対訳コーパスのためのタグ体系

塚脇　幸代
立命館大学大学院言語教育情報研究科

0. はじめに

コーパスを使う者にとって、あらかじめ付与された各種の情報は、検索においてもその後の分析においても有用である。それゆえ対訳コーパスにもタグがあれば便利ではないかと考えた。だが、各言語の品詞は各言語によって個別に与えられているのが通常で、他の言語と対比させた場合に、どちらかの言語が必要以上に細分類化されていたり、逆に求める分類がなかったりすることがある。本稿では、日本語と英語からなる対訳コーパスに付与するタグを、(1)なるべく数を抑え、(2)できるだけ多くの情報を記述でき、(3)可能な限り日英間で共通になるように、作成する試みについて述べる。[1]

1. 品詞タグ

品詞タグはメインカテゴリと2つのサブカテゴリによって記述される。メインカテゴリは必須だが、サブカテゴリは任意である。表1に、メインカテゴリの品詞タグを挙げる。斜体の部分は日英で同じ名称を使用しているタグである。

日本語のタグ	日本語のタグの説明	英語のタグ	英語のタグの説明
ADJ	*形容詞*	*ADJ*	*形容詞*
ADV	*副詞*	ADJC	形容詞の比較級
AN	形容動詞語幹	ADJS	形容詞の最上級
ANCMP	複合形容動詞語幹	*ADV*	*副詞*
AUX	*助動詞*	*AUX*	*助動詞*
AUXEQ	*助動詞相当語*	AUXBE	be 動詞(不定形)
COMP	*補文標識*	AUXBED	be 動詞(過去形)
CZ	接続詞(並立助詞を含む)	AUXBEG	be 動詞(-ing 形)
JE	終助詞	AUXBEN	be 動詞(過去分詞形)
JEEQ	終助詞相当語	AUXDO	助動詞 do(不定形)
JF	副助詞および係助詞	AUXDOD	助動詞 do(過去形)
JFEQ	副助詞相当語	*AUXEQ*	*助動詞相当語*
JK	格助詞	AUXHV	助動詞 have(不定形)
JKEQ	格助詞相当語	AUXHVG	助動詞 have(-ing 形)
JS	接続助詞	CNJC	等位接続詞
JSEQ	接続助詞相当語	CNJS	従属接続詞
N	*名詞*	*COMP*	*補文標識*
NCMP	*複合名詞*	DART	冠詞
NPRN	*代名詞*	DET	限定詞
NPRP	*固有名詞*	*N*	*名詞*
NUMB	*数字*	*NCMP*	*複合名詞*
PUNC	*句読点*	NEG	否定辞
RT	連体詞	*NPRN*	*代名詞*
RTEQ	連体詞相当語	*NPRP*	*固有名詞*
SS	接辞	*NUMB*	*数字*
SSA	形容詞化接辞	PEQ	前置詞相当語
SSAN	AN 化接辞	PREP	前置詞
SSAV	副詞化接辞	*PUNC*	*句読点*
SSN	名詞化接辞	*SYMB*	*記号類*

[1] 作業は実際の対訳データを参照しながら進めた。情報通信研究機構自然言語グループの「日英新聞記事対応付けデータ」を利用した。また、日本語の形態素解析には、奈良先端科学技術大学院大学の「Chasen for Windows」を利用した。日本語のタグは、Chasen のきめ細かい出力にヒントを得た部分もある。ここに註記して謝辞とする。

日本語のタグ	日本語のタグの説明	英語のタグ	英語のタグの説明
SSQ	数量接辞	*V*	*動詞(不定形)*
SSV	動詞化接辞	VD	動詞(過去形)
SSVN	VN 化接辞	VING	動詞(-ing 形)
SYMB	*記号類*	*VPH*	*句動詞(不定形)*
V	*動詞*	VPHD	句動詞(過去形)
VCMP	複合サ変動詞	VPHG	句動詞(-ing 形)
VN	サ変動詞語幹	VPHP	句動詞(過去分詞形)
VNCMP	複合サ変動詞語幹	VPPD	動詞(過去分詞形)
VPH	*句動詞*	VZ	動詞(3 人称単数現在形)

表 1

サブカテゴリには、メインカテゴリに記述された品詞の細分類を記述する。サブカテゴリに何を記述しなければならないかは、品詞によって異なる。英語の名詞なら、単数か複数か、代名詞なら人称と格が問題になる。サブカテゴリを2つ用意したのは、品詞によって必要となる細分類の種類と数が異なるからである。現在メインカテゴリの数は日英それぞれ38だが、サブカテゴリを設けることにより、実際には100以上の機能を区別することができる。[2]

2. 品詞タグの基本概念

日本語と英語の共通化を図るために、内容語[3]の範疇とその関係を図1のように規定している。

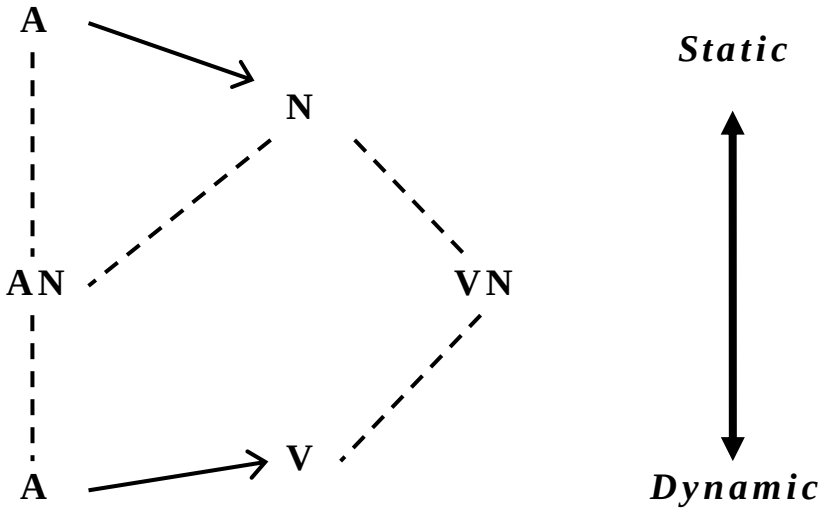

図 1

A、N、V はそれぞれ内容語の基本的な要素である。N は静的な表現を担い、V は動的な表現を担う。A は修飾要素を表し、N を修飾するものには形容詞、形容動詞、連体詞、限定詞などが属する。V を修飾するものには副詞が属する。ここで言う修飾とは文字通り飾ることではなく、依存関係、つまり係受け関係をさす。VN と AN は日本語にお

[2] British National Corpus(BNC)のタグ付けに用いられた CLAWS の品詞タグの数は、バージョンによって異なるが、CLAWS5 では 73、CLAWS6 では 160 を超え、CLAWS7 で 137 と、100 以上の品詞を設定している。あくまで数的なものだが、これらに匹敵する種類を記述できることになる。

[3] Content words, または自立語と呼ばれるもの。

いて設定される要素で、VN はＶとＮ、AN はＡとＮの両方の機能を持つ。VN はサ変動詞語幹に、AN は形容動詞語幹に与えられる。VN は動詞化接辞“*する*”が後接することによって Ｖ となり、AN は形容詞化接辞“*な*”が後接することによって形容詞となる。AN と“*な*”の連続を形容詞として扱うことにより、日本語にあって英語には無い形容動詞という範疇を吸収することができる。VN あるいは AN といった、2つの機能を持った品詞の存在は、目的言語への変換において、異なる品詞への変換の可能性を示唆する。

図2は機能語[4]の日英の対応関係を表す。機能語には言語間の差異がよく現れる。

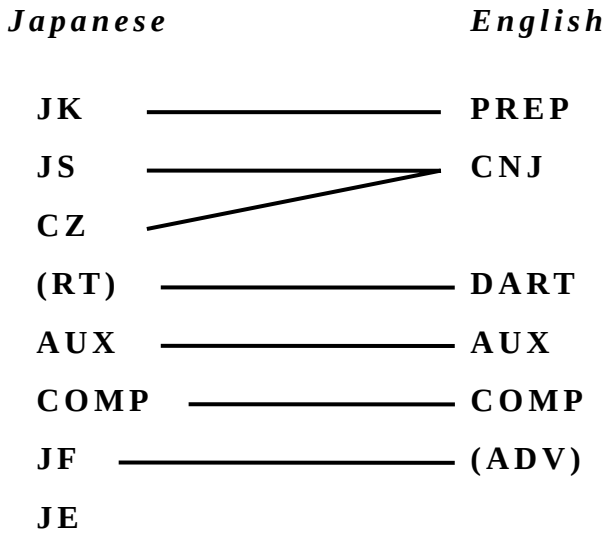

図２

冠詞は他の限定詞と区別して扱う。日本語では通常は冠詞に相当する語彙項目を訳出しないからである。特に必要なときは、連体詞となって現れる。逆に英語で語彙項目を持たないことがあるのは、日本語の格助詞である。補文標識（*COMP*）を品詞として設けたのは、補文構造を認識させるためである。関係詞、*to* 不定詞は *COMP* と記述される。日本語の助詞類で英語との対応がとりづらいのは、副助詞類（*JF*）と終助詞類（*JE*）である。副助詞は英語では副詞となって現れることも多いが、疑問の終助詞などは、その問いの対象になっているものに応じて、英語では文構造と品詞を使い分けることになる。接続詞は従来、日本語と英語とで解釈の違いがあるようだが、ここでは、等位接続詞と従属接続詞という英語の分類に従う。その意味で、日本語の並立助詞は、接続詞（*CZ*）に分類した。

図2に示した対応は、直訳が行われた場合の予測であって、実際の翻訳においては、もっと多様な変換の仕方が観察されるはずである。

3．修飾タグ

品詞間の係り受け関係を表すタグである。日本語の“*の*”、英語の“*of*”は、多くの場合、“*mod*”の修飾タグを持つ助詞あるいは前置詞と記述される。“*mod1*”は英語に固有、

[4] Function words，または付属語と呼ばれるもの。

“*mod3*”は日本語に固有なタグである。

タグ	説明	適用言語
mod	助詞あるいは前置詞が、名詞に係る場合	日本語 英語
mod1	動詞、形容詞による後置修飾	英語
mod2	動詞、形容詞による前置修飾	日本語 英語
mod3	動詞句、節による前置修飾（＝連体修飾節）	日本語

表２ 修飾タグ

4．共起タグ

“*もし*”という日本語の副詞は、接続助詞“*ば*”や、仮定形“*なら*”を伴うことがある。共起タグは、このような離れた位置にある呼応表現を関係付けるためのタグである。

タグ	説明
col_01a	一文内における最初の共起セットの最初の要素
…	
col_01*z*	一文内における最初の共起セットの*z*番目の要素
col_*nn*a	一文内における*nn*番目の共起セットの最初の要素
…	
col_*nnz*	一文内における*nn*番目の共起セットの*z*番目の要素

表３ 共起タグ

5．タグによる検索

上に説明したタグは、一形態素一行の垂直フォーマットの形態素解析結果に、表計算のシート上で記述できるように設計されている。一行を一レコードとみなしてデータベース上で検索を行うことができる。品詞タグと修飾タグのなかからひとつ以上のタグを指定すれば、原文とそれに対応する訳文を検索結果として得ることができる。“*mod3*”で記述される日本語の連体修飾節は、英文では様々なパターンで訳されていることが観察される。

6．今後の課題

現段階で設定されているタグは、まだ整理・拡張の余地がある。今回はどちらかというと日本語から英語への変換を前提にしていたが、英語から日本語への変換という見方をすれば、また新たな品詞が必要になるかもしれない。

また、このようなタグを付与された対訳コーパスを使いこなしていくためのツールも必要になるであろう。