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-directional 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

Background

- 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

Related Studies

- （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

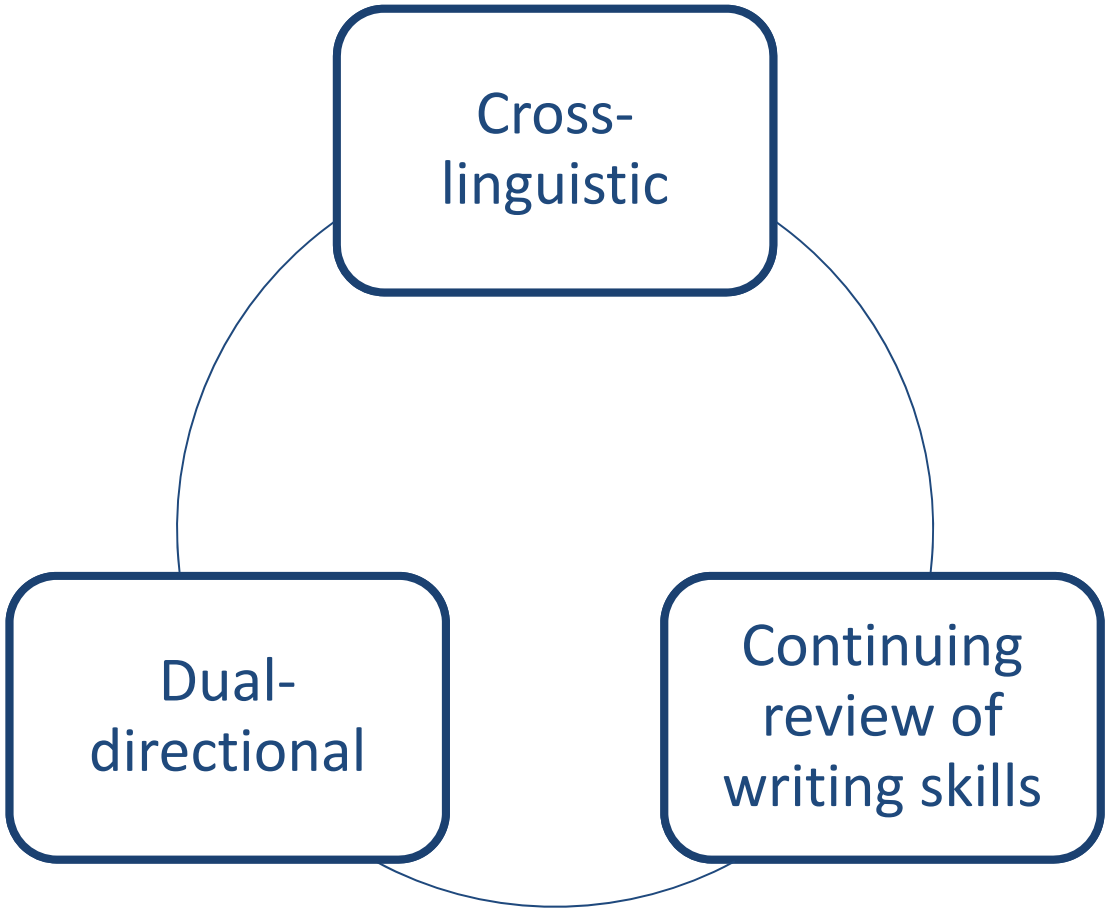
Framework of Corpus

- Extend the framework of the Japanese-English-Chinese cross-linguistic parallel learner corpus

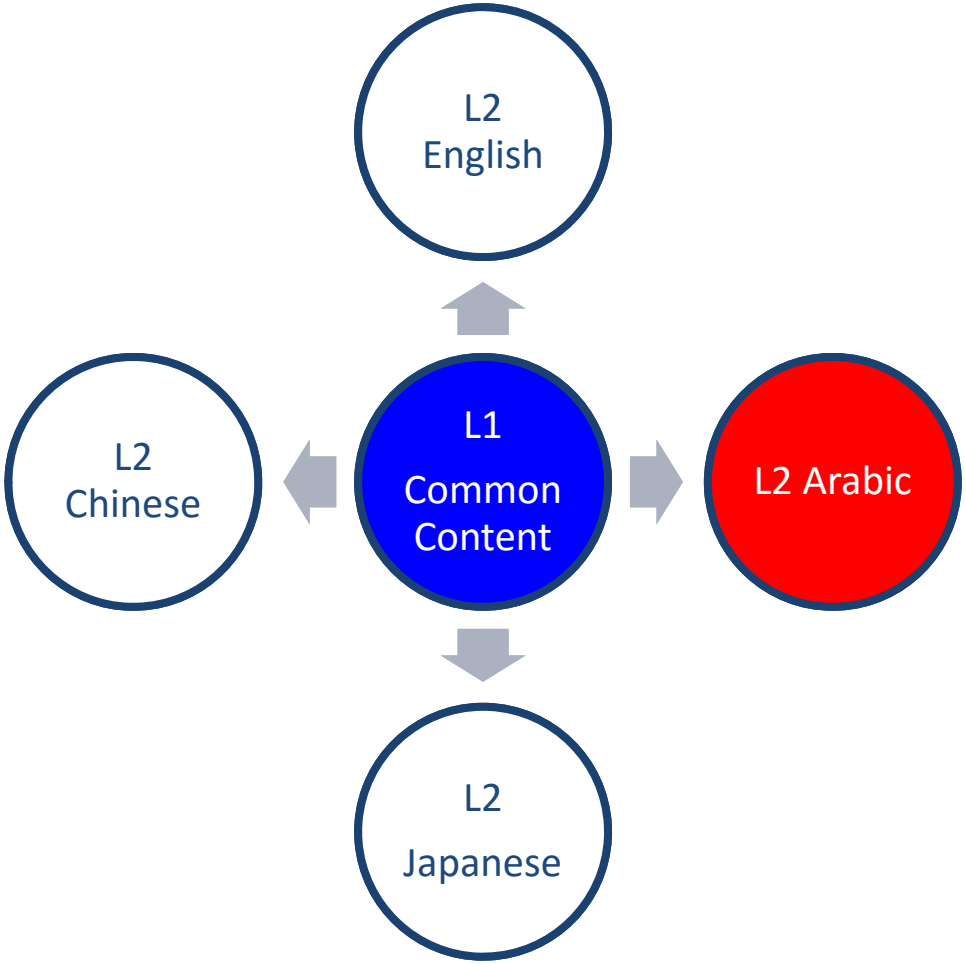
  – “Construction of a Japanese-English-Chinese Online Error Corpus and development of English, Japanese and Chinese language pedagogy taking into account learners’ native languages”

KAKEN(25284101)

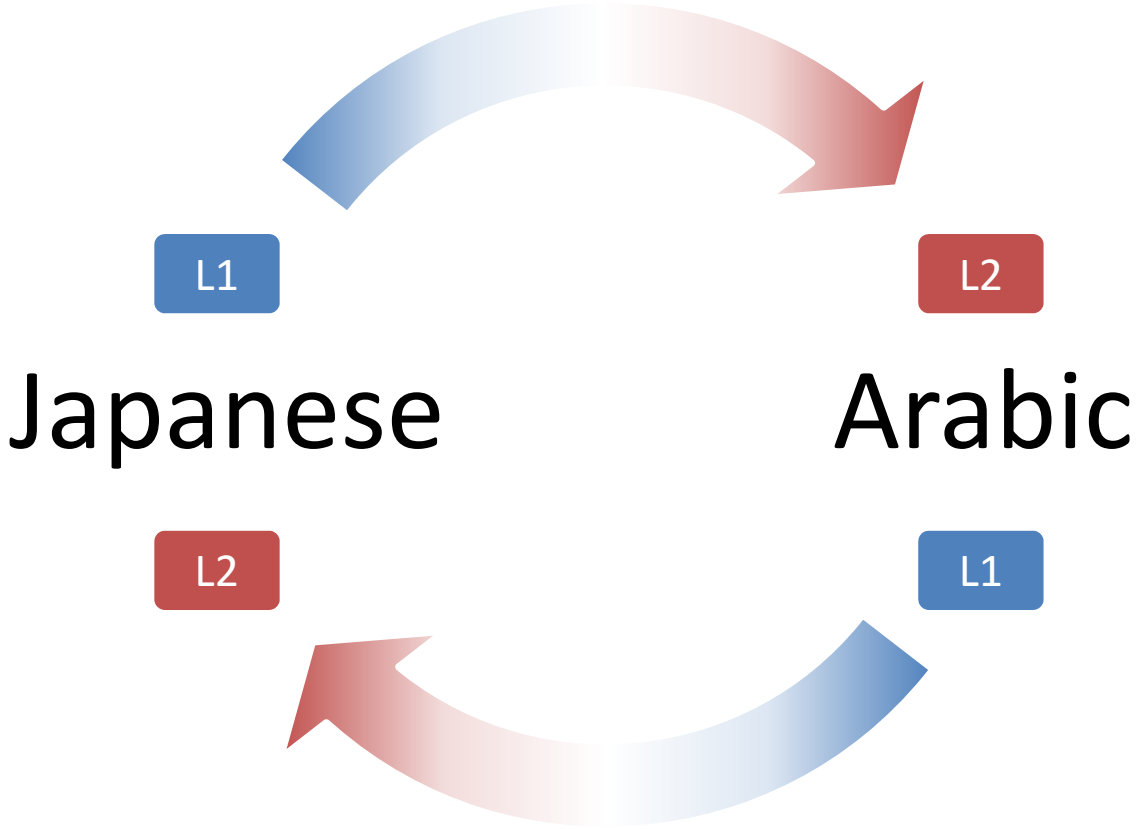
Three Characteristics of our Corpus



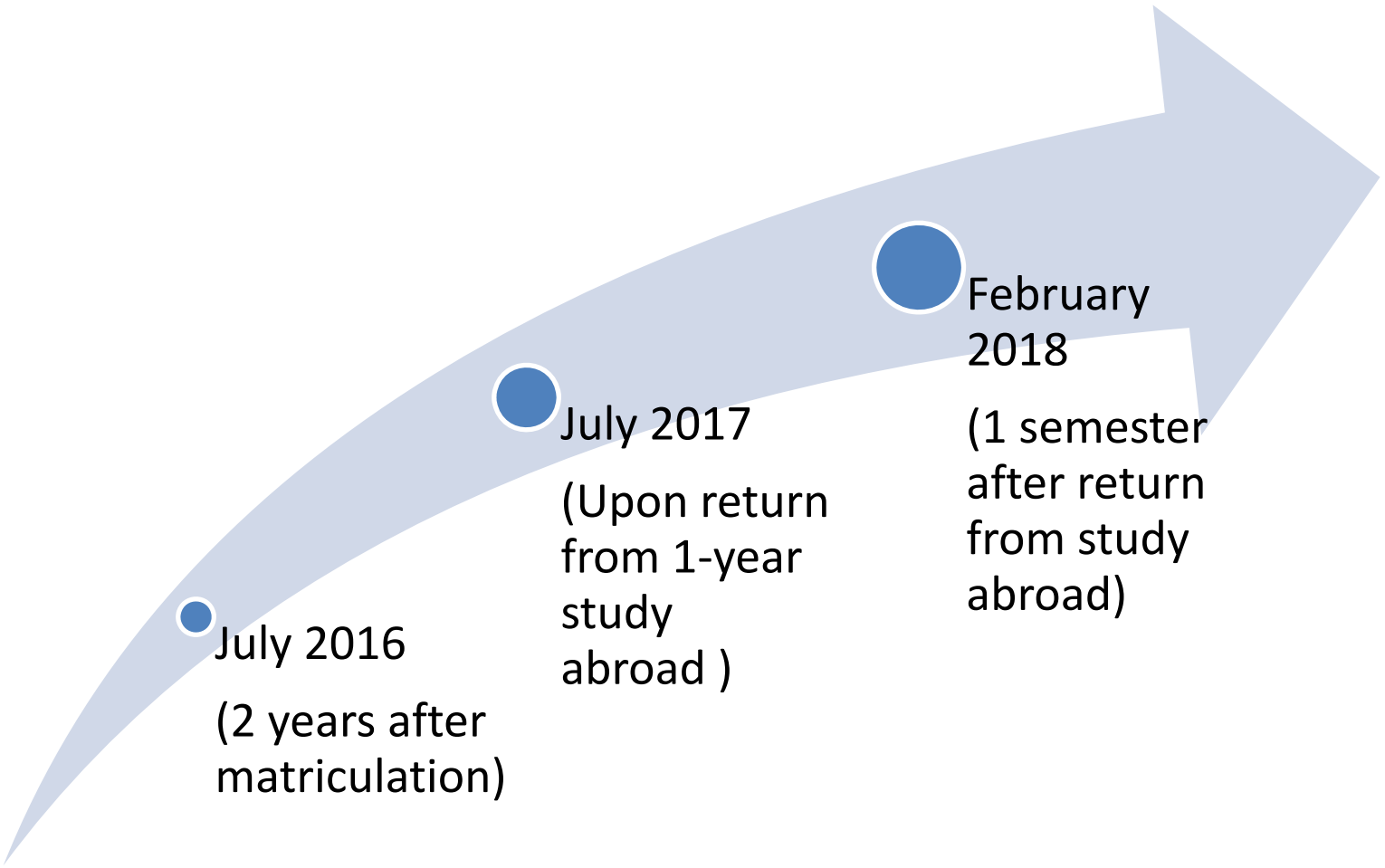
Cross-linguistic Corpus



Dual-directional Corpus



Continuing Review of Writing Skills



Corpus Design

- 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

Corpus Design

- 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

Data Sample

- 7 essays collected
  - Learner Profile
    - Undergraduates in their 3rd year
    - Majoring in Arabic Studies
  - Directions
    - Free composition
    - Around 100 words
    - Untimed
    - References permitted

Sample File

```
<?xml version="1.0"? encoding="UTF-8"?>
<!— TUFS Arabic Learner Corpus sample—>
<header>
      <id>tu_ar_2013_01</id>
      <learner_data>
            <gender>female</gender>
            <age>21</age>
            <nationality>Japanese</nationality>
            <first_language>Japanese</first_language>

            {… snip …}

      </learner_data>
      <text_data>
            <theme>Free Composition</theme>
            <production_date>01/07/2015</production_date>
            <timed>no</timed>
            <reference_use>yes</reference_use>
            <length>82</length>
      </text_data>
</header>
<body>
```

عندما كنت طالبة في مدرسة ثانوية، كنت عزفت على الفلوت الياباني، استطاعت أن عزفت بسهرة نسبيا، لأني انتميت إلى ناد موسيقي نحاسية مدرستي وعزفت على الفلوت الغربي. الفلون الياباني، اسمه "شينوبوأي"، آلة موسيقية تقليدية في اليابان. وقدم عن قارة الصين. ويستخدم في موسيقي التقليدي، مثلا موسيقي "غاغاكو" وأغنية شعبية وفنون مسرحية "كابوكي" وموسيقي العيد. أحب "شينوبوأي" لأان جرسه جميل جدا وأنا مسرورة عندما أستطيع أن أحصل على الصوت نظيفة. لا يستغنى عنه في موسيقي تقليدي ياباني.

```
</body>
```

Expected Applications

- Corpus-based learning and teaching materials
- Search interface for educational and independent learning purposes
- Automatic error correction system

Future Plans

- Further consideration of task materials
- Development of a tagset for error annotation
- Development of a search interface

References

- Abuhakema, G., Feldman, A., & Fitzpatrick, E. (2008). Annotating an Arabic learner corpus for error. The proceedings of the International Conference on Language Resources and Evaluation (LREC 2008). May26–June1,Marrakech,Morocco.

- Hassan, H., & Daud, N. (2011). Corpus analysis of conjunctions: Arabic learners' difficulties with collocations. Paper presented at the Workshop on Arabic Corpus Linguistics (WACL), 11th -12th April 2011, Lancaster University, UK.

- Alfaifi, A., Atwell, E. and Hedaya, I. (2014). Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. In proceedings of *the Learner Corpus Studies in Asia and the World (LCSAW)* 2014, 31 May - 01 Jun 2014. Kobe, Japan.