JAISTタグ付き自由対話コーパス

作成者
北陸先端科学技術大学院大学(JAIST) 白井研究室

1 概要

JAISTタグ付き自由対話コーパスは、人間同士の雑談における発話に対し、対話行為ならびに共感をタグ付けしたデータです。対話行為とは、話者の意図による発話の分類です。本コーパスでは「自己開示」「質問(YesNo)」「質問(What)」「応答(YesNo)」「応答(平叙)」「あいづち」「フィラー」「確認」「要求」の9種類の対話行為が付与されています。一方、共感は、ここでは相手に対する発話者の共感・非共感の有無による発話の分類を表わします。本コーパスでは「共感」「非共感」「その他」の3種類のタグが付与されています。

本コーパスは、自由対話を書き起こしたテキスト、具体的には国立国語研究所で公開されている『名大会話コーパス』の一部の対話に対してタグを付与しています。対話数は97、発話数は92,020です。

本データで公開しているのはタグの情報のみであり、元のテキストは含まれていません。テキストを含めた完全なコーパスを復元するためには、別途『名大会話コーパス』(無料で入手可能)を用意する必要があります。

名大会話コーパスについては以下のウェブページをご覧下さい。
http://mmsrv.ninjal.ac.jp/nucc/

2 ファイル一覧

Readme.txt
  このファイル
Readme_sjis.txt
  Readme.txtと同じ(文字コードはShift-JIS)
how_to_reconstruct.txt
  コーパスの復元方法について
how_to_reconstruct_sjis.txt
  how_to_reconstruct.txtと同じ(文字コードはShift-JIS)
recon_corpus.pl
  コーパス復元ツール
annotation/
  タグのデータを含むディレクトリ

3 コーパスの詳細

本コーパスは97個のファイルから構成されています。1つのファイルは『名大会話コーパス』に収録されている1つの対話に対応します。『名大会話コーパス』では129件の対話が収録されていますが、本コーパスではこのうち話者の数が2名である対話のみに対してタグ付けしています。

各ファイルはタブ区切りテキストになっています。文字コードはUTF-8です。フォーマットは以下の通りです。<t>はタブを表します。

  発話ID <t> 話者 <t> 発話 <t> 話者交替 <t> 対話行為 <t> 共感

以下、各フィールドについて順に説明します。

3.1 発話ID

発話のIDです。対話の先頭から順に通し番号が付与されています。

3.2 話者

話者を表すコードです。Fで始まるコードは女性、Mで始まるコードは男性を表します。話者のプロフィールは以下のウェブページで確認できます。
http://mmsrv.ninjal.ac.jp/nucc/nucc_conversant.html

本データでは原則として二者による雑談に対してタグを付与しています。
ただし、例外として、data081.txt と data128.txt は三者による雑談となっています。

また、二者以外の話者コードとして「X」が付与されていることがあります。これは、録音を担当している人や、対話の収録場所がレストランのときの店員など、雑談している人以外の人が割り込んだ発話であることを表わしています。

例外的に二者が同時に同じ内容を発話しているときがあります。
このとき、二者の話者コードをカンマ(,)でつなげて表示しています。
　(例) M015,M034

3.3 発話

話者の発話を書き起こしたテキストです。

3.4 話者交替

話者が交替しているか否かを表してます。現在の発話の話者が前の発話の話者と異なるとき(話者が交替しているとき)は1、同じとき(交替していないとき)は0となります。

3.5 対話行為

発話の対話行為です。後述の「3 対話行為」で定義された9つの対話行為のいずれかが表示されています。

一つの発話に対し複数の対話行為が該当すると考えられる場合は、対話行為をスラッシュ(/)でつなげて表示しています。

　(例) 応答(YesNo)/確認

『名大会話コーパス』では、聞きとれなかった音声は「＊＊＊」で表示されています。発話が「＊＊＊」のみであるときなど、発話の対話行為が不明であるときはアスタリスク(*)が表示されます。

3.6 共感

話者の共感・非共感の有無を表します。後述の「4 共感」で定義された3つのタグのいずれかが付与されています。

発話の共感・非共感の有無を判定できないときはアスタリスク(*)が表示されます。

3.7 発話以外のデータに対するタグ付け

『名大会話コーパス』では、
　＜録音中断＞
　＜テープ反転＞
　＜笑い＞
など、発話以外の情報が含まれていることがあります。このようなデータのときは、「話者」のフィールドは空列とし、「話者交替」「対話行為」「共感」のフィールドはいずれもアスタリスク(*)としています。

4 対話行為

本コーパスにおける対話行為は以下の9種類です。

・自己開示

自身の考えの表明や事実の列挙などの発話です。あいさつ、謝罪も自己開示に含まれるものとしています。
（例）「おもしろそう」「失礼いたしました」

・質問(YesNo)

「はい」「いいえ」などで答えられる質問をしている発話です。
（例）「間に合う？」「知ってる？」

・質問(What)

平叙文で答える必要がある質問をしている発話です。
（例）「何処いらしたの」「何で行くの？」

・応答(YesNo)

質問に対する「はい」「いいえ」などの短い肯定、否定の発話です。この対話行為は、「質問(YesNo)」「質問(What)」「確認」「要求」のいずれかの発話に応える発話にタグ付けするものとし、これら4つの対話行為を持つ発話が近くに存在しないときはタグ付けしないことにしています。また、「応答(YesNo)」は相手の「質問(YesNo)」と必ずしも対応付けてタグ付けしなくてもよい、つまり相手の発話が「質問(YesNo)」のときの応答文に必ず「応答(YesNo)」のタグを付ける必要はないとしています。
（例）「うん」「そう」

・応答(平叙)

質問に対して平叙文で答えている応答文に相当する発話です。この対話行為は、「質問(YesNo)」「質問(What)」「確認」「要求」のいずれかの発話に応える発話にタグ付けするものとし、これら4つの対話行為を持つ発話が近くに存在しないときはタグ付けしないことにしています。また、「応答(平叙)」は相手の「質問(What)」と必ずしも対応付けてタグ付けしなくてもよい、つまり相手の発話が「質問(What)」のときの応答文に必ず「応答(平叙)」のタグを付ける必要はないとしています。
（例）「え、どうもなってない」「うーん、覚えてない」

また、質問に対して、短い応答の後に平叙文が続く発話の場合、その対話行為は「応答(YesNo)」ではなく「応答(平叙)」としています。

応答(YesNo)ではなく応答(平叙)とする例
「うん、覚えてない」

・あいづち

相手の発話に対して、話の続きを促し、また、発話者が相手の話を聞いていることを示す短い発話です。
（例）「そうだよね」「うん」

・フィラー

言い淀みなど発話の合間にはさみこむ短い発話です。「でも」「それで」などの接続詞のみで意味が取れない発話はフィラーに含まれるものとしています。
（例）「うん」「あのー」

・確認

発話者がすでに知っている、またはそうだと思っていることに対して、相手にそれが正しいかを確認する発話です。または相手の発話に対して、その内容を確かめる発話です。
（例）「すごい混んでたよね」「まじですか」

・要求

相手に対して，具体的な行動を指示もしくは依頼する発話です。
（例）「帰っておいで」「ちゃんとナビしてな」

5 共感

本コーパスにおける共感のタグは以下の3つです。

・共感

相手の発話に対して共感や賛意を示す発話のとき、このタグを付与します。
（例）「そうだよね」「そらつらいね」

ただし、単に同意を示すだけの発話は共感とはみなさないものとします。

共感ではない例
「だって、帰ったらでも１１時とかじゃない。」
「あ、そか。」

・非共感

相手の発話に対して非共感や反感を示す発話のとき、このタグを付与します。
（例）「ええー」「いやー、違うと思うよ」

ただし、単に反対の事実を述べただけの発話は非共感とはみなさないものとします。

非共感ではない例
「ま、しっかりしてそうだったと言えば、そうなんだけどね。」
「でも見た目だけというのもあるしね。」

・その他

共感、非共感のいずれでもない発話のとき、このタグを付与します。

6 統計情報

対話行為のタグの分布は以下の通りです。

自己開示	53701
質問(YesNo)	6430
質問(What)	3950
応答(YesNo)	2130
応答(平叙)	7508
あいづち	9216
フィラー	4405
確認	3940
要求	751

共感のタグの分布は以下の通りです。

共感	1067
非共感	222
その他	90731

本コーパスでは、対話行為タグ、共感タグは一人の作業者が付与しています。3対話(3ファイル)についてのみ、2名の作業者がタグ付け作業を行い、両者のタグの一致率ならびにκ係数を測りました。結果は以下の通りです。

対話行為のタグ付け作業

一致率	0.77
κ係数	0.64

共感のタグ付け作業

一致率	0.28

κ 係数 0.27

7 謝辞

本データは、『名大会話コーパス』の各発話に対し、対話行為と共感のタグを
タグ付けしたものです。同コーパスを公開して下さいました姫路獨協大学の
大曾美恵子先生、名古屋大学の藤村逸子先生に深く感謝いたします。

以上