Towards construction of a learner corpus of Arabic - a preliminary study -

Go Inoue, Ahmed Ezz Karim, Ehab Ebeid, Hiroshi Sano, Keiko Mochizuki
Overview

• Expected to be the **FIRST** Arabic learner corpus of L1 Japanese learners

  – Cross-linguistic parallel learner corpus
    • Common topic and common translation task
  – Dual-direction corpus
    • L1 Japanese and L2 Arabic/ L1 Arabic and L2 Japanese
  – Continuing review of writing skills from an interlanguage standpoint
    • Collect tasks from the same learner periodically
Growing importance of Arabic language education
   - Arabic is the 5th most commonly spoken language in the world

Need for more efficient ways to teach and learn Arabic language
   - Take into account learners’ native languages

Demand for Arabic language resources
   - Lack of publicly available information on learner corpora
• (Abuhakema et al., 2008)
  – Pilot corpus of Arabic as a second language
  – Publicly unavailable
  – 9,000 words
  – L1 English
• (Hassan and Daud, 2012)
  – Publicly unavailable
  – 240,000 words
  – L1 Malaysian
• (Alfaifi et al., 2014)
  – Publicly available
  – 280,000 words
  – 43 different L1 (Japanese is not included)
Objectives

• Provide linguistic resources for Arabic education for L1 Japanese learners

• Analyze L1-interference error patterns

• Continue review of writing skills from an inter-language standpoint
• Extend the framework of the Japanese-English-Chinese cross-linguistic parallel learner corpus

  “Construction of a Japanese-English-Chinese Online Error Corpus and development of English, Japanese and Chinese language pedagogy taking into account learners’ native languages”

KAKEN(25284101)
Three Characteristics of this Corpus

- Cross-linguistic
- Dual-directional
- Continuing review of writing skills
Dual-directional Corpus

Japanese

Arabic

L1

L2

L2

L1
Continuing Review of Writing Skills

- **July 2016**: (2 years after matriculation)
- **July 2017**: (Upon return from 1-year study abroad)
- **February 2018**: (1 semester after return from study abroad)
Tasks

- Essays on a certain topic
  - Common topic between the target languages
    - Find out expressive variations (or differences) among essays
- Translation of essays, articles, and other materials
  - Common content between the target languages
    - Prevent “avoidance” occurred when writing essays
    - Highlight the differences between interlingual errors and intralingual errors
Corpus Design

• Learner profile
  – 1st to 4th year undergraduate students
  – Majoring in Arabic studies at Tokyo University of Foreign Studies
  – Most of the students had no prior experience of learning Arabic before matriculation
• Recording Format
  – XML

• Meta Data
  – Learner Data
    • Gender, age, nationality, number of years learning Arabic, number of hours learning Arabic in a week (coursework and outside of coursework), first language, experience of learning foreign languages, experience of living overseas, educational language, language proficiency test scores (English or other languages)
  – Text Data
    • Theme, date of production, timed or not timed, restriction on use of references, text length
• 7 essays collected
  – Learner Profile
    • Undergraduates in their 3rd year
    • Majoring in Arabic Studies
  – Directions
    • Free composition
    • Around 100 words
    • Untimed
    • References permitted
عندما كنت طالبة في مدرسة ثانوية، كنت عزفت على القلوت الياباني، استطاعت أن عزفت بسُمرة نسبياً، لأنني اتمنيت إلى ناد موسيقى ناحاسية مدرستي وعزفت على القلوت الغربي. القلون الياباني، اسمه "شينوبوياي"، الموسيقى تقليدية في اليابان. وقدم عن قارة الصين. ويتضمن في موسيقى التقليدي، مثل موسيقى "غاغاكو" وأنغنية شعبية وفنون مسرحية "كابوكي". موسيقى عيد. أحب "شينوبوياي" لأن جرشه جميل جداً وأنا مسرورة عندما أستطيع أن أحصل على الصوت نظيفة. لا يستغنى عنه في موسيقى تقليدي ياباني.
Expected Applications

• Corpus-based learning and teaching materials

• Search interface for educational and independent learning purposes

• Automatic error correction system
• Further consideration of task materials

• Development of a tagset for error annotation

• Development of a search interface
References

