アラビア語学習者コーパス構築に向けたタスク内容の検討

井上剛，イハーブ・エベード，佐野洋

アラビア語

- アラビア語話者人口
  - 話者数約2億4200万人
  - 世界で第5位

- 言語特徴
  - アフロ･アジア語族セム語派
  - VSO型
  - 屈折語

アラビア語教育・学習

- アラビア語学習者の増加
  - 1925年には大阪外国語専門学校1校だけ
  - 2004年には48大学まで増加
  - 2012年から東京外国語大学ではアラビア語専攻の定員倍増（15名→30名（定員））
- アラビア語教育の重要性
  - 昨今の世界情勢
  - イスラーム文化圏の拡大

アラビア語コーパス関連研究

- （Abuhakema et al., 2008 ）
  - 対象： 英語母語話者
  - 規模： 約8,000語
  - タスク: 不明

- （Farwaneh and Tamimi, 2012）
  - 対象： 英語母語話者
  - 規模： 約51,000語
  - タスク: 説明，ナラティブ，指示タイプの自由作文

アラビア語コーパス関連研究

- （Alfaifi et al., 2014）
  - 対象： 68の異なる母語話者
  - 規模： 約280,000語
  - タスク: ナラティブ，ディスカッションタイプの自由作文

- （Inoue et al., 2015）
  - 対象： 日本語母語話者
  - 規模： 約800語
  - タスク: テーマを指定しない自由作文

学習者コーパス

- コーパス構築
  - 規模(語数)
  - 対象(母語話者，対象言語)
  - 代表性(何を調査するための言語資料か)
  - データ形式
  - アクセスビリティ(公開の手段・方法)

学習者コーパスの代表性を検討

タスク内容の検討

- 学習者コーパスの代表性
  - 中間言語を分析・推定
  - 母集団性質としての代表性が重要

- タスク内容の比較検討
  - 自由作文タスク
  - 翻訳タスク

自由作文タスク

- 自由作文タスク
  - 学習者コーパス構築には一般的
  - タスク内容の統制
    - テーマ
  - タスク環境の統制
    - 時間制限
    - 参考資料の使用許可

自由作文タスクの例

- （Alfaifi et al., 2014）
  - ナラティブタイプ
    - テーマ：「a vacation trip」
    - 環境統制：40分の時間制限，参考資料の使用不許可
  - ディスカッションタイプ
    - テーマ：「my study interest」
    - 環境統制：なし

自由作文タスクの検討

- 利点
  - 文体や談話的特徴が観察できる
  - タスク作成コストが少なくて済む
- 欠点
  - 潜在的誤用が表出しない
    - 回避（Shacter, 1974）
  - 初級学習者や作文を苦手とする学習者からのデータ収集が困難

翻訳タスク

- 翻訳タスク
  - 学習者コーパス構築には一般的ではない
  - タスク内容の統制
    - 定められた内容，表現
  - タスク環境の統制
    - 時間制限
    - 参考資料の使用許可

翻訳タスクの例

- （安田ほか2009）
  - 翻訳タスク
    - 書籍から日本語文1,500文をランダムに抽出
    - 300文からなる課題セットを5つ作成
    - 学習者Webブラウザ上で英訳を行わせる
  - コーパスは実験用に構築されており，非公開

翻訳タスクの検討

- 利点
  - 潜在的な誤用を顕在化できる
  - 初級学習者や作文を得意としない学習者からも一定の産出量を見込める
- 欠点
  - 形態，統語，語彙的誤りだけしか抽出できない
  - 翻訳題材の著作権処理

ここまでのまとめ

- 自由作文タスクでは，潜在的な誤用が表出しない
  - 回避（Shacter, 1974）
- 翻訳タスクを採用することで，誤用を顕在化
  - 学習者ごとのコミュニケーションストラテジーによる個人差を吸収
- 翻訳題材は，著作権処理が必要
  - 題材を執筆する，あるいは許諾を得れば著作権処理をしなくてよいが，コストがかかる

提案するタスク内容

- 青空文庫から題材を選定
  - 著作権処理の必要なし

- 青空文庫とは
  - インターネット上の電子図書館
  - 著作権の消滅した作品を中心に電子化
    - テキスト形式，XHTML形式
  - 13,249作品収録（2015年9月3日時点）
    - うち12,997作品は著作権保護期間が終了

青空文庫を用いた翻訳タスク

- 青空文庫からタスクとして適切な作品を選定
- 選定した作品の一部を翻訳タスクとする
- 翻訳者となる学習者から許諾が得られれば，
  一般公開可能

課題

- どのような作品を選定するか
  - 作品の時代（近代 vs. 現代）
  - ジャンルの多様性
- 学習者にどのようにタスクを提示するか
  - 文章難易度の判定
  - タスク量の調整

多言語への展開

- タスクの対訳作成
  - 母語を固定して対象言語を多言語化
    - ある母語話者の第二言語における産出について，言語横断的に観察できる
  - 対象言語を固定して母語を多言語化
    - ある対象言語について母語話者ごとの特徴を観察できる

多言語への展開

- 題材の多様化
  - 任意の言語で書かれた題材からタスクを選定し，対訳を作成
    - 青空文庫にないジャンルを補う
    - 複数のリソースを複合的に利用
    - 特定の言語や文化圏に依存しない言語横断的な翻訳タスクを実現
  - プロジェクト・グーテンベルグ
    - 青空文庫の外国語版
    - 英語，ドイツ語，フランス語，イタリア語，ポルトガル語
    - 49,801作品収録（2015年9月3日時点）

おわりに

- 自由作文タスクでは，潜在的な誤用が表出しているとは言えない
- 翻訳タスクを採用することで，誤用を顕在化
- 青空文庫に収録されている作品を用いて，翻訳タスクを作成する手法を提案

参考文献

- Lewis, P., Simons, G and Fennig, C.: *Ethnologue: Languages of the World*, 17th edition. SIL International (2013).
- アラブイスラーム学院(編):日本における*アラビア語*の現状:教育と業界のニーズ, *門屋*由紀:日本の大学における*アラビア語*教育の現状とその問題:*アラビア語*教育の歴史とアンケート調査の結果から, pp.1-49, アラブイスラーム学院 (2006).
- Abuhakema G., Faraj, R., Feldman., et al.: Annotating an Arabic Learner Corpus for Error, LREC (2008).
- Farwaneh, A. and Tamimi, M.: Arabic learners written corpus: A resource for research and learning, (2012).
- http://l2arabiccorpus.cercll.arizona.edu/?q=homepage (参照2015-0903)
- Alfaifi, A. Y. G., Atwell, E., and Hedaya, I.: Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners, LCSAW (2014).
- Inoue, G., Karim, E. A., Ebeid, E. et al. Towards construction of a learner corpus of Arabic – a preliminary study –, IWALS (2015).
- Schachter, J.: An Error in Error Analysis 1. *Language learning*, Vol. 24. No 2. pp.205-214 (1975).
- 安田圭志, 喜多村圭祐, 山本誠一ほか: 多重タグ付き英語学習者コーパスの開発と英語能力自動測定への応用, 自然言語処理, Vol.16, No. 4, pp.47-63 (2009).