

## センター機関推進プロジェクト実施報告書

申請者名：廣田 輝直

タイトル：古典一次資料上における知識 DB 構築支援の試み（1年度目／2年計画）

### ・プロジェクト全体計画（概要・目的・意義など）

当センターがこれまでに蓄積してきた古典一次資料のさらなる活用のためには、将来的にスキャン画像の電子テキスト化作業を経て、そこに研究者による翻訳、研究論文等による知識を抽出し関連づけてゆく必要がある。

本プロジェクトでは、翻訳テキストに含まれる意味情報を原文テキストと密に関連づけることによって、原語の持つ意味の広がり进行分析したり、原語->翻訳語->原語とたどることによって関連概念を探し出すモデルシステムを作成する。得られた知見をもとにテキストから特定概念に言及している部分を検索したり、テキスト群から辞書を作成する支援システムとして応用をめざし、同時に、古典学者の丹念な読みの成果を、二次利用可能な形で蓄積してゆくにはどのようなシステムが必要かを明らかにする。

### ・今年度の進捗状況

当初計画の1)の電子テキストの準備については、完全に自由に使えるデータとして著作権の切れた1885年のfausbøl版PTSテキスト等についてOCR作業、エラー箇所訂正作業を行い、n-gram検索できるようデータベース化した。また種々の制限付きながらも公開電子テキストとして入手できる、タイ第5、第6結集版、スリランカ版（GNUライセンス）、ビルマ第6結集版(world tipitaka)を原文テキストとして、K.R.Norman（著作権あり）、中村元（著作権あり）、正田訳（公開）を翻訳テキストとして準備した。この作業の副産物としてromanized pali用のOCR環境が得られた。

当初計画の2)の原文-翻訳対照データの抽出については、まず、パーリ語のstemming（活用語から辞書語への変換）辞書の作成を独自に行うことにした。オープンソースプロジェクトのDigital Pali Readerのように、語形変化、連声のルールから機械的にstemming候補を生成することも行われているが、この方法では多数の意味的にありえない候補が得られてしまう。機械的な解釈には当然限界があるから、翻訳文との対照から得られた実際に解釈可能なstemmingのみを辞書形式で蓄積していく方が知識の蓄積の上で適切と考えた。PTSのPali-English Dictionaryの見出し語と活用形に一致しない単語について、手作業でstemming辞書に追加することにした。この作業は継続中である。また、原文と翻訳とを対照づけるためには、異本や版の違いになるべく影響されずに、文章中の単語位置を一意に決定する参照手法が必要であるが、XML系の既存規格では難しいため、単語境界へのアンカータグの埋め込みとにより行うこととした。

当初計画の3)については、stemming辞書の作成が完了し次第実現する予定である。

### ・公開予定の具体的な成果物

原文-翻訳-スキャン画像表示デモ（23年5月ごろ予定）

<http://pali.ioc.u-tokyo.ac.jp/>

また、本システムの応用として、倉石ノートアーカイブ（23年度内順次公開）

<http://kuraishi.ioc.u-tokyo.ac.jp/>

（派生物）

fausbøl版PTSテキスト

romanized pali用のOCR学習データ

各翻訳に対応するstemming辞書

Pali-English Dictionary見出し語データ