書き起こし・タグ付けマニュアル（日本語以外の言語）

・日本語以外の言語の書き起こしは、その言語の正書法を用いておこなう。中国語は繁体字を用いる（繁体字→簡体字の変換の方が逆の変換より容易なため)。
・保存時の使用文字コードは以下のとおり。

中国語：Big5
タイ語：Unicode(UTF-8)
韓国語：EUC

・書き起こしは、聞こえたままをできるだけそのまま表記するものとする。正書法では表記できない場合の処理は［**使用タグ**］の項を参照。
・ひとつの書き起こし単位が長くなり、エディタ上で 1 行に収まりきらなくなった場合は適宜改行してよいが、その際単語の途中では改行しないこと。
・書き起こしに句読点は使用しない。ひとつの書き起こし単位の中に短い(200ms に満たない)無音区間がある場合も、そこでコンマ「,」などをうったりはしない。
・書き起こしに数字は用いず、すべて漢字・ハングル文字・タイ文字で書く。

[使用タグ]

・音声特徴をあらわすタグとしては、以下のものを使用する[1]。タグではすべて**半角英数字大文字**のみを使用する。またタグの中に文字列を書き込む場合、文字列の前には半角スペースを入れる。

1. 文字範囲指定タイプ

このタイプのタグは、「書き起こされた文字範囲に対応する音声について、何らかの現象が起こっている」ということを示すためのものである。

1.1 (W) タグ(wrong)

概要

言い誤り[2]や発音の怠け、転訛等があった場合、その部分をこのタグで囲み、同時に「い

[1] これらのタグのうち多くは、小磯他(2001) p.55 で定義され、CSJ（『日本語話し言葉コーパス』）で使用されているものである。「音声対照言語 DB」でも多くの場合、CSJ タグの定義を踏襲したが、一部で定義しなおしたところもある。また、CSJ では使用されていないタグを独自に設定・定義したものもあるので注意。
[2] 1 人の話者の発話中に一貫して観察される「方言音」（例：中国語で、「是」の子音を、捲舌音[ʂ]ではなく、歯音[s]で発音しているような場合）は、「言い間違い」とはみなさず (W) タグは使用しない。「方言音」に対しては、後述の「(DS) タグ」を使用する。

い誤り・発音の怠け等が起こらなかったときに想定される語形」を併せ示す。このタグは、「音声上、標準的と考えられる発音から一時的に逸脱している部分」に対して使用するものであり、文法的な間違い（例：「学校を行きます」）や、事実認定の誤り（例：「日本には大統領がいる」）にはこのタグをつけない。また、本来「発音の怠け」から生じた発音形態であっても、一個人に限らず広く観察される「口語的表現」[3]については、現代語としては「言い誤り、発音の怠け」とは考えず、タグをつけない。

書式１：(W *;###)**

は、発音どおりの表記、###は、「発音の怠け等」が起こらなかったときに想定される語形を示す。中国語については、の部分はピンイン[4]で、###の部分は漢字で表記。韓国語・タイ語は、***、###ともにハングル文字・タイ文字を使用。

例：

지하(W 선;철)

※本来激音で発音されるべき철の頭子音を、誤って平音で発音してしまった

(W ปา;ปลา)

※二重子音 pl[5]の第 2 要素 l を脱落させて発音してしまった

(W lan2lou2;襤褸)

※本来lánlǚと発音すべきところを、誤ってlánlóuと発音してしまった

書式２：(W ###)

発音の間違い、怠けなどをピンイン・ハングル文字・タイ文字によってもうまく表記できない場合は前半（発音どおりに書く部分）を省略してよい。

---

[3] くだけた場面においてあらわれる口語的表現について、「小説やシナリオ等の中のせりふとして、発音どおりの表記で書かれうるもの」については、「一個人の発話においてだけでなく、社会習慣上広範囲にあらわれうる口語的表現」と考え、タグをつけない。「小説やシナリオ等の中のせりふとしても、発音どおりに書かれることは通常ない」と考えられる表現については、「言い誤り・発音の怠け」と考え、(W) タグをつける。

[4] 声調符合は数字に置き換えて入力する。第 1 声:1、第 2 声:2、第 3 声:3、第 4 声:4、軽声:0. ピンインでは軽声は声調無表記であるが、本データベースでは 0 によって軽声であることを明示的に表示する。音節が不完全に発音されている、などの理由で声調が判定できない場合のみ声調無表記となる。

[5] タイ語で、第 2 要素として流音をもつ二重子音（pr, kl など）の流音要素が脱落している場合は、発音の怠けと解釈して W タグを使用する。しかし、正書法上 ร[r]で書き表される子音が[l]または[ɹ]などで発音されている場合、これはバンコク方言では異音の範囲内と考え W タグは使用しない。同様に正書法上 ล で書き表される子音が[r]または[ɹ]で発音されている場合も W タグは使用しない。

例：

打著(W 傘)

※文脈から判断して「傘」と言おうとしていることは明らかであるが、発音は・・・・とは明らかに異なり、またピンインによっても発音を正確に表示することができないよう場合このようになる。

タグを付与する範囲

このタグは、「短単位相当単位」に対して使用する（＝ひとつの単位をひとつのタグで囲む)。各言語ごとの「短単位相当単位」については、p.119 以下を参照。言い間違い、発音の怠け等が、隣接する単位において連続して起きている場合は、語ごとにタグを打ち直す。

1.2 (?) タグ

概要

聞き取り、語彙同定に自信がない場合に使用する。「おそらく○○であろうが自信がない」場合、「○○の可能性もあるが、××かもしれない」場合、「まったく分からない」場合、すべてこのタグで対応する。

書式１：(? *)**

おそらく***といっているのであろうが、自信がない。

書式２：(? *,###)**

おそらく***といっているのであろうが、###の可能性もある。

書式３：(?)

まったく聞き取れない。この場合、聞き取れない部分に複数の語が含まれていそうであってもタグは１つでよい。

例：

讓我打了兩個禮拜工才(? 掙)回來的

※文脈から判断するとおそらく「掙」といっているものと推測されるが、発音があいまいで確証が持てない。

(? พิมพ์,ผลิต)ออกมา

※พิมพ์（印刷する）でも ผลิต（生産する）でも意味的には通用する。発音にも類似性があり（[phim]と[phalit])、いずれとも判断がつかない。

(? 都,多)吃河魚

※「都吃河魚」（みんな川の魚を食べる）でも「多吃河魚」（多く川の魚を食べる）でも意味的には通用する。発音にも類似性があり（douとduo)、いずれとも判断がつかない。

タグを付与する範囲

基本的には短単位相当単位に対して付与するが、発音が不明瞭なため単位の切れ目も不分明な場合は、よく聞き取れない部分全体に対し使用してよい。

1.3 (RP) タグ(real pronunciation)

概要

ある語について、正書法上の表記は 1 通りであっても、複数の読みがありえ、いずれで発音してもかまわない場合[6]にこのタグを使用し、実際の発音の様態を示す。(RP) タグをつけるべき語は言語ごとに決まっており、それらの語には出現するたび必ず (RP) タグを付与する。(RP) タグをつけるべき語は、p.122 以下にまとめて掲出した。