-- Instructions for transcription in KanaKanji for UD --
-- UD 用の仮名漢字表記のデータ入力について--

1. Introduction　　はじめに

Earlier versions of the Japanese database relied on transcription with the Romaji romanization script. Morphosyntactic analysis was done using the Japanese version of the MOR program. Fortunately, most corpora also had a separate %ort line with Japanese script. Using the CONVERT program in CLAN, we swapped the %ort tier with the main tier and then prepared the new main tier for analysis by the Japanese Universal Dependencies (UD) tagger. This was done for all corpora except Ishii, NINJAL-Okubo, and Noji which remain in romanization for now. We do not plan to convert Okayama immediately because of the high frequency use of dialect.
今まで日本語のデータベースではメインラインでローマ字が使われました。日本語版の MOR プログラムで形態素解析が行われました。幸いに多くのデータに日本語表記の %ort ティアが付いていました。CLAN の CONVERT プログラムで %ort ティアをメインラインと入れ替え、新しいメインラインの仮名漢字文章を Japanese Universal Dependencies (UD)というタグ付けプログラムの解析に向けて整理しました。これはほぼすべてのコーパスで行いました。例外は今のところローマ字しかない Ishii、NINJAL-Okubo、と Noji です。また、方言の高頻度の Okayama は、仮名漢字へ変換する予定がありません。

2. Basic Rules　　基礎的ルール

New corpora can now be transcribed directly in Japanese script on the main line. Afterwards, UD can be run automatically using the Batchalign "morphotag" command. When running UD, the spacing in the original is redone (retokenized), so with certain important exceptions, spacing in the input doesn't matter because it will be redone during retokenization. However, there are a few exceptions, as listed in the following guidelines.
新しいコーパスは直接日本語で書き起こします。書き起こしが完了した後、Batchalign プログラムの morphotag コマンドで UD プログラムを走らせます。UD を走らせると、各文が単語に分析され、単語と単語の間に半角スペースが入ります (retokenization)。しかし、次のような例外があります。

- Remove as many spaces as possible (since we use retokenization).
半角スペースはできる限り使いません。
- Tag special forms with @c, @i, @g, @k, @o, or @u and surround these forms with

spaces. A full listing of many of these forms is given in this document.
特殊な単語は @c や @i、@g、@o、@u でマークされます。前後に半角スペースを加えます。特殊な単語のリストは this document にまとめてあります。

- Reversing the normal convention in CHAT, the actually produced non-standard string should be placed within the [:] brackets, while the intended conventional adult form should be placed before the brackets, as in intended [: deviating]. See #9 below for examples.
CHAT の全体ルールと違って、置き換えたい単語を [:] の括弧に入れておき、置き換えの単語はその前に書きます：intended [: deviating]。詳細は #9 を参照してください。
- Make sure that there is no space after a beginning <, but a space before it
括弧 < の前は必ず半角スペースを入れます。その後はスペースを入れません。
- Make sure that there is no space before a closing >, but a space after it
括弧 > の前は半角スペースを入れません。その後は必ずスペースを入れます。
- No double spaces and no double-width Kana spaces.
全角スペースや連続の半角スペースは使えません。
- Files should pass CHECK before going into UD
UD を走らせる前に、CLAN の CHECK コマンドでチェックを行います。

3. Detailed guidelines　　ルールの詳細

1) Kana and Kanji　　仮名と漢字

Transcription of child language tends to use predominantly Hiragana and Katakana. But a long string of Kana is difficult to parse for UD, and the analysis becomes more correct, if you use Kana and Kanji alternately. Use Kanji, if appropriate. You shouldn't use spaces between the words.
子どもの発話を表記するときは平仮名・片仮名を使うことが多いですが、仮名の連続は UD (Universal Dependencies 多言語構文解析器； https://universaldependencies.org/treebanks/ja_gsd/index.html) にとって解析しにくいです。漢字や片仮名の使い分けによって、パーシング（単語分け）の信頼性が上がる。単語と単語の間はスペースが不要です。

E.g.:　　*CHI:　可愛い赤ちゃんのご飯を取っちゃ駄目!

Subsidiary verbs can be written in Hiragana.
補助動詞は平仮名でも使えます。

～てあげる、～てくれる、～てもらう、～てくる、～ていく、～ておく、～てみる、～てしまう、～ている、～てある

E.g.:　　書いてあげる、読んでくれない、仕舞っておこう

2）Names of animals and plants　　動物や植物の名称
Animals and plants should be written in Katakana or Kanji, never Hiragana.
動物や植物の名称はカタカナまたは漢字で表記します。平仮名は使えません。

E.g.:　　**キリン、ゾウさん** または **象さん**，**ウサギ** または **兎、イチゴ** または **苺、キュウリ** または **胡瓜、熊**（**クマ**は解析されない）etc.

3) Numbers　　数字
Numbers should be always written in Kanji. Arabic numerals cannot be used.
数字は原則として漢字で表記します。アラビア数字は使えません。

E.g.:　　**三十七点ゼロ度**（not「37.0 度」）

4) Baby words　　幼児語
Baby words should be in Katakana (or Kanji, if appropriate). They have to be marked by @c. Add spaces around the word (see this document).
幼児語は片仮名または漢字で表記し、@c を付けます。単語の前後に半角スペースを加えます（this document を参照)。

E.g.:　　*CHI:　**キーちゃんの ワンワン@c も ネンネ@c するよ** .

E.g.:　　**ネンネ@c、ブーブ@c、オシッコ@c、ダッコ@c** または **抱っこ@c、オテテ@i** or **お手手@c** etc.

5) Onomatopoeias　　オノマトペ
Onomatopoeias should be in Katakana, marked by @o. Add spaces around the word (see this document).
オノマトペ（擬音語・擬声語・擬態語）は片仮名で表記し、@o を付けます。単語の前後に半角スペースを加えます（this document を参照)。

E.g.:　　**ゴロゴロ@o、ピョン@o、ガンガン@o** etc.

Long onomatopoeias should be replaced (see # 8 replacement) by a double syllable version.
長く続いているオノマトペは二つ単位に［/］でお着替えます（８を参照)。

E.g.: *CHI: 〈ガンガンガンガンガンガン@o〉 [/] ガンガン@o .

Some onomatopoeias also function as baby words as in the following examples.
一部のオノマトペは幼児語としても使われます。

E.g.: *CHI: ワンワン@c が走った . 〈-- baby word @c
E.g.: *CHI: ワンワン@o と吠えた . 〈-- onomatopoeia @o

6) Interjections and greetings 間投詞・挨拶
Interjections except あっ, おっ、はい and うん have to be marked by @i (see this document).
間投詞は @i でマークします（あっ、はい、うん を除く ; this document を参照)。

E.g.: *CHI: ほら@i 止めて ！

As specific forms you should use あっあ@i aaqa "oh dear" and ううん@i uun "uh uh", "no".
特殊な形として期待外れの あっあ@i ‘aaqa’ "oh dear" と否定の ううん@i ‘uun’ "uh uh"、"no" が使われます。

Greetings have to be marked by @g. (see this document)
挨拶は @g でマークします) this document を参照)。

E.g.: *CHI: 熊さん こんにちは@g .

7) Unclear parts 不明な部分
Unclear parts are indicated by xxx (surrounded by spaces)
不明な部分は xxx で表します。前後に半角スペースを加えます。

E.g.: *CHI: ワンワン@c が xxx 走った .

&~ marks babbling (can be in Kana or Romaji; has to be surrounded by spaces)
喃語は &~ で表記します。カナでもローマ字でも使えます。前後に半角スペースを加えます。

E.g.: *CHI: &~うばば .

&+ marks false starts.(surrounded by spaces)
言い掛けは &+ で表記します。前後に半角スペースを加えます。

E.g.: *MOT: どこで &+ワン 犬を見てきたの ？

All these forms are ignored by UD and excluded from the analysis.
xxx、&~、&+ の部分は UD の解析から省かれます。

8) Repetitions and corrections 繰り返し・言い直し
For repetitions [/] and corrections [//], the repeated word(s) must be always surrounded by brackets < >.
繰り返し［/］および言い直し［//］の記号の前の単語はカッコ ＜＞ に入れます。

E.g.: *CHI: <どこ> [/] どこ行ったの ？
E.g.: *CHI: <どこから> [//] 何持って来たの ？

9) Replacements 置き換え
For replacements [:], the actually produced form is placed within the brackets, while the intended word comes before the brackets, as in intended [: deviating] . Note, that this rule applies only to KanaKanji main lines. For %ort tiers in Latin script, the normal CHAT rule of replaced [: replacement] applies. Also, spaces should be added before and after the intended word, before the deviating word, and after the closing bracket.
置き換え［: ］の場合は入れ替えらる単語を括弧の中に入れ、入れ替わる単語を括弧の前に置きます。このルールは仮名漢字のメインラインのみに当てはまります。%ort ティアのローマ字表記は一般の CHAT 原則の replaced [: replacement] に従います。さらに、入れ替えられる単語の前後、置き換えの前、そして最後の括弧の後に半角スペースを加える必要があります。

E.g.: *CHI: ここに お薬 [: おすくり] 置いておこうね .
%ort: koko ni osukuri [: okusuri] [* p] oite okoo ne .

For replacements, only one word at a time is allowed. Note that small ゃゅょ or っ cannot stand alone.
置き換えは一単語ずつで行われます。小さい「ゃゅょ」および「っ」は単独で使えません。

E.g.: *CHI: それ [: そ] は [: りゃ] 可笑しいよ .

OR: *CHI: それは [: そりゃ] 可笑しいよ .
%ort: sor [: sore] ya [: wa] okashii yo .

For small phonological deviations, we recommend using only the intended form on the main line and transcribing the deviant form only on the %ort tier.
音声的なズレの場合は、%ort ティアでその細かい情報を書き込んだ上、仮名漢字メインラインで意図された単語のみを使うことを勧めます。

E.g.: *CHI: 分かりません .
%ort: wakarimashen [: wakarimasen] [* p] .

10) Contracted forms 圧縮された形
Many contracted forms used in spoken Japanese are not recognized by UD. Again, we recommend using only the full form on the main line and writing the contracted form with a replacement on the %ort tier. If necessary, you can do a replacement (see #9) on the main line,
話しことばの砕けた単語や圧縮された形の多くが UD に認識されません。仮名漢字のメインラインで単純に標準語の単語を使い、ローマ字の %ort ティアで砕けた形とその置き換え（ [:] 、9 を参照）を表記することを勧めます。

E.g.: *CHI: この奴良く分からない . <-- full form
%ort: kono yatsu yoku wakannai [: wakaranai].<-- contracted form

Some frequent examples are listed below.
次の単語はよく使われる圧縮形です。

止めとく	→	止めておく	tomet(e) oku
結んだげる	→	結んであげる	musund(e) ageru
やんない	→	やらない	yannai [: yaranai]
つまんない	→	詰まらない	tsumannai [: tsumaranai]
食べさして	→	食べさせて	tabesashite [: tabesasete]
やらして	→	やらせて	yarashite [: yarasete]
行きな	→	行きなさい	ikina [: ikinasai] (dialect)
寒っ	→	寒い	samuq [: samui]
痛え	→	痛い	itee [: itai] (dialect)
そりゃ	→	それは or それは [: そりゃ]	sor [: sore] ya [: wa]
こりゃ	→	これは or これは [: こりゃ]	kor [: kore] ya [: wa]
どっか	→	どこか	dok(o) ka

こっから	→	ここから	kok(o) kara
こん中	→	この中	kon(o) naka
こないだ	→	この間	kon(o) aida
こんぐらい	→	このぐらい	kon(o) gurai
ち	→	家 or ち [: 家]	(u)chi
んち	→	の家 or ん [: の] ち [: 家]	n(o) (u)chi
やだ	→	嫌だ	(i)ya da
でしょ	→	でしょう	desho(o)
行こ	→	行こう	iko(o)
帰ろっか	→	帰ろうか	kaero(o) k:a

Note that ～てる (aspect; short form of ～ている) is recognized by UD analysis.
例外としてアスペクトの ～てる（～ている）が UD に認識され、そのまま使えます。

E.g.: *CHI: 食べてたよ .
E.g.: *CHI: 落ちちゃってる .

Also the following contracted or colloquial forms are recognized by UD.
そのほかに次の単語が砕けた形も UD に認識されます。
ほんと（本当）、みんな（皆）、ほか（他）、あんまり（あまり）、
いっぱい（一杯）、あと（後）、とき（時）、また（又）、ほう（方）

11) Punctuation 句読点
Punctuation should in half space (.?!,:) , not in full space KanaKanji (。etc.).
句読点（.?!,:）はすべて半角（ローマ字モード）になります。全角の仮名漢字句読点（。など）は使えません。

E.g.: *CHI: 熊さん行っちゃった .

Commas should only be used for dislocation. On the %ort tier, dislocation is marked by „ (double comma; you can find this symbol in the Special Characters Window in CLAN under "tag"). This symbol cannot be used on the main line, where only the single comma is allowed.
コンマ「,」は外置（dislocation）の表記に限定されています。ローマ字用の %ort ティアでは外置のために「„」（ダブルコンマ；CLAN の Special Characters Window で「tag」として登録）が使われますが、その記号はメインラインで使用できません。

E.g.: *CHI: こっちで良い , 車 . <-- comma
%ort: kotchi de ii „ kuruma . <-- double comma

Similarly, symbols like ‡ (vocative), “ ” (citation), and ↑ ↓ (intonation) cannot be used on the main line. You can add this kind of information on the %ort tier.
同様に、‡（呼び掛け)、“ ”（引用）および ↑ ↓（イントネーション）もメインラインで使えません。その使用が %ort ティアに限ります。

E.g.: *CHI: お母さん本がないよ .
%ort: Okaasan ‡ hon ga nai yo ↑ .

12) Lengthening of vowels 母音の伸ばし
Lengthening of vowels (for example for exaggeration) should be represented by : (half space colon).
母音の伸ばし（呼び掛けやエンファジスなどのため）は半角コロン「:」で表します。

E.g.: *MOT: アキちゃ:んこれだよ: .
%ort: Akicha:n ‡ kore da yo: .

Do not use the following symbols and letters for lengthening.
次の表記や文字では母音の伸ばしを表しません。
ー（アキちゃーん）
〜（アキちゃ〜ん）
あ（アキちゃあん）
ぁ（アキちゃぁん）
♪（アキちゃん♪）and similar.