2015 年 3 月 31 日公開

『日中 Skype 会話コーパス』について

中俣尚己

目次

１．『日中 Skype 会話コーパス』の概要

『日中 Skype 会話コーパス』は 2012 年 5 月から 2012 年 7 月にかけて，東京・実践女子大学と長沙・湖南大学で行った Skype 会話交流活動の会話を録音，文字化したものである。参加者の許諾を得た上で，文字データを公開する(http://nakamata.info/database.html)。

中国側の参加者は日本語学科の 2 年生の有志，日本側の参加者は国文学科の学部生と大学院生の有志である。固定的なペアを作り，期間中，週に 1 度のペースで，90 分を上限として日本語での会話を行った。

のべ 9 ペア，38 会話の会話を収録し，合計録音時間は 46 時間 48 分 35 秒となる。１会話あたりの時間は 1 時間 13 分 55 秒であるが，Skype を通した接触場面であるため，母語話者の対面場面よりもかなりゆっくりとしたペースで会話が進行していることには注意されたい。また，Skype における通信トラブルなどもそのまま収録している。

本コーパスの最大の特徴は，あくまでも会話交流を行うことが目的であるという真正性の高いコーパスであるという点である。また，各回ごとにトピックの指定を行っているが，それが厳密に守られているわけではなく，話題は次々と遷移していくので，接触場面の雑談データとしても利用できる。

なお，本交流活動の詳細については中俣尚己・漆田彩・小野真依子・北見友香・竹原英里(2013)「Skype を活用した日中会話交流プログラム」（『実践国文学』83 号 pp.132(25)-109(48)）にて実践報告を行っている。

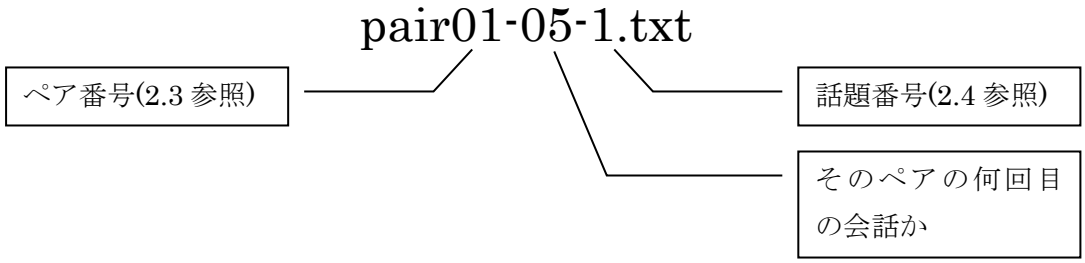
２．ファイルとファイル名について

2.1　ファイルについて

コーパスのデータは１つの会話ごとに１つのテキストファイルで提供する。文字コードは S-JIS である。

2.2　ファイルの命名規則

ファイルの命名規則は下記の通りである



2.3　参加者

参加者は下記の通りである。なお，実際の会話活動は各ペア 4 回～10 回行っているが，機器の不調で録音できなかった回もあるため，コーパスに収録した会話数はそれよりも少ない。

表１　参加者の一覧

ペア番号	日本側	中国側	収録会話数
1	修士１年生Ａ	学部２年生Ａ	3
2	学部３年生Ｂ	学部２年生Ｂ	8
3	修士１年生Ｃ	学部２年生Ｃ	6
4	学部３年生Ｄ	学部２年生Ｄ	7
5	修士１年生Ａ	学部２年生Ｅ	2
6	修士１年生Ｅ	学部２年生Ｆ	3
7	学部３年生Ｆ	学部２年生Ｇ	5
8	修士１年生Ｅ	学部２年生Ｈ	3
9	学部３年生Ｇ	学部２年生Ｉ	1

ペア番号１と５の日本人学生は同一人物である。また，ペア番号６と８の日本人学生も同一人物である。

2.4　話題

ファイル名の話題番号と事前に指示したトピックの関係は下記の通りとなっている。ただし，実際には順番を間違えて会話をしていることもあり，中身を確認した上で改めてトピック番号を付与している。また，後半の回では中国側の要望で類義語についての質問を受け付ける時間も設けており，それが含まれていることもある。

表２　話題と話題番号

1	ポップカルチャー	6	伝統・行事
2	料理	7	夏休み・夏の予定
3	家庭・家族・子供	8	大学生活
4	故郷あるいは今住んでいる場所	0	指定なし・トピック認定できず
5	敬語		

３．文字化の記号など

話者は日本側がＪ，中国側がＣであり，発話前に「Ｊ：」のように記してある。また，会話中に出現する固有名詞も「Ｊさん」のようにしている。会話参加者以外の名前が会話中に存在する場合，それが日本人名なら「Ｊ１さん」「Ｊ２さん」のように連番で付与し，中国人名なら「Ｃ１さん」「Ｃ２さん」のように連番で付与した。

また，参加者以外の人間（ルームメイトなど）が突発的に会話に入り込んでくることがある。その場合も同様にＣ１，Ｊ２のように記してある。また，どの話者か同定できない時は「？：」のように記してある。「：」は大文字で，他の箇所には使われていないので，

形態素解析などを行う場合は，正規表現で「^.+：」を置換すれば文頭の記号は一括で消せる。

また，適宜タイムスタンプを入れてある。こちらは例えば「^[0-9].+$」などの正規表現を使えば一括で消せる。その他，形態素解析を行うことを前提に，最低限の記号を付与した。記号類は全て全角である。

表３　記号について

●	聞き取れなかった箇所もしくは個人情報のため削除した箇所。●の数がモーラ数に対応。【中国語】など説明をつけた箇所もあり。正規表現は「●」「【.+?】」で削除可能。
［	［以下の発話は次の行の［以下と重なって発話されたことを表す。「[]で削除可能。
《ポーズ　３秒》	ポーズは２秒以上のものを記録した。また，Skype の通信が中断した場合は《中断》としてある。「《.+?》」で削除可能。
？	上昇イントネーション。
＜うん＞	他者の発話途中の短いあいづちなどは＜　＞で挟んでいる。「＜.+?＞」で削除可能。
{笑}	笑い声。{.+?} で削除可能。
（じゅうご）	発音のミスなどは正しい発音の後に（　）で挟んで示した。「(.+?)」で削除可能。
=ノカ=	聞き取りが可能だが不明瞭である箇所は= で挟んで示した。「=.+?=」で削除可能。

４．　形態素数など（参考情報）

本コーパスに対して、形態素解析機 Mecab　ver0.996 と電子化辞書 UniDic2.1.2 を用いて形態素解析した結果を参考までに下表に示す。分析は前節の記号のうち、●と？を除いたものを、前節の表内のやり方で削除し、「茶まめ」を用いて行った。また、日本人学生と中国人学生については、Grep を用いて該当の発話のみを取り出した後に記号を削除し、それぞれ形態素解析を行った。ただし、発話者不明の発話がわずかに含まれるため、日本人学生と中国人学生を足した数は全体と一致しない。なお、（　）内の数字は品詞情報が「空白」または「補助記号」に属するものを除いた数字である。

表４　形態素数

日本人学生	中国人学生	全体
171,262 (115,893)	152,480 (113,478)	324,716 (230,074)

謝辞

本コーパスの作成には日本学術振興会の科研費若手研究(B)「縦断型接触場面コーパスの構築とそれを用いた日本語教育のための談話研究（課題番号 26770180、研究代表者中俣尚己)」の助成を受けました。

また、Skype 会話活動の実践は湖南大学の楊昉氏と協同で行いました。楊昉氏の尽力に厚く感謝いたします。また、実践にあたっては湖南大学日本語学科長の张佩霞先生と、実践女子大学国文学科主任（当時）山内博之先生に便宜を図って頂きました。心より御礼申し上げます。山内先生にはコーパスの公開に関してもご助言を頂きました。

なお、本コーパスに関する一切の責任の所在は作成者である中俣尚己にあります。

（付録）『日中 Skype 会話コーパス』利用規約

１．制作者

『日中 Skype 会話コーパス』の制作者は中俣尚己で、公開・配布などの権利は制作者に帰属します。

２．利用範囲

『日中 Skype 会話コーパス』は研究・教育を目的とする個人のみ、自由に利用することができます。

３．譲渡・貸与・複製の禁止

『日中 Skype 会話コーパス』の一部もしくは全部を、制作者に無断で、他人に譲渡したり、貸与したり、複製したり、 公開したりすることを禁止します。ファイルの URL やパスワードを他者に教えることも、複製と見なします。 コーパスの内容を閲覧する個人個人が必ず登録を行って下さい。

４．研究成果の公開

本コーパスの全部または一部を用いた研究を行い、それを公開する時は『日中 Skype 会話コーパス』を利用したことを明記するとともに、 配布ページの URL も記載して下さい。(http://nakamata.info/database.html)　また、発表後または公開後で構いませんので、 原稿のコピーなどを制作者にお送り下さい。

５．個人情報の扱いについて

科研費による成果であり、利用実態を記録するため、また、パスワードの配布のため、 利用には氏名とメールアドレスが必要です。これらの情報は適切に管理し、本コーパスに関する連絡以外には一切使用しません。