Internal Use Only（非公開）

TR－SLT－0029

日本語の独話・対話コーパスの定量的比較分析
Quantitative analysis of Japanese different corpora

丸山 岳彦
MARUYAMA, Takehiko
柏岡 秀紀
KASHIOKA, Hideki
谷田 泰郎
TANIDA, Yasuo
熊野 正
KUMANO, Tadashi
田中 英輝
TANAKA, Hideki
Stephen Nightingale

2002年11月21日

概要

これまでに収集された複数の日本語独話・対話コーパスを対象として，数量的・言語的な側面から特徴分析を行なった．独話と対話，話しことばと書きことばの違いに関する定性的な分析と，各コーパスの量的構造に関する定量的な分析という二つの分析結果を示す．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

目次

表一覧

1 はじめに

ATRでは，これまで「バイリンガル旅行会話コーパス」という音声言語コーパス作成し，これに基づいて「対話」の音声翻訳システムの開発を進めてきた．これは「旅行会話」という限定された話題に関する対話の音声翻訳を行なうものであり，この範囲内では一定の成果を挙げてきた．このような音声翻訳技術の応用例として，「独話」の音声翻訳技術が挙げられる．独話の音声翻訳技術の例には，講演や会議の通訳や，ニュースなどの放送通訳，さらにニュース番組の字幕の自動生成などが挙げられ，幅広い応用範囲を想定することができる．

このような背景のもと，我々は，性格の異なる複数のバイリンガルコーパスを収集し，これらを統合的に利用する技術の検討を行なっている．現在までに，「独話」と「対話」，「書きことば」と「話しことば」など，異なる性格を持つコーパスが収集されている．これらの異種コーパス群を統合的に利用するためには，第一に，各コーパスが持つ言語的特徴を定量的に調査し，比較分析を行なう必要がある．そこで本稿では，これまでに収集された複数の日本語コーパスを対象として，今後の研究に有用であると思われるさまざまな数値を算出した結果と，それらをもとにして各コーパスの言語的特徴を定量的に比較・分析した結果について示す．

以下，2節では本稿で調査対象とするコーパスの概要について解説する．3節では，コーパスが備える基本的特徴である「独話」と「対話」，および「書きことば」と「話しことば」の対立ついて触れ，本稿で扱う各コーパスの性格づけを行なう．4節では，各コーパスの言語的特徴を，「文」「文字」「形態素」「文節」「節」など，さまざまな角度から定量的に比較・分析した結果について示す．

2 コーパスの概要

本稿で調査・比較分析の対象とするのは，以下の5種類のコーパスである．なお，これらの多くは日英バイリンガルコーパスであるが，今回は日本語のみを分析対象とする．

- NHK 解説番組『あすを読む』書き起こし （あすを読むと略記）
- NHK ニュース原稿 （**NHK** と略記）
- 日本経済新聞・日経産業新聞・日経流通新聞・日経金融新聞 記事コーパス （**日経新聞**と略記）
- バイリンガル旅行会話コーパス （**SLDB** と略記）
- 旅行会話基本表現集 （**BTEC** と略記）

以下では，各コーパスの概要・構成について述べる．

2.1 NHK 解説番組『あすを読む』

あすを読むは，NHK 総合テレビで平日 23 時 50 分から放映されている 10 分間のニュース解説番組「あすを読む」のエアチェックを行ない，その発話を人手で書き起こして作成したコーパスである．「あすを読む」は，時事・経済・社会問題などのテーマについて，1 人の解説委員が 10 分間の解説を行なう番組である．解説のための原稿はあらかじめ用意されているが，解説者は単にそれを読み上げるわけではなく，比較的自発的な発話が観察される点で，「話しことば」としての性格（3 節で述べる）を多く備えている．番組の収録は 1999 年 11 月から開始され，2002 年 6 月までに，327 番組分（約 54 時間）の書き起こしが完了している．

書き起こしデータは，漢字かな混じり文で表記されたテキストファイルであり，フィラー（ [] で示される）および 200ms 以上のポーズの切れ目のマーカ（/ で示される）が含まれる．また，書き起こし作業者の判断により，句読点が挿入されている．以下に例を示す．

今晩は。/
[え] 今の国会にかかっています預金保険法の改正案は来週にも成立する見通しです。/
[で] これが成立しますと、銀行が破綻した場合/
預金を全額保護する特例措置これが、二千二年の三月末まで続くことになります。/
[え] 新しい仕組みがどうなるのか、またこの間に日本の金融が再建を果たすことが... （中略）

あすを読むは，2002 年 9 月現在，以下の場所に格納されている[1]．

```
/DB/000/[230-233]/[00-99]/TXT/*.txt
```

[1] なお，/DB/000/[230-233]/[00-99]/{AUD,ENV,MOR,TRS,WIF}/ 以下は，それぞれ，音声ファイル，環境ファイル，形態素ファイル，音素転記ファイル，波形情報ファイルに対応する．中嶋ほか (2000) を参照．

2.2 NHK ニュース原稿

NHK は，NHK で放送されるニュース番組などで用いられる原稿のデータベースである．政治，経済，国際，社会，スポーツ，など多くのジャンルに区分されており，1995 年 3 月から 2000 年 4 月までのニュース原稿が収録されている．

NHK は，アナウンサーが読み上げるニュース記事の本文がその主体であり，一人の発話者が不特定多数の視聴者に向かって発話するという点で，独話の一種であると見てよい（3節で後述）．ただし，記事の中には中継記者とのやりとりが含まれる場合もあり，この点では部分的に対話が現れることになる．また，語句の振り仮名，日付，箇条書きであることを示す記号類（▼，▽など），アナウンサーへの注意書などの情報も含まれている．以下に例を示す[2]．

自由党の路線問題をめぐって、今夜、与党三党の党首会談行われました。会談の後、小渕総理大臣は、記者団に対し、「今後の連立政権の運営について基本的な考えで意見が一致せず、私と公明党の神崎代表は、自民党と公明党・改革クラブで、引き続き、連立政権を維持することで合意した」と述べ、自由党との連立を解消することを表明しました。

この映像をもとに現在の町の様子を社会部の○○記者とともにお伝えします。
アナ）○○さんきょうの噴火が起きた場所は洞爺湖温泉町のすぐ近くでしたね？
記者）この洞爺湖温泉地区の模型で説明します。洞爺湖温泉地区は有珠山の北側にあって...

神奈川県相模原市横山台の（よこやまだい）会社員、○○○○（○○○・○○○○）被告（三十七歳）で、去年二月二十六日、自宅のアパートで、当時九歳の小学校三年生の娘を殴ったりかかえあげてたたみにうちつけたりして死亡させたとして傷害致死の罪できのう（三月三十一日）起訴されました。

高速道路は▽道央自動車道の室蘭インターチェンジと長万部インターチェンジの間、国道は▽三十七号線の伊達市長和（ながわ）と豊浦町（とようらｔ）東雲町（しののめｔ）の間、▽四百五十三号線の伊達市長和（ながわ）と壮瞥町（そうべつＴ）久保内（くぼない）の間、▽二百三十号線の虻田町入江（いりえ）と洞爺村香川（かがわ）の間が通行止めになっています。

NHK は，2002 年 9 月現在，以下の場所に格納されている．

```
/data/D4L/data/NHK_News/J/j.xml
```

2.3 日本経済新聞・日経産業新聞・日経流通新聞・日経金融新聞

日経新聞は，日本経済新聞・日経産業新聞・日経流通新聞・日経金融新聞の新聞記事をデータベース化したものである．扱われている記事の範囲やジャンルはさまざまであり，一般的な報道記事から，インタビュー記事，社説，読者アンケート，図表などが含まれている．1995 年 1 月から 2000 年 12 月までの記事を収録しており，今回扱うコーパスの中では最もサイズが大きい．

[2] ○○ は人名を表す．

新聞記事の中には，記事本文以外にも，出稿元，語句の振り仮名，年齢や日付，箇条書きであることを示す記号類，表に含まれる数値などの情報も含まれている．以下に例を示す．

神奈川県茅ケ崎市東海岸南の海岸で、手製とみられる矢が背中に刺さった猫が十一月三十日夜、茅ケ崎署員らに保護された。市内の動物病院で矢を抜く治療を受け、比較的元気な様子。病院は回復を待って引き取り手を探すという。

【ニューヨーク３０日＝摂待卓】国連環境計画（ＵＮＥＰ）は三十日、産業廃棄物などの国境を超えた投棄を監視するバーゼル条約事務局長に国連本部法務部首席法務官の○○○○氏（５５）の起用を内定した。

当時、店内には閉店の準備をしていた従業員約二十人がいたが、ほかにけが人はなかった。
【図・写真】社長が撃たれたカジノクラブ（１日午前、名古屋市中区）

文　○○○（編集委員）　写真　○○○○（編集委員）

政府が決めた「行政改革大綱」の要旨は次の通り。（１面参照）
二〇〇五年までの間をめどとして行政改革を集中的・計画的に実施する。
一　行政の組織・制度の抜本改革
１　特殊法人等の改革
すべての特殊法人の事業と組織を抜本的に見直す。

東北主要企業の２０００年９月中間決算

	……２０００年９月中間期……			
	売　上　高	前年同期比	最終利益	前年同期比
＜金属＞				
東北特殊鋼	６，２８９	９．８	４８４	１７５．０
☆東洋刃物	３，８７１	１１．４	２４	—

日経新聞は，2002年9月現在，以下の場所に格納されている．

/data/D4L/data/Nikkei/J/txt[199501-200012].unl （本文）

/data/D4L/data/Nikkei/J/bib[199501-200012].unl （タイトル，各種情報）

2.4 バイリンガル旅行会話コーパス

「バイリンガル旅行会話コーパス」は，ATRが作成した音声言語データベースであり，「旅行会話」をテーマとした模擬会話を収録している．コーパス全体には，音声データベース，音声言語データベース，言語データベースが含まれるが，本稿の調査に用いるのは，計618会話から構成される「音声言語データベース」である．

「音声言語データベース」の書き起こしデータは，漢字かな混じりで表記されたテキストファイルであり，発話者（「通訳者：」「担当者：」などと示される），フィラー（［］で示される），言い誤り（（）で示される）および語句の振り仮名（⟨ ⟩で示される）が含まれる．また，書き起こし作業者の判断により，句読点が挿入されている．以下に書き起こしデータの例を示す．

通訳者：けさから頭がずきずきしてるんですよ。
担当者：［あのー］そういう頭痛⟨zutsuu⟩にはよく、なるんですか。
通訳者：ええ、頭痛⟨zutsuu⟩持ちですねえ、非アスピリン系の薬をいつも持ち歩いてるんですよ。でもちょっとなくなっちゃって。
担当者：［あのー］いつもの頭痛⟨zutsuu⟩とおんなじような症状ですか。
通訳者：いや、そうじゃないみたいなんです。今朝最後の一錠を飲んだんですけれども、いまだ変わらないんですよ。
担当者：［はー］［あのー］いつもの頭痛⟨zutsuu⟩と（比べ）比べてどんなところが違います。

「音声言語データベース」はバイリンガル会話として構成されているが，今回の調査では，日本語文のみを抜き出した「単一言語対応テキストファイル」を対象としている．

SLDB は，2002 年 9 月現在，以下の場所に格納されている．

```
/DB/SLDB/LNG/JTEXT/*.JTEXT
```

2.5 旅行会話基本表現集

BTEC は，ATR が，日本で市販されている旅行会話用の表現用例集を収集して書き起こし，作成したデータである．1 行 1 発話で記述されており，本 ID，発話 ID，テキストの三つの情報が，“\” で区切られて記述されている．以下に書き起こしデータの例を示す．

jpn001\00940\ 計算が違ってませんか。
jpn001\00950\ 日本円を扱っていますか。
jpn001\00960\ トラベラーズチェックを扱っていますか。
jpn001\00970\ ホテルのリストがありますか。
jpn001\00980\ 安くて清潔なホテルを教えて下さい。
jpn001\00990\ 駅に近いホテルを教えて下さい。

BTEC は，2002 年 9 月現在，以下の場所に格納されている．

```
/DB/phrasebook/travel/*.JTEXT
```

2.6 汎用テキストデータベースフォーマット

各コーパスの一次データに対して，以下のような加工を行ない，「汎用テキストデータベースフォーマット」とした．

- 各コーパスの文字コードを，日本語 EUC コードに変換した．
- **NHK**，**日経新聞**にあったいわゆる機種依存文字を，適切な文字（列）に置換した．
- 句点で改行し，1 行に 1 文が配置されるように整形した．ただし，括弧で囲まれた部分（直接引用部）に句点が含まれる場合，そこは改行の対象とはしていない．
- 各ドキュメント（1 番組，1 記事，1 会話，1 冊）の区切りは空行で示した．
- 原稿 ID や番組 ID，タイトルなどの情報を一次データから取り出し，各ドキュメントの先頭に # から始まる行として記述した．なお，これらの行は分析の対象から外した．

これらの加工を施した二次データ（「**汎用テキストデータベースフォーマット**」）は，2002 年 9 月現在，以下の場所に格納されている．

```
/data/D4L/data/ -
 - asu_wo_yomu/J/j.txt   NHK 解説番組『あすを読む』（あすを読む）
 - NHK_News/J/j.txt      NHK ニュース原稿（NHK）
 - Nikkei/J/j.txt        日経産業新聞（日経新聞）
 - SLDB/J/j.txt          バイリンガル旅行会話コーパス（SLDB）
 - phrasebook/J/j.txt    旅行会話基本表現集（BTEC）
```

この「汎用テキストデータベースフォーマット」を対象として，以下では分析を進める．

3 コーパスの特性に関わる基本的対立

具体的な調査・分析に入る前に，コーパスの特性に関する基本的な対立事項について述べておく．あるコーパスが備える基本的な特徴として，次のような対立を考えることができる．

- 「対話」と「独話」
- 「話しことば」と「書きことば」

以下，それぞれの対立に関して概観し，本稿で扱う各コーパスとの対応を示す．

3.1 「対話」と「独話」

「対話」と「独話」は，言語行動の様式として本質的に異なるものであり，語彙，文法，韻律など，多くの側面において差が認められる．国立国語研究所 (1961)『話しことばの文型 (1) — 対話資料による研究 —』の記述を引用する．

> 話しことばの態として，ふたり以上の言語主体のことばのやりとりという形で進行する対話と，ひとりの話し手が一方的に（多くの場合多数の）聞き手に向かってする話として進行する独話との区別が考えられる．両者は相異なる条件の上に成り立つ言語行動で，その発話の上に，かなりの差異をもつものである．表現意図の種類の出方に相違があり，また，構文についてもイントネーションについても，相違がある．また，傾向として，独話においては，表現は安定した整った構造を取り，対話においては，不安定な整わない構造を取る．対話の表現には，不整や誤用，また，それらに近接する形が少なくない．また，一般に，対話の表現は，場面に依存することが多く，音声的表出に依存することが多く，身体的表出に依存することが多い．不整・誤用も，そのことに関係することが多い．そのため，表現意図と形式との対応関係がとらえにくかったり，ひいて，文の切れ目がとらえにくかったりして，資料操作の上の困難も大きい．独話の表現は書きことばに近く，対話の表現は書きことばに遠いともいえる．(pp.16–17)

本稿では，「情報伝達の方向性」という点から，「対話」と「独話」の定義を次のように考えておく．

- **対話:** 言語を媒介として，ある特定の相手との間で，情報の伝達が双方向的に行なわれるもの
- **独話:** 言語を媒介として，ある特定の相手あるいは不特定多数の相手に対して，情報の伝達が一方向的に行なわれるもの

言語による情報伝達が，双方向的に行なわれるか，一方向でしか行なわれないかによって，「対話」と「独話」は区別されるわけである[3]．この場合の「言語」とは，後述する「話しことば」と「書きことば」の別を問わない．

[3] ただし，講義を行なっている最中に聴者から質問を受けて，それに応える場合などは，「独話」に「対話」が部分的に挿入されることになる．両者の境界は，厳密に区分され得るものとは言い難い．

「対話」

「対話」は，話し手と聞き手（書き手と読み手）との間で，ある話題に関する情報を相互に伝達すること（あるいは情報量を調整すること）をその目的とする．対話の進行にしたがって，話し手と聞き手双方が持つ情報量が動的に更新・調整されていくところが，「独話」と大きく異なる点である．「対話」の持つ特性について，片桐 (1997)「終助詞とイントネーション」の記述を引用する．

> 対話は次の二つの特徴を備える．
>
> - 動的変化
>
> 対話の参加者は対話の進行に関してあらかじめ確実な予測をすることはできない．対話参加者は対話の進行途上で起こる変化に対してその場で臨機応変に対応する必要がある．
>
> - 情報共有
>
> 対話が有効に進行するためには，常に対話参加者間の情報共有状態を保持する必要がある．対話参加者のうちの誰かが発話の内容情報の獲得・理解に失敗したまま放置されたならば，その参加者は実質的に対話から外れたことになる．
>
> 動的に変化し不確実な情報しか得られない環境下で情報の共有を実現するために，対話にはあいづち・問い返し・言い直しなど対話に特有な現象が現れる．(p.242)

当該の話題に関する情報量が対話者間で不均一であってはならないため，対話者は相手と自分の情報量の差を常にモニタしながら，その不均一な部分を埋めていく作業を行なう．例えば，ホテルの予約に関する次の対話例では，フロント係と申込者との間で，部屋の値段やベッド搬入の可否について情報の伝達が行なわれている．対話の進行にしたがって，相手の情報を獲得したことを示す「はい」や「そうですか」などの「あいづち」表現が現れている．

> **A:** ツインルームですと、一泊一万五千円になりますが、いかがでしょうか。
>
> **B:** そうですか。今、泊まってますシングルにエキストラベッドを入れてもらうことはできますか。
>
> **A:** はい、できますが、お値段のほうはツインルームと変わりませんが、よろしいですか。
>
> **B:** そうですか。そうしましたら、ツインのほうに替わります。 **(SLDB)**

料理のコース名と値段に関する次の対話例では，相手が直前に発話した内容の獲得に失敗したため，「問い直し」が行なわれている．

A: まず、ベジタリアンコースのお値段が、お一人一万六千円です。肉料理のエーコースが、二万円、ビーコースが、一万五千円となっております。

B: すいませんが、お肉料理、の方の、コースの値段をもう一度教えてもらえますか。

A: はい、申し上げます。お肉料理、エーコースが、二万円、ビーコースが一万五千円となっております。 **(SLDB)**

さらに次の例では，A が提示した情報（「御所」）を B が理解できなかったため，その情報を共有するための調整作業が行なわれている．

A: それでしたら、御所の周りはいかがでしょうか。御所の周辺はおよそ六キロでして緑も多くって、ジョギング、さぞかし楽しんでいただけると思います。

B: 御所ですか、御所ってどんなところなんですか。

A: ええ、御所というのは昔、天皇の住まいした由緒ある場所なんです。 **(SLDB)**

これらはいずれも，対話者間の情報量を調整・均一化するために，話し手と聞き手が情報量を調整している例として考えることができる．

典型的な「対話」は，**SLDB** のような旅行会話や，電話での会話，日常の雑談など，場面に応じて即興的に行なわれる「話しことば」（後述）であると考えられる．先の片桐 (1997) の「対話にはあいづち・問い返し・言い直しなど対話に特有な現象が現れる」という記述も，基本的には「話しことば」を想定したものであろう．この一方で，筆談のような「書きことば」（後述）による「対話」も存在する．「話し手と聞き手との間である話題に関する情報を相互に伝達すること（あるいは情報量を調整すること）をその目的とする」という定義に従えば，電子メールの交換による「対話」や，文通，往復書簡，誌上対談なども，この延長線上に連続的に位置づけられる[4]．ただし，対話者間で情報が交換される時間の幅が長くなるほど発話の即興性が失われ，対話者は自分の考えを整理して述べることができるため，一回の情報発信はむしろ「独話」に近づいていくことになる．

「独話」

「独話」では，情報が話し手から聞き手に対して一方通行的に伝達されるため，「対話」に見られるような聞き手によるあいづち・問い返し・言い直しなどの現象や，話し手と聞き手の間で情報量を調整するための共同行為などは成立しない．ゆえに，聞き手が話し手の発話内容の理解に失敗した場合でも，そのまま放置されるしかない．

[4] 「電子掲示板」などで多人数（不特定多数）が意見を交換し合う場合なども，書きことばによる対話の一つと見てよいかもしれない．

この問題を回避するために，話し手は，自ら発話した内容について聞き手に配慮し，時には自らの発言内容に解説を加える場合がある．

> コンピューターを利用しましたインターネットという情報ネットワークが私達の経済ですとか社会の仕組みを今大きくかえ始めています。そこで今晩は経済の動きを中心にしましてアメリカの例もみながら今後の動きを探ってみようと思います。
> なじみのない方のために手短にインターネットについて説明しておきましょう。こちらにあります。インターネットで言いますのは情報のネットワークなんです。言いかえると情報のプールと言ってもいいんじゃないでしょうか。... (あすを読む)

> 生命科学の一つの分野であるヒトゲノムの解読が国際的な政治の舞台の議題として扱われようとしています。そこでヒトゲノムの国際正理について今夜は考えてみたいと思います。
> このヒトゲノムというのは人間の遺伝情報全体を表す言葉です。三十億の遺伝文字ディエヌエーという物質でできています。この三十億のディエヌエーの並びを解読し生命の設計図を明らかにしようという研究が進められています。... (あすを読む)

それぞれの例の2段落目以降は，自らの発話した内容（「インターネット」「ヒトゲノム」）に関する解説部分となっている．「対話」のように聞き手の理解度をモニタすることができない「独話」の場合，話し手は，聞き手が発話内容を正確に獲得・理解できるように配慮しなければならない．

このように，情報伝達の方向性，および情報量の調整作業の有無という点に関して，「対話」と「独話」は大きく異なる．このような性格の違いは，「対話」と「独話」に出現する言語表現の違いとなって現れる．例えば，「対話」には，対話者間の情報量の調整・確認行為に関わる言語表現（終助詞，間投助詞，指示詞，あいづち表現，応答表現）が特徴的に現れる．

> かしこまりました。どのようなタイプのお部屋をご希望でしょうか。
> はい、シングルルームでございますね。
> それより安い部屋はないですね。 (SLDB)

「独話」にも，次の例のように聞き手に問いかけるような表現が観察されることがあるが，

> 都市ガスや水道の耐震対策はどこまで進んだのでしょうか。電気は大丈夫でしょうか。電話はどうですか。保育園や学校老人福祉施設や病院などの耐震対策はどうでしょうか。それぞれ対策が施されたのですが ... (あすを読む)

このような問いは聞き手（この場合は視聴者）からの直接的な応答を前提としたものではなく，一方通行的な問いかけでしかないため，「対話」で行なわれるような対話者間の情報量の調整・確認行為と考えることはできない．

また，韻律形式の出現傾向に関しても「対話」と「独話」には違いが認められる．「対話」では，相手に何かを尋ねる場合に

通訳を探してるんですけれども、手配はお願いできますか(↗)。 **(SLDB)**

のような「上昇調イントネーション」が多用されるが，聞き手側からの応答や情報伝達を前提としない「独話」では，「上昇調イントネーション」は非常に現れにくい．

これらの違いは，すべて，「対話」と「独話」の情報伝達様式の違いに起因するものである．

コーパスとの対応

本稿で扱うコーパス群では，旅行場面における会話例を集めた**SLDB**と**BTEC**が「対話」に相当する．一方，**あすを読む**や**NHK**，**日経新聞**は，不特定多数の相手（テレビ視聴者，読者）に向けられた伝達様式を持つものであり，「独話」として分類される．**あすを読む**には，先述したように，視聴者に向けて問いかけるような発話が観察されることがあるが，これは視聴者からの応答を前提とした発話ではないため，「独話」の一部であると見てよい．一方，**NHK**にはアナウンサーと記者とのやり取りが挿入されることがあり，この場合は部分的に「対話」が含まれることになる．**日経新聞**に含まれる社説やエッセイ風の記事では，読者への問いかけのような表現が観察されることがあるが，先の**あすを読む**と同様，応答を前提とした発話ではないため，「独話」の一部であると見てよい[5]．一方，インタビュー記事などには，インタビューの中で成立する「対話」が含まれることになる．

3.2 「話しことば」と「書きことば」

「話しことば」と「書きことば」の違いは，典型的には，その言語表現が音声言語として実現されたものか，文字言語として実現されたものか，という違いに帰すことができる．

- **話しことば:** 音声言語として実現された言語．
- **書きことば:** 文字言語として実現された言語．

「話しことば」と「書きことば」の種類について，丸山 (1996)「話しことばの諸相」の分類を引用する．

[5] 新聞記事の中で「皆さまのご意見をお寄せください」という文面を見て，読者が新聞社に意見を送ったとしても，それを「対話」と見なすことはできない．

話しことば関係　（音声による伝達）

講義・講演（独話の部分と質疑応答，読み上げか否か，OHP を使った発表）/ ゼミナールでの報告 / 式辞・祝辞 / 会議（独話と議論，大きい会議，小さい会議）/ 座談 / 窓口などでの事務的な話（銀行・郵便局・デパート・レストラン etc.）/ 日常会話一般（あいさつ，用談，雑談，さしず，けんか，感情・感覚の直接的表現など）/ 電話での会話（留守番電話，電話による情報サービス，電話会議）/ テレビ会議 / ラジオ・テレビ（ニュース・天気予報，料理番組での説明・しらせ，インタビュー・トーク番組，ドラマ）/ 駅・車内・店内などでのアナウンス / 広告放送

書きことば関係　（文字による伝達）

事務書類 / 説明書・カタログの類 / 各種広告 / 新聞雑誌の記事・論説・論文 / 小説（地の文と会話文）/ 随筆 / 日記 / メモ・ノートの類 / 手紙 / 通知 / 掲示 / 若者雑誌の文章・投稿記事 / ジュニア小説 / マンガ / 台本（台詞）/ 映画の字幕 / 文字放送 / ファクシミリ / 電子メール / パソコン通信のチャット / 電子掲示板 / 電子会議 / ポケベルでの文字会話 / 喫茶店，観光地，各種記念館におかれた感想ノート / 筆談 / 筆談通信

(pp.41–42)

「話しことば」と「書きことば」は，言語が表現される媒体によって明確に区別することができる．しかしその内容・成立に関しては，本来「話しことば」であったものが文字化されている「書きことば」（インタビュー記事など）から，本来「書きことば」であったものが正確に音読される「話しことば」（本の朗読など）まで，連続的に幅が認められる．以下では，「即興性 / 準備性」という観点から，「話しことば」と「書きことば」の違いについて整理する．

「話しことば」

畠 (1987)「話しことばの特徴 — 冗長性をめぐって —」は，「話しことば」の類型を，次の四つに分類している．

第 I 類　発話意図の形成から発話までにほとんど時間的経過のないもの
（おしゃべり）

第 II 類　予めトピックが一応決まっていて伝達内容も多少は準備されているが，言語化そのものは即興で行なわれるもの
（相談，打ち合わせ，連絡，忠告，警告，説明，脅迫）

第 III 類　かなり計画的で，時間をかけた発話であり，伝達内容は発話に先だって十分に時間をかけて決定され，訂正，追加もなされている．また言語化自体もある程度は準備されている．
（講演，会議での報告，ゼミナールでの報告，授業における教師の発言，テレビ料理番組での説明）

第 IV 類　発話の即興性がほとんどない．発話内容の決定，訂正に時間をかけただけでなく，言語化も事前に決定され，言語化の事前訂正にも時間をかけた発話．
（ニュース，大会宣言，送辞，答辞，テレビドラマ）

第 I 類から第 IV 類に進むにしたがって，発話の即興性が減少し，逆に準備性が増していくことになる．「話しことば」というと，典型的には，言い淀みや言い誤り，倒置や省略などを含む，日

常的な会話のような即興性の高い（準備性の低い）ものが想起されやすい．しかし，音声的に実現される言語を「話しことば」とする以上，即興性の低い（準備性の高い）「話しことば」も認める必要がある．例えば，ニュース報道では，アナウンサーの発話は原稿の内容，およびニュースの発話スタイルによって制限され，話し手が自由に（即興的に）発話することは許されない[6]．この点で，ニュース報道は極めて即興性の低い話しことばであるといえる．

永野 (1988)「話しことばと書きことば」は「話しことば」の特徴として，次のような点を挙げているが，

> 一つ一つの文が比較的短い．
> 同じ文や語を繰り返すことが多い．
> 動詞の連用形で文を中止することはほとんどない．
> 修飾語を用いることは比較的少ない．
> 指示語を用いることが多い．
> 終助詞が好んで用いられる．
> 漢語の用いられることが少ない．
> 文の成分の順序が正常でない場合がある．
> 言い差しの文で終わることが多い．
> 補充法を用いることが多い．
> 文の成分の一部を省略することがある．
> 敬語がよく用いられる．
> 間投詞・間投助詞が好んで用いられる．
> 文語的漢文的翻訳的な要素は少ない．
>
> (p.679)

これらは主として即興性の高い「話しことば」の特徴を捉えたものと考えることができよう．

また，「話しことば」そのものは音声であるため，そこにはアクセントやイントネーション，ポーズなどの韻律情報が含まれる．これらは，言語表現が発声行為として具現化される性質によるものであり，書記行為によって具現化される「書きことば」と大きく異なる特性である．

「書きことば」

「書きことば」は，書記という活動を経た言語表現であるため，「話しことば」と異なり，即興性という性格は極めて成立しにくい．新聞記事や小説など，典型的な「書きことば」は，事前の検討や編集，校正などの過程を経て具現化されるものであり，準備性が極めて高い．その一方で，例えば，速記や筆談，電子掲示板への書き込みやチャットなど，情報が短時間のうちに発信されることが求められる場面では，ある程度の即興性を持つ「書きことば」が現れる．また，口語調で書かれたエッセイや，インタビューを正確に書き起こした文章などは，「話しことば」をそのまま書記した「書きことば」と考えてよい．

言語表現の音声的な実現形式である「話しことば」には，韻律情報（アクセントやイントネーション，ポーズなど）が含まれるが，文字として実現される「書きことば」には韻律情報が含まれることはない．一方，「話しことば」に備わっていない「書きことば」の特性として，句読点をはじめとした各種記号類の存在を挙げることができる．

[6] NHK 以外のニュース番組では，これに当てはまらない例も多く見受けられるようである．

いわゆる「レーガノミクス」に戻ることになる？
「歳出削減と減税を柱とする点で、確かにレーガノミクスの目指した『小さな政府』への再挑戦だ。レーガン政権時代は金融引き締めと景気後退で、...　(日経新聞)

「話しことば」では，疑問文は主として上昇調イントネーションによって表示される．一方，韻律情報を持たない「書きことば」では，「？」という記号を用いることによって，その文が疑問文であることを示すことができる．また，「『」「』」という括弧記号は，その内部が引用された表現であったり，特殊な用語であったりすることを示す．句読点「、」「。」は，そこが文の構造的，音韻的な切れ目であることを表示する．

その上で、野中幹事長代理は、「私は、小渕総理大臣は、最大の仕事であるサミットの議長国としての責任を果たしてほしい」と述べ、衆議院の解散・総選挙の時期は、サミット後が望ましいという考えを改めて示しました。

さらに，**日経新聞**では，次のような記号類が用いられている．

○○○○・東京理科大教授（５３）の書斎は、ちょっとしたオフィスの雰囲気だ。最新型のパソコンやワークステーション（ＷＳ）がズラリと並ぶ。自宅と大学の研究室をＩＳＤＮ（総合デジタル通信網）回線で接続、海外の研究者との電子メールのやり取りや、学生の指導などに使う。

Q 最近、話題の懸賞金付き定期預金って何ですか？
A 一定額のおカネを預け入れると懸賞金の当たる抽選権がもらえる定期預金のことです。...

…短期金利は為替離れ？
円相場の動向がユーロ円金利先物以外の短期金利商品には影響しにくくなっている。...

ちょっと横道にそれたんですね（笑）。京都の方でよく遊んだりもしまして...

「（５３）」は年齢を表すものであり，「ワークステーション（ＷＳ）」「ＩＳＤＮ（総合デジタル通信網）」などの括弧書きは，略称や正式名称を表している．「Q」「A」などは，その言語表現の発話者を表すものであり，ある種の談話情報を表示するものである．また「短期金利は為替離れ？」は記事の表題として表されたものであり，やはり談話構造の表示に関わるものである．さらに「（笑）」などは，笑い声という非言語的情報を記号化したものである．これらの記号類は，視覚の助けを借りて補足的に意味表示を行なうものであり，「話しことば」には一切存在しない要素である．

また，**日経新聞**には図や表などが含まれるが，これらはそもそも「図」であり，「書きことば」の範疇に入れるべきものではない．

コーパスとの対応

本稿で扱うコーパス群のうち，「話しことば」として分類されるのは，実際に人間が発声した音声を収録して書き起こしたあすを読むと**SLDB**である．ただし，あすを読むは極めてよく準備された原稿をもとに発話したものであり，先の畠 (1987) の分類では「第IV類」に相当する．**SLDB**は，予め設定されたある課題を遂行するための会話例を収録したものであるが，発話そのものは即興的に行なわれているため，「第II類」に相当する．ただし，場面や発話のタイプがある程度決まっていることから，言語化自体がある程度準備されている「第III類」として考えてよいかもしれない．いずれにしても，即興的な性格を備えた話しことばではない．

一方，「書きことば」として分類されるのは，新聞記事を収録した日経新聞である．記号類の多用や表題の存在など，視覚的な情報伝達を重視した「書きことば」特有の現象が多く観察される．

NHKと**BTEC**は，実際には読み原稿や本として書かれたものであり，原則的には「書きことば」として分類される．韻律情報や，言い淀み・言い誤りなどの現象が存在しない点でも「書きことば」であるが，そもそも音声言語による伝達（「話しことば」）として実現することを前提としている点で，典型的な書きことばである日経新聞よりは話しことばに近い性格を持っている．

3.3 調査対象コーパスとの対応

以上，あるコーパスが備える基本的な特性に関して，「対話」と「独話」，「話しことば」と「書きことば」という基本的な対立について述べた．本稿で調査の対象とするコーパスと，各特性の対応について，以下に示す．

	対話 / 独話	話しことば / 書きことば
あすを読む	独話	話しことば
NHK	独話	書きことば（話しことば）
日経新聞	独話	書きことば
SLDB	対話	話しことば
BTEC	対話	書きことば（話しことば）

表 1: コーパスの特性に関わる基本的対立

以上，コーパスの特性に関する基本的な対立である，「対話」と「独話」，および「話しことば」と「書きことば」の違いについて見た．

4 日本語の独話・対話コーパスの定量的比較分析

以下では，各コーパスの言語的・数量的な特徴について比較・分析を行なった結果を示す．まず最初に，言語活動の基本的な単位である「文」の数量的特徴について示し，以下，「文字」「形態素」「文節」「節」という観点から分析を行なう．

4.1 文

(4.1.1) 文数の比較

まず始めに，各コーパスの総文数と異なり文数，異なり率，1記事/1文当たりの文数を求めた結果を，表2に示す．

	あすを読む	NHK	日経新聞	SLDB	BTEC
総文数 (1)	19,828	1,676,098	16,529,022	21,769	174,257
総文数 (2)	19,828	1,593,442	18,619,709	21,769	174,242
異なり文数	19,389	1,442,152	17,156,206	15,433	106,101
異なり率	97.78%	90.51%	92.14%	70.89%	60.89%
全記事数	327	287,315	1,758,960	618	—
平均文数 / 記事	60.64	5.55	10.59	35.22	—

表 2: 総文数と異なり文数

表2の「総文数 (1)」は，コーパス中に現れる句点「。」の数をカウントしたものである．ここでは，引用符（「〜」）に句点が入っている次のような例は2文としてカウントされている．

> 自由党の小沢党首は、「通常国会の会期は残り二ヵ月となったが、安全保障政策などの政権合意はほとんど実現していない。 小渕総理大臣の実現に向けた決意について、タイムスケジュールを含めて聞きたい」と述べました。 **(NHK)**

また，書き起こしテキスト中，1行に2文が入っている次のような「連文」も，「総文数 (1)」では2文としてカウントされている．

> ありがとう元気です。そしてあなたはいかがですか。 **(BTEC)**

上のような，引用符内部に含まれる複数の文や「連文」をまとめて1文としてカウントすると，「総文数 (2)」のようになる．**NHK** の数値が大きく減少していることから，引用符内部に複数の文を含むような長い引用表現が，**NHK** で特徴的に使われていることが分かる．一方，**日経新聞**では逆に数値が増加しているが，これは図表の1行が1文としてカウントされていることが影響している（2.3 節参照）．図表などを除去するクリーニングを行なうことにより，より正確な数

値を求めることができると考えられるが，ここでは措く．あすを読むでは数値の変動が全くないが，これはあすを読むの書き起こしテキストに引用符が用いられていないという事情による．実際には，上記のNHKの例のように複数の文を含む長い引用表現が出現することがあるが，ここではそれらを1文として扱う手段がないため，2文としてのみ扱う．さらに，SLDBやBTECのような対話では，複数の文を含む長い引用表現は観察されない．上で見たような「連文」が若干観察されるだけである．

以下では，特に断りのない限り，「総文数 (2)」の数値を総文数の値として用いる．

SLDBやBTECでは1文が比較的短く，定型的な表現や言い回しも多いため，同じ形の文が重複することがある．これは，各コーパスの「異なり文数」，および，異なり文数を総文数で割った「異なり率」を見ることによって確かめることができる．あすを読む，NHK，日経新聞の異なり率がいずれも90%を超えているのに対し，SLDBとBTECの異なり率はそれぞれ70%，60%であり，全体の3～4割は同じ文が重複していることになる．頻出する表現の上位のものを，出現頻度とともに，表3に示す．

SLDB	BTEC
かしこまりました。(543)	ありがとう。(790)
分かりました。(527)	はい。(722)
そうですか。(397)	すみません。(673)
はい。(301)	わかりました。(544)
はい、かしこまりました。(166)	かしこまりました。(388)
はい、分かりました。(151)	こんにちは。(296)
ありがとうございます。(145)	いくらですか。(284)
ありがとうございました。(142)	どういたしまして。(269)
はい、そうです。(141)	いらっしゃいませ。(246)
どうもありがとうございました。(139)	いいですよ。(229)

表3: 頻出する文 (SLDB，BTEC)

また，表2の「全記事数」とは，あすを読む全体の番組数，NHK全体のニュース記事数，日経新聞全体の新聞記事数[7]，SLDB全体の会話数に相当する．なお，BTECは「記事」に相当する単位が存在しないため，ここでは省略する．1つの話題について10分間話し続けるあすを読むと，1つのニュース記事を端的に報道するNHKとでは，当然のことながら，1記事当たりに含まれる文数にかなりの開きが見られる．

さて，先に，引用表現が複数の文を含むことによって1文が長くなるという傾向が特にNHKにおいて顕著に見られると述べた．長い引用表現とは，例えば次のようなものである．

[7] NHKと日経新聞で，記事本文が存在しないタイトルのみの記事は，記事数のカウントに含めていない．

会談の後、小渕総理大臣は、記者団に対して、「党首会談で、今後の連立政権の運営について、基本的な考えで意見が一致せず、三党の間の信頼関係の維持が困難になった。よって、私と神崎代表は、今後、自民党と公明党・改革クラブで引き続き、連立政権を維持し、国民の期待に応えるということで合意した」と述べ、自由党との連立を解消することを表明しました。

ただ、公明党の坂口政策審議会長が、「公明党は、チェックオフについて、法律で規制することには慎重に対応すべきだと考えている。仮に三党の主張が折り合わなかった場合にはどうするのか」と尋ねたのに対し、亀井氏は、「話し合いが付かなかった場合には、自民党単独で法案を提出することも検討したい。そうしたことも含めて三党で話し合いたい」と述べ、調整がつかない場合には、自民党単独で法案を国会に提出したいという考えを示しました。 **(NHK)**

また，引用部分に更に引用部分が含まれる場合もある．

自由党の小沢党首は、党本部で記者団に対し、「小渕総理大臣の入院を昨夜遅く聞いて、本当にびっくりした。三十年来の友人であり、先輩でもあり、それぞれの立場は違うが、元気になって、また頑張ってもらいたい。おとといの自民・自由・公明の三党の党首会談の半分以上は二人で話をしたが、小渕総理大臣は、連立の政策合意について、『あなたと考え方も一緒だが、自民党は大きな党で、自分の思いだけではできない』と話したり、三十年間の思い出話もした。お互いに本心をさらけ出しただけに、本当に心配している」と述べました。 **(NHK)**

直接引用のかわりに，「～という意見」「～として」などの形式で他者の発言を引用することもある．この場合も，次のように，１文が長くなる傾向にある．

その背景には、森幹事長が、小渕政権の中で、景気対策や自民・自由・公明の三党の連立政権の発足などで、党内のとりまとめにあたった実績から、小渕総理大臣の路線を引き継ぐ後継の総理大臣にふさわしいという意見が小渕派内などに出ているのに加えて、これまで、党執行部に一定の距離を置いてきた、加藤派と山崎派にも、早期に挙党体制を確立すべきだとして、森幹事長の総理大臣就任を受け入れる意見が出ていること、そして、与党の公明党が、公明党との連立に積極的な対応をしてきた人が望ましいとしていることなどが挙げられます。 **(NHK)**

さらに次のような「箇条書き」が含まれる場合にも，１文が長文化する．

オリンピックの代表に決まったのは▼百キロ級が去年の世界選手権で金メダルを獲得した二十一才の井上康生（いのうえ・こうせい）選手、▼九十キロ級がバルセロナオリンピックの金メダリストで去年の世界選手権でも優勝した三十才のベテラン・吉田秀彦（よしだ・ひでひこ）選手、▼六十キロ級が前回アトランタオリンピックの金メダリスト・野村忠宏（のむら・ただひろ）選手で、いずれもきょう福岡市で開かれた全日本選抜体重別選手権のそれぞれの階級の優勝者です。 **(NHK)**

また、北海道が維持管理している道路で通行止めになっているのは▽洞爺湖登別（のぼりべつ）線の虻田町洞爺湖温泉から壮瞥町滝之町（たきのまち）まで▽洞爺湖公園線の壮瞥町壮瞥温泉から伊達市西関内町まで（せきないT）▽洞爺虻田線の虻田町月浦から（つきうら）洞爺村洞爺町までの全線（とうやｔ）▽上長和萩原線の（かみながわ・はぎはら）伊達市西関内町から伊達市東関内町までそれに▽洞爺公園洞爺線の洞爺村洞爺町から壮瞥町滝之町までの（たきのまち）区間です。 **(NHK)**

(4.1.2) 文長の比較

次に，1文当たりの平均文字数，平均形態素数，平均文節数，平均節数を求めた結果を，表4に示す．なお，文長の測定するためには各コーパスに含まれる文字数，形態素数，文節数，節数について検討する必要があるが，これらについての詳細は，次小節以降で示す．

	あすを読む	NHK	日経新聞	SLDB	BTEC
平均文字数 / 文	49.80	78.27	45.16	21.07	13.46
平均形態素数 / 文	29.14	47.46	26.83	11.72	7.87
平均文節数 / 文	11.24	17.07	8.92	3.90	2.71
平均節数 / 文	4.68	6.01	3.26	2.05	1.50

表 4: 1文あたりの長さ

1文あたりの長さについては，**NHK**がもっとも長く，あすを読む日経新聞がそれに続く．一方の対話である**SLDB**，**BTEC**の文長は，極めて短い．これは，先に挙げた永野 (1988) が述べるように，1文が短い，文の成分の一部が省略される，言いさしの文で終わることが多い，などという対話の特性が現れているものといえる．

4.2 文字

(4.2.1) 文字数の比較

次に，各コーパスの総文字数と異なり文字数，異なり率，1記事 /1文当たりの平均文字数を求めた結果を，表5に示す．

コーパスサイズとしては，日経新聞が最も大きい．文字数だけで言えば，日経新聞は，最も小さい**SLDB**の実に1,832倍のサイズとなっている．

異なり文字数を総文字数で割った「異なり率」を比較することによって，各コーパスの総体に比べてどれだけ多様な文字が使われているかを比較することができる．例えば，**NHK**と日経新聞の異なり率は他に比べて非常に小さいが，これは（当然のことながら）両者がそのコーパスサイズに比べて少ない範囲の文字から構成されていることを示している．

	あすを読む	NHK	日経新聞	SLDB	BTEC
総文字数	987,554	124,716,496	840,781,729	458,777	2,344,795
異なり文字数	2,319	4,276	5,740	1,306	2,148
異なり率	0.23%	0.0034%	0.0007%	0.28%	0.09%
平均文字数 / 記事	3020.0	434.1	478.0	742.4	—
平均文字数 / 文	49.80	78.27	45.16	21.07	13.46

表 5: 総文字数と異なり文字数，平均文字数

また，「平均文字数 / 文」を見ることによって，1 文の長さを文字数の側面から比較することができる．ここでは **NHK** が 78 文字と極めて長く，続いて**あすを読む**と**日経新聞**が 49 文字，45 文字と同程度の長さであり，さらに **SLDB** と **BTEC** が 21 文字，13 文字と極めて短い，という分布が見て取れる．

(4.2.2) 文字種の比較

次に，各コーパスの文字種構成比率について示す．ここでは，文字の種類を「ひらがな」「カタカナ」「漢字」「英数字」「記号」の 5 種に分け[8]，それぞれの出現数，および総文字数の中で占める割合を求めた．結果を表 6 に示す．

文字種構成比率という点で見ると，**BTEC** と **SLDB** という二つの対話が非常に近い性質を持つと言える．両者とも，漢字の比率が低く，ひらがなの比率が比較的高い．これは，「旅行会話」という性質上，漢語の出現数がほかと比べて少ないこと，ひらがなが多用される文末形式（「～にしていただきたいんですけど。」「～してもらわなければいけないんですが。」など）が頻出することなどに起因すると考えられる．一方，**NHK** と**日経新聞**では，漢字の比率が高く，ひらがなの比率が低い．これは，ニュースの報道という性質上，漢語が使われることが多いためと考えられる．**あすを読む**は，漢字とひらがなの比率だけで見ると，両者の中間に位置するようである．これは，**NHK** や**日経新聞**に比べると，より即興的な話しことばであるという特性が影響していると思われる．

カタカナと英数字については，各コーパスの書き起こしに関する条件・制約に違いがあるため（例えば，**あすを読む**や **BTEC** では英数字は一切使われず，アルファベットはカタカナで，数

[8] 「きゃ」や「フィ」などは 2 文字とした．「1ヶ月」「二ヵ月」の「ヶ」「ヵ」は「カタカナ」に含めた．「英数字」はアルファベットとアラビア数字，ローマ数字，ギリシャ文字（すべて全角）を含む．「記号」は以下のもの．

、。，．・：；？！゛゜´｀¨＿ヽヾゝゞ〃仝〇ー―‐／＼～∥｜…‥‘’“”（）〔〕［］｛｝〈〉
《》「」『』【】＋－±×÷＝＜＞≦≧∞∴♂♀°′″℃￥＄£％＃＆＊＠§☆★○●◎◇◆□■△▲
▽▼※〒→←↑↓＝∪∧∨∂≒≪≫∵∫♪†‡―｜┌┐┘└

	あすを読む	NHK	日経新聞	SLDB	BTEC
ひらがな	548,565	55,930,920	274,195,239	278,419	1,433,760
	(55.5%)	(44.8%)	(32.6%)	(60.7%)	(61.1%)
カタカナ	53,059	8,324,968	79,358,960	42,712	231,962
	(5.4%)	(6.7%)	(9.4%)	(9.3%)	(9.9%)
漢字	351,689	52,001,025	374,316,123	93,909	451,074
	(35.6%)	(41.7%)	(44.5%)	(20.5%)	(19.2%)
英数字	0	450,249	23,429,439	10	0
	(0%)	(0.4%)	(2.8%)	(0%)	(0%)
記号	34,241	8,009,334	79,723,759	43,727	227,999
	(3.5%)	(6.4%)	(9.5%)	(9.5%)	(9.7%)

表 6: 文字種構成比率

字は漢数字で表記されることになっている．一方，**日経新聞**ではアルファベットやアラビア数字が多用されている），単純に比べることはできない．

記号については，**日経新聞**，**SLDB**，**BTEC** が高い比率を示している．**SLDB**，**BTEC** については，1 文が比較的短いため，句点の比率が高くなっているものと思われる．また，読まれることを前提とする**日経新聞**では，視覚的に効果の高い雑多な記号類（——，（，），・，☆など）が多用されていることにより，記号の比率が高くなっていると思われる．

ここで，句読点とそれ以外の記号の出現比率について示す．各コーパスに現れた句読点（、。）の出現数と，それが記号の総数の中で占める比率を求めると，表 7 のようになる．

	あすを読む	NHK	日経新聞	SLDB	BTEC
句読点	27,112	4,776,963	36,496,278	37,298	196,009
	(79.2%)	(59.6%)	(45.8%)	(85.3%)	(86.0%)

表 7: 記号の総数に占める句読点の比率

日経新聞の数値が低いことから，**日経新聞**では句読点以外の記号も多く用いられていることが分かる．

4.3 形態素

次に，各コーパスの形態素数の分布について示す．ここでは，日本語形態素解析システム「茶筅 Ver. 2.2.9 [9] 」を用いることにより，形態素数を求めた[10] ．

[9] http://chasen.aist-nara.ac.jp/

[10] なお，「名詞－数詞」および「記号 - アルファベット」は連結品詞とした．

(4.3.1) 形態素数の比較

各コーパスの総形態素数と異なり形態素数，異なり率，1形態素当たりの平均文字数，1記事/1文当たりの平均形態素数を求めた結果を，表8に示す．

	あすを読む	NHK	日経新聞	SLDB	BTEC
総形態素数	577,862	75,624,654	499,482,879	255,163	1,371,643
異なり形態素数	16,178	140,557	873,387	5,243	17,571
異なり率	2.80%	0.19%	0.17%	2.05%	1.28%
平均文字数 / 形態素	1.71	1.65	1.68	1.79	1.71
平均形態素数 / 記事	1767.2	263.2	284.0	412.9	—
平均形態素数 / 文	29.14	47.46	26.83	11.72	7.87

表 8: 総形態素数と異なり形態素数，平均形態素数

「異なり形態素数」は，活用形式の違いを考慮して，活用語を基本形に戻した上で異なり形態素数をカウントしたものである．それぞれの異なり形態素数を総形態素数で割った「異なり率」を見ることによって，各コーパス総体に比べてどれだけ多様な形態素が使われているかが分かる．

また，「平均文字数 / 形態素」を見ることによって，1形態素を構成する文字数を比較することができる．この点については，各コーパスにおいて，それほど大きな差は現れていない．

さらに，「平均形態素数 / 文」を見ることによって，1文の長さを形態素数の側面から比較することができる．**NHK** が47語と極めて長く，続いて**あすを読む**と**日経新聞**が29語，27語と同程度の長さであり，さらに **SLDB** と **BTEC** が11語，8語と極めて短い，という分布が見て取れる．

(4.3.2) 形態素種の比較

日本語形態素解析システム「茶筅」には，大見出しで13種類，小見出しで71種類の品詞が登録されている．各コーパスに現れた品詞（大見出し）の出現頻度（品詞構成比率）を，表9に示す．

表9で特徴的なのは，**あすを読む**，**NHK**，**日経新聞**で名詞の比率が比較的高いこと，一方の **SLDB** と **BTEC** では助動詞の比率が高いこと，同じく **SLDB** と **BTEC** で感動詞の比率が高いこと，そして**日経新聞**と **SLDB**，**BTEC** で記号の比率が 比較的高いことである．これは，報道や解説を主たる目的とする**あすを読む**，**NHK**，**日経新聞**では，報道・解説の実質的な内容を担う名詞が多く出現するために名詞の比率が高いのに対して，「旅行会話」である **SLDB** や

	あすを読む	NHK	日経新聞	SLDB	BTEC
名詞	210,212	31,046,108	228,385,183	71,466	363,051
	(36.38%)	(41.05%)	(45.72%)	(28.01%)	(26.47%)
助詞	171,334	20,979,117	116,267,772	55,248	350,982
	(29.65%)	(27.74%)	(23.28%)	(21.65%)	(25.59%)
動詞	76,250	9,223,873	45,674,932	27,640	183,591
	(13.20%)	(12.20%)	(9.14%)	(10.83%)	(13.38%)
助動詞	58,379	5,139,734	19,969,828	39,190	186,446
	(10.10%)	(6.80%)	(4.00%)	(15.36%)	(13.59%)
形容詞	5,475	498,612	3,573,896	2,510	24,430
	(0.95%)	(0.66%)	(0.72%)	(0.98%)	(1.78%)
副詞	9,121	512,554	3,218,918	5,689	23,968
	(1.58%)	(0.68%)	(0.64%)	(2.23%)	(1.75%)
接続詞	6,640	221,185	1,265,622	3,308	3,121
	(1.15%)	(0.29%)	(0.25%)	(1.30%)	(0.23%)
連体詞	10,491	469,283	1,554,758	1,330	15,756
	(1.82%)	(0.62%)	(0.31%)	(0.52%)	(1.15%)
接頭詞	2,405	420,858	4,899,081	5,018	10,693
	(0.42%)	(0.56%)	(0.98%)	(1.97%)	(0.78%)
感動詞	85	62,029	83,869	5,700	12,933
	(0.01%)	(0.08%)	(0.02%)	(2.23%)	(0.94%)
フィラー	111	30,650	26,370	65	564
	(0.02%)	(0.04%)	(0.01%)	(0.03%)	(0.04%)
その他	0	50	614	0	3
	(0%)	(0%)	(0%)	(0%)	(0%)
記号	27,359	7,020,601	74,562,036	37,999	196,105
	(4.73%)	(9.28%)	(14.93%)	(14.89%)	(14.30%)

表 9: 品詞構成比率

BTEC では一つの文が短いために句点の割合が相対的に大きくなり[11]，さらに文末に用いられる助動詞の比率が高くなっているものと考えられる．特に，**SLDB** と **BTEC** の助動詞の比率が動詞の比率よりも高いという点は，ほかのコーパスには見られない特徴である．**SLDB** と **BTEC** において感動詞の比率が高いのは，やはり「旅行会話」という特性上，「はい」「ありがとう」「どうも」などの表現が頻出するためである．また，**日経新聞**の記号の比率が高いという点は，文字種の分析で示した通り，視覚的に効果の高い雑多な記号類が多用されていることから説明できる．

さらに，**日経新聞**では，助詞や動詞，助動詞の比率が**あすを読む**や **NHK** に比べて比較的低い数値となっている．これには，次のように，文を用言ではなく名詞の形で終わらせるという新聞記事に特有なスタイルが多用されていることが影響していると思われる．

> 金融商品や各種の行政サービスの利用も可能。話回線を介さないため、電話料金がかからないうえ、定額制が一般的。

> 建物の大きさは、東西約十九メートル、南北約七メートル、高さ約十一メートル。

> 一階に畳の三室、二階にフローリングの三室があり、最大六人が入居可能。各部屋には小さなキッチン、トイレ、ベランダが付き、おふろは共同。床は段差をなくしたバリアフリー仕様で、廊下の至る所に手すりが付く。六年前に夫を亡くし、独り暮らしだった○○さんは昨年四月に入居。

> 経済運営では「財政構造改革という大変重い課題を一時たりとも忘れたことはない」と強調。

> 小渕恵三首相の年頭記者会見の要旨は次の通り。

> 政府はロシア大統領選とそれに伴う新体制の発足をにらみながら、取りあえずプーチン大統領代行との間で、平和条約締結に向けたプロセスの再構築に全力を挙げる構え。

(4.3.3) 頻出形態素の比較

各コーパスに出現した形態素のうち，頻出した上位 10 位までのものについて，出現頻度と，総形態素数に対する割合を，表 10 に示す．

[11] これは，総形態素数に占める句点の割合を見ると，**SLDB** と **BTEC** が高くなっているという事実からも裏付けられる．

	あすを読む	NHK	日経新聞	SLDB	BTEC
句点 / 総形態素数	3.4%	2.2%	3.1%	8.3%	12.6%

あすを読む	NHK	日経新聞	SLDB	BTEC
の / 助詞 - 連体化	の / 助詞 - 連体化	の / 助詞 - 連体化	。 / 記号	。 / 記号
29,844 (5.16%)	4,398,654 (5.82%)	23,206,651 (4.65%)	21,769 (8.53%)	174,257 (12.70%)
。 / 記号	、 / 記号	、 / 記号	、 / 記号	ます / 助動詞
19,828 (3.43%)	3,100,165 (4.10%)	19,961,002 (4.00%)	15,492 (6.07%)	70,993 (5.18%)
ます / 助動詞	を / 格助詞	。 / 記号	ます / 助動詞	か / 終助詞
18,455 (3.19%)	2,376,108 (3.14%)	16,529,022 (3.31%)	15,008 (5.88%)	64,850 (4.73%)
を / 格助詞	に / 格助詞	を / 格助詞	です / 助動詞	です / 助動詞
18,450 (3.19%)	2,148,212 (2.84%)	15,014,852 (3.01%)	9,373 (3.67%)	61,057 (4.45%)
は / 係助詞	た / 助動詞	は / 係助詞	の / 助詞 - 連体化	は / 係助詞
18,391 (3.18%)	2,111,070 (2.79%)	13,439,159 (2.69%)	8,068 (3.16%)	47,997 (3.50%)
て / 助詞	する / 動詞	に / 格助詞	て / 助詞	を / 格助詞
16,794 (2.91%)	2,058,769 (2.72%)	12,314,644 (2.47%)	5,355 (2.10%)	41,898 (3.05%)
に / 格助詞	て / 助詞	する / 動詞	か / 終助詞	て / 助詞
15,761 (2.73%)	1,983,499 (2.62%)	12,202,463 (2.44%)	5,298 (2.08%)	41,811 (3.05%)
が / 格助詞	は / 係助詞	が / 格助詞	に / 格助詞	の / 助詞 - 連体化
15,541 (2.69%)	1,968,858 (2.60%)	9,987,593 (2.00%)	5,186 (2.03%)	30,955 (2.26%)
する / 動詞	が / 格助詞	/ 記号	は / 係助詞	に / 格助詞
14,011 (2.42%)	1,849,562 (2.45%)	9,758,079 (1.95%)	4,912 (1.93%)	30,560 (2.23%)
た / 助動詞	。 / 記号	た / 助動詞	た / 助動詞	する / 動詞
12,969 (2.24%)	1,676,098 (2.22%)	9,612,398 (1.92%)	4,442 (1.74%)	27,151 (1.98%)

表 10: 頻出形態素の分布（上位 10 位）

いずれのコーパスにおいても，句読点や助詞などの機能語が出現頻度の上位を占めている．**NHK**でのみ句点が比較的下位に位置づけられているのは，表4にも示した通り，**NHK**の1文が他に比べて長いことを示す．敬体（デスマス調）で話されているあすを読む，**SLDB**，**BTEC**では，「です」「ます」が上位に現れているが[12]，**NHK**の場合，敬体で話されているにも関わらず，「ます」は11位，「です」は27位であった．これも，1文が長いために文末のデスマスが現れるまでが長く，出現頻度としては下位にランクされたものと考えられる．

4.4 文節

次に，各コーパスの文節数の分布について示す．文節数をカウントするため，ここでは，日本語形態素解析システム「CaboCha/ 南瓜 Ver. 0.21 [13] 」を用いることにより，文節数を求めた．

(4.4.1) 文節数の比較

各コーパスの総文節数と異なり文節数，異なり率，1文節当たりの平均文字数と平均形態素数，1記事 /1 文当たりの文節数を求めた結果を，表11 に示す．

	あすを読む	NHK	日経新聞	SLDB	BTEC
総文節数	222,930	27,196,464	166,082,917	84,811	471,563
異なり文節数 (1)	84,684	2,539,451	22,672,121	24,295	82,531
異なり文節数 (2)	82,490	2,164,764	19,192,383	22,243	77,650
異なり率	37.99%	9.34%	13.65%	28.65%	17.50%
平均文字数 / 文節	4.43	4.59	5.06	5.41	4.97
平均形態素数 / 文節	2.59	2.78	3.01	3.01	2.91
平均文節数 / 記事	681.7	94.7	94.4	137.2	—
平均文節数 / 文	11.24	17.07	8.92	3.90	2.71

表 11: 総文節数と異なり文節数，平均文節数

表11 の「異なり文節数 (1)」は，コーパス中の句点「。」を文節に含めてカウントしたものである．ある文節に句点が後続する場合と句点が後続しない場合（例えば「あります。」と「あります」）が，別々にカウントされている．これに対して「異なり文節数 (2)」は，コーパス中に現れる句点「。」を除去した上で文節数をカウントしたものである（「あります。」と「あります」は，同じものとしてカウントされている）．ここでは，「異なり文節数 (1)」を異なり文節数として採用し，異なり率や1記事 /1 文当たりに含まれる文節数を計算している．

[12] あすを読むに現れた「です」は，12位．

[13] http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

異なり文節数を総文節数で割った「異なり率」を比較することによって，各コーパスの総体に比べてどれだけ多様な文節表現が使われているかを比較することができる．例えば，**NHK** と**日経新聞**，**BTEC** の異なり率は他に比べて小さいが，これは両者がそのコーパスサイズに比べて少ない範囲の文節表現から構成されていることを示している．

また，「平均文字数 / 文節」「平均形態素数 / 文節」を見ることによって，1 文節を構成する文字数および形態素数を比較することができる．「平均文字数 / 文節」で，**SLDB** が 5.41 と高い数値を示しているが，これは「ニューヨークシティホテルでございます。」のような長い固有名詞を含む文節や，「四九八零零四五九九一九九五三一三です。」のような数詞連続を含む文節などが多いためであると考えられる．その証拠に，「平均形態素数 / 文節」では各コーパスでそれほど大きな差が現れていない．

さらに，「平均文節数 / 文」を見ることによって，1 文の長さを文節数の側面から比較することができる．ここでは **NHK** が 17 文節と極めて長く，続いて**あすを読む**と**日経新聞**が 11 文節，9 文節と同程度の長さであり，さらに **SLDB** と **BTEC** が 4 文節，3 文節となっている．

(4.4.2) 文節種の比較

文節は，その文節を構成する最後の形態素（主辞）に着目すると，「は」で終わる文節，「を」で終わる文節など，いくつかの種類に分類することができる．各コーパスに出現した文節種（主辞のみで表す）のうち，頻出した上位 10 位までのものについて，出現頻度と，総形態素数に対する割合を，表 12 に示す[14]．

あすを読む，**NHK**，**日経新聞**については，上位 6 位程度まではほぼ同じ文節種であった．助詞で終わる文節（助詞 - 連体化，格助詞，係助詞）は，いずれもほぼ同じ割合で出現していることが分かる．一方，対話である **SLDB**，**BTEC** では，疑問を表す「か - 終助詞」や，待遇表現「です」「ます」「くださる」，応答表現「はい」などが現れている点が特徴的であった．

4.5 節

次に，各コーパスの節数の分布について示す．言語表現の基本単位は句点で終わる「文」であるが，文はさらに「節」に分解することができる．「節」とは，述語を中心とするまとまりのことである．次のように一つの節から構成される文は，「単文」と呼ばれる．

> 地図を見てください。
> 一万五千円のラインをあっさりと割り込みました。
> このシンポジュームの中で災害の心理的影響とケアについて内外の専門家が意見交換をしました。 (あすを読む)

一方，二つ以上の節から構成される文は「複文」と呼ばれる（/ は，節境界を示す）．

[14] ここでは，句点は除いて文節の種類をカウントしてある．

あすを読む	NHK	日経新聞	SLDB	BTEC
の / 助詞 - 連体化	の / 助詞 - 連体化	の / 助詞 - 連体化	の / 助詞 - 連体化	か / 終助詞
29,466 (13.22%)	4,382,812 (16.12%)	23,074,479 (13.89%)	8,045 (9.49%)	64,151 (13.60%)
を / 格助詞	を / 格助詞	を / 格助詞	ます / 助動詞	は / 係助詞
18,364 (8.24%)	2,372,610 (8.72%)	14,974,478 (9.02%)	6,463 (7.62%)	47,388 (10.05%)
は / 係助詞	は / 係助詞	は / 係助詞	か / 終助詞	を / 格助詞
17,115 (7.68%)	1,927,562 (7.09%)	13,137,100 (7.91%)	4,933 (5.82%)	41,876 (8.88%)
が / 格助詞	に / 格助詞	に / 格助詞	は / 係助詞	の / 助詞 - 連体化
15,523 (6.96%)	1,913,652 (7.04%)	10,684,359 (6.43%)	4,846 (5.71%)	30,771 (6.53%)
に / 格助詞	が / 格助詞	が / 格助詞	に / 格助詞	に / 格助詞
13,226 (5.93%)	1,842,716 (6.78%)	9,982,426 (6.01%)	4,619 (5.45%)	27,564 (5.85%)
た / 助動詞	た / 助動詞	た / 助動詞	た / 助動詞	が / 格助詞
9,615 (4.31%)	1,755,960 (6.46%)	8,068,510 (4.86%)	3,846 (4.53%)	22,218 (4.71%)
ます / 助動詞	で / 格助詞	で / 格助詞	を / 格助詞	ます / 助動詞
8,840 (3.97%)	1,068,940 (3.93%)	4,703,836 (2.83%)	3,660 (4.32%)	20,715 (4.39%)
と / 格助詞	て / 接続助詞	する / 動詞	が / 格助詞	です / 助動詞
6,829 (3.06%)	675,714 (2.48%)	4,293,345 (2.59%)	3,226 (3.80%)	18,880 (4.00%)
も / 係助詞	と / 格助詞	だ / 助動詞	はい / 感動詞	くださる / 動詞
5,416 (2.43%)	585,047 (2.15%)	4,054,831 (2.44%)	3,073 (3.62%)	13,658 (2.90%)
で / 格助詞	ます / 助動詞	も / 係助詞	で / 格助詞	た / 助動詞
5,300 (2.38%)	527,396 (1.94%)	3,642,095 (2.19%)	2,319 (2.73%)	11,799 (2.50%)

表 12: 頻出文節種の分布（上位 10 位）

確かに職業意識の変化もあり / 若い世代は会社に就職しないで / やりたい / 事をやるという / フリーターが増えています。
この呼び掛けによりますと / 西暦二千年問題が原因で大きな問題が発生することは / ないと / していますけれども / 地震や風水害の備えとして用意してある / ものを点検することを / お勧めしますという / 言い方で準備の必要性を示しています。

文中に出現する節の数や種類の分布について検討するためには，何らかの方法で文に含まれる節の境界を検出する必要がある．そこで，形態素情報を参照して節境界の検出・判定を行なうスクリプトを作成した．このスクリプトは，各形態素の出現形，品詞，活用形，活用型の情報を参照して，144 種類の節境界を自動的に検出し，ラベリングすることができる[15]．自動節境界ラベリングの結果を，以下に示す．/ / で囲まれた部分が節境界であり，そこがどのような種類の節境界であるかを示すラベルが入っている．

眠っている / 連体節 /
土地の有効活用が目的ということで / 並列節デ /
既にマンションの建設や住宅などに利用されるように / ヨウニ節 /
なってまいりました。 / 文末 /
この際改正の際定期借家権についても検討されたんですけれども / 並列節ケレドモ /
いろいろ問題があるという / 連体節トイウ /
ことで / 並列節デ /
例えば / 談話標識 /
転勤で必ず戻ってくるような / 連体節ヨウナ /
場合ですとか / 並列節トカ /
あるいは / 談話標識 /
都市計画などで取り壊しが決まってる / 連体節 /
場合に限って / テ節 /
適用が認められるように / ヨウニ節 /
なっております。 / 文末 /

(4.5.1) 節数の比較

各コーパスの総節数と異なり節数，異なり率，1 節当たりの平均文字数と平均形態素数，平均文節数，さらに 1 記事 /1 文当たりの平均節数を求めた結果を，表 13 に示す．

異なり節数を総節数で割った「異なり率」を比較することによって，各コーパスの総体に比べてどれだけ多様な節表現が使われているかを比較することができる．例えば，**NHK** と**日経新聞**，**BTEC** の異なり率は他に比べて小さいが，これは両者がそのコーパスサイズに比べて少ない範囲の節表現から構成されていることを示している．

[15] ラベリングされる対象には，「主題ハ」「談話標識」「感動詞」という，正確には節境界でない要素も含む．以下，特に断りのない場合は，これら二つの要素も節境界に含む．

	あすを読む	NHK	日経新聞	SLDB	BTEC
総節数	92,705	9,579,876	60,644,993	44,627	262,050
異なり節数	67,536	4,692,310	37,736,360	22,376	108,779
異なり率	72.85%	49.00%	62.2%	50.14%	41.51%
平均文字数 / 節	10.65	13.02	13.86	10.28	8.95
平均形態素数 / 節	6.23	7.89	8.24	5.72	5.23
平均文節数 / 節	2.40	2.84	2.74	1.90	1.80
平均節数 / 記事	283.5	33.3	34.5	72.2	—
平均節数 / 文	4.68	6.01	3.26	2.05	1.50

表 13: 総節数と異なり節数，平均節数

また，「平均文字数 / 節」「平均形態素数 / 節」「平均文節数 / 節」を見ることによって，一つの節を構成する文字数，形態素数，および文節数を比較することができる．いずれも **NHK** と**日経新聞**が高い数値を示している．1 文に含まれる節の数は **NHK** が**日経新聞**の 2 倍近くになっており，**NHK** の 1 文が極めて長いことが分かる．

(4.5.2) 単文と複文の比較

次に，各コーパスにおける単文と複文の出現数と比率について，表 14 に示す．ここでは，1 文中に現れる節境界が / 文末 / のみであるものを単文，/ 文末 / 以外に一つ以上の節境界を含む文を「複文」としてカウントした[16]．

	あすを読む	NHK	日経新聞	SLDB	BTEC
単文	3,151	157,711	3,997,376	14,342	144,327
	(15.9%)	(9.5%)	(24.2%)	(65.9%)	(82.8%)
複文	16,677	1,506,556	12,542,359	7,427	29,930
	(84.1%)	(90.5%)	(75.8%)	(34.1%)	(17.2%)

表 14: 単文と複文の分布

あすを読む，**NHK** に現れる文は，それぞれ 84%，90% が複数の節を含む複文であり，単文は 16%，10% と少ない．**日経新聞**では，複文が 76%，単文が 24% となっているが，実際には単文ではない次のような「箇条書き」のようなものも多く含まれているため，実際の本文記事のみを対象とすれば，複文の比率は上がるものと思われる．

[16] 「主題ハ」「談話標識」「感動詞」という三種の非節境界要素は，除去してカウントしてある．また，引用部分に複数の文が含まれる場合は考慮していない．さらに，**NHK** と**日経新聞**では，句点のない文や図表などがあるため，単文と複文の和と表 2の「総文数 (1)」の数値とに誤差が生じている．

〈仕様〉六本入り二袋入り。/ 文末 /
〈価格〉百五十円。/ 文末 /
〈発売時期〉二月から順次発売。/ 文末 /
辛口のライム味。/ 文末 /

一方，比較的単純な言い回しの多い **BTEC** では，80% 強が単文であり，複文は 20% 弱に過ぎなかった．**あすを読む**や**日経新聞**と比べて，単文と複文の割合がほぼ逆転していることになる．また，比較的 **BTEC** よりも即興的な話しことばである **SLDB** では，複文の割合が **BTEC** よりも増え，3 割強が複文であった．

(4.5.3) 節種の比較

節は，その節を構成する最後の形態素や，文中における機能によって，「並列節ケレドモ」，「テ節」「連体節」などに分類することができる．各コーパスにおいて，何種類の節が出現したかについて，表 15 に示す．

あすを読む	NHK	日経新聞	SLDB	BTEC
125	138	144	87	102

表 15: 出現した節のバリエーション

コーパスサイズが大きくなるにつれて，より多様な表現が現れることから，コーパスサイズの大きい**日経新聞**や **NHK** により多くの種類の節が現れていることが分かる．

さらに，各コーパスに出現した節のバリエーションのうち，頻出した上位 10 位までのものについて，出現頻度と，総節数に対する割合を表 16 に示す．

頻出する節境界の種類，および割合に着目すると，**あすを読む**，**NHK**，**日経新聞**はよく似た分布を示していると言える．一方，そもそも節境界の出現数が少ない **SLDB** と **BTEC** は，/ 文末 / が極めて高い割合を示している．「感動詞」が上位を占めているのは，「はい」や「いいえ」などが頻出する旅行会話の特性によるものと考えられる．

以上，各コーパスの言語的・数量的な特徴について比較・分析を行なった結果を示した．

あすを読む	NHK	日経新聞	SLDB	BTEC
文末	主題ハ	文末	文末	文末
19,828 (21.39%)	1,778,190 (18.56%)	16,539,734 (28.46%)	21,769 (48.78%)	174,257 (66.50%)
主題ハ	文末	主題ハ	主題ハ	主題ハ
14,556 (15.70%)	1,664,267 (17.37%)	12,020,403 (20.68%)	4,418 (9.90%)	44,638 (17.03%)
連体節	連体節	連体節	感動詞	テ節
11,164 (12.04%)	1,647,936 (17.20%)	9,311,069 (16.02%)	3,690 (8.27%)	6,231 (2.38%)
談話標識	テ節	連用節	談話標識	感動詞
7,516 (8.11%)	906,707 (9.46%)	4,614,269 (7.94%)	3,306 (7.41%)	5,751 (2.19%)
テ節	連用節	テ節	テ節	連体節
6,120 (6.60%)	769,338 (8.03%)	2,283,699 (3.93%)	1,445 (3.24%)	5,595 (2.14%)
連体節トイウ	補足節	補足節	連体節	談話標識
4,209 (4.54%)	438,930 (4.58%)	1,890,002 (3.25%)	1,283 (2.87%)	3,025 (1.15%)
引用節	引用節	引用節	並列節ガ	譲歩節テモ
3,957 (4.27%)	379,013 (3.96%)	1,776,389 (3.06%)	1,064 (2.38%)	2,064 (0.79%)
補足節	談話標識	談話標識	引用節	並列節ガ
3,471 (3.74%)	260,909 (2.72%)	1,366,873 (2.35%)	974 (2.18%)	2,010 (0.77%)
連用節	連体節 - 形式名詞	並列節ガ	理由節ノデ	条件節バ
2,642 (2.85%)	219,363 (2.29%)	1,266,366 (2.18%)	781 (1.75%)	1,985 (0.76%)
連体節 - 形式名詞	連体節トイウ	並列節デ	条件節タラ	補足節
2,579 (2.78%)	195,221 (2.04%)	885,931 (1.52%)	775 (1.74%)	1,911 (0.73%)

表 16: 頻出節種の分布（上位 10 位）

5 おわりに

本稿では，「独話」と「対話」，「書きことば」と「話しことば」という，コーパスが備える基本的な特性の対立を踏まえた上で，これまでに収集された異種コーパス群の言語的・数量的な側面について，定量的に比較分析を行なった．

ここで示された結果をもとにして，さらにさまざまな言語的分析を行なうことができると考えられる．例えば，次のような観点からの分析が可能であろう．

- 係り受け間の距離（係り元から係り先までの距離）に違いはあるか．

  → 1文が長い独話の方が，対話よりも係り受け間の距離は長くなると考えられる．また，同じ独話の中でも，準備性の極めて高い**NHK**の方が，より即興的な**あすを読む**よりも係り受け間の距離は長くなると考えられる．係り受け解析ツールなどを用いることによって，これらの違いを定量的に調査することができる．

- 埋め込み構造 / 階層構造の深さに違いはあるか．

  → 引用節や連体節など，埋め込み構造の階層が深くなるほど，文構造も複雑化すると考えられる．係り受け間の距離と同様，準備性の極めて高い**NHK**の方が，より即興的な**あすを読む**よりも階層構造の深さは深くなると考えられる．これについても，係り受け解析ツールなどを用いることによって，定量的に調査を行なうことができる．

- 文末表現のパタンに違いはあるか．

  → 3節でも述べたように，独話と対話では，特に文末表現（モダリティ形式）の出現傾向に顕著な違いが認められると考えられる．文末形式の出現状況・分布について定量的に調査を行なうことで，頻出する文末表現のパタンを抽出し，独話と対話の文末表現の違いを見ることによって，発話行為のパタン，聞き手に対する働きかけのバリエーション，意図 / 目的 と言語表現の対応などの点を明らかにすることができる．

- フィラーの出現傾向について違いはあるか．

  → 特に話しことばの特徴として，フィラー（言いよどみ語）や間投詞，言い直しなどの存在が挙げられる．今回はフィラー類をあらかじめ除去して調査を行なったが，これらの要素にはどのような種類があるか，また，それらはどのような位置に現れるかなどを調べることによって，話しことばの即興性に関わる言語的特徴や，談話管理の方略などについても考察を行なうことができる[17]．

これらについては，今後の課題としたい．

[17] 丸山 (2002 印刷中) も参照．

参考文献

1. 片桐泰弘 (1997) 「終助詞とイントネーション」, 『文法と音声』. 音声文法研究会 (編), くろしお出版.

2. 国立国語研究所 (1961) 『話しことばの文型 (1) — 対話資料による研究 —』. 秀英出版.

3. 国立国語研究所 (1963) 『話しことばの文型 (2) — 独話資料による研究 —』. 秀英出版.

4. 永野賢 (1988) 「話しことばと書きことば」, 『日本語百科大事典』. 大修館書店.

5. 畠弘巳 (1987) 「話しことばの特徴 — 冗長性をめぐって —」, 『国文学 解釈と鑑賞』, 第52巻7号, 至文堂.

6. 丸山直子 (1996) 「話しことばの諸相」, 『言語処理学会 第2回年次大会チュートリアル資料』, 言語処理学会.

7. 丸山岳彦 (2002印刷中) 「話しことばコーパスに現れる「ですね」の分析」, 『さわらび』, 11号, 神戸市外国語大学.

8. 中嶋秀治・熊野正・松井知子・山本博史・隅田英一郎・柏岡秀紀・竹澤寿幸・マルコフ コンスタンティン (2000) 「TR-S-0001 データベース仕様書 (第1版)」. ATR音声言語通信研究所 テクニカルレポート.