共通教養日本語均衡コーパス(CCCJ)の概念

芝野耕司
東京外国語大学アジア・アフリカ言語文化研究所

Brownコーパスから LOBコーパスを経て，British National corpus (BNC)で一つの完成を見たコーパスは，情報環境の変化によって，大きく再検討を迫られている。この論文では，この変化に対応し，一つの言語を代表する均衡コーパスを“入力”コーパスから“収集”コーパスへの流れで再検討するとともに，新しい環境下でのコーパス構築を，“日本で学校教育を受け，日本で過ごした人の日本語能力を代表する共通教養日本語コーパス”と規定し，これを代表する共通教養日本語均衡コーパス(CCCJ)を構築について報告する。

Concept of Balanced Corpus of Common Cultivated Japanese (CCCJ)

Kohji Shibano
Research Institute of Languages and Cultures of Asia and Africa,
Tokyo University of Foreign Studies

Corpus linguistics has been established through the development of Brown, LOB and British National Corpus (BNC). However, informational environment has been drastically changed and reconsideration of the definition of corpus linguistics has emerged. In this paper, we reexamine the definition based on from “input” to “collect” corpus and then we report the development of a balanced representation of contemporary Japanese called balanced Corpus of Common Cultivated Japanese (CCCJ).

1. まえがき

Googleは，2006年夏に1兆語からなる英語の最大規模 n-gram である Web 1T [1]を発表し，2007年11月には，2,550億語からなる Google 大規模日本語 n-gram[2]を発表した。一方，自然言語処理の分野でもインターネットの普及に伴って得られるようになった大規模な言語データをもとに統計的自然言語処理[3]が広く研究されるようになった。

一方，コーパス言語学分野では，こうした趨勢に逆らい，依然として，従前の権威である British National Corpus (BNC)[4]に留まり，新たな環境への適応が進んでいないのが現状である。

筆者らは，Web 1T と BNC とを CEFR [5]との関連から比較し，大規模データの有効性を検証した[6]。また，1,200万語からなる教科書コーパスの構築[7]及びこれと Google 日本語 n-gram とを組み合わせ均衡コーパスを構築する試みを行い，こうした均衡コーパスの有用性を提示した[8]。

この研究では，こうした研究を一歩進めるとともに，大量のデータが利用可能となった現状で何を母集団とし，代表性を確保するとともに，どのように均衡をとるのか及びこうした視点に立って，「日本語」を代表する共通教養日本語均衡コーパス(CCCJ, Balanced Corpus of Common Cultivated Japanese)の構築思想及び構築したコーパスの性質に関して検証する。

2. Brown, LOB, BNC

現代のコーパス言語学は，Brown (IBM)コーパス[9]に始まる。Brown コーパスは，IBM の研究部門としては，英語の音声認識で必要とする言語データを作ることを目的とし，Brown 大学の図書館の蔵書目録を母集団とし，サンプリングによって，100万語のコーパスを作成した。Brown コーパスは，現代米語を代表するコーパスの構築を目指し，書籍・雑誌・新聞などのメディア別の割合及び文芸・情報のドメインも考慮し，コーパスを構築した。イギリス英語に関して Brown コーパスを超える試みは，Lancaster, Oslo, Bergen/IBM コーパスによって実現された。LOB/IBM コーパス[10]と呼ばれるこのコーパスでは，一つの大学の蔵書目録ではなく，英国出版総目録を母集団とするようにサンプリングを設計した。

Brown及び LOB/IBM での先駆的コーパス構築を受けて，BNC が構築された。BNC は Brown, LOB でのコーパス構築を踏まえ，蔵書目録や出版目録など生産側だけではなく，消費側の視点を導入するとともに，書き言葉以外に話し言葉を取り入れるとともに，サンプリング方法に関しても，詳細な検討を行い，結果として，もっとも権威のあるコーパスの位置を獲得した。

しかし，BNC は，インターネット以前の試みであり，電子化テキストを作成することから，コーパス構築を始める必要があった。また，1億語規模である BNC のサンプル規模は，十分ではなく，この規模では言語の個別研究において，統計的研究を行うことや用例検討を行うことが難しかった。

竹内他[6]が指摘するように，BNC は Brown の100倍のサイズであるが，ソースは Brown の 800 から BNC の 4000 と 5 倍に過ぎず，Google と比較した場合，BNC が全英語に関して，十分な代表性を有しているとはいえない。

3. 話し言葉コーパス書き言葉均衡コーパス

日本語に関しては，初期の日本語情報処理でよく参照された新聞データや青空文庫などの文芸データは，大量に存在するが，日本語全体を代表する均衡コーパスはまだない。

こうした中でも，最も広く期待されているのが，国立国語研究所による「日本語話し言葉コーパス(CSJ)」[11]と「現代日本語書き言葉均衡コーパスKOTONOHA」[12，18]である。

CSJは，3,302の講演の661時間の録音からなり，コーパス規模としては，752万語に及ぶ。しかし，CSJは開発者の前川が述べるように，均衡コーパスではない。実はそれだけではなく，「話し言葉テキスト」コーパスではない。CSJはもともと古井の音声研究の一部として構築されたコーパスであり，音声研究のためのデータ収集が目的であり，この観点からは，考え抜かれた構成となっているが対話データは，48会話に過ぎず，またそのXMLには，テキスト要素を含まない空要素となっている。すなわち，CSJ本体は，音声ファイルであり，XMLデータはその“注釈”データであり，書き起こしたテキストは音声要素に対する属性として与えられている。

KOTONOHAはその名称の通り，1億語規模の日本語全体を代表する均衡コーパスの構築を目指している。しかし，均衡部分は，6,500万語に留まり，KOTONOHAでは，生産と受容とする部分は，LOBの出版目録を生産と呼び，Brownの蔵書目録を受容としており，BNCが主張するProductionとReceptionには対応せず，両方とも，Productionと見るべきである。また，サンプリング方法に関しても，BNCの精緻さには及ばないと思われる。

また，BNCは，電子化テキストの一般流通以前，すなわち，インターネット以前の試みであったが，KOTONOHAは現在のプロジェクトであり，テキストを取り巻く状況の変化を全く無視しているといえよう。

4. Google n-gram

現在入手可能な最大の言語データは，前述のようにGoogle n-gramである。2006年夏にGoogleが発表したWeb 1Tは，英語の1兆語のコーパスをもとに1-gramから5-gramのn-gramデータであり，一般に入手が可能である。2007年11月にGoogle Japanは，日本語のn-gramデータを公開し，一般に利用可能とした。日本語n-gramデータは，MeCabで処理した2,550億語のコーパスから得られる1-gramから7-gramのn-gramデータである。配布されているデータは，圧縮し，分割されたファイルであり，それぞれのファイルには，タブ区切りでn-gramデータとその頻度とが記録されている。日本語unigramの一部を次に示す。

Unigram	頻度
…	
芝野	15946
宥和	15946
ポスティングシステム	15945
コラボモデル	15945
…	

英語1兆語のWeb 1T及び2,550億語の大規模日本語n-gramは，それぞれ展開すると，100GBにのぼり，32ビットOSの限界に突き当たる。

インターネット及びWebの発明から20年が経過し，多くの研究機関でWebデータの利用が行われている。インターネットからの情報収集には，Crawler呼ばれるプログラムが利用される。ネット上や書籍で公開されているCrawlerでも10億語規模のデータ収集は簡単に行うことができる。しかし，日本有数の大学でもその規模は，100億語規模に留まる。これは，インターネット上のデータは，静的なHTMLファイルだけではなく，

(1) 動的に生成されるページが存在すること及び
(2) HTMLファイル以外にPDF，Word，Excel，PowerPointなど多くの形式のファイルが存在することなど

を挙げることができる。結果として，大規模なインターネットデータの収集には多くの人員と時間と費用とがかかる。

研究目的には，多くの場合，現実のビジネスで利用されているデータの利用は難しい。しかし，GoogleのWeb 1T及び日本語n-gramは，実際に世界最大で世界でもっとも力のあるインターネット企業であるGoogleがその年間2兆円のビジネスの根幹である検索サービスで用いるデータを一般に提供しているという意味でも画期的なデータである。

また，このデータは，BNCの1万倍であり，大学で収集できるデータの100倍の規模であり，現在入手可能な最大の言語データである。

5. 現代の均衡コーパスの要件

JIS X 2008-1997の改正では，NTT電話帳約6,000万件，全国地名約100万件(全国町・字ファイル)に加えて，2,500冊を超える高校までの検定済み教科書の印刷物から用例の確認を行い，追加漢字の選定を行った[13]。この検討の中で，かなりの教科書にフロッピーディスクの形態での電子化テキストを提供するものがあった。

この97JISを踏まえ，その後，現行学習指導要領に基づく教科書テキストを収集し，教科書コーパスの構築を行った[7]。また，Google大規模日本語n-gramのデータと合わせ，均衡コーパスの構築を試みた[6]。その後，形態素解析をGoogleの合わせMeCabに変更し，CCCJの構築を行っている。

2.で検討したように，現代のコーパス言語学は，Brownコーパスに始まり，BNCをもって一つの規範が成立した。しかし，BNCが開発された1990年代初頭は，データ爆発の直前であった。BrownからBNCまでのコーパスは，印刷媒体から入力すべき項目を選択し，データを入力することから始めなければならなかった。例えば，1語1円で入力できたとしても，1億語のコーパスの構築には，1億円の費用がかかる。

このデータ入手上の制約から，BNCまでのコーパスが印刷媒体，言い換えれば，仮想的な“図書館”

を母集団とした“図書館”言語を代表するコーパスとなったことは致し方ない，といえよう。

1990 年代以降のインターネットを中心とする ICT 技術の進歩は，事態を一変させた。具体的には，以前は伝票から一部のデータがパンチャーによって，入力されていた。しかし，現在では，ワイヤレス POS 端末から直接入力され，伝票は逆に，プリンタから出力される。現在では，まずコンピュータに情報が入力される。すなわち，既にコンピュータに蓄積されているデータを活用することが必須となる。同時に，これは，“図書館”を代表するコーパスからの脱却が可能となったことを意味する。

語学の“世界標準”であるヨーロッパ共通言語教育参照枠(Common European Framework of Reference for Languages: Learning, teaching, assessment , CEFR)[5]では，図書館に納められている言語情報だけではなく，町にあふれる看板やメニューなどを理解できるようにすることから，言語教育を始める。言語学の目的の一つが言語教育にあるとするならば，“図書館”コーパスではなく，コーパスの代表性をまず検討し，存在する電子化データをどのように収集し，コーパス化するかを検討する必要がある。

Brown コーパスの 100 万語で始まった現代コーパスは，BNC では 1 億語まで拡大し，十分なサイズに達したと，当初は思われた。しかし，句動詞の研究などのように，複数の単語を含む用例を検討しようとすると，1 億語が十分ではなかったことが明らかとなった。BNC プロジェクトは修了したが，コーパス構築を継続しているマンハイムのドイツ語研究所などでは，収録語数を増やすことを迫られた。ドイツ語研究所が開発したドイツ語コーパスであるマンハイムコーパスでは，コーパス規模の拡大のため，継続的に新聞の情報を追加し，コーパス規模は，30 億語に達したが，言語全体を代表する“均衡”コーパスから，“新聞”コーパスへと変容した。

Brown コーパスから BNC まで基本的には，Brown コーパスで設定された設計を引き継いだ。事実，Brown コーパスでは，メディアとドメインとに分け，この二つの視点を基本に均衡コーパスの設計を行った。メディアに関して，Brown は，books 55.2%，periodicals 20.4%，press 17.6%, others 6.8%とした，一方，BNC は，それぞれ書籍 57.2%，新聞雑誌 32.5%，その他 10.3%であり，ドメインとしては，両者とも informative 75%，imaginative 25%と全く同じ比率で構成されており，BNC は，Brown の均衡概念を踏襲したといえよう。また，Brown コーパスは現代米語，BNC は現代イギリス英語であるとし,そのため，両コーパスとも，母語話者による印刷物を基本とした。しかし，これは一つの言語全体を本当に代表する均衡コーパスの設計なのであろうか？　また，“母語話者”とは誰を指すのであろうか？　事実，既存のコーパスの多くは，“図書館”コーパスとはいえようが，“言語”を代表する均衡コーパスに関しては，再考が必要ではないか。

情報環境の変化に対応し，本来のコーパスはどうあるべきなのか？　同時に，Brown コーパス以来 40 年以上を経過し，その間のコーパス言語学での利用経験を踏まえて，あるべきコーパス像再検討することが必要である。

まず均衡コーパスの満たすべき案件を構築要件と利用要件の二つに分けて検討しよう。次の条件が考えられる。まず，構築要件としては，次の条件が考えられる。

(1) 入力ではなく，収集を基本とし，何を代表するのかを明確にした上で，コーパス構築を行う必要がある。
(2) (1)の代表性を基本とし，代表性を満たすための条件として，均衡を再考する必要がある。

利用要件としては，次の条件が考えられる。

(3) 統計学の前提には，確率論があることを考慮し，均衡を重み付きの確率の集計として考える。
(4) 言語学での用例検討及び確率検討のため，できる限り多くの用例を収集する。

また，検証のための要件としては，次の条件が考えられる。

(5) 科学的な再検証性を保証するために，ソースをすべて公開することが望ましいが，著作権などの問題がある場合でも，Web での一部用例検索だけではなく，一般に入手可能なソースを集積し，コーパスを構築する。

6. 共通教養日本語均衡コーパス(CCCJ)の基本設計

それでは，日本語コーパスとして，どのようなコーパスを構築する必要があるのであろうか。まず，母集団としては，日本語話者であるのは，当然であり，教養ある日本語話者の話し，書き，読み，理解できる共通部分に関する均衡コーパスの構築が必要である。すなわち，構築する均衡コーパスの母集団として，日本で学校教育を受け，日本で暮らす人を代表するコーパスの構築を検討する。このため，教科書と書籍，雑誌，インターネットを含むメディアからのインプットによって，現代日本人の言語能力が構築されていると仮定する。図 1 に CCCJ の基本設計を示す。

すなわち，この目的を達成するための共通教養日本語均衡コーパス(Balanced Corpus of Common

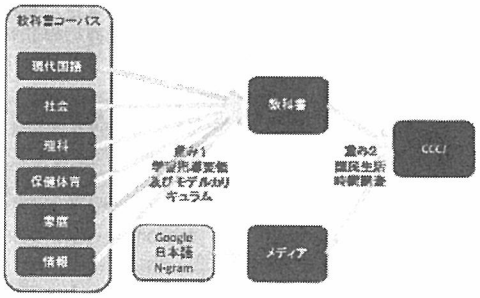

図 1 CCCJ の基本設計

Cultivated Japanese，CCCJ)の基本仮説は，言語習得理論の基本仮説であるインプット仮説[14]である。すなわち，言語能力は，インプットによって形成されるという仮説を基本とする。また，このインプットは，学校教育からのインプット及びインターネット，テレビ，新聞，雑誌，書籍などのメディアからのインプットの二つからなる，とする。

平成 21 年度学校基本調査速報[15]によれば，日本の高校進学率は 97.9%，大学進学率は 53.9%である。また，普通科高校には 72.3%の生徒が通い，大学生の過半数は，人文社会系で学ぶ。

大学における教育では，基本的には，それぞれの専門分野の教育を受けるのに対して，上記のようにほぼ日本人全員が進学する高校では，一部文化系，理科系と分かれる科目も存在するが，基本的には，同じ教養教育を受ける。また，大学・大学院進学者であっても，自分の大学・大学院での専門以外では高校までの教育で得た知識を基本とする。

これらのことから，CCCJ の推計では，高校卒業者でもっとも人数が多い普通科文系をモデル日本語話者とする。また，早稲田高校の履修例をもとに，科目毎の重みを算出する(表 1)。

表1 モデルカリキュラム

科目	単位	1年	2年	3年	コーパスデータ	重み
国語	21	5	7	9	国語	21
地歴・公民	18	4	7	7	地歴・公民	18
数学	14	5	4	5		–
理科	9	5	2	2	理科	9
英語	22	6	8	8		–
幻術	2	2	-	-		–
保健体育	9	4	3	2	保健体育	2
家庭	2	-	2	-	家庭	2
情報	2	2	-	-	情報	2
総合学習	3	1	1	1		–
合計	102	34	34	34	合計	54

一方，学校教育以外のインターネット，テレビ，新聞，雑誌，書籍からのメディアインプットと学校教育との重みについては，NHK 国民生活時間調査[17]のデータをもとに，重みを算出した。最新のデータである NHK 国民生活時間調査によれば，高校生の学校以外での学習時間は 1 時間 48 分，メディアに触れる時間は 3 時間 23.5 分であり，休日はそれぞれ 2 時間 31.5 分，4 時間 49.5 分である。また，登校日は，35 週であることから，学校教育関連の年間時間は 1932.25 時間，メディアに触れる時間は，年間で 1510.29 時間となる。これを教科書コーパスと Google n-gram の重みとした。

表2 教科書コーパス

	述べ語数	異なり語数	冊数	ページ数	単位	重み
国語	1,367,991	35,293	29	8,484	21	0.39
地歴/公民	7,839,012	56,879	75	18,571	18	0.33
理科	1,014,723	15,458	24	5,784	9	0.17
保健体育	118,250	6,266	2	306	2	0.04
家庭	1,242,876	20,210	19	3,915	2	0.04
情報	464,322	8,700	11	1,766	2	0.04
合計	12,047,174	79,172	160	38,826	54	1

また，教科書データとしては，本学で構築した現代日本語に関連する 6 教科，160 冊，1200 万語のコーパスをもととした。表 2 に教科書コーパスのデータを示す。

メディア関連情報としては，Google 大規模日本語 n-gram を用いた。

7. 共通教養日本語均衡コーパス(CCCJ)による語彙分析

日本語能力試験の 1 級では，約 8,000 語の語彙を基準として設定している。英語教育においても，最上位水準として，大学英語教育学会が設定した 8,000 語の JACET8000 があり，8,000 語水準が一つの目安と考えられる。

また，日本語能力試験は，来年度，大幅な改訂を予定しており，その際には，語彙リストについても，改訂されると思われる。

ここでは，CCCJ の分析として，語彙リストを取り上げ，分析を行う。

表3 CCCJの語彙分布

語彙	確率	累積確率	順位
の	5.33%	5.33%	1
。	3.21%	8.54%	2
に	2.94%	11.48%	3
を	2.89%	14.36%	4
、	2.76%	17.13%	5
は	2.51%	19.64%	6
て	2.18%	21.82%	7
が	2.13%	23.95%	8
た	2.01%	25.96%	9
,	1.78%	27.74%	10
ね	0.11%	50.05%	74
条	0.03%	60.02%	253
気体	0.01%	70.01%	795
始める	0.00%	80.00%	2222
建て	0.00%	85.00%	3810
切断	0.00%	90.00%	7164
分権	0.00%	92.21%	10000
カントリー	0.00%	95.00%	16665

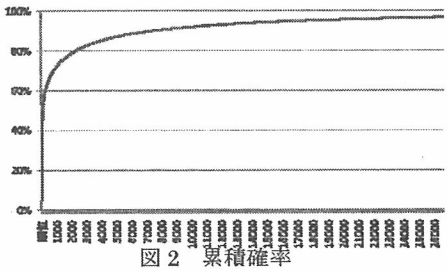

図 2 累積確率

CCCJ の語彙分布を図 2 及び表 3 に挙げる。

図 2 に示すように，語彙の累積確率分布は，8,000 語水準で飽和状態となる。表 3 に示すように，累積確率 90%の順位は，7,164 位であり，10,000 位での累積確率は 92.21%となる。このことから，日本語能力試験の 8,000 語水準は妥当であるといえよう。

表 4 に CCCJ の上位 20 語を示す。英語コーパスでは，Google 及び BNC の両方で最頻度語は，“the”であり，“of”，“and”，“to”が続き，句読点及び機能語が上位を占める。一方，“the”の確率は，BNC では約 6%，Google では約 3%と異なる。この異なりは，Google には，<p>以外の<li>や<table>などが多く含まれることによる[6]。日本語でも同様に，助詞，助動詞など

表4 CCCJの上位20語

形態素	出現確率	累積確率	順位	教科書順	Google	分野数
の	5.33%	5.33%	1	1	1	7
。	3.21%	8.54%	2	2	3	7
に	2.94%	11.48%	3	4	4	7
を	2.89%	14.36%	4	3	5	7
、	2.76%	17.13%	5	7	2	6
は	2.51%	19.64%	6	6	6	7
て	2.18%	21.82%	7	10	7	7
が	2.13%	23.95%	8	8	8	7
た	2.01%	25.96%	9	9	9	7
,	1.78%	27.74%	10	5	98	7
で	1.69%	29.43%	11	12	10	7
と	1.56%	30.99%	12	11	12	7
し	1.27%	32.26%	13	13	13	7
・	0.79%	33.04%	14	30	11	7
も	0.76%	33.81%	15	15	15	7
な	0.72%	34.53%	16	14	17	7
(	0.62%	35.15%	17	19	19	7
する	0.62%	35.76%	18	17	22	7
)	0.59%	36.35%	19	18	25	7
こと	0.50%	36.85%	20	21	33	7

の機能語及び句読点が上位を占め，英語と同様に傾向を示す。

CCCJ，教科書及び Google 順位で特徴的な点は，句読点“。”，“、”，“，”である。公用文の書き方によれば，縦書きでは，“、”と“。”を，横書きでは“，”と“。”を用いると定めている。しかし，一般には，この公的な正書法が広く認知されているわけではなく，“、”と“。”が広く利用されている。一方，教科書ではこの正書法に従い横書きである国語以外の教科では，“，”と“。”が用いられており，結果として，教科書コーパスでは，“、”より“，”が上位にくるだけではなく，ピリオドは，教科書では 5 位，Google では 98 位と大きく異なる。

表5 教科毎の語彙確率の相関

	国語	社会	理科	保健	家庭	情報
国語	100%	76%	70%	80%	72%	73%
社会	76%	100%	93%	96%	97%	95%
理科	70%	93%	100%	94%	94%	95%
保健	80%	96%	94%	100%	98%	95%
家庭	72%	97%	94%	98%	100%	96%
情報	73%	95%	95%	95%	96%	100%

縦書き，横書きに違いだけではなく，国語は，他の教科と大きく異なる。表 5 に教科毎の相関を示す。

国語以外の教科間での相関係数は，93～97%と高い数値を示すのに対し，国語との相関は，70～80%と低い数値を示す。これは，現国であっても国語で教えられているものが文学及び文芸評論であり，その上，現代日本語だけではなく，中島敦や芥川龍之介，森鴎外などの近代日本語を多く含むことによる。

品詞別の情報を見ると，一般科目での名詞が 20%台なのに対して，社会では 44%となる。これは，社会が暗記科目であるゆえんといえよう。

続いて，個別の語彙についてみてみよう。

まず，“です”“ます”について見てみよう。表 6 に分野ごとの“です”“ます”の順位を挙げる

表6 “です”“ます”

語彙	順位	Google	教科書	国語	社会	理科	保健	家庭	情報
ます	22	14	72	50	360	2812	17	293	524
です	31	16	70	40	745	4769	64	568	310

Google では，“です”“ます”の順位は 14 位，16 位と高いが教科書では，72 位と順位は一般的に低い。特に，理科では 2,817 位とかなり低く，一方，保健では 17 位，国語では 50 位となる。ここでも，国語と保健を除く他の教科との違いが大きい。保健で“です”“ます”が多く用いられる理由は，保健が内容的には，医学であり，医学者が専門家ではない高校生向けとして，患者に対するのと同様の言語使用を行っていることに起因すると思われる。

Google 語彙が教科書語彙より高い順位を示す語には，“挨拶”表現や“日常”表現がある。

表 7 に挨拶表現を挙げる。

表7 挨拶表現

語彙	順位	Google	教科書	国語	社会	理科	保健	家庭	情報
はじめまして	3720	1841	35112	21706	50093	/	/	/	/
ありがとう	516	251	5515	2309	30977	/	/	19673	/
ようこそ	1492	662	49181	34910	/	/	/	/	/
すみません	5223	2762	25783	14810	/	/	/	17169	/
ごめんなさい	5464	2828	49748	31252	/	/	/	/	/

これらの語は，理科系教科である理科，保健，

情報では用いられず，“ありがとう”を除き，教科書では，極めてまれにしか用いられない。日本語教育の最初期に教えられるこれらの語は，一方，Google では，上位に現れる。

これらの語は，理系教科では用いられず，日本語能力検定の 8,000 語水準でも“まじ”は，入らない。Google では，これらの語は，すべて 8,000 語水準に入るだけではなく，“すごい”については，1,000 語水準にも入る。当然ながら，これらの語も日本語教育では，“まじ”を除き，日常会話を考慮すると，初級で教えるべき語彙である。

表8 日常表現

語彙	順位	Google	教科書	国語	社会	理科	保健	家庭	情報
すごい	2006	999	8206	3809	29205	/	/	6676	/
かわいい	2363	1185	9957	5258	32632	/	/	3728	/
面白	2365	1141	16009	8957	24731	/	/	9015	/
かっこいい	6767	3651	34817	30080	45410	/	/	6141	/
まじ	9781	6321	16096	8349	39004	/	/	/	/

逆に，教科書コーパスで上位に位置するが，Google ではまれにしか用いられない語を見てみよう。

表9 命令形

語彙	順位	Google	教科書	国語	社会	理科	保健	家庭	情報
求めよ	1976	36166	1098	8564	12119	236	/	11998	/
述べよ	9151	47138	5782	33016	45655	1290	/	/	/
まとめよ	8055	44685	5042	16326	15130	1243	/	5196	/
答えよ	3989	35822	2392	9050	22559	559	/	12102	/
考えよ	4463	10750	2979	4011	4326	1187	/	1300	2671

表 9 に命令形，表 10 に論理表現を挙げる。

ここで検討した表現は，“図書館”コーパスでは，上位語に含まれないことが多い表現である。

表10 論理表現

語彙	順位	Google	教科書	国語	社会	理科	保健	家庭	情報
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
みずから	2957	20422	1749	12999	741	7898	##	3101	1878
ともなう	4314	21293	2641	18953	1482	3771	##	1813	2441
もとづき	7469	35628	4697	/	2437	7084	##	3747	3538
まとして	6411	12794	4451	6349	3042	3026	##	2649	/

8．あとがき

もともとの教科書コーパス[7]では，数学，英語，古典を現代日本語ではないとのことから除外したが，数学の問題文や英語の説明，古典の翻訳文には現代日本語が使われており，これらを取り入れる必要がある。また，[7]では ChaSen で形態素解析を行ったが，これを MeCab に変更し，Google と直接比較可能とした。

参考文献

[1] Alex Franz and Thorsten Brants, All Our N-gram are Belong to You (Google Web 1T 5-gram version), http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html, 2006

[2] Google. Web Japanese N-gram ver.1. (Available from 2007http://www.gsk.or.jp/catalog.html).,

[3]Manning, C. D. , Schuetze, H., Foundations of Statistical Natural Language Processing The MIT Press,1999

[4] BNC, http://www.natcorp.ox.ac.uk/

[5] Council of Europe, (2001), Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

[6] Machiko TAKEUCHI, Hiroshi SANO, Kohji SHIBANO, Evaluating CEFR Vocabulary against BNC and Google Web 1T, AAAL2009

[7] 佐野洋、于壮飛、芝野耕司，日本語教科書コーパスの構築，日本語教育国際研究大会，2009

[8] Yoshiko Muraki, Kaori Miyatake, Kohji Shibano, How should we build a word list for teaching Academic Japanese? - A straightforward approach, Eurocall 2009

[9] Kucera, H. and Francis, W. Computational Analysis of Present-Day English, Brown Univ. Pres., 1967

[10]Johansson, S., Leech, G., and Goodluck, H., Manualof Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computer, University of Oslo, 1978

[11]国立国語研究所，日本語話言葉コーパス，http://www.kokken.go.jp/katsudo/seika/corpus/

[12] 国立国語研究所，KOTONOHA http://www.kokken.go.jp/kotonoha/ex_2.html

[13]芝野耕司編著，増補改訂 JIS 漢字字典，日本規格協会，2002 年 5 月

[14] Krashen, S., The Input Hypothesis: Issues and implecations, Longman, 1985

[15] 平成 21 年度学校基本調査速報，http://www.mext.go.jp/b_menu/toukei/001/08121201/1282646.htm

[16]早稲田高校, http://www.waseda-h.ed.jp/、2009

[17] NHK 国民生活時間調査, http://www.nhk.or.jp/bunken/research/life/life_20060210.pdf, 2006

[18] 前川喜久雄，代表性を有する大規模日本語書き言葉コーパスの構築，人工知能学会誌，24 巻 5 号，pp 616-622，2009 年 9 月