第 2 章　談話データの構築

第 2 章では，〈話されたことば〉のコーパス，とりわけ英語と日本語と韓国語の既存の「話しことば」のコーパスを取り上げ，その編纂の動向を概観してゆく．そののちに，本稿における〈話されたことば〉の談話データの収集方法，話者の属性などを詳細に提示する．最後に本稿の談話データの文字化の原則について述べる．

2.1. 話しことばのコーパス

「コーパス」という名称は，亀井孝他(1996)によれば，「資料体」と提示されており，「言語分析の対象となるデータ・材料の集合体」と明示されている．

世界最初のコーパスといえる Brown Corpus(the Brown University Corpus of American English)は，書きことばのコーパスで，1961 年に公にされた[8]．今は'Corpus Linguistics'(コーパス言語学)[9]という術語も定着し，一般化されており，〈書かれたことば〉のみならず，〈話されたことば〉のコーパスも盛んに公にされている．

〈話されたことば〉の研究が盛んに行われるようになると，談話分析，社会言語学，方言学，第 2 言語習得などの様々な分野から，多量の〈話されたことば〉のデータを求めるようになった．〈話されたことば〉の音声データを集め，文字転写作業を通じ，機械的な処理ができるよう，コーパスの構築が行われるようになったのである．

Gibbon(1998:79)は，〈話されたことばのコーパス〉(a spoken language corpus)の概念を次のように述べている：

> A spoken language corpus is any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house, or by scientists in other organizations.(話されたことばのコーパスは，コンピュータが読めるような形態で扱うことができ，それを作ったところのみならず，他の機関で再利用を可能にするような十分な注釈と資料が付されている話されたことばの録音の集合体である：引用者訳)

以下，英語と韓国語，日本語のコーパスの中でも，とりわけ〈話されたことば〉のコ

---

[8] 亀井孝他(1996)参照.

[9] コーパス言語学とは，コンピュータ処理可能なコーパスに基づき言語記述を行う言語学をいう．斉藤俊雄・中村純作・赤野一郎(1998)による.

ーパスに注目し，検討してみる．

2.1.1. 英語の〈話されたことば〉のコーパス

Chafe, et al. (1991:64)は，英語における代表的なコーパスとして，1961 年にアメリカで公にされた，アメリカ英語の〈書かれたことば〉のコーパスである Brown Corpus，1961 年に出版されたイギリス英語に関する Lancaster-Oslo/Bergen Corpus(=LOB)，1960 年代から 1970 年代にイギリスで収集された英語の〈話されたことば〉のコーパスである The London-Lund Corpus(=LLC)を上げている．

斉藤俊雄・中村純作・赤野一郎(1998:3-23)によると，Brown Corpus は上でも言及したように，「世界最初の電子コーパス」であり，「15 ジャンルを代表する，各 2,000 語のテキスト 500 冊から，アメリカ英語の資料約 100 万語を集めたもの」であった．また，「イギリス英語の最初の電子コーパス」である Lancaster-Oslo/Bergen Corpus(=LOB)は「イギリスで出版された本，新聞，雑誌などから抽出した計約 100 万語のコーパス」である．

The London-Lund Corpus(=LLC)もやはり「イギリスの英語の話しことばコーパス」であり，「1953-88 年間に録音されたイギリスの教養ある話し手による日常会話，インタビュー，電話での会話，放送，講義，演説などを文字化したコーパス」である．

また，Leech & Fligelstone(1992:118)によると，LLC は Randolph Quirk が主導し，50 万語の英語の〈話されたことば〉を収集したという．また，LLC は強勢やイントネーションなどの音韻的特徴を調べることを主な目的にして転写されたものであると述べている．

英語の話しことばのコーパスは Leech & Fligelstone(1992:118)によると，LLC 以外にも，The IBMLancaster Spoken English Corpus(=SEC)があって，LLC よりは量的には少ないが，転写法的にも，音韻論的にも，文法論的(orthographically transcribed, prosodically transcribed, grammatically) にも用いられるようタグを付しているという．また Chafe,et al.(1991:65,69)が中心となった，The Corpus of Spoken American English (=CSAE)は，20 万語をデータベース化しているという．地域，社会的階級，人種，年齢別のグループに分け，サービス応対，説教，医者と患者のやりとり，法律上の告訴(弁論)，計画的/非計画的な講義，廊下での会話(立ち話)(service encounters, sermons, doctor-patient interactions, legal proceedings, planned and unplanned class lectures, hallway conversations)などの〈話されたことば〉が集められているという．

2.1.2. 韓国語の〈話されたことば〉のコーパス

韓国語では1998年から「21世紀世宗計画国語特殊資料構築」の名で，文化観光部と国立国語研究院が協力し，延世大学の서상규と임용기の責任と指導の下で，膨大な規模の〈書かれたことば〉と〈話されたことば〉のコーパスが構築され始めた．〈書かれたことば〉を문어[ムノ](文語)と呼び，〈話されたことば〉を구어[クオ](口語)と呼んでいる．日本語研究で言う「文語」，「口語」との術語の用法の差に注意せねばならない．

1998年以来毎年出されている研究報告書によると，とりわけ〈話されたことば〉においては，1998年から2004年まで構築されたコーパスは，「原始コーパス[10]」が333万文節，「形態素分析コーパス」が68万文節に達している．なお，ここで「文節」と訳したのは，韓国語では어절(語節)と言われるもので，日本語の「文節」に近い概念である．「体言+助詞」などはこれが1つの「語節」である．

談話の種類は，1人の談話と2人以上の談話とに分けられ，さらに講義，公演，放送などの公的な談話と，授業での会話，病院での会話，会議，電話での会話，主題を決めた会話，日常の会話などの私的な会話に分類されている．

また，膨大なデータの，検索や用例の抽出が容易にでき，多くの研究に用いられるよう，転写方法や体系などの自然言語処理性(machine readable)にも細心の注意が払われている．

最初は放送や講演などの談話が大半を占めていたが，2001年から2人以上の日常会話のデータの収集も顕著に増えてきた．ただ，膨大なデータを構築することに主眼が置かれているため，方言などの使用言語，話し手と聞き手の親密の度合いと年齢差，性別差，社会的な関係といった諸条件の制限は行われていない．会話をランダムに集め，データの収集の後に，会話参加者の属性，すなわち，20代，30代などの年代や性別，出身地などを尋ねる方法を取っている．しかし，1つの会話の中でも，いくつかの異なる方言話者が混在している場合や，相手との新密度，相手との正確な年齢の差などが不明な場合も散見される．聞き手と話し手の年齢や性別，使用方言，社会的関係などが重要視される談話分析の研究や社会言語学，方言学などの方面の研究においては，「21世紀世宗計画国語特殊資料構築」の〈話されたことば〉のコーパスも問題を抱えていると言わざるを得ない．しかし，韓国語においては，初めて編纂され，公刊された言語コーパ

---

[10]「原始コーパス」，「形態素分析コーパス」という名称は，韓国語の‘원시 말뭉치’，‘형태소분석말뭉치’を直訳したものである．但し，研究報告書には「原始コーパス」‘원시 말뭉치’，「形態素分析コーパス」‘형태소 분석말뭉치’の定義が行われていない．本稿は「原始コーパス」は，文字化が終ったコーパス，「形態素分析コーパス」は，形態素と単語ごとに注釈を付する作業が終ったコーパスとして推測する．

スであり，その膨大な量と精緻な形態素分析などは，韓国語研究の歴史に太い一線を画す成果である．

2.1.3. 日本語の〈話されたことば〉のコーパス

日本語においては『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese:CSJ)が国立国語研究所と通信総合研究所によって構築された．前川喜久雄(2004)の報告書によると，1999 年から 2003 年までの間，語数にして約 750 万語，時間にして 660 時間の音声データが構築されているという．CSJ は自発音声を対象としたデータベースであるため，1)まとまった内容をもつ(すなわち雑談ではない)，2)全国共通語の(すなわち語彙と文法に方言的特徴のない)，3)独話(モノローグ)音声を主要な対象としている．述べ話者は 3302 名，異なり話者は 1417 名である．また，異なり話者のうち，東京や千葉，埼玉，神奈川の「首都圏」の出生地の人は 588 名，そのうち対話に参加している異なり話者は 16 名で，東京と首都圏の話者は 12 名である．計約 661.6 時間の 3,302 の膨大なファイルの中で，本稿がまさに対象と捉えているような「自由対話」は 3.6 時間，16 ファイルにすぎず，1 人の談話である講演や朗読などが大半を占めている．

一方，日本語の談話コーパスとしては，東京外国語大学が遂行している 21 世紀 COE プログラム「言語運用を基盤とする言語情報学拠点」では，宇佐美まゆみの指導の下，2003 年収集された『BTS による多言語話し言葉コーパス――日本語会話 1(日本語母語話者同士の会話)』と『BTS による多言語話し言葉コーパス――日本語会話 2(日本人と学習者の会話)』が，2005 年 6 月公開[11]された． 公開された報告書によると，「音声学的分析や，形態素分析，構文の分析のためではなく，人間の相互作用としての「言語運用」の分析に適した形」のコーパスとして利用するとされている．

このうち「日本語母語話者同士」の会話は，友人同士と初対面の会話に分けられており，話者の属性，すなわち年齢や性別，社会的地位などの諸条件を統制せず，ランダムに収集されている．そうした「日本語母語話者同士」の会話は，計 121 会話，約 21 時間のデータである一方で，「日本人と学習者の会話」は，計 57 会話，約 7 時間の量のデータである．また，「日本語母語話者同士」の会話のうち，「断り」と「依頼」という主題がある会話，「論文指導の会話」，「電話での会話」を除く，「日常会話」は，49 会話，約 17 時間(1046 分)のデータがあるとされるが，録音時間なのか，文字化データの量なのかは判然としない．

---

[11] 宇佐美まゆみ(2005)参照

また，スピーチレベルのシフトを扱った，宇佐美まゆみ(2002)でも大量の日本語の〈話されたことば〉の会話データが用いられている．30 代の女性を基準話者と男女別の 20 代，30 代，40 代の話者との組み合わせである．異なり人数 84 名の 72 会話を収集している．1 会話当たり，3 分間を文字化，分析しているので 216 分の分析である．話者の方言に関しては宇佐美まゆみ(2002:46)では次のように条件付けている：

> To verify that the subjects spoke standard Japanese, the face sheet asked which dialect was easiest for them to speak.
> (話者が標準日本語を話したのかを確認するため，フェイスシート(質問紙)で話者にとってどの方言が最も話しやすいかを尋ねた：筆者訳)

すなわち，「最も話しやすいことばが東京方言である」とフェイスシートに会話協力者自身が答えたということを条件としているため，客観的に東京方言話者の会話であるとは言えない部分が惜しまれる．しかし，実際の自然会話のデータであり，条件が統制されたものである点，そしてそれが個人が収集した大量のデータである点，精密な文字化やコーディングが行われていた点などから，日本語と韓国語の自由会話の〈話されたことば〉のコーパスの中では，量的にも質的にも研究史的には画期的な談話データであると言えよう．

2.2. 本稿における談話データ

本稿で収集した談話データは，日本語会話 40 組，韓国語会話 40 組，計 80 組の談話データである．話者は，20 代，30 代，40 代の男女で，異なり人数で韓国語母語話者が 80 名，日本語母語話者が 80 名の計 160 名である．話者はすべて東京方言話者とソウル方言話者に制限されている．1 つの会話を 15 分間録音し，最初の 5 分間を文字化した，音声で計 1400 分，文字化転写は計 400 分の転写ファイルとなる．また，〈初対面同士の会話〉と〈友人同士の会話〉という条件別に分けている．内容はすべて直接会って話す〈自由会話[12]〉であり，年齢，性別などの諸条件が統制された談話データである．

〈自由会話〉の観点からすると，世代別に構成された話者の異なり人数の点において

[12] 2.2 で述べたように，電話を通じた会話やあらかじめ主題が決まっている会話ではなく，人と直接会って，決まった主題なしで自由に話し合う会話のことを本稿では「自由会話」と呼ぶ．韓国語の「21 世紀世宗計画国語特殊資料構築」の話しことばコーパスと，東京外国語大学が遂行している 21 世紀 COE プログラム「言語運用を基盤とする言語情報学拠点」では「日常会話」と呼んでおり，日本の国立国語研究所が行った『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese:CSJ)では「自由対話」と呼んでいる．

も，また〈話されたことば〉の文字化データの量的な面において，本研究のデータ量は先に見た日本語と韓国語の既存の話しことばのコーパスの量をはるかにしのぐ量である．さらに質的にも，条件を統制して得た会話と精緻に行った文字化は，日本語と韓国語に今日まで見られなかった，〈話されたことば〉の，高度に条件が統制された談話データだということができる．

2.2.1. 本稿の談話データの作成過程

本稿では，次のような過程を経て〈話されたことば〉の談話をデータベース化する：

1. あらかじめ年齢，性別，社会的地位などの話者の諸条件を統制した上，会話の組み合わせを構成した．
2. 作成された話者の条件と会話の組み合わせに合う，会話協力者を求める．
3. 条件に合う協力者の会話を，できるだけ雑音のない，静かでリラックスできる空間で録音・録画を行う．
4. 録音された会話を，日本語と韓国語の正書法に従って,また話しことばの特徴を生かした文字化を行う．
5. 文字転写が終ったファイルを，日本語のデータは 3 人の日本語母語話者[13]が，韓国語のデータは 3 人の韓国語母語話者が，それぞれ全データの会話録音を聞きながら，文字化したファイルの検討を行う．
6. 母語話者の検討が終ったファイルに，イントネーションなどの音韻的変化や発話単位の変化を起している部分など，コンピュータ上で処理できるよう，記号化したタグを付する．
7. タグを付したファイルの全データの 30%を，日本語と韓国語のそれぞれの母語話者が再び検討を行い，タグ付けの信頼性を検証する．
8. 他母語話者の検討した結果を，研究者が再び検討を行い，談話データとして確定する．

2.2.2. 本稿の談話データの構成

本稿の談話データは，次のような条件と構成で構築した：

---

[13] データの検討作業は，それぞれの言語の母語話者であり，言語を専攻している大学生，大学院生の方々の協力を得た．日本語と韓国語の基本的な知識を有すると見なされる方々である．日本語のデータは，東京方言話者の男性 1 人と他方言話者の女性 2 人である．3 人とも大学院生である．韓国語はソウル方言話者の大学生，男女 2 人と，他方言話者の大学院生の女性 1 人である．

〈会話における話者の条件〉

1. 話者の言語形成地の制限：

   東京方言話者とソウル方言話者の会話に限定するためである.

2. 年齢や性別の制限：

   20-23 歳の男女，30-33 歳の男女，40-43 歳の男女で会話の話者を構成する．こうすることで，幅広い年齢層の言語使用と性別による言語使用が考察できる．また，20 代，30 代，40 代ではなく，〈20-23 歳〉，〈30-33 歳〉，〈40-43 歳〉で年齢を制限しているのは，例えば 20 代と 30 代の会話であるというと，実際の年齢の差は 2 歳しかない，29 歳と 31 歳の会話も成立してしまう．これは 20 代と 30 代という世代別の言語使用を見ることの意義がないものになる．そこで 10 歳以上の年の差がある正確な世代別の言語使用を考察するため，〈20 代初〉，〈30 代初〉，〈40 代初〉で年齢の範囲を定めているのである.

〈会話の構成〉

1. 2 人の会話：

   話し手と聞き手の常に 2 名が維持される会話を獲得するためである.

2. 〈初対面同士の会話〉と〈友人同士の会話〉の場面別会話：

   言語使用の変化に大きな影響を与えると推測される親疎関係，親密度に変化をもたらすことで，親疎関係の違いによる言語使用の差異，とりわけ文末の言語表現の差異を計量することを可能にするためである.

3. 〈初対面同士の会話〉における〈同年齢の会話〉と，10 歳以上の年の差を持つ〈目下との会話〉と〈目上との会話〉：

   親疎の関係が固定されている場合の，年齢と性別の差による言語使用の差異，とりわけ文末の言語表現の違いを見るためである．友人同士の会話ではなく，初対面同士の会話に年齢の差を置いた会話を構成するのは，他の条件に影響されず，純粋な年齢や性別の差による言語使用の違いを見るためである．なお，〈初対面同士の会話〉のみこの条件を設定し，〈友人同士の会話〉に設定しないのは，友人会話では，知り合い同士であるため，年齢の差よりも，親密度の度合いや社会的地位などの他の諸条件が言語使用により大きい影響を与え得ると思われるからである.

2.2.3. 会話協力者の条件

会話協力者は次の条件で制限し，選定した：

表 4 会話協力者の選定条件

	日本語会話	韓国語会話
母語	日本語母語話者	韓国語母語話者
方言	東京方言話者	ソウル方言話者
言語形成地	東京，埼玉県，神奈川県，千葉県	ソウル[14]
学歴	短期大学以上	専門大学(日本の短期大学に準じる)以上

まず，日本語会話は日本語母語話者，韓国語会話は韓国語母語話者の協力者を選ぶ．また，各言語の方言の違いから来る言語使用の違いを防ぐため，日本語は共通語の基礎を成すと言われる東京方言の話者，韓国語は標準語の基礎を成すソウル方言の話者に限定する．生まれと育った地域，すなわち言語形成期を過ごした地域が日本語母語話者は東京，埼玉県，神奈川県，千葉県に，韓国語母語話者はソウルに，会話協力者を限定するが，生まれた地域が他の地域であっても，4 歳，5 歳，6 歳から言語形成時期を東京方言の地域とソウル方言の地域に在住であった協力者も上記の条件を満たすものとする．また言語使用において，ある程度のいわゆる教養を身につけている協力者を選定するため，短期大学卒業以上の学歴を条件としている．

会話協力者の年齢は次の通りである．韓国では一般的には数え年が適用されるが，日本語母語話者と韓国語母語話者の年齢を統一するため，ここではいずれも満年齢に統一する：

表 5 会話協力者の年齢

20 代初め	20 歳～23 歳まで
30 代初め	30 歳～33 歳まで
40 代初め	40 歳～43 歳まで

2.2.4. 会話の構成

上記の条件を満たす会話協力者を〈初対面同士の会話〉と，〈友人同士の会話〉の 2 種類の場面に分ける．年齢別と性別に組み合わせ，2 人 1 組とする．初対面の会話は同じ年齢の 2 人の会話と，30 代を中心に 10 歳以上の差を置いた目上との会話，そして目下との会話に構成される．友人同士の会話は 20 代と 30 代の同年齢の 2 人の会話に構

[14] 1 名のみ水原．なお，筆者の判断ではソウル方言と全く区別がつかない話者であった．

成する．会話の組み合わせを以下の表で示す．日本語と韓国語とも，下記の表の会話の組み合わせが各 2 組ずつ構成されている：

表 6 初対面の会話の組み合わせ：目上，目下との会話と同年齢の会話を別々に提示

目上と目下との会話	
20 代 － 30 代	30 代 － 40 代
男 － 男	男 － 男
男 － 女	男 － 女
女 － 男	女 － 男
女 － 女	女 － 女

同年齢同士の会話	
20 代 － 20 代	30 代 － 30 代
男 － 男	男 － 男
男 － 女	男 － 女
女 － 女	女 － 女

表 7 初対面の会話の組み合わせ：目上，目下との会話と同年齢の会話を 1 つの表に提示

		20 代		30 代		40 代	
		男性	女性	男性	女性	男性	女性
20 代	男性	●					
	女性	●	●				
30 代	男性	●	●	●		●	●
	女性	●	●	●	●	●	●

表 8 友人同士の会話の組み合わせ

20 代 － 20 代	30 代 － 30 代
男 － 男	男 － 男
男 － 女	男 － 女
女 － 女	女 － 女

こうした組み合わせにより，日本語の会話 40 組と韓国語の会話 40 組，計 80 組の会話を構成した．

初対面の関係の会話と友人関係の会話において，20 代同士と 30 代同士の同年齢同士は，生まれた年が同じである協力者同士である．こうした設定は韓国の場合，生まれた年が 1 年でも早ければ，目上の人として待遇する習慣があるからである．こうした点に細心な注意を払い，両言語において同年齢同士の会話においては 1 歳の年の差もないように設定した．

また，初対面の会話において，20 代と 30 代，30 代と 40 代の 10 歳以上の年齢の差を持つ 2 人の会話を設定したのは，初対面の同じ年齢の相手に対する言語使用と目上の相手に対する言語使用を，それぞれ比較，考察するためである．

2.2.5. 会話協力者の異なり人数

会話に参加している協力者の数は，のべ人数ではなく，すべて異なり人数である．同一の協力者が他の会話に重ねて参加することはない．

会話に参加した協力者の，年齢別，性別の異なり人数の分布は以下の通りである：

表 9 日本語の会話協力者数

	初対面の会話 28 組		友人会話 12 組		計 40 組
	男性	女性	男性	女性	
20 代	10 人	10 人	6 人	6 人	32 人
30 代	14 人	14 人	6 人	6 人	40 人
40 代	4 人	4 人	なし		8 人
計	28 人	28 人	12 人	12 人	80 人

表 10 韓国語の会話協力者数

	初対面の会話 28 組		友人会話 12 組		計 40 組
	男性	女性	男性	女性	
20 代	10 人	10 人	6 人	6 人	32 人
30 代	14 人	14 人	6 人	6 人	40 人
40 代	4 人	4 人	なし		8 人
計	28 人	28 人	12 人	12 人	80 人

日本語母語話者と韓国語母語話者を合わせ，異なり人数 160 人の協力者[15]が参加している．協力者の出身地別の分布を以下の表で示す：

表 11 会話協力者の出身地別分布

日本語母語話者	
言語形成地	人数
東京	54 人
埼玉県	9 人
神奈川県	7 人
千葉県	10 人

15 このうち，日本語母語話者の中で，30 代の男性 1 人が生まれは東京であるが，4 歳から 10 歳まで福岡に在住である．韓国語母語話者のうち，40 代の男性 1 人の出生が忠清道で，6 歳からソウル在住，20 代の男性 2 人が出生が全羅道，4 歳と 5 歳からソウル在住である．この他の日本語母語話者 79 人は，東京，埼玉県，神奈川県の生まれと生育であり，韓国語母語話者 77 人はソウル，1 人が水原の生まれで育ちである．

韓国語母語話者	
言語形成地	人数
ソウル	79 人
水原	1 人

2.2.6. 会話の録音および録画

上記の組み合わせによる 80 組の会話の録音および録画に関する詳細を以下の表で示す[16]：

表 12 録画および録画の諸条件

	日本語の会話	韓国語の会話
地域	日本の東京	韓国のソウル
場所	大学の研究室．会社の会議室． 会話協力者の自宅	大学の研究室．会社の会議室，接客室． レストランの個室
時間	15 分間	15 分間
機器	Sony MD Recoder MZ-B100 SONY Digital Video Camera Recorder DCR-PC110 SONY DAT TCD-D10	Sony MD Recoder MZ-B100 SONY Digital Video Camera Recorder DCR-PC110
日時	2003 年 6 月：20 代友人同士の会話 3 組 2003 年 9 月：30 代友人同士の会話 3 組 2003 年 11 月～2004 年 3 月： 初対面同士の会話 22 組 2004 年 5 月：20 代友人同士の会話 3 組 2004 年 6 月：20 代初対面同士の会話 6 組 2004 年 12 月：30 代友人同士の会話 3 組	2003 年 8 月：20 代友人関係会話 3 組 2004 年 11 月：友人同士の会話 9 組 初対面同士の会話 28 組

会話の録音の進行は宇佐美まゆみ (1997) に倣う．録音を行う前に名前，性別，生年月日，出身地などを問う簡単な「フェイスシート」を記入してもらう．また初対面の会話協力者に，話題は自由に話すように指示した．友人関係の会話協力者は大学の研究室や自宅，レストランなどで会ったときのように，普段おしゃべりするような雰囲気で話すように指示した．

会話の録音は実験者不在で 15 分間行い，Sony MD Recoder MZ-B100，および SONY DAT TCD-D10 で録音，SONY Digital Video Camera Recorder DCR-PC110 で録画を行った．Recorder MD を用いて行った，録音データは，静かで外の雑音が入らない，リラックスできる場所で録音を行ったため，音の判別，発話の判別など，言語研究のデータとしては全く問題のない，きれいに音声が識別できるものばかりである．

16 録音機材は 2 機以上を常に併用した．

会話の録音，録画の前，話者の名前，年齢，性別，言語使用の状況などを，フェイスシートに記入してもらい，録音終了後，アンケートを行った．フェイスシートとアンケートは，会話の分析を行う際に研究者が 2 次的な会話データとして用いるものである．録音そのもの，及びフェイスシート，アンケートは個人情報などを守るため，公開はしない．

録音終了後，会話の感想を尋ねるアンケートでは，「録音，録画はどのぐらい意識したのか」などを，5 段階で評価してもらった．アンケートのうち，一例を上げる：

●録音，録画はどのぐらい意識しましたか？
1. 全く意識しなかった　2. ほとんど意識しなかった 3. やや意識した
4. かなり意識した　5. 非常に意識した
●上の質問に 4，5 で答えた方のみ答えてください．
録音，録画を意識することで，どのぐらい話せましたか？
1. 全く自然に話せた　2. かなり自然に話せた 3. まあまあ自然に話せた
4. かなり話せなかった　5. 非常に話せなかった

「録音，録画はどのぐらい意識したのか」の質問に対しは，録音を意識し，「全く話せなかった」と「かなり話せなかった」の 5，4 のレベルにチェックされた会話は，他の対話者との組み合わせを変えた上での再録音を行うことを前提とする．録音，録画の機械を意識し，話者の普段通りの話し方ができなかった場合は，「自由会話」として認められないためである．

しかし，本稿のすべての会話の録音において，2 つの質問に対して最も悪い 5，4 のレベルにチェックされている会話は 1 つもなかったため，再録音は行っていない．

2.2.7. 談話データの構築量

80 組の会話においてそれぞれ 15 分間，会話を録音し，会話最初の部分からの 5 分間を文字化する．音声データと文字化データの量(単位:分)を以下の表に示す：

表 13 音声データの量(単位:分)

	音声データ		計
	日本語会話	韓国語会話	
初対面同士の会話	420 分	420 分	840 分
友人同士の会話	180 分	180 分	360 分
計	600 分	600 分	1200 分

表 14 文字化データの量(単位:分)

	文字化データ		計
	日本語会話	韓国語会話	
初対面同士の会話	140 分	140 分	280 分
友人同士の会話	60 分	60 分	120 分
計	200 分	200 分	400 分

こうして得られた談話データを，以下の文字化システムに従い，文字化を行う．

以下，「文字化データ」は，録音した会話の文字起しの作業のみならず，コンピュータで処理できるよう，タグを付する作業を経たデータである．

2.3. 文字化

〈話されたことば〉の談話を研究するためには，線条性を保ちながら，瞬間的に消え去る話されたことばを，視覚化し，留めて置かなければならない．こうした〈話されたことば〉の視覚化は，談話の文字化を通してのみ得ることができる．実際に現れる言語のあり方を描写し，本質を捉えるのに，文字化はこれ無しではやり遂げることのできない，絶対不可欠な過程として存在するのである．

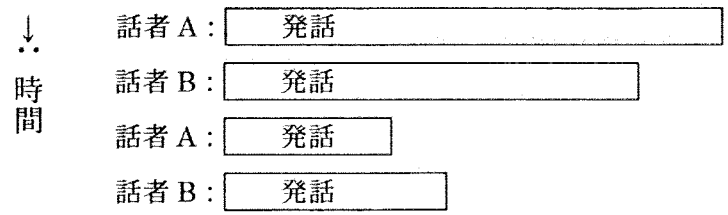
会話録音から得られたデータを，〈複線的文字化システム〉[17]に従い文字化を行う．複線的文字化システムは，文字化の方法のみならず，本稿が提示する日本語と韓国語の文字化における表記法，文字化における注釈の方法までを含めた文字化システムとして，本研究が提起するものである．

2.3.1. 複線的文字化システム

談話分析の既存の文字化システムは，発話順に上から下へ文字化していく，いわば"単線的な文字化システム"であったといえる．すなわち turn の展開を以下の図のごとく認識しているのだと言えよう．例えば，メイナード (1993)，宇佐美まゆみ(1995,2004)) などをはじめとする研究は談話を次のような型で示している：

[17] 金珍娥(2003)参照.

図 1 従来の文字化方法－単線型：垂直に表される turn の展開 (→：時間の流れ)



上の図の文字化方法は，垂直ではなく，水平に表すと，実は次の図 2 のように単線的に理解しているのだということがわかる：

図 2 単線型：水平に表わされる turn の展開 (→：時間の流れ 発話 ：turn の流れ)

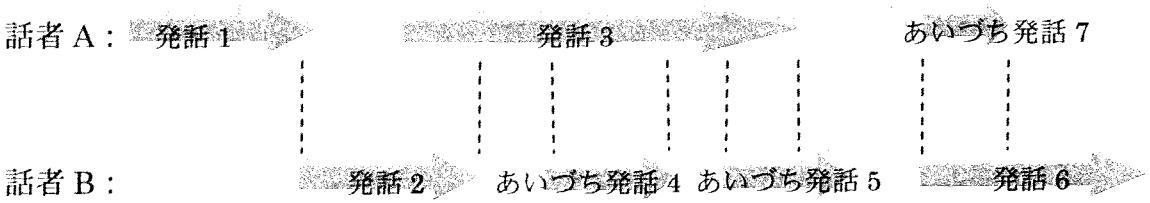
こうした文字化の方法を〈単線的文字化システム〉と呼ぶことにする．

このような単線化のシステムは，いわゆる「文」の範囲を越えられず，音声言語独特の在り方の特徴を十分に表すことができない．例えば，turn の流れを話者ごとに変わっていく型で把握すると，話者 A の発話に話者 B が重ねて発話を行ったり，あるいは話者 A の発話に割り込み，A の発話を区切りながら，「はい」と「そうですか」と言った場合，従来の文字化の方法である単線型なら，相手の turn を区切ったり，相手の発話に重なっている発話は位置づけ得ないことになる．話者交代(turn-exchanging)や割り込み，重なった発話などを明確に記述することもできない．データ処理の利便性の影に隠れた，せっかくの話されたことばならではの固有の特徴を描き出すことができない．

先行諸研究の限界を超えるためには，談話の〈流れ〉という言語の本質的な部分に目を向けざるを得ない．そこで，本研究では文字化にあたって，つとにソシュール(1940:146)によって示唆され，また，南不二男(1987:7)，ノマ ヒロキ(1996a:17)，Noma (2005:66)などが力説する言語の線条性を考慮に入れ，時間の流れと共に談話の流れが見えるよう，複数の直線の文字列による構造化を行う．

談話は常に話者が図式的に互いに交代しながら「単線的」に進んで行くものではなく，複数の話者の発話が同時に進んだり，途中で重なったりしながら，並行して「複線的」に進んでゆくものである．まさにこうした複線的な構造こそが談話，とりわけ自由会話のような談話の本質的な在り方なのだと言わねばならない．単線的文字化システムは談話の構造のいわば最も大切なものを見失ってきたのである：

図 3 複線的文字化システムの turn の展開－複線型　(→：時間　発話　：turn の流れ)



このように〈話されたことば〉の本質を考え，録音した談話データの発話は複線的に文字化をおこなう．こうした文字化システムを**〈複線的文字化システム〉**と呼んでおく．

どの発話と相手の発話が重複しているのか，どの発話で割り込まれているのかなどを，こうした複線的文字化システムは明確に判断することができる．

時間軸に沿って横へ直線に繰り広げる文字列の複線的な構造化は，2 人の会話だけではなく，3 人以上の話し手による談話の文字化においても，発話の重複や割り込みの位置，話者交代が行われる位置などを精密に表記することができるのである．複数の話者の発話においても，時間の流れに沿って談話の流れを同時に文字化＝構造化しうる，こうした複線的文字化システムも，本稿を含む一連の研究において提起したい試みの 1 つである．

2.3.2. 複線的文字化システムにおける日本語の表記法

日本語の表記法は，基本的にはまず「内閣告示」に従う．また，〈話されたことば〉の特徴を生かすため，新たな表記法を立てる部分もある．これらを総合すると次のごとくである：

1. 漢字仮名交じり文を原則とする．
   - 専門用語，固有名詞などに片仮名書きの習慣が強いものは片仮名を用いる．
     例：ト書き．
   - 慣用的に漢字仮名交じりと平仮名の両方表記が用いられるものは，原則的には漢字仮名交じりとする．
     例：例えば．即ち．
2. 仮名遣いは，内閣告示(1986)の「現代仮名遣い」に基本的に従う．
   - 漢字より平仮名の表記が，より慣用的になっている単語は平仮名で表す．

例：もらう．まいる．いただく．くださる．

・助動詞の類は平仮名で表す．

例：やってみる．取っておく．

・形式名詞の類は平仮名で表す．

例：こと．もの．ところ．

3. 外来語の表記は内閣告示(1991)の「外来語の表記」に基本的に従う．

・語形やその書き表わし方については，慣用が定まっているものはそれによる．

例：アジア．アパート．ガソリン．

・書き表し方にゆれがあるものや個人の名前，個別の地名，名称などは原音に最も近い，提示されている仮名で表記する．

例：Stubbs, Micle：マイケル・スタッブズ

4. 送り仮名のつけ方においては，内閣告示(1973)の「送り仮名のつけ方」に基本的に従う．

・言い間違った単語，最後まで言わず，意味の判別ができない単語などは，漢字を用いず，ひらがなで表わす．

例：がっこ，学校が．ご，ご．

5. 使用する漢字は内閣告示(1986)の「当用漢字表」を参考にする．

・簡易字体は現在慣用されているものの中から採用する．

・新字と旧字のゆれがあるものは新字を用いる．個人の名前や名称などは例外とする．

6. 数詞は「いち，に，さん，し…」の漢字語数詞はアラビア数詞の「1，2，3，4…」で表す．「ひとつ，ふたつ，みっつ…」の固有語数詞は漢字交じりの「一つ，二つ，三つ…」で表す．

7. 一般的にローマ字を用いる略語などはローマ字を用いる．

例：URL．ISBN．

8. 一般に記号を表すものも基本的には記号で表す．

例：10%．

話されたことばの特徴を生かすために次のような表記法を用いる：

1. 単語で決まっていない促音や拗音，長音が音声上入っている場合は，促音「っ」，拗音「ん」，長音「ー」で表す．伸ばす音が長い場合は「ー」を一音節の相当分で複数用いる．どれほど音を延ばしているのかを表記するためである．フィラー(間つなぎことば)や間投詞，他に強調している単語などに多く用いる：

例：えーっと．うーんとね．はっ．へーー．おもっしろい．ってゆーか．

2. 長音を表わす「う」，「い」の平仮名が発音上，消えている場合は表記しない．

例：でしょ．だろっ．

3. 長音を表わす「う」，「い」の平仮名を長音記号の「ー」で表わすことが〈話されたことば〉では一般的になっているものは，長音記号の「ー」を用いる．

例：ちょーおかしいでしょ，

4. 書きことばとは異なる形で現れる，いわゆる話しことばの形は，異形態もしくは短縮形として認め，文字化を行う：

表 15 日本語の標準語形と異形態

標準語形	異形態
してしまった．	しちゃった．
いいじゃない．	いいじゃん．
というか．	ちゅーか．てゅーか．てゆーか．ってゆーか．
いいです．	いいっす．
みせて．	みして
やってくるの．	やってくんの
明日かもしれない．	明日かもしんない
だめになってしまう．	だめんなっちゃう．
ほんとうは．	ほんとは
うん．	ん．う．
そう．	そ．そっ．

5. 「笑い」はあいづち同様の言語表現[18]として扱う．原音に最も近いつづりで平仮名で表記する．「はは」の場合は応答詞の「はい」を「は」という発話もあるので，区別を行うため，すべての笑い声には後に「(笑)」を表記する．

例：ははは(笑)．へへへ(笑)．くくく(笑)．

6. 息を吸う音や舌を打つ音などは準言語音[19]として認め，原音に最も近いつづりで平仮名で表記する．後部に「(息)」と「舌打」と表記する．

例：すー(息)．つっ(舌打)

7. 咳，あくびなどの音は言語音として認めない[20]ので，「(咳)」，「(あくび)」と表記する．

18 本稿の 4.1.2 の項を参照．
19 本稿の 4.1.2 の項を参照．
20 本稿の 4.1.2 の項を参照．

2.3.3. 複線的文字化システムにおける韓国語の表記法

韓国語の表記法は，基本的には韓国の文教部公示が制定した「正書法」，「標準語規定」，「外来語表記法」に従う．また，話されたことばの特徴を生かすため，新たな表記法を立てることもある．総合すると次のごとくである：

1. 漢字は用いず，ハングルで記すことを原則にする．
2. 正書法に合わせて「分ち書き」を行う．
3. 外来語は「外来語表記法」に従うことを原則にする．
   - 表記法に定まっていない，団体名，商品名などは原音に近く記す[21]．
     例：아르방스，르나포프
   - 一般的にローマ字を用いる略語などはローマ字を用いる．
     例：URL. ISBN.
   - 一般に記号を表すものも基本的には記号で表わす．
     例：퍼센트는 10%.（프로はハングルで 10 프로と記す）
4. 数字は，漢字語数詞はアラビア数字で，固有語数詞はハングルで表す．
   - 공일학번，칠삼년생　など漢字語数詞と読み方が異なる場合はハングルで記す．
   - 유월，시월は漢字語数詞と読み方が異なるので，月を表す일월，이월　などはハングルで記す
5. 表記と発音が異なるもののうち，濃音化，激音化，口音に鼻音化，終声の初声化，n の挿入など発音変化の規則によるものは正書法に従い表記する：
   例：갈 거에요.［갈 꺼에요］(行きますよ．)
   제가 할게요.［제가 할께요］(私がします．)
   어제 거로 주세요.［어제 꺼로 주세요］(昨日と同じものでください．)
   막막하죠.［망마카죠］(気が遠くなりますよ．)
   이렇게［이러케］(このように)
   문이 안 열려요.［무니 안 녈려요］(ドアが開けません．)
6. 해요 [hɛjo] 体の語尾-요 [jo]，丁寧化の語尾-요 [jo] /-이요 [ijo] は唇の狭く，突き出した[요 jo]で発音されることはなく，唇の広い[여 jɔ]で発音されるのがほとんどである．これらは正書法に従い'요'で記す．
   例：집에요?(家にですか．)

---

21 「外来語表記法」の原則は原音主義（英語：프라그(プラグ)，チェコ：프라하(プラハ)）であるが，本稿では話者の原音に近く表記する．

어제 왔었어요.(昨日来ました.)

아니요.(いいえ/違います)

なお，아니요は아니の敬体である．아니요は 3 音節で発音するものは아니요，2 音節で発音するものは아뇨と表記する．

話されたことばの特徴を生かすため，次のような新たな表記法も立てる：

1. 助詞や語尾の母音ㅗ[o]が[u]に発音されるものは，発音通りㅜ[u]に記す．

   例：밥은 먹구(먹고) 왔어요.(ご飯は食べて来ました.)

   저두.(저도)(私も)

   집으루 가요.(집으로)(家へ)

2. 助詞-는の-ㄴへの縮約，母音 이の縮約など，縮約による発音の変化は発音通りに記す．

   例：그게 어딨는(어디 있는) 건가.(あれはどこにあるんだろう.)

   재밌어요.(재미있어요)(面白いですよ.)

3. 母音ㅟ[wi]と母音ㅓ[ɔ]が合さって母音が短縮された[ujɔ]は，コンピュータ上では処理できない文字であるため，「ㅜㅕ[22]」で表わすことにする．

   例：수ㅕ요.(쉬어요)(休んでいます.)

   바꼈ㅕ어요.(바뀌었어요) (変わりました.)

4. ソウル方言では一般に에と애は区別されず，[ę]程度に発音される．ここでの表記は基本的には正書法に従う．

5. 次のものは異形態として認め，それぞれ異形態の形で文字化を行う：

表 16 韓国語の標準語形と異形態

	標準語形	異形態
用言	같아요	같애요
	그렇지, 그렇죠	그쵸, 그죠, 그치, 글치, 그릇치
	기다렸어	기달렸어
	쉬었거든요	슀ㅕ거든요
	사귀어요	사겨요
	예쁘다	이쁘다
	재미있어서	재밌어서
	(힘) 세지	쎄지

22 「21 世紀世宗計画国語特殊資料構築」(1998)では，「쉬'어요」のように「'」で表わしているが，本稿では過剰な記号使用を避けるため，ハングル字母で表わすことにする．

	낳다	나요, 났어요, 나서, 나나야지
	놓아	나
	가르쳐	가르켜, 가리켜[23]
	줄였으면	쫄였으면
	가지고	가지구, 갖구
	갖고	갖구
接続詞	그래 가지고	그래 가지구,글 갖구
	그래 갖고	그래 갖구
	그리고	그리구, 그러구
	그러면	그믐, 그믄, 금
	그러니까	그니까, 그닌까, 그니깐, 그닌깐, 근까, 긍까,닌까, 니까, 그르니까, 그르닝까
	뭐라고 그러지	모라 글지
	아니면	아님
	근데	은데, ㄴ데, 근,
語尾	-더라도	-더래도
	-더라니까	-더래니까, 드래니까
	-이라던지	-이래든지
	-니까	-닌까, -니깐, -닌깐,
	-거든요	-거던요
	하려고요	할려구요, 할라구요
	할까	하까
	한다던가	한데던가[24]
	하나 보지?	하나 부지?
	어떡하냐.	어떡허냐
副詞	좀, 조금	쫌, 쪼끔, 쬐끔, 쪼금
	되게	디게
	뭐	모, 머
	제일	젤
	때문에	땜에
	다음에	담에
	그냥	기냥
	다른	따른
	처음부터	첨부터
	이런 데	요론 데
	아무튼	암튼
	너무	넘

23 가리키다は標準語では「指す」の意で，가르치다は「教える」の意で用いる．

24 日本語では「するとか」の意．引用の「するという言ったんだ」の意は한대던가で記す．

	별로	별루
	확실히	학실히
	이렇게	이케, 힉케, ㅎ케[25]
名詞	저희	저이
	갓난쟁이	깐난쟁이
	성질이	승질이
Umlaut	먹이다	멕이다
	맡기다	맽기다

6. 間投詞の類は発音に最も近い表記を行う.

あいづち詞：아. 어. 에. 네. 예. 녜. 아휴. 아유.허ㄱ. 헤. 하. 호.

7. 「笑い」はあいづち同様の言語表現として扱う. 原音に最も近いつづりで平仮名で表記する. 後部に「(笑)」を表記する.

例：하하하(笑). 헤헤(笑). ㅋㅋㅋ(笑).

8. 息を吸う音や舌を打つ音などの準言語音として認め, 原音に最も近いつづりで平仮名で表記する. 後部に「(息)」と「(舌打)」と表記する.

例：스읍(息). 스(息). 쯔읍(息). 쯧(舌打). 쯧(舌打). 쯧쯧(舌打)

9. 咳, あくびなどの音は言語音として認めないので, 「(咳)」, 「(あくび)」と表記する.

10. 伸ばす音が長い場合は「ー」を一音節の相当分で複数用いる. どれほど音を延ばしているのかを表記するためである.

2.3.4. 文字化の注釈方法

文字化に必要な記号類は基本的にすべて1バイトの半角ASCII文字を用いる.

1. 文の終わりに用いる記号は次の通りである：

表17 文の終わりに付すタグ

順序	1	2	3	4
記号	~ イントネーションが上がった文	? 疑問文	; 相手の発話, ポーズによる文の切れ目. 発話単位	. 文の終り

表で提示した機能は, 上の表の「順序」の順に各記号を付する：

25 힉あるいは無声化がさらに進み, 母音が脱落し, [hk^{h}e]となったもの.

① イントネーションが上がった疑問文：

いつですか~?.

② イントネーションが上がらない疑問文：

何してるの?.

③ 叙述文で終わっている文：

私ですよ.

④ 1 つの文の中に 2 秒以内のポーズが生じ，発話単位として分かれている場合：

A:日曜日;　アイちゃんから電話があって;　東京に戻ったんだって.
B：　うん.　あー.

⑤ 〈文〉の終わりにはすべて「.」を付す．結果的に「.」の数はすべての文の数と一致する.

⑥ あいづちを含む間投詞の類のみで終っている発話も，1 つの文として判定し，発話の後ろに「.」を付する.

⑦ あいづち，笑い声は文として認め，「.」をつける．しかし，準言語音である「息を吸う音」や「舌を立たす音」など (表記を参照) は文として認めず，「.」はつけない.

⑧ 文中で，確認などのためイントネーションをあげる場合は，疑問文ではないので「?」は用いず，該当の単語の後ろに「~」をつける:

例: 父の故郷が仙台~ですから.

⑨ 相手の発話より話し手の文が非意図的に終了した文には「;」記号を記す：

例： A：それだったら;.
B：　その感覚があるんですよ.

⑩ 助詞や語尾など，文法的には前の要素にくっつき，離せないところで，音声が切れて，ポーズが入る場合がある．こうした場合は「ˆ」で表し，音声が切れ，ポーズが入っていることを示す.

例：それが仕事;ˆで.

2. 次の場合に「,」を用いる

・倒置文[26]の前：

例：いつ帰るんですか,韓国に.

・言い直しの前：

例：もうおわ,終わりですか?.

・項目やことがらの列挙の後ろ：

26 第 4 章の「倒置文」の項を参照.

例：家に帰るか,食事に行くか,映画に行くか,早く決めてよ.

・文頭や文中のあいづちなど，間投詞の前後：

例： A:あ,そうですね.
B： 名前が,え,え,面白いですね.

・文中での笑い声のその前後：

例：徹夜して,はは(笑),そのまま帰っちゃったの.

3. 文の中に「直接引用」が含まれた場合は「‘ ’」で括る.

例：お母さんが‘早く今帰ってきて’っていうんだもん.

「間接引用」においては韓国語ではあいまいであり，その形を定めるのが非常に困難である．本稿では日本語と韓国語を統一させるため，明らかな直接引用にのみ記号を付す.

4. 本，映画などの題名には「“ ”」を用いる.

例：“世界の中心で愛を叫ぶ”見た?

5. 音を引いたり，延ばした場合は「-」を用いる．引く時間が長い場合は発話の音節数ほど「ーー」のように用いる．その後，文が終了した場合は「.」をつける.

例： A：まったく新しいっていうーー. あのー，ちょっとー.
B： あーー,そうなんですか.

「それでー」，「行きましてー」などは，音を引く現象として判断する．しかし，それが文を終わらせる「言い淀み」なのかどうかは話者でなければ，分析者には客観的かつ明確には区別できない．音を延ばすことは明確に判断できるので，「音を延ばす」現象として記号を記する.

6. 聞き取り不能の箇所は発話の長さに合わせて「***」を用いる.

7. 笑いながら話している発話，すなわち笑いを含んでいる単位には末尾に「@」をつける．文の最後まで笑いが含まれている場合は「@」の後,「.」をつける.

例：それは変じゃないって言っても@,うん,だめですね.

例：昨日行ったところもまったく同じだったよ@.

8. 音声で認識できる笑い声はその音声に最も近い表記をする．この場合，後ろに「.」をつけ，1つの文として扱う：

例： A：そんなことありえない． ははははは(笑).
B: ははははは(笑).

9. 発話と発話の間のポーズが2秒以内の場合は上の 1. で提示している「；」を付する．ポーズが3秒以上になると，「(…秒沈黙)」と記する.

10. 本稿の文字化の特徴は「複線的システム」であり，重複，ラッチング(音声的に間がなく,つながって発話される 2 つの文)などは記号ではなく，視覚的に表す：

［例］

30 代男	あ, そうですか.　　いろいろそうですか. あ, まあだったら. うーーん.
30 代男	都内ん中ででいろいろ行ってますけど. はは(笑).　行ってますけど.

11. 論文上で引用し，公開される発話の部分においては，会話協力者のプライバシーのため，「30 代男」，「20 代女」のように世代と性別の属性で表す．会話の中で現れる人名などは仮名を用いる.