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。

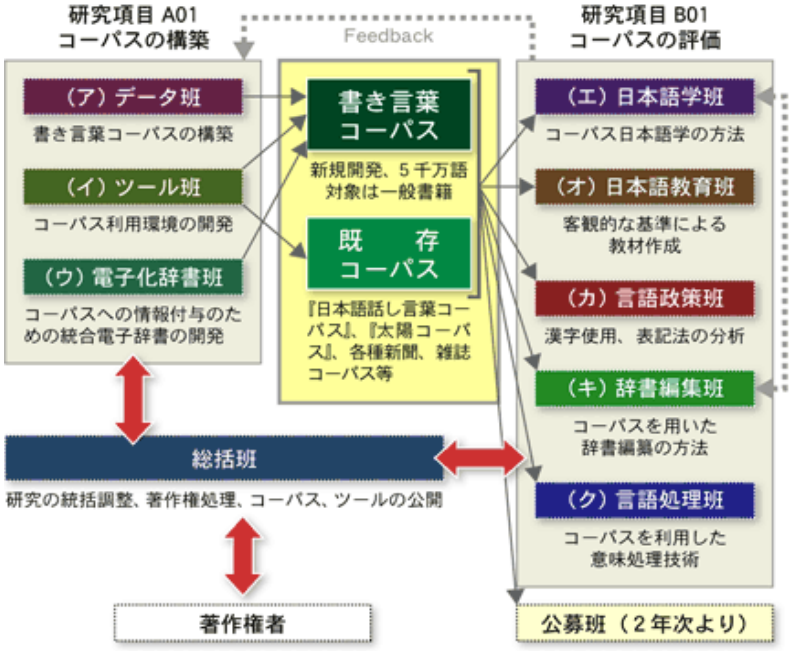

図１．本特定領域研究の構成

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。

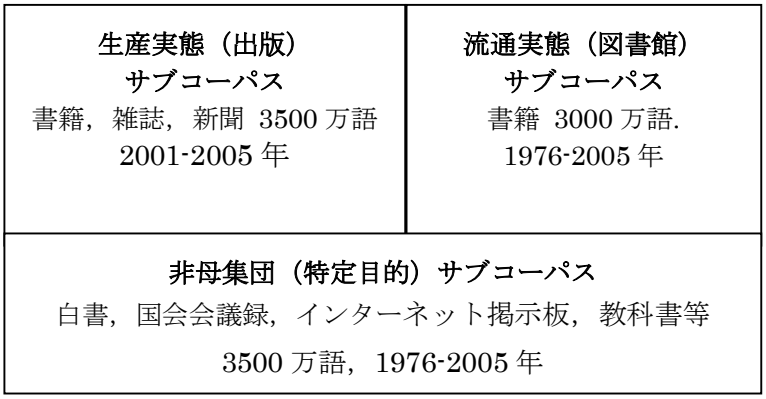

図２.『現代日本語書き言葉均衡コーパス』を構成するサブコーパス

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。

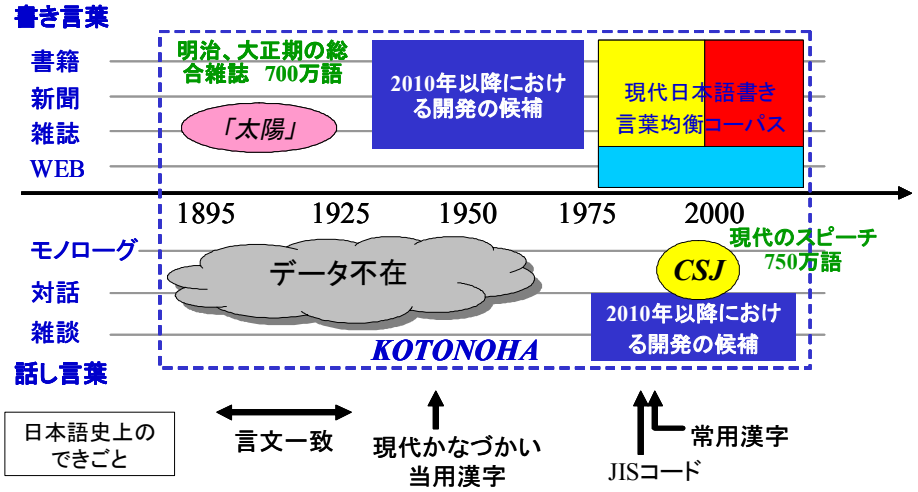

図３．国立国語研究所のコーパス整備計画 KOTONOHA

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。

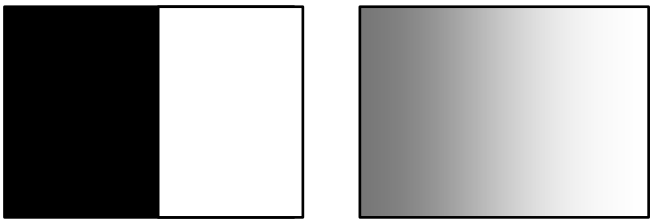

図４．文法性判断の離散性と連続性

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