。これにより文字境界毎に異なる確率値を文脈に応じて与えることができる。

(岡野原 2006) は、条件付確率場 (Lafferty 2001)により文字境界毎に単語境界であるか否かのラベルを付与する手法を提案した。文字列$x = x_1 \cdots x_n$が与えられているとして、ラベル系列$y = y_1 \cdots y_n$（但し$y_i \in \{B, I\}$）を付与する条件付確率値を次のような式を用いて推定する：

$$P(y \mid x) = \exp(\sum_{1 \le i \le |y|} \sum_k \lambda_k f_k(x, y_{i-1}, y_i, i)) \Big/ Z(x)$$

但し
$$Z(x) = \sum_{y' \in \{B,I\}^{|y|}} \exp(\sum_{1 \le i \le |y|} \sum_k \lambda_k f_k(x, y_i, y_{i+1}))$$

このように推定された確率値は系列全体の確率値である。各文字境界に対する確率値は次の式に示す周辺確率を用いる：

$$P_i = \sum_{y_1, \cdots, y_{i-1}, y_{i+1}, \cdots y_n \in \{B,I\}^{|y|-1}} p(y_1, \cdots, y_{i-1}, y_i = B, y_{i+1}, \cdots y_n \mid x)$$
$$= \frac{1}{Z(x)} \frac{\alpha_{i,B} \cdot \beta_{i,B}}{e_{i,B}}$$

ここで $e_{i,y}$は位置 i に依存するコストで次式で定義される：

$$e_{i,y} = \exp(\sum_k \lambda_k f_k(y, x, i))$$

また、α_i, β_iは条件付確率場のモデル推定時に用いられる forward-backward コストであり、次式のように再帰的に定義される：

$$\alpha_{i,y} = \sum_{y' \in \{B,I\}} (\alpha_{i-1,y'} \cdot \exp(\sum_k \lambda_k f_k(y', y, x, i)))$$

$$\beta_{i,y} = \sum_{y' \in \{B,I\}} (\beta_{i+1,y'} \cdot \exp(\sum_k \lambda_k f_k(y, y', x, i)))$$

このモデルでは全系列の情報を周辺確率として利用する。最大エントロピーモデルに基づく手法が、局所的な文字列が同じ場合にまったく同じ確率値を付与するのに対して、系列全体の文脈に応じて異なる確率値を出力することができる。

２．３　その他の関連研究

(工藤 2005) は、単語単位の辞書引きによる条件付確率場に基づく形態素解析器の形態素周辺確率を用いた単語分割の一般化について提案している。確率的単語分割モデルが文字単位の冗長解を保持するモデルであるのに対して、工藤の手法は単語単位の冗長解を保持するモデルであるといえる。形態素解析器 MeCab は -l2 -a オプションを用いることにより、上に述べた解析結果を出力することができる。

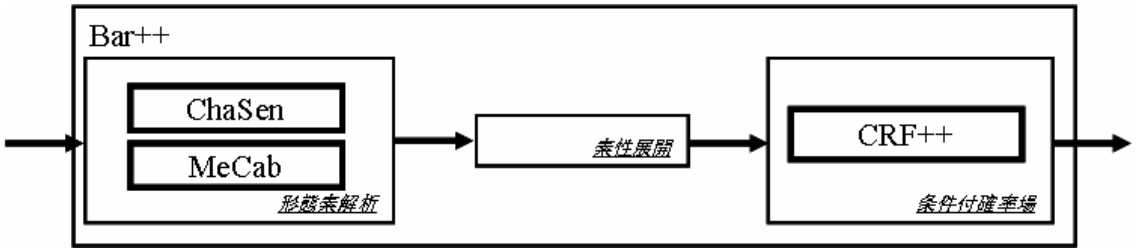


図２：Bar++の構成

３．Bar++ の構成

Bar++ は確率的単語分割を行うためのツールである。(岡野原 2006) の手法と同様に確率値を文字単位の条件付確率場が出力する周辺確率を用いる。図２に Bar++の構成を示す。内部で形態素解析器 ChaSen もしくは MeCab を呼び出し、あらかじめ形態素解析器の一次解を得る。この形態素解析器結果を基に、素性展開モジュールでは、文字境界単位に、文字、字種、文字位置がどのような単語に含まれるか、文字位置がどのような品詞に含まれるかなどの情報を展開する。最後に CRF++ による条件付確率場のモデルを呼び出し、文字境界単位に単語分割確率情報を付与する。

学習対象の単語分割済コーパスとして、RWCP テキストコーパスに対して、UniDic 品詞体系で形態素情報を付与しなおしたものを用いた。入出力の文字コードとして EUC-JP と Shift_JIS の２つに対応している。

４．応用事例

４．１　言語モデルとしての利用

(Mori 2004) では、音声処理などのための言語モデル（単語 n-gram モデル）を構成するために確率的単語分割コーパスを利用する方法を提案している。確率的単語分割コーパス x_1^N から単語列 w_1^n（表層文字列を x'^L_1 であると仮定する） の n-gram 頻度を得るために以下の式で算出する：

$$f_r(w_1^n) = \sum_{\{i|x_i^{i+L-1}=x_1^L\}} P_i \left[\prod_{m=1}^{n} \left\{ \prod_{j=b_m}^{e_m-1} (1-P_j) \right\} P_{e_m} \right]$$

ここでb_mは単語列w_1^n中 m 番目の単語の開始文字、e_mは単語列w_1^n中 m 番目の単語の終了文字を表す。この n-gram 頻度は得たい単語列の境界と確率的単語分割コーパスの境界とが一致する期待頻度を計算していることになる。図３に n-gram 頻度の計算手法を示す。

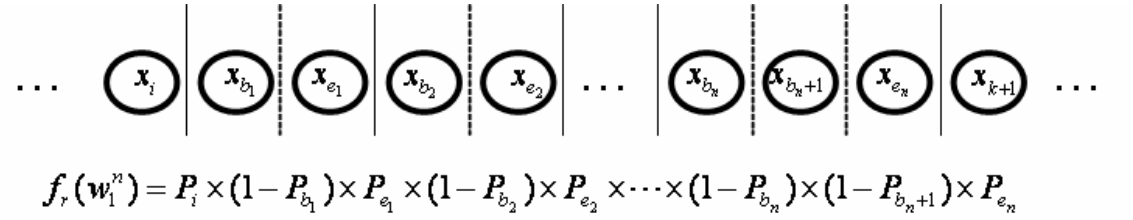


図３：確率的単語分割コーパスからの n-gram 頻度の取得（森 2006b)より

尚、単語 0-gram 頻度は、$f(\bullet) = 1 + \sum_{1 \le i \le N} P_i$ によって定義される。

（森 2006a）では、この統計的言語モデルを用いて仮名漢字変換のための言語モデルの作成手法を提案している。確率的単語分割コーパスを接尾辞配列などの形で持つことにより、コーパスを辞書引きするだけで頻度情報を得ることができる。言い換えると確率的単語分割コーパスを確率情報つき辞書として活用する手法を提案している。

４．２　全文検索

(岡野原 2006)は、確率的単語分割モデルを用いた全文検索手法について提案している。全文検索では文字により索引を作成する手法と単語により索引を作成する手法がある。文字索引の場合、検索漏れはなくなるが、無関係の単語（他の単語の部分文字列など）が結果に含まれることがある。提案手法では、検索対象にあらかじめ文字間に単語境界確率を付与し、検索時には全候補の境界の確率値に基づきソートすることにより、より関係のある文脈から順に出力する方法を提案している。また確率的単語分割は文字単位に確率情報を持つために、文字索引による検索手法に組み込みやすいという利点についても示している。

４．３　用語抽出、コロケーション抽出

(Asahara 2004) は、Support Vector Machines により決定的に文字単位の語境界情報をラベルづけにより未知語を抽出することを報告した。確率的単語分割モデルも同様に文字単位の解析モデルであるため、同様に未知語抽出に適用することができる。４．１節に述べた手法により 1-gram 頻度を得ることにより期待頻度順に未知語候補をランキングすることも可能である。

また n-gram 統計を用いてコロケーション抽出を行う場合でも同様に期待頻度を用いることが可能である。本手法を用いると t-score などの統計量を得る際に形態素解析の一次解を用いるのに比べて、形態素解析誤りの影響が少ないと考える。

４．４　誤入力検出

大規模コーパスを構築する際に誤入力の問題は避けられない。本プロジェクトでもコーパスの大部分をＯＣＲ読み取りや手入力により作成する。確率的単語分割ツールは、単語分割が困難な部分である誤入力出現位置では、0.5 に近い確率値を出力する。この特性を利用して誤入力検出ツールを構成できると考える。予備実験として、本プロジェクトの一部のデータに対して誤入力の検出を今回作成したツールを用いて行った。コーパスを確率的単語分割ツールにより解析し、確率値が 0.4〜0.6 である部分を 0.5 から近い順に出力し検証を行った。人手による評価を行ったところ出力した約 １％ が誤入力であった。精度が低い理由として、誤入力以外の単語分割が困難な部分として、辞書にない単語や、助詞などの形態素解析器が本質的に誤りやすい箇所が出力されることがわかった。今後、大規模コーパスからの頻度に基づく生成モデルを相補的に用いることにより精度向上を達成したいと考える。

５．おわりに

本稿では今年度作成した確率的単語分割ツールとその背景にある理論について紹介した。確率的単語分割の枠組自体は、森、岡野原、工藤らにより提案されたものであることを断っておくとともに、この場を借りて３氏に感謝の意を表したい。

本プロジェクトの開始当初、多くの言語研究者がコーパスを用いる場合、まずコーパスに付与されているラベルを取り除いてから文字列検索という話を聞いた。確かに言語処理ツールの精度は形態素解析器のような基本的なツールでも 98% 程度であり、これは 50 単語に１つは間違える計算である。今回見てきた枠組みではラベルが付与されるか否かの問題を２値の問題ではなく、確率的な連続値を用いる手法である。コーパス研究における基本操作である検索や統計量取得において、言語研究者にとって有用な枠組であると考えており、今後この枠組に基づいたさまざまなツールを整備していきたい。来年度以降の開発項目として、既存の検索ツールとの融合、コロケーション抽出部分の精緻化、Unicode 対応などを考えている。

いくつかの利用方法についてはシンポジウム当日にデモを行う。是非現在の言語処理技術を体験していただきたい。

文献

Masayuki Asahara and Yuji Matsumoto (2004). “Japanese unknown word identification by character-based chunking” In Proc. of COLING-2004.

John Lafferty, Andres McCallum, and Fernando Pereira (2001). "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data" In Proc. of ICML 2001.

Shinsuke Mori and Daisuke Takuma (2004). "Word n-gram probability estimation from a Japanese raw corpus" In Proc. of ICSLP 2004.

岡野原大輔，工藤拓，森信介 (2006) 「形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用」第１回ＮＬＰ若手の会.

工藤拓 (2004) 「Conditional Random Fields を用いた日本語形態素解析」『情報処理学会研究報告』SIGNL-161.

工藤拓 (2005) 「形態素周辺確率を用いた分かち書きの一般化とその応用」言語処理学会全国大

会 NLP-2005.
森信介 (2006a) 「無限語彙の仮名漢字変換」『情報処理学会研究報告』SIGNL-172.
森信介，倉田岳人，小田裕樹 (2006b) 「最大エントロピー法による単語境界確率の推定」『情報処理学会研究報告』SIG-SLP-63.

関連 URL

確率的単語分割ツール『Bar++』	http://barpp.sourceforge.jp/
形態素解析器『ChaSen』	http://chasen.naist.jp/
形態素解析器『MeCab』	http://mecab.sourceforge.net/
条件付確率場ツール『CRF++』	http://chasen.org/~taku/software/CRF++/

タグ付きコーパス検索ツールの開発

谷口雄作[†]
新保仁（分担者）
（奈良先端科学技術大学院大学情報科学研究科）

Development of Search Tool for Annotated Corpus

Yuusaku Taniguchi
Masashi Shimbo
(Graduate School of Information Science, Nara Institute of Science and Technology)

１．はじめに

本稿では，現在開発を行っているタグ付きコーパス検索ツール「茶杓」について述べる．本システムは，言語処理・言語学やその関連分野の研究者を対象に，柔軟な検索と容易な操作を実現することを目的として構築されたものである．

同様のツールとしては，我々のグループで既に開発が進んでいるコーパス管理・検索ツール「茶器」(松本，2006）がある．このツールは個々の利用者が独自のコーパスを持ち，その検索と修正を行うことを目指して実現されたものであり，各ユーザがそれぞれコーパスを操作できる環境を事前に用意する必要があった．本システムは Web ブラウザを用いてコーパスを操作する Web アプリケーションであり，このような事前の用意を必要とせず，操作を行うことが出来る．また，コーパスの検索機能に特化することで，不特定多数のユーザへのコーパス公開にも有用である．本稿では本システムの機能の紹介を行い，最後に今後の方針について述べる．

２．機能紹介

本節ではタグ付きコーパス検索ツール「茶杓」が備えている機能を紹介する．現在本システムが行える検索機能は主に文検索・単語列の頻度統計・文集合内の非連結頻出系列パターンの獲得である．これらの機能について，インターフェースの機能も交えて紹介する．

本システムは，文字列または単語列を入力として与えることでコーパスに含まれる文を検索する事が出来る．検索された文は KWIC（KeyWord In Context）形式にユーザに表示するもので，文だけでなく，品詞や活用型などのコーパスにタグ付けられた情報を同時に閲覧することができる．

[†] yuusaku-t@is.naist.jp

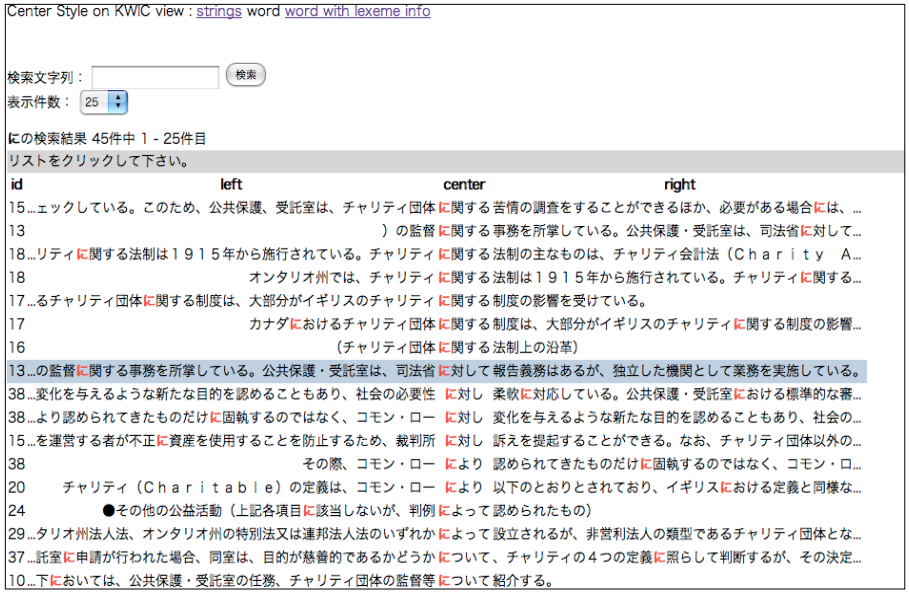


図１　検索結果の KWIC 表示

文字列による検索では，KWIC 表示の中央に表示される語句を，入力文字列とその該当する単語の単位に切り替えることができ，また単語ごとにソートすることも可能である．図１は単語区切りに表示を行った例である．また，検索された各文の前後の文章の表示，入力文字列に該当する単語の検索，同種の品詞を持った単語に対して色別表示なども行うことができる．

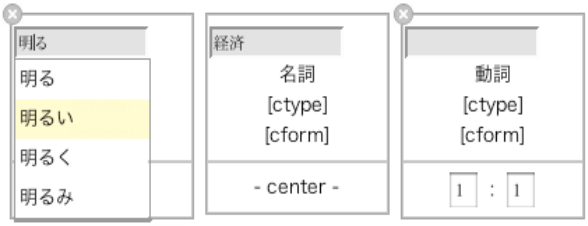


図２　単語列入力インターフェース

単語列による検索では，各語がもつ任意の情報を指定して連続あるいは非連続の単語列を検索することができる．例えば図２は表層形に「経済」を持った名詞を中心に，前に形容詞の「明るい」後に動詞を含む文を検索する際の条件指定の様子を示している．また，このインターフェースには表層形の入力補助としてオートコンプリート機能を備えている．

文の検索では結果件数が多い場合にどのような単語がどの程度出現しているかを瞬時には把握しにくい．本システムは単語列に対する頻度統計を取るができる． 図３は入力として「名詞＋銀行」を与えた場合の例である．最左列(Count)に出現頻度が表示されている．

	Center Lexeme				Right side Lexeme 1			
Count	Surface	POS	C type	C from	Surface	POS	C type	C from
39	の	名詞-固有名詞-地域-一般			銀行	名詞-固有名詞		
26	中央	名詞-固有名詞			銀行	名詞-固有名詞		
23	大和	名詞-数			銀行	名詞-固有名詞		
19	民間	名詞-固有名詞			銀行	名詞-固有名詞		
16	都市	名詞-固有名詞			銀行	名詞-固有名詞		
14	共同	名詞-一般			銀行	名詞-固有名詞		
14	開発	名詞-一般			銀行	名詞-固有名詞		
13	信託	名詞-一般			銀行	名詞-固有名詞		
12	三菱	名詞-数			銀行	名詞-固有名詞		
8	信用	名詞-一般			銀行	名詞-固有名詞		
7	輸出入	名詞-固有名詞			銀行	名詞-固有名詞		
7	や	名詞-固有名詞-地域-一般			銀行	名詞-固有名詞		
7	大手	名詞-固有名詞			銀行	名詞-固有名詞		
5	住友	名詞-数			銀行	名詞-固有名詞		
5	世界	名詞-固有名詞			銀行	名詞-固有名詞		
5	全国	名詞-固有名詞			銀行	名詞-固有名詞		
4	地方	名詞-固有名詞			銀行	名詞-固有名詞		
3	救済	名詞-一般			銀行	名詞-固有名詞		
3	する	名詞-サ変接続	特殊・ダ	連用形	銀行	名詞-固有名詞		
3	省	名詞-固有名詞			銀行	名詞-固有名詞		
3	投資	名詞-一般			銀行	名詞-固有名詞		

図３　単語列に対する頻度統計

これらの機能によって抽出された単語列は，文中のひとつの固まりとして，より大きなパターンの中で使われることがある．例えば「名詞＋を＋動詞」の周辺に出現するパターンには図４のようなものも起こると考えられる．

最後に作業を終えたら・・・しましょう

危険を犯したら・・・になるだろう

最新の機器を使えば

図４　非連結頻出系列パターン例

本システムは，コーパスに含まれる特定の文集合から，このような非連結頻出系列パターンを探し出すことが出来る．

これらの検索はタグ付きコーパスを関係データベースに格納して行っている．また本システムは言語に依存しておらず，英語・日本語・中国語のコーパスが利用できる．

３．システムの詳細

タグ付けコーパスを格納するデータベースについて検索に関連するテーブルの概説を行う．本システムが利用するデータベースの各テーブルはコーパス管理ツール「茶器」が使用するテーブルの仕様に準拠している．図５に本システムが利用する各テーブルの関係を示す．

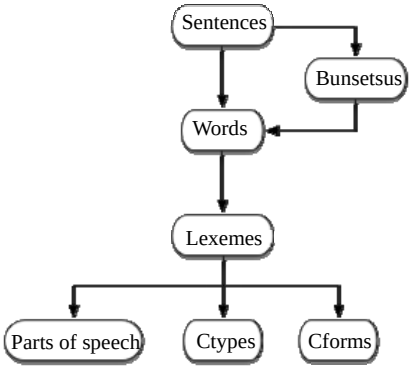


図５　検索対象となるテーブルの関係図

各テーブルはテーブル間の関係を示す ID を持っている．例えば文内の単語情報を受け持つ Words テーブルには各文のテキストを持つ Sentences ・文節の情報を受け持つ Bunsetsus ・語彙の情報を持つ Lexemes の各テーブルの ID を持っており，Lexemes は各単語の情報として Part of speech（品詞）・ Ctypes（活用型）・ Cforms（活用形）を持つ．

Words テーブルに登録されたあるアイテムを指定すると，そのアイテムがどの文のどの文節のどの単語を持ったアイテムであるかを特定することができるといった仕様になっている．

４．おわりに

本稿ではインターネットを介してタグ付きコーパスの検索を行うことができるツール「茶杓」について説明した．今後の予定として，依存構造の検索や統計機能の充実，プログラムから検索結果を呼び出す Web API 及びユーザインターフェースの開発に力を入れて行きたい．

文献

松本裕治，浅原正幸，橋本喜代太，投野由紀夫，大谷朗，森田敏生（２００６）「タグ付きコーパス管理/検索ツール「茶器」」，言語処理学会第 12 回年次大会論文集，pp. 460−463.

「日本語コーパス」用 Yahoo!知恵袋データについて

岡本真（総括班研究協力者：ヤフー株式会社）
木戸冬子（総括班研究協力者：ヤフー株式会社）
佐古智正（ヤフー株式会社）

Yahoo! Chiebukuro Data for “Japanese Corpus”

Makoto Okamoto (Yahoo Japan Corporation)
Fuyuko Kido (Yahoo Japan Corporation)
Toshimasa Sako (Yahoo Japan Corporation)

1. 目的

Yahoo! JAPAN を運営するヤフー株式会社は、「現代日本語書き言葉コーパス」（以下、BCCW）に対して、知識検索サービス Yahoo!知恵袋に投稿された質問・回答形式のテキストデータ（以下、Yahoo!知恵袋データ）を提供する。本稿は、この Yahoo!知恵袋データの性格や特徴について情報提供を行い、コーパス構築に資するものとしたい。

2. Yahoo!知恵袋データとは

Yahoo!知恵袋データは、知識検索サービス Yahoo!知恵袋に一般のインターネット利用者が実際に投稿した質問文と回答文で構成されるテキストデータである。Yahoo!知恵袋（図 1）は、「日ごろ疑問に思っていることを他の参加者に尋ねたり、他の参加者の疑問に答えることで、お互いに知恵や知識を教えあい、分かち合うことを目的としたサービス」（Yahoo!知恵袋ガイドライン）であり、一般には知識検索サービス（Knowledge Search）と称されている。

Yahoo!知恵袋は一般のインターネット利用者が質問と回答という形式でコミュニケーションと問題解決を図るサービスである。

コーパスの観点からはインターネット上で書かれた日本語のテキストデータとして大きな可能性を秘めている。そこで、Yahoo!知恵袋に実際に投稿された質問・回答を抽出し、BCCW への組み込みが可能なテキストデータに変換した。このテキストデータは、独立行政法人国立国語研究所とヤフー株式会社の間で締結した契約に基づき、無償で独立行政法人国立国語研究所に提供される。これが Yahoo!知恵袋データである。

図 1 Yahoo!知恵袋
（ http://chiebukuro.yahoo.co.jp/ ）

3. Yahoo!知恵袋データのスペック

3.1 基本属性

Yahoo!知恵袋データは、2004 年 4 月 7 日に Yahoo!知恵袋がベータ版（テスト版）として公開されてから、2005 年 10 月 31 日にベータ版の運用を停止し、正式版に移行するまでの

1 年 7 ヶ月間に Yahoo!知恵袋に投稿された質問・回答で構成されている。このうち回答はベストアンサーとベストアンサー以外に分かれている。Yahoo!知恵袋では、質問者は自分の質問に寄せられた回答のうち、もっとも納得した回答をベストアンサーに選択し、質問を終了する仕様となっている。このため、一口に質問・回答といっても、以下のような構造となっている。

- 質問
  - 回答（ベストアンサー）
  - 回答（ベストアンサー以外）
  - 回答（ベストアンサー以外）

なお、Yahoo!知恵袋では、Yahoo! JAPAN 利用規約や Yahoo!知恵袋ガイドラインに違反する投稿を随時削除しており、これらの削除済みの投稿は Yahoo!知恵袋データには含まれていない。

表 1　Yahoo!知恵袋データのデータ量

投稿数	総計		16,595,744
	質問		3,116,009
	回答	小計	13,479,735
		ベストアンサー	3,116,009
		ベストアンサー以外	10,363,726
文字数	総計		1,662,711,657
	質問		304,110,512
	回答	小計	1,358,601,145
		ベストアンサー	423,880,732
		ベストアンサー以外	934,720,413
投稿者数	総計		242,336
	質問		165,064
	回答	ベストアンサー	119,263
		ベストアンサー以外	165,061

3.2 データ量

このような構造を持つ Yahoo!知恵袋データは、約 24 万の投稿者によって投稿された約 166 万件の質問・回答で構成されている（約 16 億字相当）。詳細を表 1 に示す。

なお、投稿者数は Yahoo! JAPAN ID 数である。Yahoo! JAPAN ID は一人の人物が複数取得することが可能であるため、約 24 万という投稿者数はそのまま約 24 万人の人間を意味するわけではない。同一人物による複数の Yahoo! JAPAN ID の使い分け状況は調査していないが、物理的に存在する人物として考えた場合、投稿者数は 24 万を下回る可能性があることに注意する必要はあるだろう。

4. Yahoo!知恵袋データの著作権処理

一般にインターネット上のテキストデータは著作権処理が困難を極めるため、本件のようなコーパスへの組み込みは難易度が高いとされている。だが、Yahoo!知恵袋データでは、以下に列挙する処理を行うことで、著作権処理の問題をクリアしている。

Yahoo!知恵袋データを構成する質問・回答の著作権は投稿者本人に帰属するが、Yahoo! JAPAN 利用規約では第 7 条のように定めている。

> （第 7 条）
> ユーザーは Yahoo! JAPAN に対して、当該コンテンツを日本の国内外で無償にて非独占的に使用する（複製、公開、送信、頒布、譲渡、貸与、翻訳、翻案を含む）権利を許諾（サブライセンス権を含む）したものとみなします。また、ユーザーは著作者人格権を行使しないものとします。

また、Yahoo!知恵袋の利用者に対しては、投稿者が必ず同意する Yahoo!知恵袋ガイドラインにおいて、

> 投稿内容の利用について
> 本サービスに投稿された文章等は、今後 Yahoo! JAPAN およびその提携先が出版する書籍等に使用させていただく場合があります。
> 本サービスに投稿された文章等は、Yahoo! JAPAN およびその提携先が、出版、公衆送信、放送、DVD 化等を行う目的で無償にて自由に利用（複製、公衆送信、譲渡、翻案および翻訳等を含む）させていただきますので、あらかじめご了承のうえご投稿ください。
> 投稿された文章等についての著作者人格権を Yahoo! JAPAN およびその提携先に対して行使しないことについても同意いただいたものとみなします。

と掲示し、加えて Yahoo!知恵袋ヘルプにおいて、以下の告知を行っている。

> ■研究機関への研究データの提供について
> Yahoo! JAPAN では投稿者の Yahoo! JAPAN ID を暗号化するなど、個人を特定することができない情報に処理した上で投稿内容、投稿日時などの投稿に関する情報を大学、独立行政法人等の研究機関に提供します。
> Yahoo! JAPAN が提供する情報によって、当該大学、独立行政法人等が投稿者が誰であるかを知ることはありません。

これらの措置により、Yahoo!知恵袋データにおける著作権処理は適正に行われており、本件のような利用についても投稿者の同意が得られている。

5. Yahoo!知恵袋データのテキストとしての特徴

日本語のテキストデータとしてみた場合、Yahoo!知恵袋データは以下の特徴を備えている。

5.1 質問・回答の照応関係

テキストはすべて質問、あるいは回答であるため、質問は明確な疑問文や同意を求める付加疑問文の形式をとり、回答は定義や同意を示す形式をとることが多い。この特徴は文と文との照応関係を調査・検討する際に有用と考えられる。

5.2 テキスト長

Yahoo!知恵袋データの質問・回答あたりのテキスト長の平均値は表 2 のようになっている。

表 2. 1 文あたりの平均文字数

質問	98 文字
回答	101 文字
ベストアンサー	136 文字
ベストアンサー以外	90 文字

同じ回答であっても、ベストアンサーとベストアンサー以外とでは、平均文字数に大きな開きがあるように、質問者が納得する回答は一定程度の文字数を要するものであることがうかがえる。なお、Yahoo!知恵袋データを構成する質問・回答が投稿された期間では、質問は全角 200 文字以内、回答は 400 文字以内に記述を制限しており、上記の平均値はこの字数制限のもとで記述されている。

5.3 文体の特徴

一般にインターネットで記述されるテキストは、通常の書き言葉よりくだけた表現になるといわれているが、実際に Yahoo!知恵袋に投稿された質問・回答でもその傾向はみてと

れる。ただし、Yahoo!知恵袋の場合、いわゆる掲示板サービスで多用される独特な用語・用法は少数にとどまる。また Yahoo!知恵袋の仕様によりアスキーアートの投稿を実質的に禁止しているため、テキストデータではあるものの、実質的にはアスキーアートであるという質問・回答もごく一部に限られている。

6. ヤフー株式会社としての期待

記述の通り、Yahoo!知恵袋データの提供にあたって、ヤフー株式会社は金銭的な対価を一切受けない契約を独立行政法人国立国語研究所と取り交わしている。ヤフー株式会社としては、Yahoo!知恵袋データが「日本語コーパス」に組み込まれることで、本特定領域研究はもとより、今後の日本語研究に資すること、そしてそこで得られた知見が役立てられ、結果的にインターネットにおいて一般の利用者が投稿するテキストが向上することを期待している。関係者の方々には、「日本語コーパス」の構築と利用にあたって、この趣旨をご理解いただければ幸いである。

なお、「日本語コーパス」の構築に限らず、ヤフー株式会社は Yahoo!知恵袋データの提供や Yahoo!知恵袋を題材にした学術研究の支援に取り組んでいる。すでに成果報告されている研究については本稿末に記載してある参考文献をご参考いただきたい。

参考文献

川浦康至、三浦麻子、大瀧直子、地福節子、岡本真「知識共有コミュニティを創り出す人たち」（人工知能学会第 20 回全国大会、2006-06-09）
http://www.jaist.ac.jp/jsai2006/program/paper-163.html
川浦康至、三浦麻子、大瀧直子、地福節子、岡本真「知識共有コミュニティを創り出す人たち（2）:「質問タイプ」から見た参加行動」（日本社会心理学会第 47 回大会、2006-09-18）
http://db1.wdc-jp.com/cgi-bin/jssp/wbpnew/master/detail00.php?submission_id=2006-E-0067
川浦康至、三浦麻子、大瀧直子、地福節子、岡本真「知識共有コミュニティを創り出す人たち（3）:「回答者」データから見るコミュニティ内の「知識」」（日本社会心理学会第 47 回大会、2006-09-18）
http://db1.wdc-jp.com/cgi-bin/jssp/wbpnew/master/detail00.php?submission_id=2006-E-0176
ヤフー株式会社「ホントはしたい人助け　～知識共有コミュニティ「Yahoo!知恵袋」にみる、インターネット上の利他的行動」（「Yahoo! JAPAN ネット生活予測レポート」3、2006-09-29）
http://docs.yahoo.co.jp/info/report/data/003.pdf
Asako Miura, Yasuyuki Kawaura, Naoko Otaki, Setsuko Jifuku and Makoto Okamoto, People Who Create Knowledge Sharing Communities, New Frontiers in Artificial Intelligence (Lecture Notes in Computer Science)（2007-01）

関連 URL

Yahoo! JAPAN：http://www.yahoo.co.jp/
Yahoo! JAPAN 利用規約：http://docs.yahoo.co.jp/docs/info/terms/
Yahoo!知恵袋：http://chiebukuro.yahoo.co.jp/
Yahoo!知恵袋ガイドライン：http://chiebukuro.yahoo.co.jp/docs/guidelines.html
Yahoo!知恵袋ヘルプ：http://help.yahoo.co.jp/help/jp/chiebukuro/

計画班研究発表

3月18日（日）第二日目　13:15〜16:10

『現代日本語書き言葉均衡コーパス』の基本設計について

▶山崎 誠

セグメントとリンクに基づくコーパス・アノテーション・ツールの設計と実装

▶徳永 健伸、乾 健太郎、野口 正樹、三好 健太、飯田 龍、小町 守

単独ラベラによる大規模アクセントラベリングとそれを用いた統計的アクセント結合処理の実装

▶峯松 信明、黒岩 龍

因子分析を用いた程度副詞と述語等の共起関係の研究試論

▶服部 匡

日本語教育における語彙シラバスの作成について

▶山内 博之

国語教育と語彙指導

▶鈴木 一史

共起関係およびコロケーションに関する研究の流れ
— 計量言語学分野、自然言語処理分野および辞書データなどを中心に —

▶荻野 綱男、荻野 孝野

語彙概念構造辞書の構築による意味役割分析

▶竹内 孔一

『現代日本語書き言葉均衡コーパス』の基本設計について

山崎誠（データ班班長：国立国語研究所研究開発部門）†

Basic Design of the "Balanced Corpus of Contemporary Written Japanese"

Makoto Yamazaki (Dept. Lang. Res., National Institute for Japanese Language)

１．現代語研究と言語資料

過去の言語や自分の母語でない言語の研究にとってそうであるように、現代語研究にとっても言語資料は必須のものである。しかし、現代語の言語資料は、体系的に捉えられたことが少なく、まして、現代語の言語資料を組織的に整備するという試みはほとんど行われなかった。例外として挙げるならば、国立国語研究所の語彙調査をはじめとする記述的調査研究、奥田靖雄氏らを中心とする言語学研究会の活動がある。現代語研究のデータが未整備である理由としては、素材となる言語事実が身近に存在するため、適宜それを拾ってくれば用が足りた、あるいは、内省による観察が可能であることが考えられる。インターネットが普及した現在、使用例を探すことは以前に比べて格段に容易になった。

ただし、言語事実の量的な側面を観察したり、ジャンルによる分布を調べたりするためには、内省や身近な素材の観察ではなく、その目的に沿って組織的に集めた言語資料を活用しなければならない。田野村(1994)は、次のように指摘している。

「ある言語形式の用法に関わる統計的な偏りの現象は内省だけでは正確な把握が困難であるが、大量の用例を調査・分析することでその実相が浮かび上がってくる。」

身近で手に入りやすい素材を集めただけのデータは、過去の事実の積み重ねがただちに歴史として成立しないように、言語資料としてのコーパスとは呼べない。コーパスとして成立するためには、収集の目的と方法が必要である。Francis(1992)は、言語研究を目的とせず収集された資料をコーパスの定義から外れるとしてしりぞけている。

２．現代日本語書き言葉均衡コーパス

国立国語研究所は、1948 年の創立当初より、現代日本語の書き言葉、話し言葉の実態を明らかにする調査研究を行ってきた。この実態解明はそれ自体を自己目的とするものではなく、国民あるいは日本語を母語とする人々の言語生活の向上を図ることを目指したもので、さまざまな個別具体的な目標と方法論とに裏打ちされたものである。書き言葉の語彙調査で言えば、基本語彙の選定という目的のために、調査単位が設計され、統計理論に基づくランダムサンプリングが採用された。我々が平成 17 年度から構築を開始した『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese 、以下 BCCWJ と略す)は、目的の一つにその流れを受け継ぐものであり、加えて、文法研究、文字表記研究、文体研究などの日本語学研究、常用漢字の見直し等の国語政策のための基礎資料の提供、国語教育や日本語教育における活用、言語情報処理技術の精度向上など関連する諸領域からの多面的な要請を受けて設計するものである。

† yamazaki@kokken.go.jp

３．BCCWJ の設計方針

３.１ 代表性とは

コーパスについて、その条件が語られるとき、「バランスのとれた」「偏りのない」「代表性(representativeness)のある」という表現がよく用いられる。これらはほぼ同一の内容を指していると言ってよいだろう。コーパスにおける代表性とは、対象となる言語を分析するために過不足のない状態であることを指すが、狭義にはその実現を科学的に保証する統計的代表性のことである。すなわち、母集団からのランダムサンプリングによって得られた標本を母集団の縮図とみなすことである。したがって、縮図に現れた日本語を観察することによって、母集団における日本語の状況を高い精度をもって推測できる。BCCWJ も統計的代表性を基本方針の一つとして掲げる。

ちなみに、代表性に関して否定的な意見を紹介したい。

「文法研究において帰納的方法が重要なものであることはいうまでもないが、現実には、①客観的なものだけでは、用例がそろわない②いくら資料(corpus)の範囲を広げても、具体言語の一部でしかないことには変わりがない③帰納的方法では、どうしても具体言語の現象を説明することに終わってしまう傾向が強いなどの点が、問題になってくる。」北原（1989）

「一般論としては，間違いなく，特定の種類のテキストだけを利用できるよりも多様なテキストを利用できるほうが望ましい。〔中略〕しかし，だからと言って，各種のテキストをブレンドしたコーパスを用いることがそれほど重要なのかということになると，筆者は，まったく懐疑的である。と言うのも，どのような資料をどのような比率でどれだけの量ブレンドすれば日本語の“平均像”を反映するコーパスが得られるのかという問いは，誰も確かな答を知らない難問だからである。〔中略〕そもそもある言語の“平均像”なる概念を想定することができるのかということからして問題であり，この点についても筆者は否定的な見解に傾く。〔中略〕性質を異にする複数の資料を恣意的な基準でブレンドしたコーパスを使うよりも，むしろそれぞれの資料について個別に分析を行い，資料の種類ごとの異同を見るという方法のほうが正確な知見をもたらす可能性もあると思われる。」田野村(2000)

北原(1989)の指摘は、資料の範囲を従来よりも格段に広げた場合には「用例がそろわない」状況は改善されると思われる。田野村(2000)の指摘する「各種テキストをブレンドしたコーパス」の持つ意味への疑問は、言語体系と言語使用を適切に結び付けるモデルが確立していないことを示唆する。言語使用、とくに量的な側面を言語体系の中でどのように位置付けるか、換言すれば定量的分析から定性的分析への関連付けはコーパスを活用した言語研究の最も重要な課題である。

３.２ BCCWJ の前提となる基本方針

BCCWJ を構築するに当たり、大前提となる方針は以下のとおりである。設計に関する諸項目は以下の方針に沿って具体化していく。

(1)現代日本語の書き言葉の縮図となるコーパス

これまで国立国語研究所が行ってきた語彙調査の手法を生かし、コーパスがその母集団の統計的な縮図になるように設計する。それにより、母集団における言語的諸特性の分布が縮図において過不足なく再現でき、母集団における分布を高い精度で推測できるようにする。

(2)幅広い目的に供するコーパス

BCCWJ が活用される研究領域としては、日本語学、言語情報処理、心理学、日本語教育、国語教育、辞書編集、国語政策などが考えられる。特に日本語学、国語政策への応用のために、現代語の幅を広く捉え、言語変化を観察できるようテキストを収集する。

(3)公開可能なコーパス

いくら大規模なコーパスであっても広く利用されないのではその価値が著しく下がる。過去の国語研究所の語彙調査は、その結果である語彙表を主たる成果として発表しており、語彙研究としては公開であったが、文脈を考慮した文法研究にとっては非公開であった。BCCWJ では、文法研究や文体研究にも利用できるよう、著作権処理を施して誰にでも使えるよう公開する。公開の形態は、無償公開（オンライン）と有償公開（オンライン、DVD）を予定している。コーパスが学界の共有財産となることによって，これまではほとんど行われなかった追試を可能にし，より実証的な研究が推進されると期待される。

(4)既存のコーパスとのスムーズな接続

国立国語研究所は、『日本語話し言葉コーパス(CSJ)』を 2004 年に公開した。翌年、現代語確立期の総合雑誌「太陽」を対象とした『太陽コーパス』を公開した。『CSJ』では「短単位」「長単位」[1]の２種類の解析単位を用いている。また、『太陽コーパス』は、書き言葉特有のルビや字体に関する情報や文章の構造に関する情報を積極的に付与している。これらの既存のコーパスと仕様をそろえることにより、BCCWJ をより上位のコーパス KOTONOHA[2]の中に適切に位置付けることができる。

３.３ BCCWJ の基本設計

現代日本語の書き言葉の全体像は、どのようにして把握されるだろうか。書き言葉は、我々のまわりには、「書かれた言葉」として、客観的に観察可能な状態で存在する。したがって、この書かれた言葉全体を書き言葉の全体像とみなすことができる。書かれた言葉はその媒体（新聞、雑誌、書籍等）によって分類することができ、母集団を把握するのにも便利である。しかし、このように定義した書き言葉を構成する媒体を網羅的に把握することは非常に困難である。そこで、次のような考えのもとに、媒体を選択した。

(1)母集団の範囲を確定でき、かつ影響力の大きい媒体。

(2)必ずしも母集団の範囲を確定する必要がない媒体あるいはジャンルで、特定の研究目的のために収集するもの。

(1)は、媒体の要素となる個々のテキストのリストが存在するものである。BCCWJ では、出版目録や図書館の所蔵目録を利用する。出版目録と所蔵目録とは、それぞれ書き言葉の異なる側面を捉えたものであるから別個のデータとして取り扱う。BCCWJ では、前者を出版物の「生産実態」、後者を「流通実態」と位置付ける。影響力の大きい媒体としては、新聞、雑誌、書籍が挙げられる。

(2)には２つのタイプがある。一つは、白書のようにその一部が(1)に含まれる可能性もあるが、それだけでは量的に少ないため、集中的に収集するもの、もう一つは、Web 掲示板

---

1 「短単位」、「長単位」は、元は国立国語研究所の語彙調査で使われた調査単位。具体的な規則の集合として規程される。近似的な表現として、短単位は形態素の１回結合までに相当する単位、長単位は文節に相当すると言える。

2 国立国語研究所の言語コーパス整備計画の総称。書き言葉、話し言葉の両方にわたり、明治以降の日本語を対象とする。

のように母集団の確定が困難なため、特定のデータをその例として採用するものである。いずれも、特定の研究目的を指向したデータである。

以上を踏まえて設計した現代日本語書き言葉均衡コーパス(BCCWJ)の基本的な構成を字図１に示す。

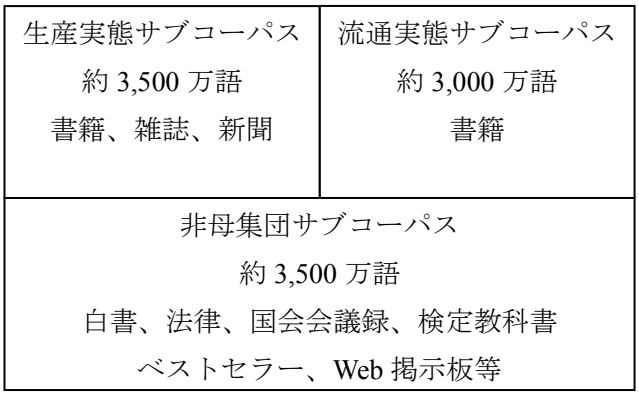


図１　BCCWJ の基本構成

以下、具体的に個々の項目について解説する。

４．BCCWJ の具体的設計

４.１ 生産実態サブコーパス

書き言葉が生産される時点における実態を捉えることを目的とする。母集団を既存のデータにより確定することができ、かつ、言語生活に大きな影響力を持つ媒体である書籍、雑誌、新聞を収録対象とする。

書籍は、国立国会図書館の所蔵データである J-BISC より抽出した 2001 年～2005 年発行の刊行物 317,117 冊[3]を母集団とする。雑誌は、2001 年～2005 年の間に社団法人日本雑誌協会に加盟していた出版社の発行する雑誌 1,259 タイトル、総数 55,779 冊を母集団とする。新聞は、2001 年～2005 年に発行された、全国紙 5 紙（朝日、毎日、読売、日経、産経）、ブロック紙 3 紙（北海道、中日、西日本）、地方紙 8 紙（河北新報、新潟日報、京都新聞、神戸新聞、中国新聞、高知新聞、愛媛新聞、琉球新聞）を母集団とする。

書籍に比べて雑誌、新聞の母集団が小さめの算定[4]になっているのは、母集団の確定に時間がかかり、また、資料の入手という点において困難が予想されることから、実際にコーパスを構築した際に誤差が少ない範囲で母集団を決定したためである。

４.２ 流通実態サブコーパス

生産実態で把握した書き言葉のうち、書籍は比較的長い間世の中に流通する。流通の主な舞台は書店と図書館であるが、所蔵リストが整備されている図書館所蔵データを母集団とした。具体的には、東京都内の 51 自治体[5]の公共図書館の所蔵図書である。これは、東京都立中央図書館より ISBN 総合目録の提供を受け、そのデータを利用した。51 自治体の共通所蔵の分布は図２のとおりである。51 館すべてに所蔵されている書籍は、異なりで 2,305

---

3 漫画、写真集、人名録のように言語表現を主体としないものなどを除いた数。

4『雑誌新聞総合かたろぐ』（メディア・チサーチ・センター発行）記載されている雑誌数は約 18,000、新聞等は約 4,000 である。

5 ISBN のデータが得られなかった世田谷区、檜原村、島嶼部を除いている。

冊、逆にどれか一つの館に所蔵されているものは、異なりで 1,146,418 冊である。実際には、これらの中間に母集団を求めることになるが、例えば、生産実態の書籍部分の冊数 317,117 に一番近いところを選ぶと、18 館以上に所蔵される書籍（321,502 冊）となる。

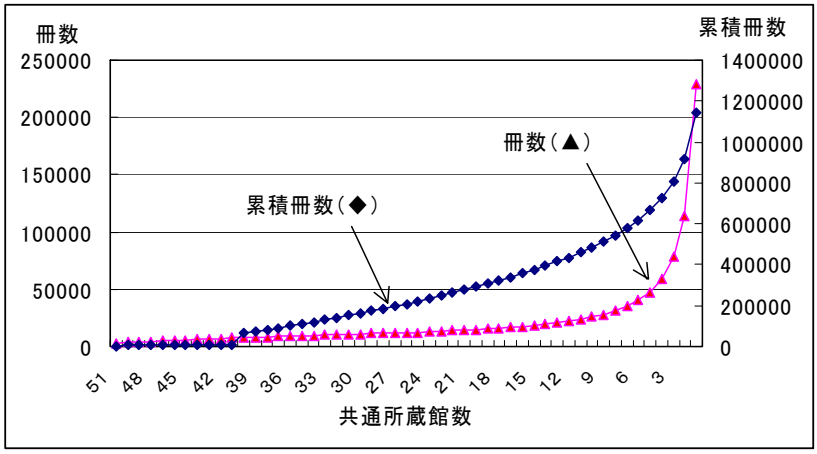


図２　51 自治体共通蔵書の分布

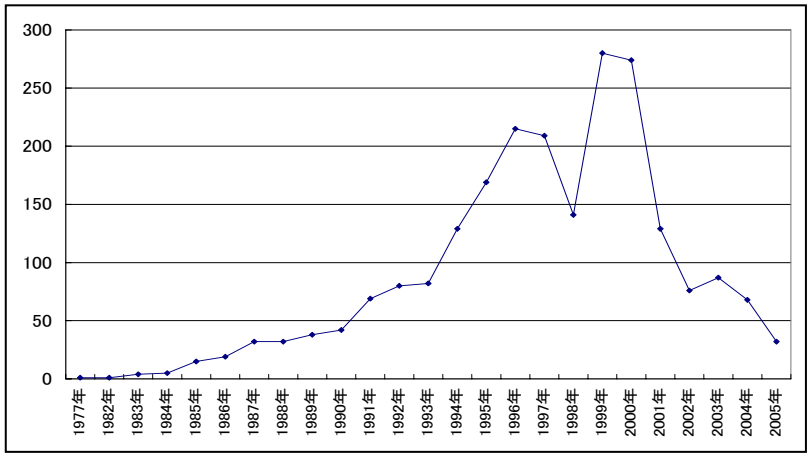


図３　51 自治体共通蔵書の年代分布

図３は、51 自治体すべてに所蔵されている 2,305 冊の年代分布を調べたものである。80 年代の最初のころが少ないのは、古い本は除籍されてしまった可能性と ISBN が付与されていないためデータに現れてこない可能性の二つが考えられる。2001 年以降の冊数が落ち込んでいるのは、東京都の予算や図書館間の分担収集など図書館行政的な理由が想定される。流通実態サブコーパスの母集団はこれから確定するが、そこに含まれる年代の分布は、図３の曲線に近いものになるだろう。

４.３　非母集団サブコーパス

生産実態、流通実態のサブコーパスは、その中に異なる媒体や異なるジャンルを包含した汎用的なデータであるが、それが故に、これらのサブコーパスではとらえにくい資料やジャンルが出てくる。例えば、政府系の白書は生産実態にも流通実態にも存在するはずであるが、分析に必要なだけの量が得られない。同様にベストセラーだけを分析するという

こともできない。これらの要請にこたえるために、特定の資料やジャンルを集中的に格納する部分が非母集団サブコーパスである。特に、国立国語研究所の研究活動に必要な公的な性格の資料などをここに収める。必ずしも母集団からのランダムサンプリングによらないため非母集団サブコーパスと呼ぶ。

４.４ コーパスの規模

コーパスの構築に当たって、最も重要な数値目標はコーパスの規模、すなわち収録語数である。コーパスの規模は、開発期間や予算と直結するものだからである。また、収録語数やテキストの種類は、研究目的の成否にもかかわる。所与の目的のためには、どれくらいの量の語数が必要かが分かれば、それに沿って計画を進めるのが効率的である。

BCCWJ は、全体で１億語超を目標とする。各サブコーパスの規模は、図１に示したとおりである。ここでいう「語」とは、短単位で数えたときのものであり、句読点などの記号は含んでいない。

生産実態サブコーパスの収録語数は、約 3,500 万語であるが、これは以下のような手順で決定した。なお、詳細については丸山他(2007)を参照。

(1)書籍、雑誌、新聞の母集団を確定する。
(2)それぞれの母集団に含まれる文字数を推計する。
(3)書籍、雑誌、新聞の量的構成比に従って、それぞれの必要サンプル[6]数を決定する。
(4)書籍、雑誌、新聞の必要サンプル数にそれぞれの平均サンプル長を掛け、総語数を計算する。

流通実態サブコーパスの収録語数は、約 3,000 万語であるが、これは、上記の生産実態サブコーパスの書籍の総語数に合わせたものである。非母集団サブコーパスの規模は、それを構成する個々のデータの合計である。

４.５ サンプルの長さ－固定長と可変長－

サンプルの長さは、研究目的と結び付く重要な問題である。サンプルの長さについては、検討しなければならないことが２つある。まず、サンプルの長さを固定とするか、可変とするかという問題である。周知のように、Brown Corpus は各テキストが 2,000 語という固定長であった。それにならったとされる LOB Corpus も各テキストの長さは 2,000 語で固定長である。一方、British National Corpus のように、テキストの長さを決めず、短いテキストと長いテキストが混在しているものもある。

統計的な厳密さを求めるならば、固定長が有利なわけであるが、文脈を考慮した文法研究や談話研究、文章構造の研究のためには可変長が適している。この２つはどちらが優れているというものではないので、BCCWJ では、サンプルの長さとして固定長のサンプルと可変長のサンプルの両方を作成することにした。以下に、それぞれのサンプルの特徴を述べる。

(1)固定長サンプル

固定長のサンプルは、句読点などの記号類を除く 1,000 字から構成される。1,000 字は、

6 本稿では、特に断りのない限り「サンプル」を母集団から抽出される個々のテキストの意味に用いる。統計理論における用法と異なることに注意されたい。

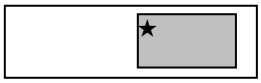
短単位に換算するとおよそ 590 語である[7]。BCCWJ では、これまでの語彙調査の経験を踏まえ、生産実態サブコーパスから固定長サンプルを 1,000 万語取得するという目標を立てた。生産実態サブコーパスには、書籍、雑誌、新聞を格納し、これらの媒体における統計的に厳密な調査結果が同時に得られることを意図したものである。従来の語彙調査は対象となる媒体が一つであったので、二つ以上の媒体を同じ条件で比較することができなかった。固定長サンプルは、流通実態サブコーパス及び非母集団サブコーパスの一部にも存在する。

(2)可変長サンプル

可変長のサンプルは、テキストの長さではなく、内容を基準にして判断する。具体的には、新聞、雑誌の１記事、書籍における章・節などのまとまりが１サンプルに該当する。ただし、無制限に長いサンプルが出来ると、コーパスの分析に影響を与えるため、上限を１万字とする。１万字以内に章や節が収まらない場合は、空行などを手がかりに適宜それより小さいまとまりを１サンプルとする。可変長サンプルはすべてのサブコーパスに存在する。

現実的には、固定長と可変長のサンプルを全く別々に取得するのはコストがかかりすぎるので、１回のサンプリングで当たった同一箇所から固定長と可変長の２つのサンプルを取得することにする。そのため、固定量サンプルと可変長サンプルの関係はおよそ次のようになる。

A：可変長サンプルが固定長サンプルを包含する場合



B：固定長サンプルの一部が可変長サンプルの外に出る場合

C：固定長サンプルが可変長サンプルを包含する場合

図４　固定長サンプルと可変長サンプルの関係

Aは、可変長サンプルが 1,000 字を超える場合で、かつ、固定長サンプルがその中に収まる場合である。書籍、雑誌のサンプルはこのタイプが多いと推測される。図４の★は、ランダムに決められる「サンプル抽出基準点」を表すが、この位置を含む後続の 1,000 文字が節や章などのまとまりを超える場合は、B のタイプになる。C の固定長サンプルが可変長サンプルを包含する場合とは、例えば短い新聞記事などが当たった場合で、固定長サンプルは複数の記事にまたがる 1,000 字となり、可変長サンプルはその中の１記事が該当することになる。

４.６ 文字を抽出単位とすること

もう一つの検討事項は、サンプルの長さを測る単位の問題である。英語のように分かち

[7] １短単位を 1.7 文字として計算。

がちの習慣が確立している言語では、文章の長さを測る単位は分かち書きによって実現された「語」でよいのであるが、日本語のように表記からただちに語を切り出すことができない言語の場合はそれが使えないため、他の客観的な測定法を考えねばならない。国立国語研究所の従来の語彙調査は、統計的に厳密なサンプリングを行っていたが、それは、言語表現の固定長ではなく、エリアサンプリングという手法（佐竹 2001)で、書き言葉を載せる媒体の物理的計測によるサンプリングであり、言わば面積の固定長であった。

BCCWJ では、サンプルの長さを測る方法として、文字を採用する。文字は、分かち書き言語における語と同じ程度の正確さで言語表現の長さを測ることができるからである。コーパスの設計に当たって、エリアサンプリングを採らず、文字サンプリングを採用したのは、表１に示すようなメリットが望めるからである。

表１　サンプリング方法の比較

	エリアサンプリング	文字サンプリング
母集団の把握	○（全ページを数えれば済む）	×（推定する必要がある）
抽出比の正確さ	○（正確）	○（ただし、母集団文字数の推定に依存）
サンプル間の量的統一	×（ばらつきがある）	○（一定に保つことができる）
データ量の把握	×（調査後でないと分からない）	○（最初に目標値を設定できる）

４.７ 媒体の構成比率

生産実態サブコーパスにおいては、４.１節に述べたように、書籍、雑誌、新聞の構成比率を決定する必要があった。詳細は、丸山他(2007)にゆずるが、結果だけを示すと、次のようになる。

書籍：雑誌：新聞＝74.14：16.06：9.80

この比率に従って、固定長部分の 1,000 万語を配分すると、書籍 741.4 万語、雑誌 160.6 万語、新聞 98.0 万語になる。語数が決まったことにより、その分量を確保するためのサンプル数が決まる。１語は 1.7 文字とみなして、語数に 1.7 を掛け、1,000 で割れば必要サンプル数が出る。必要サンプル数は、書籍 12,604 サンプル、雑誌 2,730 サンプル、新聞 1,666 サンプルである。必要サンプル数に各媒体の平均サンプル長を掛ければ、生産実態サブコーパスにおける可変長部分の分量が推定できる。計算の過程は省略するが、結果は、書籍が 2891.5 万語、雑誌が 481.8 万語、新聞が 98.0 万語である。可変長部分は合計で 3471.3 万語である。

なお、この構成比の算出は、各媒体の構成要素である１冊の本、１冊の雑誌、１日分の新聞の異なりにより計算したものである。新聞、雑誌は、対象とするタイトル（要素数）が書籍に比べ少ないが、発行部数を考慮すれば、比率は変わってくるはずである。構成比の算出に発行部数を考慮しなかったのは、書き言葉が生産される原点における使用実態を把握するためである。目的は異なるが、総務省が毎年発表している「情報流通センサス」では、原発信情報量と発信情報量という用語で異なりと延べの関係を区別している。

４.８ 資料の年代

BCCWJ では、現代語の範囲を 1976 年～2005 年の 30 年間に設定した。この間には書き言葉をめぐる状況にいくつかの転機があり、その変化を観察できるようにするためである。一つは、1981 年に告示された「常用漢字表」である。それまで使われていた「当用漢字表」に 95 字の追加があり、それらの使用実態がどのように変化したのかを評価するためである。

もう一つは､1990 年代のパソコンの普及､90 年代後半からの携帯電話の普及にともなって、漢字を仮名漢字変換により入力する機会が増え、漢字の使用実態は変化したのではないかと推測されることである。

生産実態サブコーパスの部分は、2001 年～2005 年の資料を対象とする。この部分は、将来的に同じ仕様で拡張が可能なように期間を短めに設定している。一方、流通実態サブコーパスは、図書館の蔵書は比較的長い期間所蔵されると推測されるため、1976 年～2005 年の 30 年を対象とする。

５．電子化形式

電子化の形式は、文字入力仕様とタグの仕様に分けられる。文字入力仕様は、文字コードに Unicode（UTF16-LE）を、文字集合に JISX0213:2004 を用いる。異体字についての明確な包摂基準を持ち、多くの漢字を詳細に区別することができる。

タグの仕様は､『太陽コーパス』の設計を取り込み、拡張させた形式をとる。タグの設計に当たっては、できるだけ一般的な概念に一致する、作成者、利用者の双方にとって分かりやすい､汎用性のある形式にする｡タグは､以下の 4 つに類別される｡詳細は､山口他(2007)を参照されたい。

(1)書誌情報

サンプリングで獲得される、媒体(新聞、雑誌、書籍など)、書名、著者名、出版年、ジャンル、出版社などの情報を管理できるようにする。著者については、氏名だけでなく、言語分析に有用な生年、性別、出身地などの情報を含めて管理する。これらは、コーパス本体とは別にデータベース化し、サンプル本文内部に付与された ID とリンクして参照できるように設計する。

(2)文書構造情報

記事、見出し、段落、引用、文などの枠組みを用意し、テキストを構造化して表現する。これらの枠組みは、以下の目的を満たすものとして策定したものである。

①ある一定のまとまりをもつ文書（記事を想定）が有する、階層性を表現できること。例えば、sample > article > cluster > paragraph > sentence 等の階層性を持ったタグの親子関係や、cluster の入れ子構造により、文書の論理構造を明示的に示せるようにする。

②title、 abstract、caption など文書内での特別な役割を持つ要素を抽出できること。文書内容把握、文書要約、文書タイプの分類などに活用が期待できる。

③title、 caption、 quotation など、文体・語彙・文法的に差異が見られることが知られる要素を区別すること。それぞれの部分を比較したり、抽出、排除が可能なようにする。

(3)文字情報

文字の読みに関するルビ、原文の誤植及びそれを訂正した情報、文字集合に含まれない文字や記号（外字)、囲み文字、数式・化学式などに現れる上付き・下付き文字などの情報を付与する。

(4)サンプリング情報

サンプリング時に決定するサンプル抽出基準点（乱数による縦横交差点から決まる 1 文字）の情報を付与する。

６．形態論情報

形態論情報とは、言語表現をあらかじめ設計された言語単位に分割する際、分割された言語単位に付与する言語情報のことである（言語単位の設計も含む)。主として、形態素解析ソフトを使って自動的に情報付与がなされるため、語あるいは形態素に関する情報（見

出し語の形、読み、表記、語種、品詞、活用形など）である。形態論情報の基本的な設計方針は、以下のとおりである。

(1)コーパスに基づく用例収集、各ジャンルの言語的特徴の解明に適した単位を設計する。
(2)『日本語話し言葉コーパス』（ＣＳＪ）と互換性のある形態論情報を設計する。
(3)国立国語研究所の語彙調査等における知見を活用する。

解析単位は、目的に応じて２種類の単位を採用した。一つは、用例収集や基本語彙の選定に適している短単位、もう一つは、特徴語の抽出や語構成の分析に適している長単位でである。いずれの単位も国立国語研究所の語彙調査で用いられたものをもとにしている。付加情報としては、短単位・長単位の両方に、代表形（辞書の見出しに相当）、 代表表記（代表形に対して付与した漢字等の国語の表記）、品詞、活用型、活用形などの情報を付与する。

形態素解析は茶筌を利用して自動的に行われるため、解析精度を高めるには、辞書の整備が不可欠である。解析用の辞書は、千葉大学・伝康晴氏らの開発となる UniDic を用いる予定であるが、そこに搭載する短単位のデータベースを国立国語研究所との共同で整備している。この辞書の特徴は、見出し語の認定を言語学的な立場で行うことである。従来の形態素解析システム用辞書は、同じ語であっても語形や表記が違えば、別の語と扱っていた。UniDic では、見出し語のもとに、その見出し語と同じ語の範囲にある語形や表記の情報を盛り込むことによって、言語学的な分析に耐える結果を出すことを目指す。形態論情報の詳細は、小椋他(2007)を参照されたい。

７．今後の予定

BCCWJ を構成する各サブコーパスのうち、生産実態サブコーパスについては、仕様が確定した。流通実態サブコーパスは、母集団の設定の最終段階にある。非母集団サブコーパスは、白書については電子化を終え、近く試験公開の予定である。Web の掲示板（Yahoo! 知恵袋）はサンプリングの方針がほぼ決まり、年度内に収録語数 500 万語分を抽出する予定である。

参考文献

Francis, W.N.(1992) Language corpora B.C. in Svartvik(1992) “Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991”, pp.17-32, Mouton de Gruyter

小椋秀樹、小木曽智信、小磯花絵、他(2007)『現代日本語書き言葉均衡コーパス』における短単位の概要，本予稿集収録

北原保雄(1989)日本語文法理論（井上和子編(1989)『日本文法小事典』所収、pp.5-54）

佐竹秀雄(2001)研究対象の量とサンプリング、「日本語学」20-5，pp.74-83

田野村忠温(1994)丁寧体の述語否定形の選択に関する計量調査－「ません」と「ないです」－「大阪外国語大学論集」11，pp.51-66

田野村忠温(2000)「用例に基づく日本語研究－コーパス言語学－」,「日本語学」19-5，pp.192-201

丸山岳彦、柏野和佳子、山崎誠、他(2007)『現代日本語書き言葉均衡コーパス』におけるサンプリングの概要，本予稿集収録

山口昌也、高田智和、北村雅則、他(2007)『現代日本語書き言葉均衡コーパス』における電子化フォーマットの概要，本予稿集収録

セグメントとリンクに基づく コーパス・アノテーションツールの設計と実装

徳永健伸 * (分担者：東京工業大学 大学院情報理工学研究科)
乾健太郎 † (分担者：奈良先端科学技術大学院大学 情報科学研究科)
野口正樹 (協力者：東京工業大学 大学院情報理工学研究科)
三好健太 (協力者：東京工業大学 工学部情報工学科)
飯田龍 (協力者：奈良先端科学技術大学院大学 情報科学研究科)
小町守 (協力者：奈良先端科学技術大学院大学 情報科学研究科)

A corpus annotation tool based on segments and links — Design and implementation —

Tokunaga, Takenobu (Tokyo Institute of Technology)
Inui, Kentaro (Nara Institute of Science and Technology)
Noguchi, Masaki (Tokyo Institute of Technology)
Miyoshi, Kenta (Tokyo Institute of Technology)
Iida, Ryu (Nara Institute of Science and Technology)
Komachi, Mamoru (Nara Institute of Science and Technology)

1 背景

近年，自然言語処理の分野では，コーパスに基づく統計的手法が研究の中心となっている．これまでに，様々な情報を付与したコーパスが構築されており，それらは統計的手法を用いた解析モデルの学習データや解析システムの評価用テストセットに用いられるなど，統計的自然言語処理の研究にとって必要不可欠な存在となっている．また，コーパスに付与される情報は，多様化，複雑化しており，このために様々なアノテーションツールの開発が行われている．

従来のアノテーションツールは，コーパスに付与する情報に応じて開発されることが多く，そのデータフォーマットなどはアノテーションツールごとに異なった形式を採用している．したがって，複数の異なる情報を同一のコーパスに重畳的に付与しようとすると，付与する情報に応じてツールごとにデータフォーマットなどの変換作業が必要となる．

これまで，様々なアノテーションツールとともに付与される情報の記述形式がいくつか提案されている．例えば，テキストに対し構文や意味の情報を付与した GDA (Global Document Annotation) (Hasida 2005) や，CES (Corpus Encoding Standard) (Ide 2000) が提案されている．

GDA は，図 1 (a) に示すように統語的依存関係をはじめ，代名詞等の照応，共参照， 多義語の語義など様々な情報を XML の形式で表現する．CES は，コーパスのメタデータ，文章の構成などの情報を SGML の形式で表現している (図 1 (b) 参照)．しかし，これらのアノテーション形式は情報の交換を目的としており，アノテーションツール内でこのような表現形式を必ずしも使う必要はない．

また，より汎用性を目指したアノテーションツール Tagrin (高橋，乾 2006) では，異なる情報を複層的に付与することを考慮し，スタンドオフ形式を採用している．たとえば，図 2 において最初の np はテキストのオフセット 1.110 (1 行目の 110 文字目) から 1.113(同 113 文字目) までの属性として定義されており，この文字列には “9” という識別子 (id) が付与されている．

大規模で多様な情報が付与されたコーパスを構築することを考えると，様々なアノテーションで利用できるツールを容易に提供できるように基本となるデータ形式を整備する必要がある．このようなフレームワークのひとつに AGTK (Annotation Graph Toolkit) (Cotton and Bird 2002) がある．AGTK

*take@cl.cs.titech.ac.jp
†inui@is.naist.jp

```
<su>
  <np sem="time0">time </np>
  <v sem="fly1">flies </v>
  <adp>like
    <np>an arrow</np>
  </adp>.
</su>
```

(a) GDA

```
<cesDoc version="3.9">
  <cesHeader version="2.0">
    ...
  </cesHeader>
    <text><body>
        <div> [optional]
          <p>
          <p>
          ...
```

(b) CES

図 1: アノテーション形式の例

```
# A-ID:950101003
村山富市首相は...
また，一九九五年中の衆院...
EOT
950101003
np 1.110, 1.113 id="9"
np 2.41, 2.45 id="12"
...
```

図 2: Tagrin のデータ形式

では，グラフを基本データ構造として，句構造や依存構造など付与する情報ごとにデータに対する操作を提供している．

本論文では，種々のアノテーション情報を抽象化し，それに対する基本操作を定義することで多様な情報を付与する枠組みを提案する．また，この枠組みを利用したアノテーションツールの実装例として，述語項構造のアノテーションツールと句構造のアノテーションツールを紹介する．

2 データ形式の抽象化

データ形式の抽象化およびアノテーションツールの実装にあたり，アノテーションごとに作業のしやすいインターフェースを採用することは，スムーズなアノテーションにもつながると考えられる．そのため，データ形式の抽象化に当たり，付与する情報とインターフェースを区別して扱うことにする．すなわち，生成や削除といった付与する情報に対する基本操作は，付与する情報に依らず共通化し，インターフェースは付与する情報に応じて適切なものを実装する．そして，インターフェースと基本操作を結びつけることによって柔軟な設計と実装が可能になる．

具体的には，あらゆる情報をセグメントとリンクの 2 つのオブジェクトのみを用いて形式化し，付与する情報に関するより詳細な属性や制約についてはアノテーションごとに各設定項目を設定ファイルに定義する．

以下，各オブジェクトと設定項目の役割と表現方法について説明する．

2.1 セグメント

アノテーションの対象となるドキュメント中の任意の文字列とその文字列に付与されたラベルの対をセグメントと呼ぶ．セグメントにおける文字列はドキュメント中の開始位置と終了位置のオフセット値のペアで表現し，ラベルは設定ファイルで定義する．セグメントオブジェクトは表 1 に示す属性を持つ．

2.2 リンク

リンクは 2 つのセグメントオブジェクト間の関係を表すオブジェクトで，2 つのセグメントオブジェクトの組 (source, destination) で表現する．また，ラベルは，セグメントオブジェクト間の関係を

表 1: セグメントの属性

属性名	属性値
segmentID	セグメントオブジェクトの ID
documentID	作業中のドキュメントの ID
label	ラベル
start	文字列の開始位置 (オフセット値)
end	文字列の終了位置 (オフセット値)

表す情報で，設定ファイルで定義する．リンクオブジェクトは表 2 に示す属性を持つ．

表 2: リンクの属性

属性名	属性値
linkID	リンクオブジェクトの ID
documentID	作業中のドキュメントの ID
label	ラベル
source	リンク元のセグメントオブジェクトの ID
destination	リンク先のセグメントオブジェクトの ID

2.3 設定ファイル

アノテーションによって付与される情報は，その目的によって異なるため，設定ファイルでは，ラベルをキーとして，付与する情報が持つ属性-属性値を定義する．表 3 に品詞タグ付けにおける定義の例を示す．たとえば，label の値が “動詞-接尾-サ変” である場合，このセグメントオブジェクトは，セグメントの属性として，conjugation が “true”，conjugationtype が “サ変” という値を持つことを表す．

表 3: セグメントの属性 (例:品詞タグ付け)

属性名	属性値
label	動詞-接尾-サ変
conjugation	true
conjugationType	サ変
...	...

表 4 に述語項構造のアノテーションにおける定義の例を示す．この表から label の値が “ガ格” であるリンクオブジェクトは，リンクの性質として，有向性 (directed=“true”) であるが，推移性 (transitive=“false”) はなく，リンクの方向が動詞から名詞句へ向かうという性質を持つことがわかる．

また，属性 directed を “false” にすることで無向リンクを表現できるので，セグメントオブジェクトの集合は推移的な無向リンクで接続されたセグメントオブジェクトの集合として表現できる．

3 ツールの実装例

この節では，本論文で提案した枠組みを用いて既存のツールと同様の機能を再実装した例について述べる．アノテーションごとに基本操作を実現する機能を実装している．

再実装したいずれのツールも，ドキュメントの表，セグメントの表，リンクの表を持つデータベースをバックエンドに持ち，インタフェースと通信しながら作業をする構成になっている (図 3 参照)．

表 4: リンクの属性 (例:述語項構造)

属性名	属性値
label	ガ格
directed	true
transitive	false
srcLabel	動詞
dstLabel	名詞句
...	...

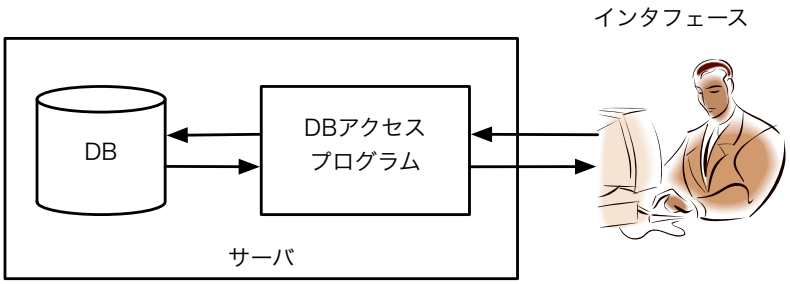


図 3: システム構成

3.1 Tagrin

Tagrin (高橋，乾 2006) は情報抽出や照応解析など多様なコーパス作成を目的としたアノテーションツールである．提案手法を用いることで，任意の文字列をセグメントで表し，文字列の関係をリンクで表すことができる．

例えば，述語に対して格関係の情報を付与する述語項構造のアノテーションの場合，ユーザは対象とする述語を選択した後に，格関係にある文字列を選択し，選択された述語と文字列との間の関係を付与する．再実装したツールでは，述語ごとに格関係を付与する操作を次のように実現する．

1. リンク元となるセグメントをキーボードの shift キーを押しながらクリックで選択する．もしくは，矢印キー [←, →] で選択する．
2. リンク先となるセグメントをクリックで選択する．もしくは，文字列を選択する．
3. リンクのラベルに対応するキーを押す．

共参照関係についても同様に，共参照関係にある要素に対応するセグメントを生成し，それらを共参照リンクで結ぶことによってアノテーションが可能である．この場合，設定ファイルにおいて，共参照リンクは，無向リンクとして定義しておく．

また，これらの機能が使いやすいようにインターフェースを実現するため，設定ファイルにインターフェースの情報として，表 5 に示す属性を加えて実装した．各ラベルに入力キーを対応付けることによって，ラベル名を人手で入力する必要は無い．また，文字列を選択してリンクを作成した場合には，リンク先に対応するラベルを持つセグメントを自動的に作成する．

図 4 に実装したインタフェースのスナップショットを示す．

3.2 eBonsai

eBonsai (市川他 2005) は句構造のアノテーションツールである．eBonsai のアノテーション作業は，入力文に対してパーザが出力した複数の句構造候補から正しい句構造を選択するタスクが中心になる．提案手法を用いると，非終端記号をセグメントで，親子関係をリンクで表現することにより eBonsai と同等の機能を実現できる．

表 5: リンクの追加設定項目

属性名	属性値
label	セグメントオブジェクト間の関係
...	...
keybind	リンク作成時の入力キー
srcfcolor	リンク元セグメントの文字色
srcbcolor	リンク元セグメントの背景色
dstfcolor	リンク先セグメントの文字色
dstbcolor	リンク先セグメントの背景色

eBonsai での基本操作は，不要なセグメント，リンクを削除することである．再実装したツールではこれらの操作を行うために次の 2 つの機能を実装した．

1. 複数の候補セグメントが存在する文字列 (非終端記号) から正しいセグメントを選択し，それ以外のセグメントを削除する．
2. 複数の候補リンクが存在するセグメントから正しいリンクを選択し，それ以外のリンクを削除する．

また，階層構造が重要になるため，候補を句構造表示するインターフェースを採用している．図 5 は実装したインタフェースのスナップショットである．

4 まとめと今後の課題

本論文では，種々のアノテーション作業時に保持するデータをセグメントとリンクに抽象化し，それらに対する基本操作を定義することで多様な情報を付与する枠組みを提案した．また，この枠組みを利用したアノテーションツールの実装例として，述語項構造のアノテーションツールと句構造のアノテーションツールを紹介した．

現在の実装では，データに対する基本操作とインタフェースの関係を設定ファイルによって完全に記述できるようにはなっておらず，一部，インタフェースのプログラム中に直接記述している．今後，この点を改善し，設定ファイルの記述能力をより高めて，異なるアノテーション間の移行をより容易におこなえるようにする予定である．

また，現状では，作業者のアノテーションの精度は，作業者のアノテーションスキル (学習具合) に依存している．コーパスの精度を高めるためには，作業者がアノテーションをする際の意志決定の支援を実現することも重要である．現在の枠組の上にこのような機能も実現する必要がある．

参考文献

Scott Cotton and Steven Bird. (2002). An integrated framework for treebanks and multilayer annotations. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1670–1677. ELRA.

Koiti Hasida. (2005). Global document annotation (gda). http://i-content.org/gda.

Nancy Ide. (2000). Corpus encoding standard. http://cs.vassar.edu/CES/.

高橋哲郎，乾健太郎. (2006). アノテーションツール “Tagrin” の紹介. **言語処理学会第 *12* 回年次大会予稿集**.

市川宙, 野口正樹, 吉田恭介, 橋本泰一, 徳永健伸, 田中穂積. (2005). 構文木付きコーパス作成支援統合環境：eBonsai. **言語処理学会第 *11* 回年次大会予稿集**.

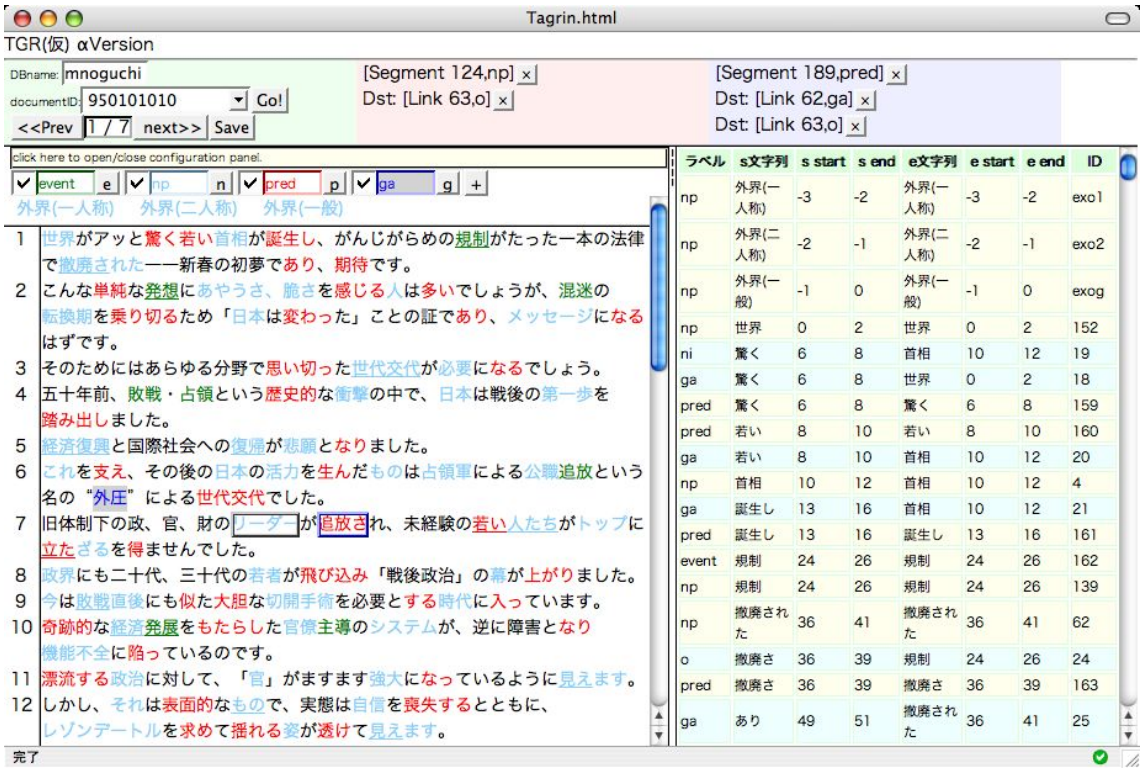


図 4: Tagrin の実装例

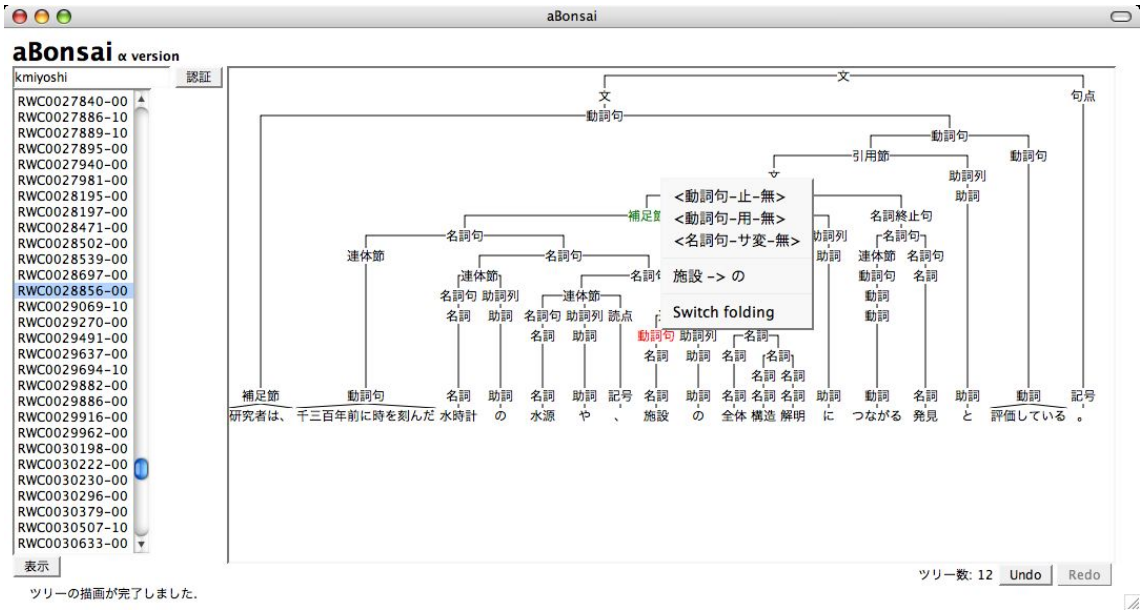


図 5: eBonsai の実装例

単独ラベラによる大規模アクセントラベリングとそれを用いた統計的アクセント結合処理の実装

峯松 信明（電子化辞書班分担者: 東京大学新領域創成科学研究科）†
黒岩 龍（電子化辞書班協力者: 東京大学情報理工学系研究科）‡

Development of a large-scale accent corpus labeled by a single labeler and its use for statistical learning of word accent sandhi

Nobuaki Minematsu (The University of Tokyo)
Ryo Kuroiwa (The University of Tokyo)

1 はじめに

日本語テキスト音声合成システムを構築する場合，一般的には，次のような言語処理系／波形生成系が必要となる。1) 形態素解析を行ない，形態素境界を特定し，各形態素の読みやアクセント型（各形態素を単独で読み上げた際のモーラ列・音素列とアクセント型）の情報を得る，2) 無声化，連濁などの音韻処理，アクセント結合・イントネーション・継続長・パワーに関する韻律処理を行なう，3) 得られた情報を基に，波形生成を行なう。

従来筆者らは，アクセント句境界が与えられた場合に，句内のどのモーラをアクセント核として波形生成するのか（アクセント結合問題）を，各形態素（自立語／付属語）のアクセント属性を定義し，規則によってアクセント変化を記述することでシステム構築を行なって来た [1, 2]。しかし，全ての事象を規則で網羅することには限界があり（例えば [1] では副次アクセントや，付属語連鎖への対処が行なわれていない），アクセントラベリングが施された大規模なコーパスを用いた機械学習・統計学習で解決を図ることも行なわれるようになってきた [3]。

コーパスベースの方法論を検討する場合，当然，大規模コーパスは必須のものとなるが，現時点で，高品質のアクセントラベリングが施され，研究目的で自由に利用可能なコーパスは存在していない。これらの現状を鑑み本研究では，1) アクセント結合処理モジュールを構築可能な，大規模かつ高品質なアクセントラベリングが施されたコーパスの構築と，2) それに基づく（かつ，従来の規則ベースのアクセント結合処理を踏まえたハイブリッド的な枠組みを持つ）統計的なアクセント結合処理モジュールの構築を試みる。

2 特定ラベラによる大規模アクセントラベリングコーパス

2.1 ラベリング対象とする言語事象

日本語東京方言で文を発声する際，文は幾つかのまとまりに分かれ，各々の内部では音の高さ（ピッチ）が連続的に変化する。まとまりが始まる箇所ではピッチの上昇が見られ，その後，まと

† mine@k.u-tokyo.ac.jp, ‡ kuroiwa@gavo.t.u-tokyo.ac.jp

まりの内部では上昇は無く，ゆるやかに下降してゆく。まとまりの内部には，語彙に依存した比較的急激なピッチの下降箇所が概ね高々1つ存在する。このようなまとまりをアクセント句と呼び，急激な下降の箇所をアクセント核と定義するのが一般的である。

しかし，実際の発声におけるピッチ変化を観察すると，明確なアクセント句の分離が困難な場合も多く，複数の句が影響を及ぼし合い，融合する現象も見られる。これについて，日本語話し言葉コーパス [4] では，後続アクセント句の句頭上昇が見られない程度まで融合が起きていれば1つのアクセント句として扱うものとするが，同程度の融合が見られても双方が核を持つアクセント句である場合は,「アクセント句には1つの核しか存在し得ない」との大前提に従い，複数のアクセント句として扱っている。

アクセント核についても，発声者や聴取者にとって知覚されているにも拘らず，実際のピッチ変化に明確に現れていない場合がある [4]。更には，発声者等の知覚も必ずしも明確ではない。

このようにアクセント句・核は明確に（物理的に）定義できるものでないが，ラベリング作業を行なう（行なわせる）場合，何らかの定義が必要となる。本研究では下記の言葉でこれらを定義し，ラベリング対象とした。

アクセント句境界 ピッチの句頭上昇が見られる箇所をアクセント句境界とする。視覚提示される話速（約7モーラ/秒）に合わせて自然に読んだ場合を想定する。休止が入った場合も，通常は境界が生じる。

アクセント核 ピッチが急激に下降する箇所の直前のモーラ。句内の出現回数は制限しない。意識的に当該箇所で急激にピッチを下げ，それ以外では同じピッチで平坦に（イントネーションを除去して）読んだ場合に，違和感が生じない位置。

上記のラベリング対象は，文読み上げ時の事象である。これとは別に，個々の自立語を単独で発声する場合のアクセント型もその対象とした。後述するように本ラベリングは特定のラベラによる作業となる。単独発声時のアクセント型は，アクセント辞典等に掲載されているが，アクセント感覚の個人差を含まない高品質のコーパス構築を念頭に置き，単独発声時に対するラベリングも作業項目とした。

最終的に，本研究におけるラベリング対象は，1) 文発声時のアクセント句境界とアクセント核位置，及び，2) 文中の全自立語に対する単独発声時のアクセント核位置，である。なお，付属語については，単独発声を想定すること自体が困難であり，また，不合理とも考えられるため，これらに対する単独発声時のアクセントラベリングは行なわないこととした。

2.2 ラベラ・ラベル検査者の選定

アクセント感覚には個人差があるため，本研究では特定ラベラに全ラベリングを依頼することとした。作業量が膨大となるため，誤ラベリングも免れ得ない。そこで，付与されたラベルを検査する（誤りが含まれる可能性のある文を選定する）検査者をラベラとは別に用意した。東京生まれ・育ちであり，合唱部に所属する比較的音感の鋭い大学生6名に対して，日本語アクセントに対する教育を施した上で，試験により選抜し，最終的にラベラ1名，検査者1名（今後増員する可能性あり）を選定した。

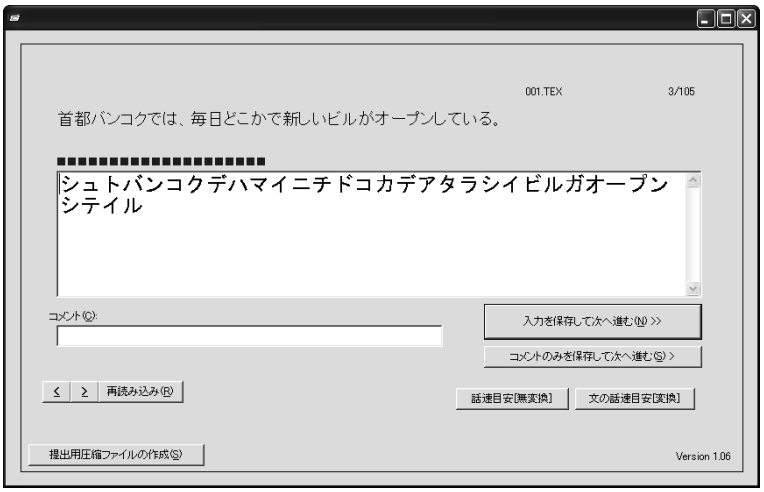


図 1: ラベリング作業用インタフェース

2.3 使用した文セット

新聞記事読み上げ音声コーパス（JNAS）で使用されている文（毎日新聞記事から抽出した 16,178 文および ATR 音素バランス文 503 文）を用いた。既に音声コーパス用に使用されている文であり，全文に凡そ正しい読みが与えられていること，砕けた文が少なく扱いやすいことなどが選定理由である。但し，不自然な読みや誤った読みも散見されたため，事前に全文の読みを確認し，適切でないと思われる箇所に対しては，適宜，修正を施した。

2.4 形態素解析

文中の自立語に対するラベリング作業を行なうこと，及び，ラベリング結果を利用する際の利便性のため，全文に対して形態素解析を行なった。形態素解析の品詞体系は UniDic に準拠した。形態素解析作業は，辞書として UniDic を用いて自動的に行なった後，用意した読みと異なった読みが付与されたものを中心に手修正した。従って，全て正確に形態素解析が行なわれている訳ではないが，読みについてはラベリング作業に用いたものと完全に同一になっている。

2.5 実際のラベリング作業

ラベリング作業では，音声の録音・聴取はせず，テキストに直接ラベルを付与させた。これは，音声を経ると極めて作業時間が長くなるためであり，また，ラベラの言語的内省としてのアクセント情報を引き出すことを重視するためである。作業用に，図 1 に示したインタフェースを用意し，単純な操作でラベリング作業ができるようにした。

単独発声時の自立語のラベリングでは，完全な単独発声では曖昧性が残るため，名詞には助詞「が」を後続させた状態でラベリングをさせ，形容詞には名詞「こと」が後続することを想定してラベリングをさせる（「こと」自体はラベリングの対象としない）などの工夫を行なった。この作業により，

- **シュ’ト／バ’ンコクデ’ハ／マ’イニチ／ド’コカデ／アタラシ’イ／ビ’ルガ／オ’ープン／シテイル**

	A	B	C	D	E	F	G	H	I	J	K
1	まだ	マダ	未だ	マダ	副詞	*	*	マ’ダ／	2	1	1
2											
3	正式	セイシキ	正式	セイシキ	名詞-普通名詞-形状詞可能	*	*	セイシキ	4	-1	-1
4	に	ニ	だ	ダ	助動詞	助動詞-ダ	連用形-ニ	ニ／	1	*	-1
5											
6	決まっ	キマッ	決まる	キマル	動詞-一般	五段-ラ行-一般	連用形-促音便	キマッ	3	-1	-1
7	た	タ	た	タ	助動詞	助動詞-タ	基本形-一般	タ／	1	*	-1
8											
9	わけ	ワケ	訳	ワケ	名詞-普通名詞-一般	*	*	ワ’ケ	2	1	1
10	で	デ	で	デ	助詞-格助詞	*	*	デ’	1	*	1
11	は	ハ	は	ハ	助詞-係助詞	*	*	ハ／	1	*	-1
12											
13	ない	ナイ	無い	ナイ	形容詞-非自立可能	形容詞-ア段-無イ+ない	基本形-一般	ナ’イ	2	1	1
14	の	ノ	の	ノ	助詞-準体助詞	*	*	ノ’	1	*	1
15	で	デ	だ	ダ	助動詞	助動詞-ダ	連用形-一般	デ	1	*	-1
16											
17	カネ	カネ	金	カネ	名詞-普通名詞-一般	*	*	カネ	2	-1	-1
18	の	ノ	の	ノ	助詞-格助詞	*	*	ノ	1	*	-1
19	カ	チカラ	力	チカラ	名詞-普通名詞-一般	*	*	チカラ’	3	3	3
20	も	モ	も	モ	助詞-係助詞	*	*	モ／	1	*	-1
21											
22	十分	ジュウブン	十分	ジュウブン	副詞	*	*	ジュウブ’ン／	4	3	3
23											
24	知っ	シッ	知る	シル	動詞-一般	五段-ラ行-一般	連用形-促音便	シッ	2	-1	-1
25	て	テ	て	テ	助詞-接続助詞	*	*	テ	1	*	-1
26	い	イ	居る	イル	動詞-非自立可能	上一段-ア行	連用形-一般	イ	1	-1	-1
27	た	タ	た	タ	助動詞	助動詞-タ	基本形-一般	タ	1	*	-1
28											
29	首都	シュト	首都	シュト	名詞-普通名詞-一般	*	*	シュ’ト／	2	1	1
30											
31	バンコク	バンコク	バンコク	バンコク	名詞-固有名詞-地名-一般	*	*	バ’ンコク	4	1	1
32	で	デ	で	デ	助詞-格助詞	*	*	デ’	1	*	1
33	は	ハ	は	ハ	助詞-係助詞	*	*	ハ／	1	*	-1

図 2: 形態素解析結果と対応付けたアクセントラベル

のような文発声ラベリングと、

- **シュ’トガ**
- **アタラシ’イ**
- **スル**

のような形態素単独発声ラベリングが得られる。なお,［／］がアクセント句境界位置を,［’］がアクセント核位置を表している。ラベル検査者から誤りの可能性を指摘された文については，ラベラに確認させ，必要に応じて，訂正させた。

2.6 進捗状況と分析

2007 年 1 月現在，4,166 文（JNAS の先頭 40 ファイル）について，文発声・形態素単独発声の双方のラベリングが完了している。今後，用意した全ての文に対してのラベリングを目指して進めていく。形態素解析結果に対して，単独発声/文中アクセント型を追記したものが図 2 である。I, J, K 列にラベリング結果が示されている。数値の意味については，第 3.2 節にて説明する。作業が完了した全文に対して，アクセント句を構成する形態素数，及び，アクセント句を構成する品詞列の上位 10 種類を表 1，表 2 に示す。4 形態素以下のアクセント句が 90%以上を占めていることや，品詞列は 11 位以下の低い出現率のものの合計がおよそ半数に達することなどが読み取れる。以降の節では，構築したコーパスを用いたアクセント結合処理モジュールについて検討する。

表 1: アクセント句を構成する形態素の数

形態素数	出現数	出現率
1	5079	17.4%
2	9829	33.6%
3	7902	27.0%
4	3972	13.6%
5	1586	5.4%
6	554	1.9%
7 以上	303	1.0%

表 2: アクセント句を構成する品詞列（上位 10 種）

品詞列	出現数	出現率
[名][助]	5273	18.0%
[名]	2639	9.0%
[名][名][助]	2180	7.5%
[名][接尾][助]	1409	4.8%
[動][助動]	792	2.7%
[動]	788	2.7%
[名][名]	758	2.6%
[名][接尾]	739	2.5%
[動][助]	571	2.0%
[名][助][助]	541	1.9%
上記以外	13535	46.3%

3 CRF を用いたアクセント結合

3.1 条件付確率場

観測データ $\boldsymbol{x}$ に対する出力ラベル $\boldsymbol{y}$ を学習するに際し，条件付確率場（CRF）[5] は $(\boldsymbol{x}, \boldsymbol{y})$ 内での連続する変数の組（y_{t-1} と y_t, y_t と x_t など）の関係についての独立した特徴（素性）f を列挙し，各素性 f の重要度を θ_f，$(\boldsymbol{x}, \boldsymbol{y})$ 内で素性 f が満たされている箇所の数を $\phi_f(\boldsymbol{x}, \boldsymbol{y})$ とおいた上で，入力 $\boldsymbol{x}$ に出力 $\boldsymbol{y}$ を割当てることの確信度として，$\sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y})$ を考え，これを，

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y} \in Y} (\exp \sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y}))}$$

として確率分布とする。正解データを与えることによる学習は，この確率をできるだけ大きくするような重要度 θ_f を探る作業となる。本研究では，CRF++[6] を用いてアクセント結合処理を実装した。

3.2 学習・推定の内容

アクセント句境界は事前に与えられているものとし，各句中のアクセント核位置について学習・推定した。これらの情報は，前節で構築したコーパスより得られる。アクセント句境界位置情報を使用してアクセント句に区切り，句に含まれる各形態素の文中アクセント型を学習・推定の対象とした。例えば「音声合成」であれば「オンセー」（無核）と「ゴーセー」（1 モーラ目に核）という文中アクセント型を，単独発声時の「音声（オンセー）」と「合成（ゴーセー）」から推定する枠組みである。なお，形態素境界に核が生じているもの（「流れ・を（ナガレ・オ）」「出来・ない（デキ・ナイ）」など）では，当該アクセント核は前側の形態素に属するものとした。このようにして形態素解析結果，単独発声/文中アクセント型をまとめた結果が図 2 であり，列 H が文中アクセントラベル済みテキストを形態素に対応付けたもの，列 K が形態素の文中アクセント型，列 J が形態素の単独発声アクセント型である。なお，“–1” はアクセント核が存在しないことを表し，“*” は単独アクセント型ラベリングの対象となっていない形態素であることを表す。その他の数値は核のあるモーラの位置を表している。

但し，形態素内にアクセント句境界が存在している文や複数のアクセント核を持つ形態素を含む文は文ごと除去した。最終的に，学習用 3,581 文（25,692 アクセント句），推定用 527 文（3,533 アクセント句）に分けて使用した。JNAS の 40 ファイルを，35 ファイルと 5 ファイルに分割した。

3.3 単独アクセント型を与えない学習

形態素単独発声アクセント型が与えられていない状況（[3] と同様の条件）を想定し，観測素性として，前後 2 形態素を含めた 5 形態素について，[基本形/基本形読み/書字形/品詞/活用型]（以下，“基本形等” と記す。），[品詞]，[品詞 (大分類のみ)]，[活用型 (大分類のみ)]，[活用形 (大分類のみ)]，[モーラ数] のそれぞれと当該形態素の文中アクセント型との関係を与えた。また，遷移素性として，連接する形態素の文中アクセント型同士の関係を与えた。CRF++で学習・推定した結果が表 3 上段である。82.1% のアクセント句について正しい推定が得られている。{ 名詞, 動詞, 形容詞, 形状詞 }+{ 助詞, 助動詞 } の 2 語で構成された単純なアクセント句を対象とした場合は正解率が高く（85.9%），逆に名詞の連続（複合名詞）を含む句では率が低くなる（77.0%）。なお，複数の核を持つアクセント句については，最初の核（主アクセント）が一致していれば正解と見なした。全ての核が一致した場合のみを正解とすると，最大で 3%程度正解率が低下する。

3.4 単独アクセント型を与える学習

前節での素性に加え，観測素性に [単独発声アクセント型] を加えて同様の学習を行なった。推定結果は，表 3 の 2 段目に示している。単独型未使用時と使用時では多くの違いが見られ，特に単純な句においては 85.9% から 94.3% になるなど顕著な違いが見られる。しかし，複合名詞を含む句では正解率に向上が見られない。単純な句では単独型がそのまま文中型となることが多いのに対し，複合名詞においては型変形が頻出することによると考えている。

3.5 簡易変化ラベルの学習

以上の型推定は，文中アクセント型を直接学習・推定の対象としていたため，単独発声型と文中型がいずれも “1” である場合と “2” である場合は別々の事象として扱われる。そこで，単独型

表 3: CRF による文中アクセント型推定の結果

	すべての句		単純な句		複合名詞を含む句	
直接学習（単独型なし）	2833 /3533	82.1%	703 /822	85.9%	530 /688	77.0%
直接学習（単独型あり）	3081 /3533	87.2%	775 /822	94.3%	523 /688	76.0%
簡易変化学習	3137 /3533	88.8%	791 /822	96.2%	553 /688	80.4%
簡易変化学習（組合せ素性使用）	3214 /3533	91.0%	790 /822	96.1%	578 /688	84.0%
規則適合変化学習（組合せ素性使用）	3238 /3533	91.7%	792 /822	96.4%	589 /688	85.6%

表 4: CRF による文中アクセント型推定の結果（許容度 3 以上を正答とする場合）

	アクセント句単位					
	すべての句		単純な句		複合名詞を含む句	
規則適合変化学習（組合せ素性使用）	3307 /3533	93.6%	808 /822	98.3%	605 /688	87.9%

から文中型への「型変化」の様子を学習・推定の対象とすることで，類似する現象を共通のものと捉えられるようにした。具体的には次のようなラベル（“簡易変化ラベル” とする。）を使用して学習・推定を行なった。

単独型が有核の場合，文中型が有核であれば，単独型からの変化量を表すラベル（“[0]”; “[+1]”, “[+2]”, ...; “[−1]”, “[−2]”, ...）を学習対象とし，文中アクセント型が無核であれば，無核を示すラベル（“non”）を学習対象とした。一方，単独型が無核あるいは，値を持たないものに対しては，文中型を直接の学習対象とした。推定結果より，機械的に文中型に相当する数値に復元する。表 3 の 3 段目に結果を示す。

単独型から文中型への型変化を推定の対象とした場合では，文中型を直接推定する場合と比較し，全体的に正解率の向上が見られる。相対的な変化を学習の対象とすることで，効率の良い学習ができたものと言える。

3.6 隣接形態素組み合わせ素性を用いた学習・推定

ここまでに用いた観測素性は，当該形態素または周辺形態素についての品詞等のそれぞれと出力ラベルの関係のみであった。結果として，[1] の複合単語アクセント結合規則のように，名詞が連続した場合にのみ起こる現象，例えば，当該形態素が名詞で 1 つ後の形態素も名詞である場合に文中アクセント型が無核になることを学習する場合，2 つの観測素性

- 当該形態素の品詞が名詞である場合には無核になる
- 1 つ後の形態素の品詞が名詞である場合には無核になる

についての重要度 θ_f を上げる処理がされる。しかし，これは名詞の連続以外に対しても影響を与えてしまい，不都合である。そこで，特定の組み合わせに特化した学習のために，これまでの観測素性に加え，[当該形態素の品詞/**前**の形態素の品詞], [当該形態素の品詞/**後**の形態素の品詞], [当該形態素の品詞, **前**の形態素の基本形等], [当該形態素の品詞, **後**の形態素の基本形等], [当該形態素の基本形等, **前**の形態素の品詞], [当該形態素の基本形等, **後**の形態素の品詞] のそれぞれと当該形態素の文中アクセント型との関係を観測素性として使用し，相対変化ラベルの学習・推定をした。

品詞と品詞の組み合わせ以外にも，当該形態素の品詞と前後形態素の基本形等，当該形態素の基本形等と前後形態素の品詞について組み合わせることで，例外的なアクセントについても学習がされやすいように配慮した。その結果が表 3 の 4 段目である。名詞連続を含むアクセント句について特に顕著な改善が見られ，組み合わせの学習が効果を発揮していると言える。

3.7 アクセント結合規則に適合させた学習・推定

以上のように，アクセントの変化を学習の対象とするなど，[1] の規則で可能な処理と類似した結果が得られる学習を行なうことで，高い正答率が得られた。よって，規則で表現されている処理をさらに忠実に取り入れた形で学習することにより，規則的現象をより的確に捉えることができるものと考えられる。そこで，[1] の規則を基に検討し，以下のような相対変化ラベルを導入し，これを学習・推定の対象とすることにした。

- 単独発声アクセント型が有核のもの（“−1” 以外の値を持つもの）に対しては，1) 文中アクセント型が無核（“−1”）の場合，その旨を示すラベル（“non”），2) 文中アクセント型が単独発声型と同じである場合，その旨を示すラベル（“same”），3) 文中アクセント型がモーラ数と同じである（末尾に核がある）場合，その旨を示すラベル（“morae”），4) 文中アクセント型が単独発声型より 1 小さい場合，その旨を示すラベル（“same−1”），5) 文中アクセント型が 1 型の場合，その旨を示すラベル（“one”），6) 文中アクセント型がモーラ数より 1 小さい（末尾の 1 つ前に核がある）場合，その旨を示すラベル（“morae−1”），7) その他の場合は，単独発声型からの変化量を表すラベル（“[0]”; “[+1]”, “[+2]”, . . .; “[−1]”, “[−2]”, . . .）
- 単独発声アクセント型が無核のもの（“−1”）に対しては，1) 文中アクセント型が単独型と同様に無核（“−1”）の場合，その旨を示すラベル（“samenon”），2) 文中アクセント型がモーラ数と同じである（末尾に核がある）場合，その旨を示すラベル（“morae”），3) 文中アクセント型が 1 型の場合，その旨を示すラベル（“one”），4) 文中アクセント型がモーラ数より 1 小さい（末尾の 1 つ前に核がある）場合，その旨を示すラベル（“morae−1”），5) その他の場合は，文中アクセント型（“1”, “2”, . . .; “−1”）
- 単独発声アクセント型を持たないものに対しては，文中アクセント型（“1”, “2”, . . .; “−1”）

これらは，アクセント変化の結果として発生しやすい現象を示すラベルから順に記しているため，複数に該当しうるアクセント句については，最も上にあるものを相対変化ラベルとして使用する。

また，学習・推定のデータに，1) 単独発声アクセント型が無核である場合，その旨を示すラベル（“non”），2) 単独発声アクセント型がモーラ数と同じである（末尾に核がある）場合，その旨を示すラベル（“morae”），3) 単独発声モーラ数より 1 小さい（末尾の 1 つ前に核がある）場合，基本形読みの末尾 2 モーラ（“オイ” など），4) 単独発声型が上記に当てはまらない場合，その旨を示すラベル（“else”）とするラベル（“単独型種類ラベル” とする）を用意し，単独発声アクセント型の種類による結果の分岐を行ないやすくすることを目指した。

また，学習・推定の際に使用する観測素性に，当該形態素についての [単独型種類ラベル], [出現形読みの 先頭のモーラ], [出現形読みの 第 2 モーラ], [出現形読みの 単独発声核位置の前のモーラ], [出現形読みの 単独発声核位置のモーラ], [出現形読みの 単独発声核位置の後のモーラ], [出現形読みの 末尾の前のモーラ], [出現形読みの 末尾のモーラ] のそれぞれと当該形態素の文中アクセ

図 3: 許容度判定プログラム

ント型との関係を加えた。出現形の読みの各モーラを用いたのは，[1] の規則における音節内移動則（アクセント核が音節の先頭のモーラに移動する現象）などの影響を学習させるためである。

上記のように，出力ラベルや観測素性に工夫をして学習・推定をした結果が，表 3 の 5 段目である。すべての場合に渡って，4 段目のものからの向上が見られる。

3.8 不正解に分類された形態素の許容度に関する調査

統計的なアクセント学習に各種の工夫を重ねることにより，アクセント推定は，91.7% のアクセント句に対して正しい結果が得られており（主アクセント核のみについて考える場合），単純なアクセント句に限れば正答率は 96.4% に及ぶ。しかし，このことは，単純なアクセント句であっても 3.6% の句ではアクセント推定結果と正解ラベルが一致しないことを意味する。それらの不一致が本質的な誤りであるか，アクセントの揺れの範囲内として許容できるものなのかによって，結果の意味するところは異なる。そこで，アクセント推定結果と正解ラベルが一致しなかったアクセント句について，アクセントデータベースを作成する際にラベリングを行なったラベラに，許容度を判定させた。

許容度の判定には図 3 のプログラムを使用し，3.7 節の実験において正答とならなかった（主アクセント核のみについて考える場合にだけ正答となったものを含む）アクセント句 374 句に対して，1（とても不自然）から 4（とても自然）の 4 段階で許容度を回答させた。この結果，3 以上の許容度を得られたものは正答と見做すこととすると，表 3 の下段のようになる。単純な結合では，98.3% ものアクセント句について正しい結果または許容できる結果が得られていることとなった。音声として出力する場合にはさらに許容できる幅が広がると考えられるため，完全に近い精度であると言うことができるだろう。ただし，許容できないと判定されたアクセント句がわずかながら存在したのは事実であり，何らかの方法で改善することが望まれる。

3.9 考察

統計的な学習にアクセント結合規則から得られる知識を導入し，正答率を高めることに成功した。

隣接する形態素の品詞を組にして観測素性に加えたことによる改善からは，[1] のアクセント結合規則で結合する品詞によって規則が使い分けられているように，品詞がどのように連続するかが文中でのアクセントを考える上で重要であることが分かる。その影響は、当該形態素や前後の形態素の品詞を個別に考えるのでは不十分となるほどに大きいものである。

アクセント結合規則に細かく適合させた学習において正答率の向上が見られた。これは規則が完全であるとは言えないものの，アクセント結合処理の重要な部分を押さえていることが分かる。今回の実験では，アクセント結合規則への細かな適合をする前とした後のみの比較をしたため，どの部分が特に大きな影響を与えたのかについては不明であり，この点は今後の課題といえる。

また，各種手法による改善をしても，名詞の連続を含むアクセント句の約 12% では，推定結果に許容できない誤りがあり，改善が必要である。複合名詞では，文脈や係り受け関係によって結合の仕方に違いが出ることがあり，揺れも多い。例えば,「政権交代」が「セーケンコータイ」「セーケンコータイ」のいずれにもなり得るが,「配達証明」（ハイタツショーメー）,「合併失敗」（ガッペーシッパイ）のように一方しか許容されないものも多い。ここに掲げた複合名詞はいずれも単独発声では無核となる名詞で構成されているにもかかわらず結果が異なるため，これらを的確に判別するのは難しい。また,「ソ連邦解体後」「日米交換船」のようなアクセント句では係り受けの影響を受ける可能性が高い。このような問題は規則に基づくアクセント結合処理でも解決できていない問題であり，新たな手法が必要になると考えられる。

4 まとめ

コーパスベース及び規則ベースのハイブリッド型アクセント結合処理モジュールの構築を念頭に置き，高品質なアクセントラベリングが施されたコーパスを構築すると共に，CRF を用いた統計的アクセント結合処理を実装した。従来の規則から得られる知見をコーパスベースの統計的手法に組み入れることで，従来手法や単純な統計的処理に比して精度に大きく改善が見られた。複合名詞の問題など未解決の問題があるものの，高い推定精度を得ることができた。

参考文献

[1] 匂坂芳典, 佐藤大和: “日本語単語連鎖のアクセント規則”, 電子通信学会論文誌, vol.J66-D, no.7, pp.847–856, 1983.

[2] N. Minematsu, R. Kita, and K. Hirose, “Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion,” Trans. IEICE, vol.E86-D, no.3, pp.550–557, 2003.

[3] 長野徹, 森信介, 西村雅史: “N-gram モデルを用いた音声合成のための読みおよびアクセントの同時推定”, 情報処理学会論文誌, vol.47, no.6, pp.1793–1801, 2006.

[4] 「日本語話し言葉コーパスの構築法」, 独立行政法人国立国語研究所, 2006.

[5] J. Lafferty, A. McCallum, F. Pereira: “Conditional Random Fields: Probablistic Models for Segmenting and Labeling Sequence Data”, Proceedings of the 18th International Conference on Machine Learning, pp.282-289, 2001.

[6] 工藤 拓: “CRF++: Yet Another CRF toolkit” http://chasen.org/˜taku/software/CRF++/

因子分析を用いた程度副詞と述語等の共起関係の研究試論

服部匡　（分担者：同志社女子大学学芸学部）

A Factor Analysis Approach to Cooccurrence Restrictions Between Degree Adverbs and Predicates

Tadasu Hattori(Doshisha Women's College of Liberal Arts)

１　目的

電子化データを用いた発見的研究の新たな可能性を探るため、程度副詞と述語の間の共起関係を、新聞記事のデータに因子分析の手法を適用することによって分析し、それに基づいてそれぞれの程度副詞の特徴づけを行うことを試みる。

コーパスを用いた共起特性の研究では表現の出現頻度が問題にされることが多いが、今回は頻度は問題にせず表現が 1 度でも出現するかしないかに注目する。一般に程度副詞とされる語類の中から、予備的な分析を経て次の 13 語を対象として選定した。

きわめて、とても、大変、非常に、かなり、ずいぶん、相当、少し、多少、だいぶ、やや、結構、なかなか

２　データについて

使用したデータは、毎日新聞の 1993 年から 2005 年までの 13 年分、およそ 16 億字(見出しと本文)からなる。[1] 以下の手順によって程度副詞から文末までの部分を切り出した。

手順１：　「**程度副詞**−−−。」という文字列をすべて抜き出す。ただし、−−は 10 字以内の文字列であり、この部分は以下で言う「**述語等**」に当たるべきものである。

ここで、述語等については、多義語の区別や異表記のまとめは行っていない(「いる(要る/居る」のようなものは区別する)。程度副詞については異表記・異形をまとめている。

手順２：　「程度副詞と、意味的にそれがかかわるもの」という関係を含んでいない文字列を排除する。

「かかわる」範囲がどこまでかについては判定が困難な例もあり、恣意的になることを避けるため分析せず必ず文末までの文字列全体を取ることにした。

手順３：10 回以上出現（表で横の合計が 10 以上）の述語等のみを残す。この結果表１のような共起度数表を得る（下略)。

---

1) 新聞記事のデータは著作権者である毎日新聞社の許諾を得て研究用に使用している。

表１：共起度数

	きわめて	非常に	大変	とても	相当	ずいぶん	かなり	大分	けっこう	なかなか	多少	やや	少し	合計
、違う	0	0	0	0	0	0	0	11	0	0	0	0	0	11
あいまい	4	5	0	0	0	0	3	0	0	0	0	0	0	13
あいまいだ	22	5	0	1	0	0	1	0	1	0	0	0	0	31
あった	0	2	0	0	13	13	60	2	12	0	1	0	2	105
あったという	0	0	0	0	2	0	8	0	0	0	0	0	0	10
ありがたい	0	13	30	17	0	0	0	0	0	0	0	0	0	60
ありました	0	2	0	0	0	2	11	0	9	0	2	0	2	28
あります	0	1	2	1	2	2	24	0	21	0	0	0	4	60
ある	0	8	0	3	29	13	134	0	83	0	5	1	8	284
あるという	0	0	0	0	2	0	12	0	6	0	0	0	1	21
あると思う	0	1	0	0	4	0	4	0	1	0	0	0	0	10

手順４：共起度数表で、１以上の値を１に変換する。その結果、次のような行列を得る（下略）。

表２：共起の有無

	きわめて	非常に	大変	とても	相当	ずいぶん	かなり	大分	けっこう	なかなか	多少	やや	少し
、違う	0	0	0	0	0	0	0	1	0	0	0	0	0
あいまい	1	1	0	0	0	0	1	0	0	0	0	0	0
あいまいだ	1	1	0	1	0	0	1	0	1	0	0	0	0
あった	0	1	0	0	1	1	1	1	1	0	1	0	1
あったという	0	0	0	0	1	0	1	0	0	0	0	0	0
ありがたい	0	1	1	1	0	0	0	0	0	0	0	0	0
ありました	0	1	0	0	0	1	1	0	1	0	1	0	1
あります	0	1	1	1	1	1	1	0	1	0	0	0	1
ある	0	1	0	1	1	1	1	0	1	0	1	1	1
あるという	0	0	0	0	1	0	1	0	1	0	0	0	1
あると思う	0	1	0	0	1	0	1	0	1	0	0	0	0

つまり「共起例があるかないか」の二値データになる。共起の量的側面（頻度）を捨象するのは乱暴なようであるが、用例というのは「あるかないか」が最も重要と思われる。

表 2 において、述語等は 555 個、副詞は　15 個あり、掛けると 8325 の欄が存在する。そのうち１（共起例あり）の欄は 2526 個，0（共起例なし）の欄は 5799 個である。 なお述語等のうち丁寧体のものは 65 である。

３　因子分析

３．１　分析の概要

Gorsuch(1983)によれば、因子分析の通常の目的は、「概念化の補助手段として、諸変数間の相互関係を簡潔ながらも正確な仕方で要約すること」である。

それぞれの程度副詞（があるものと共起するかしないか）を変数と見、述語等をケースと見て、変数の背後にある共通因子を抽出する。主因子法・プロマックス回転（斜交回転）によって４因子を抽出した(初期の固有値はそれぞれ 3.07, 1.87, 1.34, 1.19 である)。

「ずいぶん・大分・少し・かなり・相当・多少・やや」との共起は第1因子に規定されるところが大きいことになる。一方、「非常に・とても・大変」の3語との共起は第2因子に強く規定されている。また、「結構・なかなか」との共起は第3因子に規定され、「きわめて・非常に」との共起は第4因子に規定される。それぞれのグループの副詞には意味的に共通点があるのは明らかだが、その解釈は、因子得点を参照して検討することとする。

表３：因子相関行列

	1	2	3	4
1	1.00	**-0.41**	0.22	0.16
2	**-0.41**	1.00	0.12	-0.06
3	0.22	0.12	1.00	**0.32**
4	0.16	-0.06	**0.32**	1.00

表４　：因子パターン行列

	因子			
	1	2	3	4
ずいぶん	**0.757**	0.120	-0.139	-0.238
大分	**0.620**	-0.005	-0.117	-0.196
少し	**0.528**	-0.004	-0.013	-0.053
かなり	**0.497**	-0.187	0.229	0.270
相当	**0.487**	0.045	0.252	0.031
多少	**0.441**	0.035	-0.077	-0.140
やや	**0.355**	-0.169	-0.078	0.210
非常に	0.138	**0.723**	-0.064	***0.391***
大変	-0.037	**0.612**	-0.030	0.016
とても	-0.022	**0.611**	0.236	-0.121
けっこう	0.011	0.030	**0.687**	-0.251
なかなか	-0.162	0.077	**0.561**	0.039
きわめて	-0.230	0.075	-0.142	**0.647**

３．２　因子の解釈

それぞれの因子に対する各ケース（述語等）の因子得点を表にして示し、因子の解釈を行っていく。因子得点とは、抽出された各因子の持つ傾向を、各々のケースがどの程度強く有しているかを示すものである。

まず、第１因子の因子得点の高い述語等をそれぞれ 30 位まで示すと表 5 のようになる。意味的に、基準との差異、基準からの変化、物事の存在を表すものが多い。

一方、低い因子得点を示す述語等は(表は略すが)、全体に、差異や存在としては解釈しにくい。例えば「困難(だ)」「便利(だ)」「元気(だ)」といった表現固有の意味の中に差異や存在という要素が含まれているとは言いにくい。

表５：第１因子の因子得点（高得点順）

	ずいぶん	大分	少し	かなり	相当	多少	やや	非常に	大変	とても	けっこう	なかなか	きわめて	因子得点
変わる	1	1	1	1	1	1	1	1	0	0	0	0	0	2.71
違う	1	1	1	1	1	1	1	1	1	0	1	0	0	2.69
減った	1	1	1	1	1	1	1	0	0	0	0	0	0	2.63
異なっている	1	1	1	1	1	1	1	0	0	0	0	0	0	2.63
異なる	1	1	1	1	1	1	1	0	1	0	0	0	0	2.55
あった	1	1	1	1	1	1	0	1	0	0	1	0	0	2.50
増えている	1	1	0	1	1	1	1	1	0	0	0	0	0	2.40
変わった	1	1	1	1	1	1	0	0	0	0	0	0	0	2.35
違ってくる	1	1	1	1	1	1	0	0	0	0	0	0	0	2.35
変わってきた	1	1	1	1	1	0	1	1	0	0	0	0	0	2.25
違っていた	1	1	1	1	0	1	1	0	0	0	0	0	0	2.22
違った	1	1	1	1	1	0	1	0	0	0	0	0	0	2.17
増えた	1	1	1	1	1	0	1	1	0	1	0	0	0	2.16
ある	1	0	1	1	1	1	1	1	0	1	1	0	0	2.12
狭い	1	1	1	1	1	0	1	1	1	0	0	0	1	2.08
違いがある	1	1	1	1	0	1	0	0	0	0	1	0	0	2.00
楽になった	1	1	1	1	1	0	1	0	1	1	0	0	0	2.00
安い	1	1	1	1	1	0	1	1	1	1	0	0	1	1.99
増えてきた	1	1	1	1	1	0	0	1	0	0	0	0	0	1.97
長い	1	0	1	1	1	1	1	1	1	1	1	0	1	1.95
違うようだ	1	1	1	1	0	1	0	0	0	0	0	0	0	1.94
違います	1	1	1	1	0	1	0	1	0	1	0	0	0	1.94
進んでいる	1	1	0	1	1	0	1	1	0	1	1	0	0	1.91
回復した	1	1	0	1	0	1	1	0	0	0	0	0	0	1.91
悩んだ	1	1	1	1	1	0	0	1	0	1	0	0	0	1.88
進んだ	1	1	0	1	1	0	1	0	0	0	0	0	0	1.85
差がある	1	1	0	1	1	0	1	0	0	0	0	0	0	1.85
残っている	1	0	1	1	1	1	0	0	0	0	1	0	0	1.84
改善された	1	1	1	1	0	0	1	1	0	0	0	0	0	1.84

次に、第２因子の因子得点の高い述語等を 30 位まで示すと表６のようになる。人の感情や評価に関わるものが多い。対照的に、因子得点の低いもの(表は略す)は、「低くなった」「減った」「異

なっている」など、それ自体の意味の中には感情や評価の要素を含まない、いわば、客観的な事柄を意味するものが多い。

表６：第２因子の因子得点（高得点順）

	ずい										けっ	なかな	きわめ	
	ぶん	大分	少し	かなり	相当	多少	やや	**非常に**	**大変**	**とても**	こう	か	て	**因子得点**
参考になった	1	0	0	0	0	0	0	**1**	**1**	**1**	0	1	0	**1.27**
楽しかった	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
面白かった	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
楽しみです	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
おいしい	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
難しいのです	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
難しいものです	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
難しい問題です	0	0	0	0	0	0	0	**1**	**1**	**1**	1	1	0	**1.26**
良かった	1	0	0	0	0	0	0	**1**	**1**	**1**	0	1	1	**1.25**
難しかった	0	0	1	0	0	0	0	**1**	**1**	**1**	1	1	1	**1.18**
厳しかった	0	0	1	0	0	0	0	**1**	**1**	**1**	1	1	1	**1.18**
素晴らしい	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	0	**1.18**
魅力的だ	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	0	**1.18**
興味深かった	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	0	**1.18**
興味深い	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	1	**1.16**
便利	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	1	**1.16**
難しいですね	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	1	**1.16**
元気	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	1	**1.16**
魅力的	0	0	0	0	0	0	0	**1**	**1**	**1**	0	1	1	**1.16**
勉強になった	1	0	0	0	0	0	0	**1**	**1**	**1**	0	0	0	**1.15**
勉強になりました	1	0	0	0	0	0	0	**1**	**1**	**1**	0	0	0	**1.15**
参考になりました	1	0	0	0	0	0	0	**1**	**1**	**1**	0	0	0	**1.15**
うれしかった	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	0	**1.14**
危険です	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	0	**1.14**
喜んでいる	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	0	**1.14**
喜ばれた	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	0	**1.14**
重要である	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	1	**1.13**
満足している	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	1	**1.13**
多いのです	0	0	0	0	0	0	0	**1**	**1**	**1**	1	0	1	**1.13**

次に、第3因子に対する因子得点の高い述語等をそれぞれ30位まで示すと表7のようになる。概ね、対象や事態が容易でない感じ、一筋縄で行かぬ感じ、軽視できない感じなど、いわば「手ごたえ」の大きさを表すものであり、得点の低いもの(表は略)はそれに遠いもののように思われる。

表７：第３因子の因子得点（高得点順）

	ずいぶん	大分	少し	かなり	相当	多少	やや	非常に	大変	とても	けっこう	なかなか	きわめて	因子得点
深刻だ	0	0	1	1	1	0	0	1	1	1	1	1	1	2.29
強かった	0	0	1	1	1	0	0	1	1	1	1	1	1	2.29
難しい	0	0	1	1	1	0	1	1	1	1	1	1	1	2.29
重い	0	0	1	1	1	0	1	1	1	1	1	1	1	2.29
つらい	0	0	1	1	1	0	1	1	1	1	1	1	0	2.29
怖い	0	0	1	1	1	0	1	1	1	1	1	1	0	2.29
大変です	0	0	0	1	1	0	0	1	0	1	1	1	0	2.28
迫力がある	0	0	0	1	1	0	0	1	0	1	1	1	0	2.28
面白い	0	0	0	1	1	0	0	1	1	1	1	1	1	2.27
忙しい	0	0	0	1	1	0	0	1	1	1	1	1	0	2.27
勇気がいる	0	0	1	1	1	0	0	0	1	1	1	1	0	2.22
高い	0	1	1	1	1	0	1	1	1	1	1	1	1	2.12
いい	0	1	0	1	1	0	0	1	1	1	1	1	1	2.10
少ない	0	1	0	1	1	0	1	1	1	1	1	1	1	2.10
深い	1	0	0	1	1	0	1	1	0	1	1	1	1	2.03
厳しい	1	0	0	1	1	0	1	1	1	1	1	1	1	2.03
うるさい	0	0	1	1	1	0	0	1	0	0	1	1	0	2.02
厳しいものがある	0	0	0	1	1	0	0	1	1	0	1	1	1	1.99
むずかしい	0	0	1	1	0	0	1	1	1	1	1	1	1	1.90
うまい	0	0	1	1	0	0	0	1	1	1	1	1	0	1.90
説得力がある	0	0	0	1	0	0	0	1	0	1	1	1	1	1.89
大変だ	0	0	0	1	0	0	0	1	0	1	1	1	0	1.88
役に立つ	0	0	0	1	0	0	0	1	1	1	1	1	1	1.88
広い	0	0	0	1	0	0	1	1	1	1	1	1	1	1.88
おもしろい	0	0	0	1	0	0	0	1	1	1	1	1	0	1.88
しんどい	0	0	0	1	0	0	0	0	0	1	1	1	0	1.81
大変	0	0	1	0	1	0	0	1	0	1	1	1	0	1.68

しっかりしている	0	0	0	0	1	0	0	1	0	1	**1**	**1**	0	**1.66**
きつい	0	0	1	1	0	0	0	1	1	0	**1**	**1**	0	**1.61**

最後に、第 4 因子に対する因子得点の低い述語等を 15 位まで示すと表 8 のようになる。口頭語的文脈で生じ易く評価的にプラスのものが多い。因子得点の下位 50 形式中で丁寧体のものは 16 形式あるのに対して、上位 50 形式中で丁寧体のものは 1 形式しかない。因子得点の高いもの(表は略)は上記とは逆の傾向が見られる。

表８： 第４因子の因子得点（低得点順）

	ずいぶん	大分	少し	かなり	相当	多少	やや	*非常に*	大変	とても	けっこう	なかなか	**きわめて**	**因子得点**
楽になりました	1	1	1	0	0	0	0	*0*	0	0	0	0	**0**	**-1.79**
お世話になった	1	0	0	0	0	0	0	*0*	1	0	0	0	**0**	**-1.57**
迷惑をかけた	1	0	0	0	0	0	0	*0*	1	0	0	0	**0**	**-1.57**
おいしかった	0	0	0	0	0	0	0	*0*	1	1	1	0	**0**	**-1.53**
気に入っている	0	0	0	0	0	0	0	*0*	1	1	1	0	**0**	**-1.53**
好きだ	0	0	0	0	0	0	0	*0*	1	1	1	0	**0**	**-1.53**
好きです	0	0	0	0	0	0	0	*0*	0	1	1	0	**0**	**-1.47**
楽しいです	0	0	0	0	0	0	0	*0*	0	1	1	0	**0**	**-1.47**
好き	0	0	0	0	0	0	0	*0*	0	1	1	0	**0**	**-1.47**
、違う	0	1	0	0	0	0	0	*0*	0	0	0	0	**0**	**-1.37**
お世話になりました	0	0	0	0	0	0	0	*0*	1	1	0	0	**0**	**-1.36**
うれしく思う	0	0	0	0	0	0	0	*0*	1	1	0	0	**0**	**-1.36**
驚いています	0	0	0	0	0	0	0	*0*	1	1	0	0	**0**	**-1.36**
感激しました	0	0	0	0	0	0	0	*0*	1	1	0	0	**0**	**-1.36**

４　内省等に基づく分析との関連

当分析の結果を渡辺(1987)の、内省に基づいた分類と比較する。

表９： 渡辺実氏の分類

		比較構文	計量構文	判断構造	評価	表現性	語例
発見系	**とても類**	×	○	発見	±	驚嘆	とても・はなはだ・すこぶる・大変・極めて・非常に・**ずいぶん**
発見系	**結構類**	×	○	望外発見	＋	脱懸念	けっこう・なかなか・わりに・ばかに・やけに
比較系	**多少類**		○	潜在比較	－	反期待	多少・すこし・ちょっと・やや・いささか・かなり
比較系	**多少類**	○		潜在比較	±	反期待	同上

この分類は当分析の結果とよく一致していることが分かる。つまり、「とても類」は第２因子

に、「結構類」は第3因子に、「多少類」は第1因子に対応する。唯一の重要な例外は「ずいぶん」であるがこの語については他にも渡辺氏の分類と一致しない事実がある(服部(1996))。

なお、渡辺氏は別の観点から程度副詞（を含む様々な表現）を下記のように「わがこと系」と「ひとごと系」に分類することも提案されている（1991)。

ひとごと系　―　ずいぶん、ずっと、よほど、いっそう、はるかに、いたって、大分、相当、かなり、なかなか、わりに、けっこう、ごく

わがこと系　―　大変、非常に、極めて、はなはだ、とても、ばかに、やけに、ちょっと、少し、いささか

こちらは、当分析の結果には直接の対応を見出しにくいようである。

５　まとめ

述語等との共起の有無のデータを基に、程度副詞の共起傾向に関わる因子を抽出し、因子得点と述語の意味的（文体的）特性を参照しながら意義付けを試みた。

今回の方法は、程度副詞の性質の一側面（文末にある述語等との関係）を問題とするものにすぎない。例えば、ある副詞が主文中によりもむしろ特定の形式の種類の節の中に生じやすいといった大局的な要因に関連した傾向は無視している。

しかし、このような単純な方法で、従来の内省等による分析の一部とかなり一致する結果が得られたことは興味深いことであり、今後このような方法が未知の現象の解明につながる発見手順として利用される可能性を示すものである。

今後は、従来あまり研究されていない語類に同様の方法を適用することにより、共起傾向に関する事実を発見することを試みたい。仮にそれが原理的には内省等による方法で到達しうる結果であったとしても、現実の研究者の有する時間は有限である以上、こうした方法を発見手順として用いることは無意義とは言えないと思われる。

謝辞

因子分析の適用に関して、柳井晴夫・聖路加看護大学大学院教授よりご助言を頂いた。お礼申し上げる。

参考文献

Gorsuch,Richard L.(1983) *Factor Analysis,* second edition, Lawrence Erlbaum Associates.

服部匡(1996)　「程度副詞と比較基準　―「多少」、「少し」を中心に―」　『同志社女子大学学術研究年報』47:Ⅳ, pp269-284

服部匡(2006)　「多変量解析の手法を用いた、程度副詞－述語の共起関係の分析」　「日本語コーパス」日本語学班研究会発表資料

渡辺実（1987)　「程度副詞の体系」　『上智大学国文学科紀要』　4, pp1-16

渡辺実（1991)　「「わがこと・ひとごと」の観点と文法論」　『国語学』　165, pp1-14

日本語教育における語彙シラバスの作成について

山内博之（実践女子大学文学部）

An Attempt to Construct a Lexical Syllabus for Japanese Language Education

Hiroyuki Yamauchi (Faculty of Literature, Jissen Women's University)

１．研究の目標

日本語教育のための語彙シラバスを作成する。

２．当面の研究課題

実質語の難易度を設定する。

３．難易度を設定する方法

paradigmatic に対立する語を集めて、その中で相対的に難易度を決定する。

「診療所」と「激怒する」の難易度の違い　　⇒　わかりにくい･･･
「病院」と「診療所」の難易度の違い　　⇒　わかる！
「怒(おこ)る」と「激怒する」の難易度の違い　⇒　わかる！

『分類語彙表』の「1.2650-21」
医院　診療所　クリニック　病院　避病院　救急病院　サナトリウム　療養所　産院　養老院　老人ホーム　ホスピス　養護施設
↓
↓（難易度の設定）
↓
難易度１：病院　老人ホーム
難易度２：医院　診療所　救急病院
難易度３：産院　ホスピス　養護施設　クリニック
難易度４：療養所　養老院　サナトリウム
難易度５：避病院

『分類語彙表』の「2.3012-04」
怒(おこ)る　怒(いか)る　頭に来る　青筋を立てる　怒り狂う　腹立つ　腹が立つ　小腹が立つ　腹を立てる　むかっ腹を立てる　立腹する　激怒する　憤怒する　激高する
↓
↓（難易度の設定）
↓

難易度１：怒(おこ)る　頭に来る　腹が立つ
難易度２：怒(いか)る　立腹する　腹を立てる
難易度３：激怒する　怒り狂う　腹立つ　むかっ腹を立てる
難易度４：青筋を立てる　小腹が立つ　憤怒する　激高する

仮説：paradigmatic に対立する実質語の場合、出現頻度と難易度が概ね比例する。

４．言語能力とは：コントロールできる言語の範囲の広がり

(1) paradigmatic な広がり：難易度１→ → → 難易度５

私はビールを飲みます。	私はビールを飲みます。	おいしいビール
↓日本酒	↓いただく	↓うまい
↓冷酒	↓がぶ飲みする	↓ほろ苦い
↓どぶろく	↓痛飲する	↓まろやかな
↓ひれ酒	↓口にする	↓口当たりのいい
↓般若湯	↓すする	↓のど越しのいい

(2) syntagmatic な広がり：単語→単文→複文→段落→複段落

私、飲む、ビール、おいしい。
私はおいしいビールを飲みました。
私はおいしいビールを飲むと、すぐに酔っ払ってしまいます。
私はおいしいビールを飲むと、すぐに酔っ払ってしまいます。でも、おいしいビールって、体の健康にも心の健康にもいいと思うんです。だから、→→→

５．語彙シラバスの作成

(1) paradigmatic な広がりを支える語彙のグループ（文レベルの教育に寄与する）

《ビール》グループ
1.4350　飲料・たばこ
06　酒（さけ）　ささ（酒）　酒（しゅ）　般若湯　美禄　百薬の長　御酒（ごしゅ）　洋酒　中国酒　地酒　清酒　生酒　原酒　吟醸酒　新酒　古酒　うま酒　美酒　銘酒　銘醸　薬酒
07　燗酒　燗冷まし　熱燗　ぬる燗　冷や酒　冷酒
08　食前酒　アペリティフ　寝酒　ナイトキャップ　朝酒　迎え酒　やけ酒　罰杯　振る舞い酒　升酒　コップ酒
09　合成酒　混成酒　醸造酒　蒸留酒　みりん　酒塩　焼酎　リカー　ホワイトリカー　泡盛　濁酒　濁り酒　どぶろく　もろみ［～酒］　甘酒　白酒　ひれ酒　卵酒　甘露酒

10　果実酒　梅酒　りんご酒　ぶどう酒　ワイン　赤ワイン　白ワイン　ロゼ　シャンパン　シェリー　ウイスキー　モルト　バーボン　ブランデー　コニャック　ウォッカ　ラム　ラオチュー　麦酒　ビール　生ビール　黒ビール　スタウト　地ビール　缶ビール　発泡酒　リキュール　キュラソー　サワー
11　カクテル　オンザロック　ハイボール　水割り［～酒］　ポンチ　フルーツポンチ
12　屠蘇（とそ）　お屠蘇　神酒（みき）　お神酒　祝い酒

《飲む》グループ
2.3331　食生活
02　召し上がる　聞こし召す　頂く
04　暴飲する　痛飲する　がぶ飲みする　酔いつぶす　満喫する　愛飲する　飲み飽きる
12　飲む　飲ます・飲ませる　すする　飲み干す　満を引く　独酌する　晩酌する　酌み交わす　杯を傾ける　献杯する　返杯する　乾杯する　飲酒する　禁酒する
2.3393　口・鼻・目の動作
02　飲む　飲ます・飲ませる　飲み込む　飲み下す　嚥下（えんか・えんげ）する　飲み干す　飲みかける
03　らっぱ飲みする　がぶ飲みする　あお（呷）る　仰ぐ［毒を～］　やる［一杯～］　引っ掛ける

《酔っ払う》グループ
2.3003　飢渇・酔い・疲労・睡眠など
03　酔う　酔っ払う　微醺を帯びる　酩酊する　酔いしれる　陶酔する　飲みつぶれる　酔いつぶれる　沈酔する　盛りつぶす　乱酔する　泥酔する　悪酔いする　船酔いする　麻酔する
04　覚める［酔いが～］　さえる［頭が～］

《おいしい》グループ
3.5050　味
01　おいしい　うまい　まずい　美味　あごが落ちそう　味のよい　後味がよい／が悪い　のど越しがよい　口当たりのよい　すっきり［～としたワイン］
02　甘い　甘ったるい　甘美
03　酸い　酸っぱい　甘酸っぱい
04　塩辛い　しょっぱい　甘辛い　塩の利いた　辛（から）い　激辛
05　苦い　ほろ苦い
06　渋い　えがらっぽい　えぐい・えごい
07　ひりひり　ぴりっと　ぴりぴり
08　あっさり　さっぱり　淡泊　淡淡　大味　小味　しつこい　くどい　こてこて　こってり　濃厚
09　こくのある　まったりした　芳醇　リッチ　まろやか　マイルド　脂っこい　あくが

強い　スパイシー
10　気の抜けた

(2) syntagmatic な広がりを支える語彙のグループ（談話・段落レベルの教育に寄与する）

【飲酒】という話題を構成する類語グループ群
《ビール》グループ、《飲む》グループ、《酔っ払う》グループ、《おいしい》グループ、《居酒屋》グループ、《ジョッキ》グループ、……

(3) 語彙シラバスの全体像

収録語数　　：約 10000
類語グループの数：《ビール》《飲む》《居酒屋》など、約 500　→　難易度の決定
話題グループの数：【飲酒】【食事】【趣味】など、100～200

６．まとめ

(1) 収録語の決定：『分類語彙表』から、均衡コーパスにおける出現頻度により抽出する。
(2) 類語グループの作成：①文レベルの教育　②語の難易度の決定
(3) 話題グループの作成：談話・段落レベルの教育

７．今後の課題：単位の不一致の解消

◇パイロットコーパスを用いた出現頻度調査
「1.2650-21」
難易度１：病院（252）　老人ホーム
難易度２：医院（5）　診療所　救急病院
難易度３：産院　ホスピス（1）　養護施設　クリニック（7）
難易度４：療養所　養老院　サナトリウム
難易度５：避病院

「2.3012-04」
難易度１：怒(おこ)る（50）　頭に来る　腹が立つ
難易度２：怒(いか)る　立腹する　腹を立てる
難易度３：激怒する　怒り狂う（1）　腹立つ（1）　むかっ腹を立てる
難易度４：青筋を立てる　小腹が立つ　憤怒する　激高する

文献

国立国語研究所(2004)『分類語彙表－増補改訂版』大日本図書.
牧野誠一(監修)(1999)『ACTFL－OPI 試験官養成マニュアル（1999 年改訂版）』アルク

国語教育と語彙指導

鈴木一史（東京大学教育学部附属中等教育学校）

Teaching Japanese Language and Enriching Students' Vocabulary

SUZUKI Kazufumi（Tokyo University Secondary Education School)†

1. はじめに

生徒が学習すべき語彙について、確固たる枠組みがない。学習漢字が一応それに当たるが、語彙の範疇は漢字だけではない。どの程度の語彙を持つことが義務教育終了段階で必要か、先行調査は多いものの、確定されてはいない。このことは、語彙調査を続けている井上[1]氏に、「国語科学習基本語彙及び語彙指導に関する研究発表はほとんど行われておらず、憂慮すべき事態といわざるを得ない。」と言わしめることにもなる。語彙の確定と語彙教育・語彙指導の研究は両輪となっている。

本発表では、語彙指導の現状を見通すことで、大規模コーパスの教育への有用性と課題について述べる。

2. 語彙と他領域との関係

2.1 授業形態と語彙指導

本校（東京大学教育学部附属中等教育学校）は平成 15-16 年度、教科の研究開発として、「国語力向上プログラム」に参画した。テーマとして「伝え合う力の育成」、重点項目としては「正確に読み、正確に書く生徒の育成」を掲げた。その中で行った語彙指導の授業形態について述べる。

平成 15 年度から少人数授業の授業形態をとっている。少人数学習の目的は多様な生徒への対応と個々に応じた授業である。また国語力の向上として、少人数授業を、文法学習・語彙指導と表現指導とに分け、授業に特色を持たせた。その成果を定期テストと全国テストを踏まえ、領域別に分析した。

16 年度 1 年生（59 回生）における前期末試験の得点分布を見てみると、＜表 1 ＞のように標準偏差が文法・語彙指導の分野で狭くなり、表現の分野で広がっている。

＜表 1＞

標準偏差 (100 点満点換算)	59 回生(1年)	58 回生(2年)		57 回生(3年)
一斉授業・読解	12.9	14.9	一斉・読解	14.6
少人数・表現	13.1	16.5	一斉・文法・語彙	17.5
少人数・文法	11.7	12.2	一斉・古典	15.8

（平成 16 年 9 月下旬実施　1・2・3 学年対象）

† suzuki-j@hs.p.u-tokyo.ac.jp

[1] 井上一郎　「語彙力の発達とその育成」明治図書　2001.4

上記データにより、一斉授業よりも少人数授業の文法・語彙指導は生徒の成績の幅が狭まり、少人数の表現指導は逆に幅が広がってしまうことが１・２年において顕著である。つまり、少人数学習において文法学習などの知識重視型内容を持つ学習は、当初の目的である「個々に応じた指導」に適合し、学習効果を挙げていると考えられる。さらに、少人数授業を行わない３年になると、成績に広がりが生じていることが伺える。また、３年生では領域別の授業を行ったにもかかわらず、文法･語彙指導の分野で広がりが大きくなっていることから、語彙・文法という「領域」が生徒の成績の幅を狭くしているのではなく、少人数学習という「形態」が個別的対応により生徒の知識の定着を図り、成績の幅を縮めているのではないか。

次に、成績自体の伸長を全国テストから分析する。国語科では全国テストを年度末と年度当初に行った。領域別に全国比を算出してみると、以下の＜表２＞のような結果が得られた。56回生は少人数授業を経験しておらず、57回生は少人数授業を２年生の時に１年間、58回生は少人数授業２年目、59回生は入学初期のテスト結果である。テストの種類によって若干の結果は異なるものの、全体的に言語事項の部分で全国比が飛びぬけている。

＜表２＞

全国比（全国を 100 とした場合）	CRT(2004,3,4)			NRT(2004,4,12)		
	56 回生	57 回生	58 回生	57 回生	58 回生	59 回生
話すこと･聞くこと	108	108	115	110	114	112
書くこと	120	112	125	126	125	134
読むこと	120	113	113	135	132	141
言語事項	121	121	128	145	171	145

＜表３＞

読書力テスト(2004,3) 全国比	56 回生	57 回生	58 回生
読字力	125	124	157
語彙力	120	127	147
文法力	115	119	132

学年ごとに分析すると、56 回生は言語事項も高いが、全領域にわたって高い数値が得られた。つまり、少人数によって、文法的言語事項だけが増したわけではないと考えられる。しかし、少人数授業を受けた生徒（57 回生）は言語事項が他の領域より突出して高い。これは少人数制の授業によって文法･語彙知識が定着していることの証明であろう。また、語彙・文法の成績に比例して、「読むこと」の成績も若干高くなっている。これは文法･語彙が読みに影響を与えていることを示唆している。しかし、「書くこと」はそれほど高くはない。「正確に読み、正確に書く」という目標の達成はできなかったということだろうか。そこで、＜表１＞の標準偏差を考慮に入れると、「書くこと」つまり表現の領域においては、少人数でも成績に広がりが見られるため、成績の上位者と下位者の差が大きくなっている。従って、「話すこと･聞くこと」という音声言語の領域より成績は高いものの、言語事項のような「定着」には至っていない。

また、異なった観点の成績のデータも以下に示す。上記テストと同時期に行ったものであるが、「読書力」という特化した能力をみるテストであり、観点も上記のテストとは異なっている。しかし、一年間の指導により、全体的な「読書力」が全国平均よりも大きく上回ってきていることが読み取れる。

＜表4＞

		語彙	表現	文法	読解0	CRT国語	CRT社会	CRT数学	CRT理科
語彙	相関係数	1	.120	.194(*)	.286(**)	.129	.192(*)	.078	.213(*)
	有意確率（両側）		.195	.035	.002	.161	.036	.398	.020
表現	相関係数	.120	1	.421(**)	.362(**)	.287(**)	.156	.236(**)	.186(*)
	有意確率（両側）	.195		.000	.000	.002	.091	.010	.043
文法	相関係数	.194(*)	.421(**)	1	.470(**)	.462(**)	.236(**)	.481(**)	.268(**)
	有意確率（両側）	.035	.000		.000	.000	.010	.000	.003
読解0	相関係数	.286(**)	.362(**)	.470(**)	1	.320(**)	.355(**)	.337(**)	.323(**)
	有意確率（両側）	.002	.000	.000		.000	.000	.000	.000
CRT国語	相関係数	.129	.287(**)	.462(**)	.320(**)	1	.430(**)	.504(**)	.450(**)
	有意確率（両側）	.161	.002	.000	.000		.000	.000	.000
CRT社会	相関係数	.192(*)	.156	.236(**)	.355(**)	.430(**)	1	.485(**)	.548(**)
	有意確率（両側）	.036	.091	.010	.000	.000		.000	.000
CRT数学	相関係数	.078	.236(**)	.481(**)	.337(**)	.504(**)	.485(**)	1	.591(**)
	有意確率（両側）	.398	.010	.000	.000	.000	.000		.000
CRT理科	相関係数	.213(*)	.186(*)	.268(**)	.323(**)	.450(**)	.548(**)	.591(**)	1
	有意確率（両側）	.020	.043	.003	.000	.000	.000	.000	

＊　相関係数は 5% 水準で有意（両側）。　＊＊　相関係数は 1% 水準で有意（両側）。

2.2　語彙テストと他領域テストとの相関

中等教育学校 1 年生はどのような学習能力を持っているのか。国語科の学習領域どうしの関係はどのようになっているのか。また、他教科との関連性はあるのか。

1 年の入学当初 6 月に行った外部テスト（CRT）と前期末に行った定期テストとを重ねて検討することで、1 年生の学習傾向と能力を分析する。分析方法は、定期テストについては、「語彙」「文法」「読解」「表現」の各項目で個別に採点し得点を出す。

前期課程 1 年生 120 人に、前期末テストを 10 月上旬に実施した。配点は「語彙」20 点、「表現」30 点、「文法」50 点、「読解」100 点、である。時間は「語彙」「表現」「文法」合わせて 50 分。「読解」50 分。内容は、「語彙」は、共起問題、類義語問題、意味把握問題を出した。各問題の語句は教科書教材からとった。「表現」は書くときに注意すべき事柄、聞くときの聞き方と注意点、創作韻文の特徴である。「文法」は主語述語、修飾語被修飾語など係り受けの問題。「読解」は教科書教材の「はちどりの不思議」の内容把握を中心とした読解問題である。

外部テストである CRT は 4 月下旬に実施。「国語」「数学」「理科」「社会」をそれぞれ 100 点満点 50 分で実施。範囲は小学校での既習範囲である。

すべての項目の相関関係を分析したものが以下の表 4 である。相関係数はピアソンの相関係数に従った。

① 当初、語彙はすべてに高い相関を示すと思われたが、実際には「読解」に弱い相関の有意性が認められるだけであった。
② 「文法」がすべての項目に対して高い相関があり、有意であった。
③ 「読解」もすべての項目に高い相関と有意差を示している。
④ CRT は 4 教科それぞれが高い相関と優位性を示している。
⑤ 特に「国語」は他のすべての項目に高い相関と有意である。

①の「語彙」の相関が低いことについて。「語彙」の問題として、共起問題、類義語問題、意味把握問題を出したが、二つの問題点が考えられる。一つは語彙能力を図る上で、共起問題と類義語問題で適切かどうか。もう一つは、問題量として少なすぎるため、適切な値ではない可能性があること。今後の検討課題である。

問題の語句は教科書教材から選択したために、「読解」との相関は高く出た。やはり文章を読むときには語句の理解が重要であることが示せた。しかも、辞書的な意味ではなく、共起語や類義語を知るという高度な語彙能力も要求される。

②について。文法は品詞よりも文の構成を中心とした問題であった。一文単位での文の理解はすべての文章を読む際の基本的能力と考えられる。

③について。教科書教材の内容理解を中心とした読解問題は、すべての学習に関連している。国語の読解力が高い学習者はすべての学習に対して高い能力を持つといえる。ことばや文字を読むことがすべての学習の基本にあると考えられる。ただ、因果関係は不明であり、今後の課題として共分散構造分析の必要がある。

④について。CRT の 4 教科それぞれが高い相関と優位性を示していることは、まだ学習項目が未分離であることを示しているのではないか。それぞれが密接に絡み合って、学習者の理解を促進していると考えられる。中学で行う総合学習的要素が、学習者の中で働いている。

⑤について。CRT の「国語」は当然といえば当然であるが、校内の定期テストとの相関も高い。これは、校内のテストが外部テストとも対応し、独自のカリキュラムでありつつも、学習指導要領に沿った学習と評価がなされていると考えられる。

3. シソーラスを使った試み

近年語彙指導の重要性が言われる中、均質でない新しい指導方法の登場が見られる。

塚田（筑波大学）[2]は語彙指導と読書指導の関連性の中で、マッピングの有効性を指摘している。文章を読んだ後、ある語彙を中心にしたマッピングを行う。そのことにより、読解が進み、文章の理解が深まる。

マッピングは連想ゲーム的なブレインストーミングの方法である。一つの言葉を中心にして、連想する言葉を次々に連ねていく。マッピングは自己の内面の語彙体系を探る上で、大変有効である。同じ語からはじめても、かなり個人差が生じる。しかし、系統や方向性として、同じカテゴリーに含まれる言葉も存在する。2004 年に卒業研究で筑紫さんはこのマッピングを使ったコミュニケーションを調査した。この資料によると、「道」という言葉

[2] 塚田康彦 「語彙力と読書」東洋館出版社 2001.7.15

から連想するものについて、「通学路」というカテゴリーと「将来の職業（進むべき道)」というカテゴリーが共通項として本校の生徒に見られる。このように自己の語彙体系を把握するという観点からはマッピングは有効であるが、語彙を広げるにはいたらない。

語彙を増やしていくための指導として、漢字練習があげられる。漢字検定などの外部的な指導漢字体系に基づいて、学年に応じた級ごとの漢字を練習させる。外発的動機としては、獲得した級について、学内の成績に反映させるやり方がある。また、普段の授業における指導としては、新出漢字などを中心とした漢字練習であり、校内定期テストへの出題という形で成績に反映し、生徒はテスト勉強として漢字を勉強・練習することで、語彙を広げていく。これが外的・内的な漢字練習の実態であろうか。

しかし、漢字以外の語彙についてはどうか。先述したように、体系的語彙が明確でないために、漢字練習のようにはいかない。従って現在、漢字以外の語彙指導としては、言葉遊び等によって、語彙を膨らませようとする指導に終始してしまう。実際に教科書には次のようなものが掲載されている。

「次の言葉を、和語と漢語に分けよう。
言葉　ゆっくり　淡黄色　黄色　機織り　牧場（ぼくじょう）　牧場（まきば）
美しい　大砲」
「例にならって、□の中に漢字を入れ、しりとりを完成させよう。
役□→□合→合同→同□→□約」（「国語１」光村図書）

「ア　おお牧場はみどり　草の海　風が吹く
おお牧場はみどり　よく茂ったものだ
イ　そもそも国政は、国民の厳粛な信託によるものであって、その権威は国民に由来し、その権力は国民の代表がこれを行使し、その福利は国民がこれを享受する。
ウ　打ったランナー、セカンドへ。ボールはセンターがキャッチ。バックホーム。いい球だ。ランナースライディング。タッチ、タッチ。タッチアウト。
アの中の和語、イの中の漢語、ウの中の外来語のうち、他の種類に置き換えられるものはありますか。また、置き換えるとどのように印象が違いますか。」（「新しい国語1」東京書籍）

現行の教科書には、このように言葉遊びとして、語彙を増やそうという方法が見られる。また、言葉遊び的な副教材も存在する。

しかし、これでは体系としての語彙の獲得は望めない。と同時に、内的体系、ソシュールに従えば個人的言語の共時的選択のなされていないことになる。そこで、共時的選択の幅を広げるために、シソーラスを用いる。これによって、言葉遊びから同属言語や言語範疇意識を伴った語彙の拡充がなされると考える。

たとえば「騒々しい」という言葉は「うるさい」という言葉にほとんどの場合置き換えられてしまっていて、「騒々しい」を使うことは少ない。今までの国語科の学習の中で「騒々しい」を使う場合、前述のように、しりとりやクロスワードパズルのようなゲーム的なものであった。しかし、これでは活性化はしない。つまり、自分で文章を書いたりスピーチをしたりするときに使う意識は生じない。活性化するとは、自分が書いた文章で「騒々しい」という言葉が、いかにも当てはまるという状況を作ることである。あくまでも生徒自

身の文章でなければならない。

そして、自分の文章の中の言葉に対して、さらに適切な言葉はないかを探させる。このときにシソーラスを使うことで、同じ範疇にある言葉の中から、自分の概念と対応する言葉を探し出す活動が生じる。さらに、シソーラスは単一の語彙を表示せず、範疇として表すために、「騒々しい」だけでなく、同じ範疇にあるさまざまな言語を「枠」として捉えることができる。

学習指導要領で「語句についての指導事項」については、「慣用句，類義語と対義語，同音異義語や多義的な意味を表す語句の意味や用法に注意すること」となっているため、実際の授業としては、多様なことばに触れることを目標とした。

単元として「シソーラスを使った語彙拡充」を設定した。語彙の獲得は様々な要因が挙げられる。生徒の発達段階に応じて、身体言語を伴った語彙の獲得は、発達心理学のほうで詳しく述べられるところであるが、様々に獲得した語彙をどのように広げていくかについて、本授業では一つの試みを行っている。それがシソーラスによる指導である。

また、ことばの選択は言語体系的選択と個人文脈的選択とがある。よって、この両者を生徒個人の中で結合させるために、他者との相互関係の場面においての言語使用の選択をさせる必要がある。つまり学習形態としてグループが必然的に必要である。

以上のことにより、語用に焦点を当てて授業を組んだ。ここにおいて、決定された文章ではなく、ことばの選択が可能な文章を提示し、ことばどうしの関係性を思索させることで、語彙の拡充が図れると考える。

言いかえにおいて、次の三点に注意させる。一つは品詞を変えて言いかえる。二つは漢語と和語の言いかえをさせる。三つはカタカナ語・擬音を使うことを指示する。品詞については文法的知識を活用することによって、思考の柔軟性を図る。漢語と和語の差異は生徒にとって意識的ではないながらも、ことばの変換に大きな役割を果たす。擬音については、ことばでうまく説明しにくい生徒が、擬音を多用することで相手に伝えようとすることから、擬音でもよしとする。

単元の目標として、以下の二つを設定した。

①シソーラスを使うことで、ことばの微妙なニュアンスを捉えることができる。

②様々なことばの中から自分の考えに一番合うことばを捜そうとする。

三年になってより、文法事項や漢字の学習など言語事項についての学習を経てきている。授業者の目的として、理解するだけでなく使える言語事項を目指して授業を行ってきた。文法では、活用などを理解するだけでなく、自分でいろいろな文章から文法事項を見つけてきたり、文法法則にそぐわないことばたちについて理論づけようとしたりしてきた。

しかし、文章を書く際には、それらの獲得した語彙を必ずしもうまく使っているとはいえず、自分の考えと文章の内容とが乖離している場面も多く見受けられる。さらに、一度書いた文章をさらによいものにしようという意図の下、書き直すという行動があまり見られない。これは、どのように獲得した言葉を使うかという、語用の問題が不十分であると考えられる。

そこで、もともとある文章を同じような内容になるように言いかえることで、ことばのニュアンスの違いを捕らえられることを企画した。違うことばに言いかえられない生徒には、シソーラスやコンピューターなどを使って、類語を参考にすることによって、生徒自身が考えることばに近いものを探す。

単元名　　ことばの力を育てる

「ことばの力を育成するための国語力向上に向けて

～シソーラスを使った語彙拡充単元～」

単元の指導計画

第１次　生徒自身が作った文章について、違うことばに置き換えられるかどうかを検討し、それぞれのことばのニュアンスを考える。(2 時間)

第２次 置き換えた言葉たちについて、実際の本（文章）の中でどのように使われているかを探し、自分たちの使用状況と比較する。(2 時間)

	指導事項	学習活動	留意事項	時間
導入	・ことばをいろいろと変えても同じような意味になることを知らせる。	・本字の学習として、ことばをさまざまに変換させることを知る。	・年間計画の中で、どの部分をやっているのか､何をこれから行うのかについて確認させる。	5
展開	・ プリント配付 ・ 個人でプリントに書かれている文章の一部を文章の大意を変えずに言いかえるように指示。 ・ 同じ文を選んだグループごとに集まり､文意が変わっていないか確認させる。 ・ グループ内でことばのニュアンスの違いについて書き出せる。	・ 個人個人で配られたプリントに書かれている文を読み、どの文の言いかえが可能か考える。 ・ 自分なりに言いかえのたくさんできそうな文を選び、かえてみる。 ・ 選んだ文章ごとにグループを作り、どのようなことばにかえられたかお互いに見合う。 ・ グループごとに六つ以上のことばに言いかえて、文章を作り直す。 ・ 他にもないかどうかを辞書類を使って確かめる。 ・ それぞれのことばについて、どのようにニュアンスが違うかを説明する。 ・ それぞれのことばについてどのようなシチュエーションで使うかについて考え、文脈を考えさせる。	・ 大意を変えずに言いかえるとはどういうことかについて､例を挙げながら説明する。 ・ ことばの言いかえには､品詞を変える、和語と漢語の変換、カタカナ語に直す､擬音を使うなどがあることを指示。 ・ 品詞分類や変換がうまくいかない生徒に対して､期間巡視をしつつ、変換の種類について、個別指導をする。 ・ 例文は十以上を用意するも､一つの文を三人以上は選択するように調整。 ・ グループにそれぞれの文を持ち寄った後､文意についてチェックする。 ・ ニュアンスの違いの説明について、文章として説明したり、長い文章を作って使い方を説明したりすることを説明。 ・ グループごとに回り､書き方のわからないところに具体的な指示をする。	30

まとめ	・グループごとにことばの言いかえとニュアンスの違いを確認させる。 ・ニュアンスの違いが正当かどうか聞き取る。	・班ごとに元の文章と言いかえた文章と使い方の違いについて発表する。 ・それぞれのニュアンスの違いを聞き、その違いは正しいかどうかについて意見交換する。 ・次回は実際の文章でどのように使われているかを確かめることを確認する。	・発表の仕方を指示。元の文章を言い、変えた文章を書き、それぞれがどのように違いがあるのかを説明させる。 ・発表された内容について、自分なりに検討し、自分の使い方とあっているか考えるようにさせる。 ・ノートにメモをさせることで、他者が考えたことばの種類や広がりを自分のものとさせる。	15

4. 「ひまわり」を使った学習

シソーラスを使った学習を行ってきたが、シソーラスだけでは限界があると感じ始めた。シソーラスの使用によって、さまざまな言葉の言いかえは可能になった。しかし問題点もある。それは、書き換えが単語レベルであるということと、しっくり来る言葉がなかなか見つからないということである。

単語レベルの書き換えの有効性も多くある。しかし、文章は単語の書き換えだけでは全体とのバランスや文脈での使用方法など問題も多い。また、学習者は自分の文章を直すときに、どんなにシソーラスを使っても、はじめに自分で書いた表現が一番いい、つまり「しっくりくる」という状況から抜け出せない。

そこで、それらの問題を解決するために、「ひまわり」を使い、文やフレーズで言い回しを獲得する学習を組み込んだ。

学習者ははじめに「夢中になったこと」と題して 600 字程度の作文を書く。その後、夢中を中心として、さらに深めたい表現を「ひまわり」を使って検索する。授業の実践は以下のとおりである。

1　年間指導計画における位置と学習形態

1 年生は教科書として「新編　新しい国語」（東京書籍）を使用している。年間計画としてこの教科書教材に沿って学習していく。しかし、本校では学習形態を一斉授業と少人数授業と二形態とっているために、必ずしも順番に行われるわけではなく、相互の連携を取りつつ進めている。

一斉授業は主に「読解」を扱い、週に二時間、読むことの学習に当てている。少人数授業は週に一時間を「文法・語彙」指導。もう一時間を「表現」指導、としている。

「主題を考えよう」の読解教材である「少年の日の思い出」は一月に学習する教材である。語句の確認、漢字の確認を経て、内容読解の学習に入り、学習者は主題について読み深めていく。

その後、表現に関わる語彙の学習として、類語について考え調べ、自己の表現を豊かにする学習へと進む。本時の授業はこの調べ学習の段階と自己の表現への移行の段階である。

2　学習指導要領の目標との関連

本単元の学習は学習指導要領の以下の目標と関連している。

「B　書くこと
ア　身近な学習の中から課題を見つけ、材料を集め、自分の考えをまとめる。
イ　伝えたい事実や事柄、課題および自分の考えや気持ちを明確にすること。
ウ　自分の考えや気持ちを的確に表すために，適切な材料を選ぶこと。」
また、内容に関しては言語事項の以下の項目と関係している。
「イ　語句の辞書的な意味と文脈上の意味との関係に注意すること。
ウ　事象や行為などを表す多様な語句について理解を深めるとともに，話や文章の中の語彙について関心をもつこと。」
さらに、次年度の学習項目の先取りとして、
「慣用句，類義語と対義語，同音異義語や多義的な意味を表す語句の意味や用法に注意すること。」にも配慮している。

3　教材研究

自分が夢中になったことについて、さまざまなことばを使って書き表すことは、語彙の広がりとともに夢中になった感じ方や深さなど、概念の広がりにつながる。

「青空文庫」はインターネット上に作られた主として文学作品のデータベースである。これら使うことで教科書教材だけでは収まらない、多様な文集や語彙に触れることになる。

全文語彙検索ソフトとして「ひまわり」を使用した。これは国立国語研究所が作成したソフトで、テキストベースのものであれば、語句などの情報を検索することができ、学習者は自ら自分でことばを拾い集めてくることができる。

4　学習者の実態

学習者は 1 月に「少年の日の思い出」を学習し、読解として深い読みを経験している。しかし、その中に出てくる語彙については、なかなか自分のものとならず、表現に結びついていかない。

1 年生の入学時および前期テストの成績による分析によれば、語彙学習と表現学習が結びついていないことが問題として挙げられる。

そこで本単元では読解教材として行った文章と語彙とを、類語を用いることによって自分の表現に役立てることを目標とする。

5　単元目標

①　「夢中」について類語を理解し、使われている文を探すことで、語のニュアンスをつかむことができる。
②　自分の文章に調べた語を生かすことができる。

6　本時の目標

①　「夢中」について、類語を理解する。
②　さまざまなことばが、文中でどのように使われているかを把握する。

7　評価

①　類語を調べ、意味を書き出すことができたか。
②　「青空文庫」を検索ソフトを使い、「夢中」について、多様な使い方があることを理解したか。

8　本時の学習

導入　7 分
①今日の授業と PC の使い方について説明する。

②「少年の日の思い出」の「とりこ」の意味を確認する。

展開　40 分

①　「とりこ」の類義語を書き出し、意味の違いを考えてみる。
　　類語辞典やネット上の類語検索を使いって探す。

②　それぞれの類義語について、文章中でどのように使われているかを書き出す。
　　コンピューターの検索ソフトを使い、さまざまな文中での事例を探す。
　　補助プリントを配布し、使用例と使用状況について書き出す。

③　友達の文例と合わせて、言葉でニュアンスがどのように違うかを考える。
　　グループ学習により、情報を共有する。

④　自分が夢中になったときのことを書くために、どのようなことばを使えばよいかを考える。

まとめ　3 分

①　自分の文章に生かすことを指示。プリントの提出。

5.　課題

日本語コーパスを国語教育でどう活用するかについて、二つの課題がある。一つは学習指導語彙の選定のように、学習内容の枠を決める必要がある。もう一つは、その内容をどう教えるかという学習方法についての検討が必要である。

学習内容については、教科書コーパスの作成などにより、学習者に必要な語彙を学習者の読解力や他の学力との関係を考えながら決めていく必要がある。内容については、今までの語彙指導の不備や欠点をどのように補っていけるか。その方法として、コーパスや語彙表や検索ソフトなどの使用が考えられる。

共起関係およびコロケーションに関する研究の流れ

----計量言語学分野、自然言語処理分野および辞書データなどを中心に----

荻野綱男（日本大学文理学部）　　荻野孝野（日本システムアプリケーション）

A review on the researches in co-occurrence and collocation

----Focussing　on the field of quantitative linguistics,
natural language　processing and dictionary data----

Tsunao OGINO (Nihon University　College of Humanities and Sciences)
Takano OGINO(Japan System Application Co.,LTD.)

1．はじめに

我々は、結合価やコロケーションについて、辞書に記述するレベルや形式を検討するために、本テーマに関する研究文献の調査を行った。これは、結合価が今までどういうレベルで使われ、また今後、結合価に関して何が要求され、どういう記述が望ましいかなどを検討するためのものである。

まず、ここで語と語の結びつきの度合いにより、調査対象とした結合価がどのレベルのものであるか、位置づけをする。

① 形態的に可能な接続であるが、意味的には比喩などを除いて、稀有な結びつき
例 1　鉛筆が歩く

② 形態的にも意味的にも結びつき可能な表現で、全体の意味も各単語の意味の合成で通じるもの
例 2　子供が歩く

③ 形態的にも意味的にも結びつき可能な表現であり、全体の意味も各単語の意味の合成で通じるものであるが、先行する単語が後続する単語に対し形態的、意味的な制約をするもの
例 3　○決して　行か｛ない｝　　×決して　行く

④ 類義の範囲の単語が共起し、全体で類義の意味を合成しているもの
例 4　｛顰蹙、失笑｝を買う

⑤ 構成する各単語の合成でも全体の意味は出てこないもの
例 5　青菜に塩（＝しょんぼりしている）　　気を落とす（＝がっかりする）

本調査の対象は、②③④の範囲であるが、調査対象範囲として理工系学会文献（一部境界領域の学会文献も含む）を主としたため、多くの論文において②の範囲が中心となった。共起関係を意味的制約の段階からみて、ここでは意味マーカレベルの選択制限の範囲である、いわゆる結合価でとらえられるレベルの研究論文が主となっている。

日本における理工系分野の言語研究は、機械翻訳や情報検索など日本語文を対象にした自然言語処理の開発に大いに貢献してきた。これらの自然言語処理分野で、仮名漢字変換をはじめとする日本語文の処理に結合価を導入した様々な研究がある。

結合価を日本語解析の研究に取り入れた初期の文献として、石綿敏雄の研究があげられる（石綿敏雄　1975)。1970 年代初期から海外の結合価研究に着目した石綿は、「結合価」という言葉の歴史的背景を、「結合価の考えは、マイナー（1781)、ハイゼル（1908)、ベハーゲル(1924)、ビューラー(1934)などにその萌芽が見られるが、結合価（Valenz)ということばが使われたのは、ドイツのテニエルが最初である。」と紹介している（石綿敏雄　1983)。

一方で、言語学分野とは全く異なる工学分野において、1960 年半ばから 1970 年半ばにか

けて九州大学工学部栗原俊彦教授を筆頭とする九州大学の自然言語処理グループによって、「格助詞、体言の意味分類、体言の意味役割」に着目した文分析の研究が行われていた（栗原俊彦、吉田将、鶴丸弘昭、藤田毅　1977)。これらの研究はその後の自然言語処理の土台を築き上げ、その研究成果はメーカーの研究グループによって、ワープロ第１号機（1978）として実用化された（古瀬幸広 1992)。ワープロの基盤技術となった仮名漢字変換システムは、今日、一般人が簡単にキーボードで電子データを作成、蓄積、通信、検索できるという部分において、情報化社会の進展に多大な貢献をしてきたといえる。また、日本語入力の簡便化によってWebデータや新聞データなど大規模な電子データが蓄積され、言語事象の分析にも有効に活用できる時代となっている。

ここでの文献調査では、結合価の果たしてきたこれらの歴史的背景を踏まえ、(1)結合価そのものを言語データとして蓄積している研究、(2)それらの利用研究、に着目し調査を行った。

2.　文献調査

2.1 文献調査の概要

2.1.1 調査対象

調査対象は、過去３０年ぐらいまで遡って、自然言語処理関連の学会誌、結合価や共起に関する電子データや書籍とした。今回の調査対象は、学会論文を中心とし、論文の記述言語の種別と研究対象の言語の組み合わせで、以下の範囲を対象とした。

①日本語で書かれた日本語を対象にした研究
②英語で書かれた日本語を対象にした研究
③日本語で書かれた英語を対象にした研究
④日本語で書かれた日本語・英語以外を対象にした研究
⑤英語で書かれた英語を対象にした研究

表１　調査対象とした学会誌（一部書籍）と関係論文数

	調査対象	区分	対象文献年度	抽出論文数
(1)	日本語で記載された国内文献	①②④		
(1.1)	計量国語学会		1970～2006	10
(1.2)	言語処理学会大会		1996～2006	105
(1.3)	言語処理学会論文誌		1994～2006	22
(1.4)	情報処理学会自然言語処理研究会（NL）		1960～2006	143
(1.5)	情報処理学会NL以外の研究会		1960～2006	11
(1.6)	情報処理学会論文誌		1980～2006	21
(1.7)	書籍類			17
(2)	英語で記載された国内外文献、日本語で記載された英語を対象とした文献[注1)]	②③⑤		
(2.1)	Computational Linguistics	(②⑤)	1986～2006	9
(2.2)	Proceedings of COLING	(②⑤)	1986～2006	30
(2.3)	その他（日本語記載の英語対象文献、英語で記載された国内文献）	(②③)	1996～2006	3

注1)　表１の（２）は、二つの海外学会を中心とする調査であるが、一部に、国内の学会誌で英語を対象に書いたもの、日本語を対象にして英語で書いたものを含む。

調査は、表１の(1)(2)の両方について行ったが、紙面の都合上 (1)を中心に述べる。
(2)を含んだ形での詳細報告については、特定領域研究「日本語コーパス」辞書編集班平成１８年度報告書を参照いただきたい。

2.1.2 調査の形式および観点

調査は以下のように行った。

A　書誌情報および利用分野などの区分情報の記載

表 2 の文献調査表の形式で、書誌情報および「利用分野、結合価のレベル」などの観点から分類した情報を記入する。

○表に記載する項目：

①学会区分：調査した学会論文誌名や研究会名
②著者名　③表題　④書誌情報　⑤発行年月日　⑥ページ
⑦ファイル名
⑧キーワードなどあった場合
⑨該当論文かどうか：第一段階としてキーワード抽出したものも作業過程では記載したので、手作業段階で該当しないと判断したものは×を入れる。作業対象外となる。
⑩言及のレベル
以下の三つのレベルにわけ、論文が言及している範囲について○を入れる。
・格助詞レベル
・体言の意味分類レベル：体言の意味分類や意味マーカ部分まで言及しているもの
・深層格〈関係子〉レベル：「体言、格助詞」の組み合わせが「主体、対象、道具、」といった意味的役割を表現している範囲まで言及しているもの
⑪調査対象品詞（係り先のレベル）
結合価や共起関係の検討において、係り先である用言の範囲が「動詞、形容詞、形容動詞」のどの部分までかについて記入する。
⑫論文の記述レベル
以下の四つの区分を立てて、該当する範囲の番号を記入する。
１：結合価の概論や説明　２：辞書開発あるいはデータ
３：利用研究　４：ツール開発
⑬利用分野
⑫の区分で「３:利用研究」に分類されたものについて、実際にどういった利用分野なのかを自由に記載してもらう。作業段階では利用分野の項目名に多様性が出てしまうが最終的に同義のもの類似のものを判定し、項目をグループ化して集計している。
⑭対象とする言語

B　要約

Aの⑨の作業で該当論文と判断されたものすべてについて、内容を読み要約を作成した。国内文献では、表２の調査表に⑮として要約を入れた。

2.2 文献調査結果の数量的検討

2.2.1　記述レベルについて

本調査では、結合価の記述レベル(言及レベル)を大きく、以下の 3 段階でとらえ、それぞれの論文がどのレベルまで扱った論文であるかを区分した。

① 格助詞のレベルで扱ったもの
② 格助詞に前接する体言の意味分類のレベルまで扱ったもの

③　{体言＋格助詞} が担う［主体、道具、場所］という意味役割まで扱ったもの

各論文の傾向は表 3 に見られる通りである。全体の傾向でみると、国内文献では、 ①格助詞レベル、②体言部分の意味レベルまで触れたものが多く、③意味役割まで言及しているものは少ないといえる。

表 2　文献調査表とその作業見本

参考文献					言及のレベル			係先タイプ	区分			要約
著者名	表題	文献	年月日	結合価、コロケーション、共起などにかかわる文献	（1）格助詞レベル	（2）体言部分の意味	（3）深層格、関係子	品詞による区分	区分（1：概論や説明、2：辞書開発、3：利用研究、4：ツール開発）	利用分野	対象言語（1：日本語、2：英語、3日英、4その他（ ）	
藤井 敦, 秋山 典丈, 徳永 健伸, 田中 穂積 (東工大)	動詞の多義性解消における格の弁別能力と集中度の有効性について	言語処理学会第1回年次大会 (1995)	1995	○	○	○		1	3	多義性判定	1	

表 3　論文別にみた記述レベルの状況(国内文献)

	①格助詞レベル	②体言部分の意味レベル	③深層格、関係子レベル
(1.1) 計量国語学会	8	2	4
(1.2) 言語処理学会大会	58	38	19
(1.3) 言語処理学会論文誌	15	15	9
(1.4) 情報処理学会自然言語処理研究会 (NL)	67	50	25
(1.5) 情報処理学会研究会報告その他	7	5	3
(1.6) 情報処理学会論文誌	11	7	4
(1.7) 日本語書籍	9	9	2
合計	175	126	66

2.2.2 利用分野について

A 全体の傾向

国内日本語記述文献では、表 4 に示す通り、「構文解析、意味解析」といった要素技術と「機械翻訳、情報検索」といったシステム利用とがあるが、その中でも「機械翻訳」への利用が多いことがわかる。これは、理工系分野の学会では、学会別、年代別でもほぼ同じ傾向が出ている。

表 4　利用分野別論文数(国内文献)

利用分野	論文数
機械翻訳	65
構文解析	28
情報検索	28
格解析	18
照応・省略解析	12
仮名漢字変換	10
言い換え	10
意味解析	9
情報抽出	9
シソーラス構築	9
文書分類	8
語義判別	7

利用分野	論文数
多義性判定	7
音声認識	6
単語分類	6
自然言語処理	5
音声処理	5
キーワード抽出	4
テキスト処理	4
共起表現抽出	3
形態素解析	3
辞書作成	3
質問応答	3
自動要約	3

利用分野	論文数
校正システム	2
呼応表現抽出	2
固有名詞抽出	2
情緒表現解析	2
深層格抽出	2
大規模辞書	2
トピック抽出	2
複合名詞解析	2
用言の分類	2
格フレーム獲得	2
要約	2
その他	29

B 年度別にみた論文数

図 1 からわかるように、1996 年、1999 年あたりに論文数の極端な増加が見られる。これを時代別利用分野でみると、1990 年代に入って、「機械翻訳、情報検索、構文解析」などへの結合価利用が盛んになったことがわかる。これは、「IPAL 動詞辞書」(情報処理振興事業協会 1987)、「EDR 電子化辞書」(日本電子化辞書研究所 1995)、「日本語語彙大系」(池原悟,宮崎正弘,白井諭,横尾昭男,中岩浩己,小倉健太郎,大山芳史,林良彦 1997) など、結合価分析に利用することができる大量の言語データが整ってきたことにも関連するかと思われる。

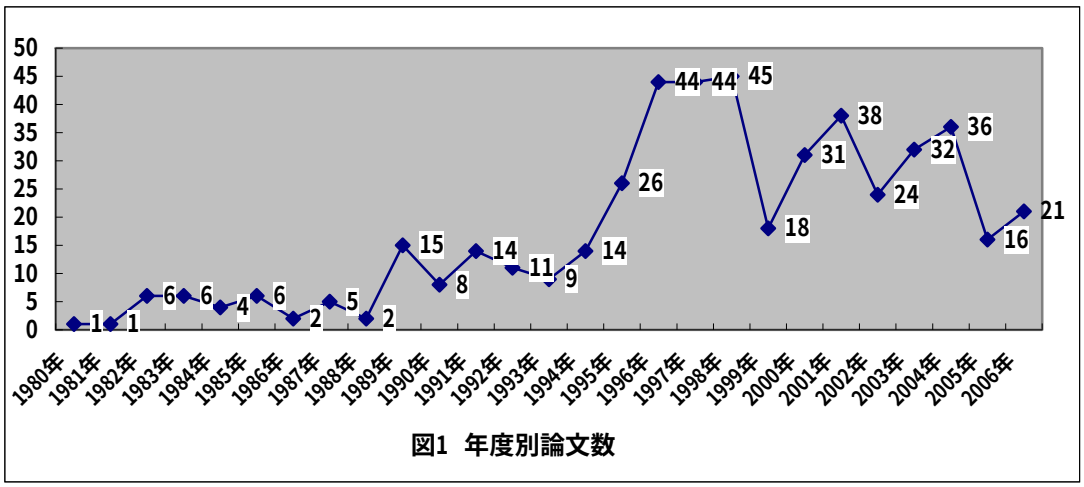


図1 年度別論文数

2.3 共起関係およびコロケーションに関する研究概観

2.3.1 結合価に着目したデータの蓄積

「1 はじめに」でも触れたように「結合価」は言語学分野におけるドイツ文法を中心に 1900 年代初期から着目されていたと報告されている。その後、日本においても言語データを形式的に検討する言語研究の中に取り入れられ、石綿敏雄によって用言（和語動詞、形容詞、形容動詞）900 語について具体的な結合価記述の試み（石綿敏雄　1975）がなされた。その後、結合価の研究は、当時進展し始めた自然言語処理の流れに沿って、機械翻訳などの研究者や電子化辞書開発者から注目され、自然言語処理の分野において計算機用日本語基本動詞辞書ＩＰＡＬ（情報処理振興事業協会　1987）（動詞約 800 語）など、結合価を組み込んだ自然言語処理のための辞書開発へと広がっていった。やがて、機械翻訳の訳語選択のための辞書として、「日本語語彙大系」（池原悟他　1997）が開発された。その後、荻野孝野らは、大量かつ実際の文例に基づいて結合価データを収集する必要性を実感し、ＥＤＲコーパス及び共起辞書の係り受けデータを用い、約 12,400 概念に相当する規模の表層レベルの結合価を格助詞別単語事例の形式で「日本語動詞の結合価」（荻野孝野他　2005）としてまとめた。また、黒橋禎夫、河原大輔は、黒橋らの作成した日本語構文解析システムを用い、係り受け関係を認定し、用言別に大量の「体言＋格助詞」データを自動的に収集している（河原大輔,黒橋禎夫　2002)。新聞記事２６年分、2,600 万文から約 18,000 用言の格フレームの自動構築や、Ｗｅｂから大量の言語データの自動的収集（河原大輔,黒橋禎夫　2006)を行い、共起関係の検討などを行っている。データそのものは一般に公開され、利用が可能である。

2.3.2　結合価の自然言語処理での利用

結合価データは、

①係り受け関係にあいまい性がある場合の判定
②格助詞選択が不適切な文の校正
③必須の格要素が省略されているときの省略格の推測
④同じ表記の単語に対して複数の表記や訳語が対応しているときの判定

などに利用されてきた。

具体的には、動詞が必要とする格関係は、「きしゃのきしゃがきしゃできしゃする」の文でよく紹介される仮名漢字変換システムの同音異義語の選択をはじめ、機械翻訳の訳語選択、文脈理解における照応関係や省略された格の推定などの言語データとして活用されてきた。

参考までにいくつかの先行研究を利用分野別に概観する。

Ａ　仮名漢字変換における同音異義語の処理

自然言語処理における結合価を用いた解析システムの代表的なものとして、まず同音異義語の処理を含む仮名漢字変換があげられる。

仮名漢字変換における同音異義語の判定は、1980 年代半ばに連語の共起情報に意味分類を導入してその関係から判断する手法（本間茂,山階正樹,小橋史彦　1986）や格文法を用いた判定（大島義光,阿部正博,湯浦克彦,武市宣之　1986）が提案された。開発の初期段階においては構文解析レベルから進んで、こうした意味解析部分の導入で精度の上昇がみられたことなどが研究論文レベルで報告されている（大島義光他 1986）が、その後仮名漢字変換に関する論文は 1980 年代後半から急激に減少している。これは仮名漢字変換の開発が各社実用段階に入り、それぞれの開発機関がいかに精度のよい仮名漢字変換をするかが製品開発に直結するため、研究発表としては封じられた段階であると想像される。

B　結合価を導入した機械翻訳における訳語選択

機械翻訳の分野では、まず訳語選択向上のために作成した「日本語語彙大系」（池原悟他 1997）の研究があげられる。これは、日本語側の格助詞パターンとそれぞれの格助詞部分に導かれる体言を意味分類で表現し、英語側文型パターンを対応させたものである。

例えば「掛ける」という日本語には「物を壁などに吊るす」「椅子に腰掛ける」「お金を掛ける」などいくつかの語義があり、その語義によって訳語も異なってくるが、格助詞の組み合わせ及び体言部分の意味特徴によって適切な訳語が選択できることが示されている。

表５　「掛ける」の結合価に対応する英語文型

が	を	に	英語文型
N1	N2	N3	
主体	絵、額、衣類	住居	N1 hang N2 on N3
人間	腰	椅子	N1 sit down on/in N3

機械翻訳の利用研究に関する論文がどの時代でも大変多いことがわかる。結合価を用いた訳語選択の評価については「結合価文法による動詞と名詞の訳語選択能力の評価」（金出地真人，徳久雅人，村上仁一，池原悟　2003)などがある。

C　構文解析における係り受け先の決定

黒橋禎夫らは、格助詞関係が同じになった並列句の係り受けの多様性の決定に体言の意味分類の近似性を導入している（黒橋禎夫,長尾眞　1994)。これは結合価の第２段階レベルの意味特徴を利用したシステムと言える。これは、例 6 に示すように、動詞が複数出現し、かつ動詞にかかる体言が複数出現して並列構造になっている場合、どの部分の体言とどの部分の体言が同格で同じ動詞にかかっているかを判定するものである。

例 6　プログラム言語は　問題解決の　アルゴリズムを　表現できる　A 記述力と
計算機の　B 機能を　充分に　C 駆動できる　D 枠組みが　E 必要である。

「A 記述力と B 機能を」が並列構造になって「C 駆動できる」にかかるのか
「A 記述力と D 枠組みが」が並列構造になって「E 必要である」にかかるのかを、体言部分の意味の類似度の近さによって判定している。

D　文脈処理における照応関係や省略関係の格の推測

自然言語処理は文の範囲を超えて文脈の解析に進み、指示代名詞による照応関係の推測（飯田龍,乾健太郎,松本裕治　2005）や、省略されている格の推測などにも結合価データの利用が考えられている。特に照応関係の推測や省略格の推測には、格助詞レベルだけでもこれが基本形の格助詞パターンであるという基準データがあれば、解析が進むものと思われる。

3　最後に

1980 年代から現在に至るまでの、結合価およびコロケーションにかかわる既存研究を広く把握するために文献調査を行った。自然言語処理において、結合価は、仮名漢字変換や機械翻訳などに大いに活用され、また現在はより解析のレベルを深めた文脈処理や、新しい方向である Web から意味解釈を含めた情報抽出など次のステップに研究が進んでいる。今後どういう部分に活用されるかの予想をたて、どのレベルの情報を記載すれば使えるかを検討するために、今回の抽出論文をさらに詳細に検討して、今後の研究方向に活用できればと思う。

謝辞

国内文献の調査では、衛藤純司さん（ランゲージウェア)、木村睦子さん、金丸敏幸さん

（京都大学大学院人間・環境学研究科）、十河則子さん、小笠原あゆみさん、大島玲子さん、海外文献の調査では、和泉絵美さんにご協力をいただきました。謝意を表します。

参考文献

飯田龍、乾健太郎、松本裕治(2005)．“照応性判定を含む名詞句照応解析の実験と分析”.自然言語処理研究会,2005-NL-169,PP.93-100

池原　悟、宮崎正弘、横尾昭男(1993).“日英機械翻訳のための意味解析用の知識とその分解能”.情報処理学会誌 Vol.34 NO.8. PP.1692-1704

池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩己、小倉健太郎、大山芳史、林良彦（1997）.日本語語彙大系.岩波書店

石綿敏雄（1975）．“日本語の生成語彙論的記述と言語処理への応用”.電子計算機による国語研究Ⅶ.国研報告 54. 秀英出版

石綿敏雄 荻野孝野(1983).“日本語用言の結合価”．『朝倉日本語新講座３　文法と意味１』.PP.226-272.朝倉書店

石綿敏雄（1983）.“結合価文法の理論的背景”．『朝倉日本語新講座３　文法と意味１』.PP.81-109.朝倉書店

大島義光、阿部正博、湯浦克彦、武市宣之(1986)．“格文法による仮名漢字変換の多義解消”.情報処理学会論文誌,Vol.27 No.7, PP.679-687

荻野孝野、石崎俊（1985).『用言に関する結合価の自動抽出データ資料』. 電子技術総合研究所

荻野孝野、小林正博、井佐原均(2003)．『日本語動詞の結合価』.三省堂

金出地真人、徳久雅人、村上仁一、池原悟(2003)．“結合価文法による動詞と名詞の訳語選択能力の評価”.自然言語処理研究会,2003-NL-153,PP.119-124

河原大輔、黒橋禎夫(2002)．“用言と直前の格要素の組みを単位とする格フレームの自動構築”.自然言語処理,9(1),PP.3-19

河原大輔、黒橋禎夫(2006)．“高性能計算環境を用いたWebからの大規模格フレーム構築”．自然言語処理研究会,2006-NL-171,PP.67-74

絹川博之、木村睦子(1980).“日本語文構造解析による自動インデックス方式”. 情報処理 21[3]. PP.200-207

栗原俊彦、吉田将、鶴丸弘昭、藤田毅(1977)，“言語と思考のシュミレーション”，講座情報社会科学『言語と情報』．学研

黒橋禎夫、長尾眞(1992)．“格フレーム選択における意味マーカと例文の有効性について”. 自然言語処理研究会,92-NL-91,PP.79-86

黒橋禎夫、長尾眞(1994)．“並列構造の検出に基づく長い日本語文の構文解析”. 自然言語処理,1(1),PP.35-58

古瀬幸広(1992).“ワープロ発明者の知られざる末路”.新潮５

柴谷方良,影山太郎,田守育啓(1982)．『言語の構造』. くろしお出版

情報処理振興事業協会（1987).『計算機用日本語動詞辞書 IPAL』. 情報処理振興事業協会

情報処理振興事業協会（1987).『計算機用日本語基本動詞辞書ＩＰＡＬ　解説編』. 情報処理振興事業協会

電子技術総合研究所推論システム研究室（1985）．『用言に関する結合価の自動抽出データ資料』． 電子技術総合研究所

独立行政法人情報通信研究機構.『EDR電子化辞書』. http://www2.nict.go.jp/kk/e416/EDR/J_index.html

中岩浩巳,白井諭,池原悟(1997)．“日英機械翻訳における語用論的・意味論的制約を用いたゼロ代名詞の文章外照応解析”.情報処理学会論文誌　Vol38 No.11. PP.2167-2178

本間茂、山階正樹、小橋史彦(1986)．“連語解析を用いたべた書きかな漢字変換”.情報処理学会論文誌,Vol.27 No.11, PP.1062-1068

語彙概念構造辞書の構築による意味役割分析

竹内 孔一 (岡山大学)[1]

Semantic Role Labeling Based on Lexical Conceptual Structure Dictionary

Koichi Takeuchi (Okayama University)

1 はじめに

岡山大学の研究グループでは日本語コーパスから語彙概念構造辞書 (Lexical Conceptual Structure, 以下 LCS) を半自動で構築する研究を行っている．LCS とは動詞の意味を統語構造との対応をとりつつ動詞間の含意関係を整理して記述する枠組みであり，言語学で主に研究され現在も発展中の理論的枠組みである．これに対して記述できる範囲を限定した形で電子化した辞書 (約 1200 語の動詞を対象) が竹内 (2004) によって公開されており，この辞書を元にしてコーパスによる自動的な拡張をおこなうことを考えている．本年度はまず作成しようとする LCS 辞書の体系がどの程度言語処理で有効であるかについて述語と項との関係を抽象化してとらえる意味役割付与システムを作成し，その精度をとおして LCS 辞書の有効性を評価したので報告する．評価の枠組みを固めておくことで，これから自動的に LCS 辞書を拡張した場合に比較による辞書の良さが評価できる．

次にテキストコーパスを利用した LCS 辞書の構築法について現段階で検討している方法について記述する．既存の LCS 辞書を利用する枠組みでどのような問題があるかを整理し具体的なモデル化について議論する．

2 LCS 辞書評価の枠組み

LCS 辞書とはどのような情報を持っているのか具体的に説明した後意味役割との対応関係について述べて評価の枠組みについて説明する．

LCS は動詞の意味を項の間の関係で記述するとともに構造的に記述することで動詞間の関係を記述するものである．例えば「あげる/渡す」という動詞の LCS は [x CAUSE [BECOME [y BE AT z]]] という式で表される．これは「x が y を z にある状態にするという変化結果を引き起こす」ことを示しているとともに「y が z にあるように変化する」と「y が z にある」が含意されていることを LCS 内の部分構造が示している．

一方，意味役割は文の述語と項との関係を抽象化したラベルであり，これを文に対して付与することができると情報の集約を行うことができる．例えば以下の例文の場合，

- 彼が　彼女に　プレゼントを　あげる/渡す

「彼」が Agent,「彼女」が Goal,「プレゼント」が Theme となり，これによって「あげる/渡す」のどちらの表現も Theme(「プレゼント」) が Goal(「彼女」) に移動するように Agent(「彼」) が仕向けたことを示している．移動や変化という概念を中心にこのような抽象的な役割分析を行うことで事実関係の集約が可能となる[2]．

上記で示した LCS と意味役割の記述的対応関係から明らかなように LCS 辞書には動詞がどのような意味役割を持っているのかを記述している．そこで LCS 辞書を元に動詞に対応した項の意味役割を付与するモデルを作成し，その精度で LCS の体系が実際どの程度のうまく辞書化できているかを測定する．

3 意味役割付与タスクとモデル化

3.1 問題設定

言語処理においてどの程度の意味役割の種類が必須であるかはまだ定まっていない．そこで先行研究とデータ分析から主要な意味役割を Agent, Experiencer, Instrument, Theme, Source, Goal,

[1]koichi@cl.it.okayama-u.ac.jp

[2]例文ではプレゼントが結果として今どこにあるのかという事実関係.

Location, Time, Scene, Path, Reason, Opponent の 11 種類に定めた (詳細は下村他 (2006) 参照). この 11 種類の意味役割は名詞に対して決定されるのではなく，文における述語の持つ動作・状態をある典型的な骨格 (フレーム) として考えたときに仮定できるものである．LCS 辞書はこの典型的な振る舞いの型を動詞ごとに書いた言語資源であるので LCS 辞書を用いて表層格表現と意味役割を結びつける表層深層対応規則を作成することができる．図 1 には「発足する」という動詞の LCS 辞書から意味役割フレームが 2 種類対応する可能性を示している．それぞれのフレームは文脈と名詞のタイプによってどちらかが選択される．これらのフレームの違いを文で表現すれば「彼が野球チームを発足する/ 野球チームが発足する」である．つまり LCS はある動詞が選択できる意味役割フレームのセットを与えており，ここから名詞の概念や文脈を利用してどのフレームが文で表現されていたかを推定するのが意味役割タスクである．ただし上記の意味役割セットの中には Time や Location といった個別の動詞に依存しない付加詞も決定する必要がある．これは LCS 辞書とは別に名詞の概念と文脈により推定するモデルを別に用意する必要がある．

3.2 意味役割付与モデル

意味役割付与モデルの全体像を図 3 に示す.
処理手順としては以下のようになる.

- 述語と項の組み合わせを取得 (イベント単位の取得)
- LCS 辞書からの意味役割フレーム候補の取得
- 名詞の概念体系を利用した意味役割フレームの選択
- 付加詞の意味役割分析

以下順に説明する.

図 1: LCS 辞書から表層深層対応規則の生成

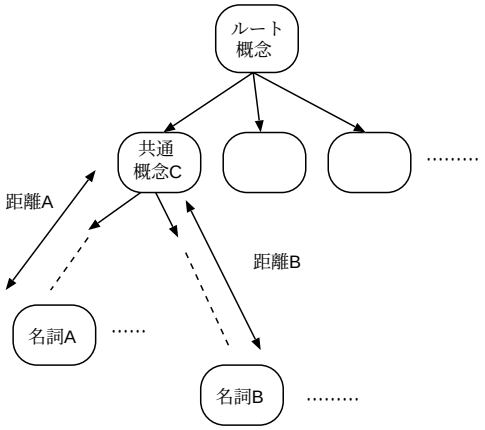


図 2: EDR 辞書を利用した名詞の概念距離計算

述語と項の組み合わせの取得 まず入力文に係り受け解析器 (Cabocha) を適用し動詞，サ変名詞と係り関係にある項を同定する．一つの述語 (動詞) に対して項が複数あり一つの事象単位をあらわすことから，ここではイベント単位と呼ぶことにする．イベント単位はガヲ二格といった格助詞だけを対象にするのではなく，図 4 に示すように「彼女たちの (Agent) 送金」のように接続助詞「の」によるサ変名詞との関係も対象にする.

LCS 辞書からの意味役割セット候補の取得 LCS 辞書そのものは式で表現されているだけで (2 節参照) 意味役割ラベルが存在しない．しかし LCS は意味役割と対応関係があることから，LCS 辞書からあらかじめ意味役割ラベルの形式に変換した制約付き表層深層対応規則を作成する．図 1 に例を示す．図中の「意味役割との対応」が表層格に対応した意味役割を示しており「格の制約」がそのフレームを選択する場合の名詞に対する制約である.

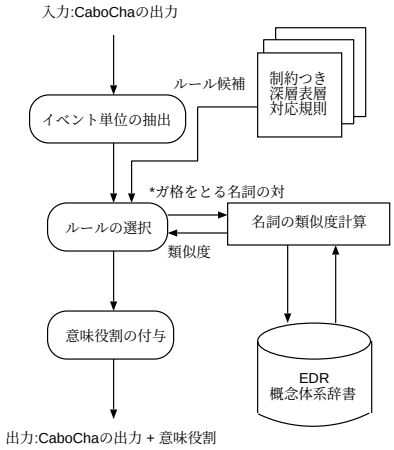


図 3: 意味役割付与システムの全体図

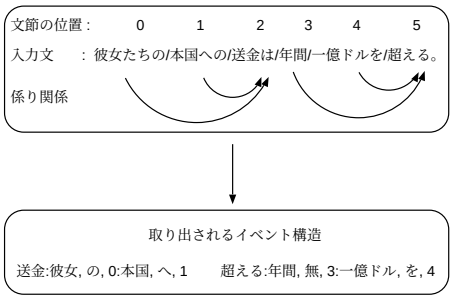


図 4: イベント単位の抽出例

名詞の概念体系を利用した意味役割フレーム選択 名詞の概念体系辞書を利用して表層格に対応する名詞の分類に従って意味役割フレームの選択を行う．例えば「研究部門が発足する」が入力された場合，LCS 辞書の対応規則 (図 1 参照) によってガ格の名詞が人か物かのどちらに近いかによって意味役割フレームが選択される．この近さを EDR 概念体系辞書の構造を利用して測定する．この例文では「部門」が人に近いのか物に近いのかで判断する．具体的には EDR 概念辞書で「部門」と「人」,「部門」と「物」との共通概念からの距離の和をそれぞれ比較して距離の短いほうの概念を正解とする (図 2)．この際入力単語 (ここでは「部門」) には複数の概念パスが存在するがここではもっとも比較対象 (「人」,「物」) に対して近い概念を利用して計算する (詳細な計算方法は下村他 (2006) を参照).

付加詞の意味役割分析 意味役割の中で深層格 Instrument，Reason，Location，Scene，Time および表層格におけるデ格やマデ格は付加詞であるため動詞で決定すべきものではない．ここでは簡易的に各意味役割に対して名詞概念が対応すると仮定して処理を行う．具体的には上記の深層格に対してそれぞれ，Instrument:道具，Reason:現象，Location:場所，Scene:イベント，Time:時間と概念を設定して名詞がどの概念に対応しているか EDR 概念体系辞書を利用して決定する．例えば「飛行機で到着する」という文の場合,「到着する」のデ格は LCS 辞書では記述されていないため「飛行機」を EDR 概念体系で「道具」「現象」「場所」「イベント」のどの概念に近いかを判定する (下村他 2006)．これにより付加詞による意味役割を付与する．

4 実験と考察

実験対象は毎日新聞 95 年版のうち京都コーパスに収録されている 100 文とし，人手により意味役割を付与したタグ付きコーパスを作成した．ただし LCS 辞書が小規模であるため LCS 辞書に載っている動詞のみを実験の対象とする．評価はイベント単位と単語単位で行った．実験結果を表 1 に示す．イベント単位で 7 割程度の精度がでており，他の実験結果と単純な比較はできないが，肥塚ら (2007) の FrameNet による意味役割推定に比べて高い精度を示している．このことから本提案モデルの有効性は示せたと考えられる．表 2 に実験で誤ったものを原因別にまとめたものを示す．もっとも主要な誤り要因は名詞概念のカテゴリ分けであり，ついで機能語の判別，その次に LCS 辞書の不備が挙げられる．この誤り分析の結果から LCS 辞書はかなり正確に機能していることがわかる．よって意味役割付与タスクにおいて限定的にではあるが既存の LCS 辞書が有効であることを示すことができた．

5 LCS 辞書の構築の構築に向けた現段階の考察

前節の実験結果に示したように既存の LCS 辞書の体系と分類は有効である一方で人手での構築は大変なコストがかかるため，バランスの取れた書き言葉コーパスから統計的手法を利用して獲得し

表 1: 名詞単位, イベント単位での結果

	正解/全体	精度
名詞単位	219/268	81.71 %
イベント単位	117/165	70.90 %

表 2: 誤り解析

原因	誤り数/誤り全体	割合
名詞のカテゴリ分け	30/49	61.2 %
機能語	9/49	18.3 %
LCS の格の不備	4/49	8.1 %
定義	2/49	4.08 %
その他	4/49	8.16 %

たい．LCS は動詞の格パターンをベースにどのような意味が含意されているかを移動・変化という観点から分類している体系であるが，同時に格パターンと移動・変化という概念が近い動詞をクラス化していることになる．動詞がもつ概念的意味が似ている場合テキストコーパス上での表現も似ていると仮定できるならばコーパスにおける動詞の出現分布から動詞をクラスタ化してそのクラスと既存の LCS 辞書との比較から未知の LCS を獲得できるはずである．

動詞のコーパスにおける振る舞いから分類する際問題となるのは動詞だけの分類だけでなく名詞に対する分類も必要である．以下に例を示す．

- 彼が 道路/道 を 尋ねる/聞く/質問する
- 彼が 不満/愚痴 を 言う/もらす

これらはどちらもガ格とヲ格をとる他動詞であるが「尋ねる」のグループと「言う」のグループは LCS 辞書の観点からはそれぞれ異なる分類になる．例を見てわかるようにそれぞれの動詞グループのヲ格には特徴的な名詞が来ることが理解できるため，名詞の分類が動詞の分類に重要な影響を与えることがわかる．しかしながらどのような名詞グループがどのような格パターンに現れるかは先見的にはわからない．つまり，動詞と名詞を同時クラスタリングする手法が必要となる．

これを具現化するにはいくつかの手法がある．もし LCS 辞書に例文が付与されていればその事例をたよりに SVM などの統計的手法に基づく学習によりクラスタ化することができる．その際の名詞のグループは分類語彙表などの体系を利用して正解例文から拡張するなどの方法が必要である．しかし大量な例文作成は簡単ではない．正解を必要としない手法として Aizawa(2002) の同時共起クラスタリング手法がある．これはグラフ理論をベースにした手法で複数のカテゴリを分類する際の統計的指標の良さから全体のクラスタリングを行う．現在この手法による LCS 辞書の構築を研究している．

6 まとめ

約 1200 語の LCS 辞書と EDR 電子化辞書による概念体系を利用して意味役割付与システムを作成し，その精度から LCS 辞書の評価をこころみた．意味役割付与の精度は高く，誤りの多くは名詞のカテゴリー分析に関するものであった．LCS 辞書に起因する誤りは相対的に少なく LCS 辞書は精度良く構築されていると判断できる．今後，大量のコーパスを利用した LCS 辞書の構築について実行していく予定である．

参考文献

Akiko Aizawa (2002) “A method of Cluster-Based Indexing of Textual Data,” in *Proceedings of COLING 2002*, pp. 1–7.

下村拓也、竹内孔一 (2006)「名詞の概念体系を利用した規則に基づく意味役割付与システムの構築」，自然言語処理研究会，*175-NL-2006*，pp.13–20.

竹内孔一 (2004)「語彙概念構造による動詞辞書の作成」，第 *10* 回言語処理学会年次大会，pp.576–579.

肥塚真輔、岡本紘幸、斎藤博昭、小原京子 (2007)「日本語フレームネットに基づく意味役割推定」，自然言語処理，第 14 巻，第 1 号，pp.43–66.

書　名　特定領域研究「日本語コーパス」平成18年度公開ワークショップ（研究成果報告会）予稿集
発行日　平成18年3月15日
発行者　文部科学省科学研究費特定領域研究「日本語コーパス」総括班
http://www.tokuteicorpus.jp/
連絡先　〒190-8561　東京都立川市緑町10-2　独立行政法人国立国語研究所研究開発部門内
電　話　042-540-4300（代表）
文書管理番号　JC-G-06-01