複層意味フレーム分析の簡略版を使った意味役割タグづけの現状

タグづけデータから派生する言語資源の紹介を中心に

黒田 航* 李 在鎬* 渋谷 良方*,** 井佐原 均*

*NICT **京都外国語大学

複層意味フレーム分析 (MSFA) を使ったタグづけデータからは次の三つの言語資源/語彙資源が派生:

A. フレームのデータベース (Frame List)

B. フレーム要素 (=状況ごとの粒度の細かい意味役割) のデータベース (FrameElement List)

C. 有意味な部分文字列 (語や連続/不連続な句) のデータベース (Word/Phrase List)

これらに加えて MSFA Lite 版を使った場合,

D. フレームを実現するパターン/構文のデータベース (Pattern List) を実装する予定

三つの公開サイトの特徴と統計

ite	Data Type	Format	N(MSFA)	N(Frames)	N(FrameEle	N(M-W-P)	Note
OCAL Hiki	Free	Lite, 1st	67	546	2342	1303	in Progress
OCAL Hiki1	Protected	Lite	238	1122	7107	4389	in Progress
OCAL Hiki2	Protected	1st	69	821	3121	1685	to be Converted int
Total			374	2489	12570	7377	

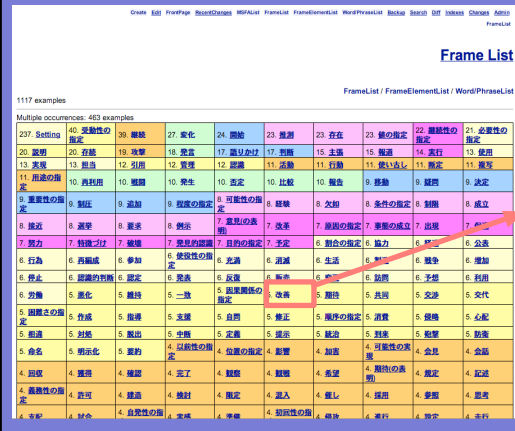
開発速度の向上と語彙資源としての有用性を目指して 2007年中旬に MSFA Lite に移行

対訳データの日本語文に付与されたフレーム記述を英語文に“転用”する作業も試行中

(A lot of) Things to Do

1. インターフェイスの改良 (特に頻度順提示以外のソート法の実装)
2. 検索機能の実装
3. FEClass List の追加
4. Pattern List の追加
5. フレーム (とその副作用としてのフレーム要素の)の階層化
6. 言語ごとのフレーム(要素), 語句リストの分離

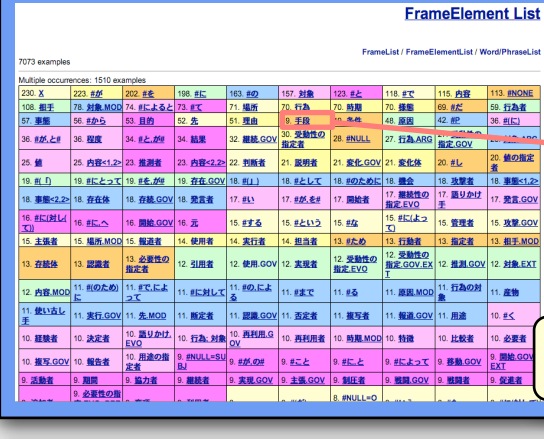
A. Frame List (部分)



Frame-to-MSFA の画面でフレーム要素の抽象化が必要

Frame-to-MSFA の画面でフレームの下位類のための情報が落ちている (改良中)

B. FrameElement List (部分)



パターンの要素(#要素)とフレームの要素(非#要素)が混在していて煩雑 (改良予定)

C.Word/Phrase List (部分)

Word/PhraseList で現在データベース化されているのは (i) MSFAの形態素列に現われた形態素と (ii) 単一のフレーム要素を構成する形態素列の二種類だけ. 現在, これらに加えて (ii) (修飾語句と一緒になったフレーム要素を構成する) 複合的な形態素列のデータベース化を実装中 (助詞 (MARKER=#要素)つきの名詞句はこれに含まれる)