アノテーションツール入出力データ仕様書

河原 大輔
京都大学

2012 年 12 月 14 日

1 用語

- 形態素: 文を構成する最小単位。「単語」に相当する。内容語 (名詞、動詞、形容詞など) と付属語 (助詞、助動詞など) に大別される。
- 基本句 (旧名称: タグ単位): 1 つの内容語と 0 個以上の付属語からなる。
- 文節: 1 つ以上の基本句からなる。
- 関係タグ: 基本句間のさまざまな関係を表す。
- KNP: 日本語構文・格解析システム。アノテーションツール入出力データは、KNP による出力である。

2 データフォーマット

- * から始まる行は文節に関する情報を表す。
- + から始まる行は基本句に関する情報を表す。
- # から始まる行はコメント行を表す。
- EOS 行は 1 文の情報の終わりを表す。
- それ以外の行は形態素に関する情報を表す。

3 文節行と基本句行

書式: {*,+}_数字 {PIAD}_⟨feature1⟩⟨feature2⟩...
※ "_" はスペースを表す。以下でも同様。

- 文節行および基本句行は、頭から順番に 0 番、1 番、... と暗黙的に番号がふられている。
- "*" や "+" の後の数字は係り先の文節番号あるいは基本句番号を表す (この数字は-1 以上の整数)。-1 はその文節あるいは基本句が文の root(文末) であることを示す。

- 数字の直後のアルファベットは係り受けのタイプを示し、P が並列、I が部分並列、A が同格、D が通常の係り受けを表す。
- feature は、その文節あるいは基本句の各種情報を表す。

4 形態素行

書式: 表記_読み_原形_品詞_品詞番号_品詞細分類_品詞細分類番号_活用型_活用型番号_活用形_活用形番号_意味情報_⟨feature1⟩ ⟨feature2⟩...

- 品詞細分類、活用型、活用形がない場合は “*” であり、それに対応する番号は 0 である。
- 意味情報にスペースが含まれる場合には、ダブルクオート (”) で括られる。
- 内容語の形態素には、⟨ 内容語 ⟩ もしくは ⟨ 準内容語 ⟩ という feature が付与される。付属語の形態素は、それ以外の形態素である。

5 関係タグ

- 関係タグは、基本句 feature の ⟨rel⟩ として表す。⟨rel⟩ は常に 2 つの属性 type, target を持ち、関係先が文章中にある場合には属性 sid, id を持つ。これらの属性を次のように表す: ⟨rel type=”関係タイプ” target=”関係先基本句の表記” sid=”関係先基本句の文 ID” id=”関係先基本句の番号”/⟩。ただし、関係先が文外の場合は、属性 sid と id は持たないものとする。
- 同じ関係タイプのタグが複数ある場合には、⟨mode⟩ によって、それらのタグ間の関係を表す。⟨mode⟩ は、属性 rel(関係タイプ) を持ち、値として次の 3 つのうちのいずれかをとる: AND, OR, ？。
- 「関係先基本句の表記」(⟨rel⟩ の target 属性) は、関係先基本句における内容語の形態素の原形に相当する。
- アノテーション時のメモは、⟨memo text=”メモの内容”/⟩ で記述する。

6 コメント行

書式: #_S-ID:文 ID_コメント文字列

- コメント文字列中には、KNP の実行日として「DATE:2012/12/14」のような文字列が含まれる。また、KNP のバージョンとして「KNP:4.0-20121016」のような文字列が含まれる。
- この文に対して一度以上アノテーションが行われた場合には、アノテーションの修正日として「MOD:2012/12/14」のような文字列が記述される。

7 例

「太郎は京都大学に行った。」に対するデータを示す。

```
# S-ID:950101001-001 KNP:4.0-20121016 DATE:2012/12/14 SCORE:-11.95722
* 2D <文頭><人名><ハ><助詞><体言><係:未格><提題><区切:3-5><主題表現><格要素><連用要素><正規化代表表記:太郎/たろう><主辞代表表記:太郎/たろう>
+ 3D <文頭><人名><ハ><助詞><体言><係:未格><提題><区切:3-5><主題表現><格要素><連用要素><名詞項候補><先行詞候補><SM-人><SM-主体><正規化代表表記:太郎/たろう><解析格:ガ>
太郎 たろう 太郎 名詞 6 人名 5 * 0 * 0 "人名:日本:名:45:0.00106 疑似代表表記 代表表記:太郎/たろう" <人名:日本:名:45:0.00106><疑似代表表記><代表表記:太郎/たろう><正規化代表表記:太郎/たろう><漢字><かな漢字><名詞相当語><文頭><自立><内容語><タグ単位始><文節始><固有キー><文節主辞>
は は は 助詞 9 副助詞 2 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* 2D <SM-主体><SM-場所><SM-組織><BGH:大学/だいがく><組織名><ニ><助詞><体言><係:ニ格><区切:0-0><格要素><連用要素><正規化代表表記:京都/きょうと+大学/だいがく><主辞代表表記:大学/だいがく>
+ 2D <文節内><係:文節内><地名疑><体言><名詞項候補><先行詞候補><正規化代表表記:京都/きょうと>
京都 きょうと 京都 名詞 6 地名 4 * 0 * 0 "代表表記:京都/きょうと 地名:日本:府" <代表表記:京都/きょうと><地名:日本:京都府:市><正規化代表表記:京都/きょうと><品曖><ALT-京都-きょうと-京都-6-4-0-0-"代表表記:京都/きょうと 地名:日本:京都府:市"><品曖-地名><漢字><かな漢字><名詞相当語><自立><内容語><タグ単位始><文節始><固有キー>
+ 3D <SM-主体><SM-場所><SM-組織><BGH:大学/だいがく><組織名><ニ><助詞><体言><係:ニ格><区切:0-0><格要素><連用要素><名詞項候補><先行詞候補><正規化代表表記:大学/だいがく><解析格:ニ>
大学 だいがく 大学 名詞 6 普通名詞 1 * 0 * 0 "代表表記:大学/だいがく 組織名末尾 カテゴリ:場所-施設 ドメイン:教育・学習" <代表表記:大学/だいがく><組織名末尾><カテゴリ:場所-施設><ドメイン:教育・学習><正規化代表表記:大学/だいがく><漢字><かな漢字><名詞相当語><Wikipedia エントリ-京都大学:2-3><自立><複合←><内容語><タグ単位始><文節主辞>
に に に 助詞 9 格助詞 1 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* -1D <BGH:行く/いく|行う/おこなう><文末><時制-過去><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><係:文末><提題受:30><主節><格要素><連用要素><動態述語><正規化代表表記:行く/いく?行う/おこなう><主辞代表表記:行く/いく?行う/おこなう>
+ -1D <BGH:行く/いく|行う/おこなう><文末><時制-過去><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><係:文末><提題受:30><主節><格要素><連用要素><動態述語><正規化代表表記:行く/いく?行う/おこなう><用言代表表記:行く/いく?行う/おこなう><主題格:一人称優位><格関係 0:ガ:太郎><格関係 2:ニ:大学><格解析結果:行く/いく:動 6:ガ/N/太郎/0/0/1; ニ/C/大学/2/0/1; デ/U/-/-/-/-; 時間/U/-/-/-/-; ノ/U/-/-/-/-><rel type="ガ" target="太郎" sid="950101001-001" id="0"/><rel type="ニ" target="大学" sid="950101001-001" id="2"/>
行った いった 行く 動詞 2 * 0 子音動詞カ行促音便形 3 タ形 10 "代表表記:行く/いく 付属動詞候補(タ系) ドメイン:交通 反義:動詞:帰る/かえる" <代表表記:行く/いく><付属動詞候補(タ系)><ドメイン:交通><反義:動詞:帰る/かえる><正規化代表表記:行く/いく?行う/おこなう><品曖><ALT-行った-おこなった-行う-2-0-12-10-"代表表記:行う/おこなう"><品曖-動詞><原形曖昧><移動動詞><かな漢字><活用語><表現文末><自立><内容語><タグ単位始><文節始><文節主辞>
。 。 。 特殊 1 句点 1 * 0 * 0 NIL <英記号><記号><文末><付属>
EOS
```