KIT Speaking Test Corpus
書き起こし・タグ付与ガイドライン
Ver. 1

1　単語の文字化の際の注意

1.1　使用可能な記号

- コンマ・ピリオド・クエスチョンマーク：[,] [.] [?]
  - 句・文字区切りは書き起こし作業者の判断で，適切な場所にコンマ・ピリオド・クエスチョンマークを挿入する。きちんとした疑問文の形になっていなくても，何かを確認したり，尋ねたりしているような文脈でかつ語尾が上昇調ならクエスチョンマークを挿入する。
- アポストロフィ：[’]
  - 所有格，助動詞の短縮形など通常アポストロフィを必要とする箇所に使用する。
- ハイフン：[-]
  - 複合名詞など必要な箇所に適宜使用する。
- ダブルクォーテーションマーク：[“ ”]
  - 映画や本のタイトル，誰かの発話の引用など，引用全般に使用する。

1.2　使用禁止の記号

- エクスクラメーションマーク：[!]
  - 文末が強調されていたり，感嘆文であっても [!] を付与してはならない。
- シングルクォーテーションマーク：[‘ ’]
  - アポストロフィとの混同を避けるために，使用を禁止する。
- コロン・セミコロン：[:] [;]
  - 通常の書き言葉ならコロン・セミコロンが置かれるような文脈でも，本書き起こしでは使用しない。
- 括弧類：[( )] [[ ]] [< >]

  - タグを付与する際を除いて，あらゆる括弧の使用を禁止する。
- 数字類
  - 数字に関しては，すべてスペルアウトすること。
- 短縮表記
  - No. (number), vol. (volume), est. (established) などの短縮表記は一切禁止とし，すべてアルファベットでスペルアウトする。

1.3 スペースの挿入

- スペースを挿入しなければならない箇所
  - 句，文区切り記号（コンマ・ピリオド，クエスチョンマーク）の直後
  - タグが隣接する場合
    - 例：<F>ah</F> <F>mm</F> <F>mhm</F>
  - アルファベットを一字ずつ発音している場合
    - ただし，TOEIC，SNS のような略語はスペースを挿入しない
- スペースを挿入してはならない箇所
  - 単語とそれを挟み込むタグの間
    - 例：<F>ah</F> <F>mm</F> <F>mhm</F>

1.4 スペリング

原則としてアメリカ英語のスペルを使用する。英語はなるべく聞き取ったとおりに書き起こす。発音が日本語英語的であっても，文脈からどの単語を発話しているのかわかればそれをスペルアウトする。また，[r] を [l] で発音するような発音上の誤りは無視してよい。

2 タグの付与方法

本書き起こしで使用するタグは，表 1 のとおりである。付与する順序は以下のとおり。

- <?><CO, TO, RE><laughter><R, SC><JP><F>want</F> ...

2.1 フィラー

- フィラーとは，次の発話を考えているときなどに発せられるつなぎの音である。代表的なフィラーとして，以下のものが挙げられる。

  (1) ah, eh, er, em, erm, mm, mhm, uhu, ah-huh, oh, hum, wow, you know, well, let me see

- 上記のどれにも該当しないような音でも，フィラーと判断したら，なるべくその音を忠実に文字化する。フィラーはすべて <F></F> タグで囲む。**複数のフィラーが連鎖し**

表 1　使用するタグ一覧

タグ	用途
<F> </F>	フィラー・あいづち・感動詞
<R> </R>	繰り返し（聞き取りに自信がある）
<R?> </R?>	繰り返し（聞き取りに自信がない）
<SC> </SC>	自己訂正（聞き取りに自信がある）
<SC?> </SC?>	自己訂正（聞き取りに自信がない）
<TO> </TO>	タイムアウト
<RE> </RE>	レコーディングエラー
<nvs> </nvs>	非言語音
<CO> </CO>	途中で中断した発話
<?> </?>	聞き取りに自信がない語
<??> </??>	まったく聞き取り不可能な語
<H pn="X"> </H>	固有名詞・差別用語など
<JP> </JP>	日本語
<.> </.>	2 秒〜3 秒のポーズ
<..> </..>	3 秒以上のポーズ
<laughter> </laughter>	笑いながらの発話

ているときは，ひとつひとつのフィラーに別々に <F></F> タグを付与する。

(2)　<F>ah</F> <F>mm</F> <F>mhm</F>

- 日本語らしいフィラーを発しているときは，以下のように <F></F> タグと <JP></JP> タグを二重で付与する。入れ子の順番は必ず <JP><F></F></JP> にすること。

  (3)　<JP><F>etto</F></JP>

- 同じフィラーが繰り返されている場合は，繰り返しタグを付与する。

2.2　繰り返し

- 同じ語・表現を繰り返している場合，最初に発せられた方に <R></R> タグを付与する。

  (4)　When <R>he</R> <R>he</R> he was a child ...

  (5)　<F>Oh</F> <R>there is a</R> there is a station and this town located between the rivers.

- 後続する語の断片にも <R></R> を付与する。

  (6)　When he <R>wa</R> <R>wa</R> was a child.

  (7)　<R>I wan</R> I want to build a kind of rotary in the station.

- 語の断片の表記は，次の単語が分かっている場合はスペルをそれに合わせて表記する。

  (8) This book is <R>inter</R> interesting.

- 音の切れる場所よっては，文字化すると表記が異なってしまうものがある。その場合は**その音に近い表記にする**。表記上は後続の語と異なるが，断片と判断できれば <R></R> タグを付与する。

  (9) I think he is a very <R>ka</R> kind person.

  (10) My <R>pe</R> parents don't allow me to live in Tokyo.

- <R></R> 内でフィラーが発生することもある。

  (11) <F>Oh</F> <R>there <F>mm</F> is a</R> there is a station and this town located between the rivers.

- <R></R> の対象になる箇所の聞き取りに自信がない場合は <R?></R?> を付与する。

2.3 言い直し・自主訂正

- **話者が最終的にある表現に決定するまでの言い淀み**部分には <SC></SC> タグを付与する。<SC></SC> タグ付与部分と後続部分は同じであってはならない。

  (12) He <SC>don't</SC> doesn't know anything about this.

  (13) <SC>He passed the exami</SC> he will pass the examination <SC>last</SC> next year.

- 次の例のように正しい表現→誤った表現の順の言い直しであっても，後ろに来る方の表現を「話者が最終的に適切だと判断した表現」であると考え，通常通り前の部分を言い淀みとし，<SC></SC> タグを付与する。

  (14) He <SC>doesn't</SC> don't know anything about this.

- 言い淀みが複数ある場合は，区切りだと判断したところで <SC></SC> タグを括り直す。

  (15) It would be a kind of trash so <SC>it's a</SC> <SC>it's waste</SC> it's a kind of waste.

- 表現 A→ 表現 B→ 表現 A のように，最初の言い淀みが，最終的に話者が適切と判断したものと同じであっても，間に別の言い淀み B が入っているため，最初の表現 A にも表現 B にも <SC></SC> が付くことになる。

  (16) He <SC>doesn't</SC> <SC>don't</SC> doesn't know anything about this.

- 単なる繰り返し（<R></R>）なのか，言い直し・自己訂正（<SC></SC>）なのか判断に迷う例については，以下のように対処すること。

- <R></R> 付与を優先させる場合：(17) は，"vividly"（副詞）の言い淀みとして "vivid"（形容詞）を発話したのか，"vivid" は "vividly" の単なる音の断片なのか判断できない。このような場合は，(17) のように原則として <R></R> 付与を優先させる。

(17) It's difficult to keep the plants <R>vivid</R> vividly.

- <SC></SC> を優先させる場合：**動詞の人称や時制，名詞の単複**に関しては，<SC></SC> を優先させる。また，"I'm" も "I am" も発音上の違いだけで，基本的な意味は変わらないため，<SC></SC> を付与する。

(18) I <SC>close</SC> closed the door when I left school.

(19) <SC>Many line</SC> many lines are passing the station so this must be a big town.

(20) <SC>I'm</SC> I am a high school student.

- 以下のように，<R></R> が <SC></SC> に内包される例もある。

(21) <SC>It's <R>pla</R> planning</SC> it's planned by my teacher.

- ただし，以下のような場合は <R></R> を <SC></SC> に内包させない。

(22) It's <R>pla</R> <SC>planning</SC> planned by my teacher.

- 以下のように，<SC></SC> が <SC></SC> に内包される例もある。(23) は "it's planned" に対する言い淀みが "it's planning" で，"it's planning" 内で "planned" から "planning" への自己訂正が入っている。

(23) <SC>It's <SC>planned</SC> planning</SC> it's planned by my teacher.

- 次のように，<SC></SC> 内でフィラーが発生することもある。

(24) <SC>It's <F>er</F> planning</SC> it's planned by my teacher.

- <SC></SC> の対象となる箇所の聞き取りに自信がない場合は <SC?></SC?> を付与する。

2.4 中断した発話

- 文が途中で終わっている発話は，全体を <CO></CO> で囲む。**</CO> タグの直後には必ずピリオドを置くこと。**

(25) <F>Oh</F> O K. So it's getting dark but is it O K for you to come out? <CO>Is that</CO>.

- 学習者はしばしば，"so" と言いかけてそのまま言葉に詰まってしまう。そのような場合も，以下のようにタグ付けする。このとき，"so" の後のフィラーは中断された発話の一部とはしないこと。

(26) <CO>So</CO>. <F>Um</F>.

2.5 タイムアウト・レコーディングエラー

- 文が回答時間切れで中断されている発話は，全体を <TO></TO> で囲む。また，受験者が録音開始前に回答を始めたために冒頭部分が録音されていない発話は，全体を <RE></RE> で囲む。

2.6 固有名詞

- 固有名詞のうち，コーパス公開時に支障が出るような箇所には，表2にしたがい固有名詞タグを付与する。“ ”の内部には，以下の表で指定する固有名詞コードからタグ付与の対象に適したものを選び，代入する。特に受験者など個人を特定できる恐れのある語には必ず付与する。主に対象となるのは個人名・学校名・企業名であるが，それらが直接言及されていなくても，前後の文脈から特定可能な場合は，適当な範囲にタグを付与する。また，差別的発言や誹謗・中傷が出現した場合も，発言全体にこのタグを付与する。タグが必要かどうか迷った場合はとりあえず付与しておく。有名人（政治家・作家・歴史上の人物・芸能人など）の名前や本・映画の題名などには必要ない（ただし，それらを対象にした誹謗・中傷には付与すること）。評価・批判と判断されるものには付与しないが，判断に迷った場合は付与すること。
  - 対象となるもの：受験者とその家族・友人の名前やニックネーム，所属する会社・学校名など，本人の特定につながると判断されるもの
  - 対象とならないもの：受験者とその関係者の特定につながらないもの（ペットの名前など）や，映画・本などのタイトル，および芸能人・作家・政治家等の有名人の名前

表2　固有名詞コード

対象	用途
1. 個人名	<H pn = “name1”>...</H>
2. 学校名	<H pn= “school name1”>...</H>
3. 会社名	<H pn= “company name1”>...</H>
4. その他	<H pn= “others1”>...</H>

- コードの最後には通し番号を付け，同じ種類コードが付く複数の語の区別を可能にしておく。一発話内で同じ語が2回以上出現する場合も想定されるが，必ず一つの語には一つの番号を付与しつづけること。一発話内で，ある語が一度しか出現せず，かつ同じ種類のコードが付く語が他に出現しない場合でも，番号を付与すること。

(27) I am studying at <H pn= “school name1”>K University</H>. Before that, I used to study at <H pn=”school name2”>F University</H>. Two years ago, I moved to <H pn=“school name1”>K university</H> to do more specific research.

2.7 聞き取りに自信がない箇所・不可能な箇所

- 文脈推測等により文字化できるが聞き取りに自信がない箇所には <?></?> を付与する。

  (28) <?>They</?> should be very beautiful.

- 何を言っているのかまったく聞き取れず，文脈推測も無理で，文字化不可能な場合は，<??></??> タグを空で付与する。

  (29) <??></??> should be very beautiful.

2.8 日本語の使用

- 日本語をそのまま使用している場合は，<JP></JP> タグを付与する。

  (30) <F>Mm</F> <R>I</R> I don't like <JP>osechi</JP>.

2.9 ポーズ

- 発話中にポーズがあれば，2 秒〜3 秒のポーズには <.></.> タグを，3 秒より長いポーズには <..></..> タグを付与する。

2.10 非言語音

- 笑い・ため息・咳・あくびなどの非言語音には，表 3 にしたがいタグを付与する。

表 3 非言語音タグ

タグ	用途
<nvs>laughter</nvs>	笑い・照れ笑い
<nvs>sigh</nvs>	ため息
<nvs>cough</nvs>	咳
<nvs>yawn</nvs>	あくび

2.11 笑いながらの発話

- 笑いながら発話している場合は，その該当範囲を <laughter></laughter> タグで囲む。

  (31) It's a kind of <laughter><JP>mama-chari</JP></laughter>.

参考文献

The NICT JLE Corpus 書き起こし・基本タグ付与ガイドライン ver.2.1.3 (https://alaginrc.nict.go.jp/nict_jle/src/readme_transcription.pdf)