JAECS

Japan Association for English Corpus Studies

英語コーパス学会 2023 年度春季研究会

英語コーパス学会 2023 年度春季研究会は，下記の内容で実施いたしました。

The JAECS Spring Forum 2023 will be held online on May 13th, 2023 (Sat.) JST.

・開催日時：2023 年 5 月 13 日(土) 13:00～16:20

・参加費用：無料（No registration fee）

・場所：オンライン（Online）

プログラム（Schedule）

13:00～13:03 開会式（Opening）

13:03～14:55 Part 1

コーパスと CEFR 研究会担当（Organized by the JAECS SIG on Corpora and CEFR）

司会進行（Chair）：投野由紀夫（Yukio Tono）

1. Linking lexicographic and CEFR resources (13:10～13:40, Video presentation)

・Kris Heylen (Dutch Language Institute)

・Ilan Kernerman (Lexicala by K Dictionaries)

・Carole Tiberius (Dutch Language Institute)

発表概要（Abstract）

The Common European Framework of Reference for Languages (CEFR) promotes the development of empirically based datasets for 30 languages of Europe according to graded proficiency levels, by grading word difficulty prevalent in native and additional language learning, in production and reception tasks, for text readability analysis and vocabulary testing. We will present ongoing work to better support the creation of relevant teaching and training materials by linking CEFR difficulty-graded word lists with lexicographic data, cross-lingualizing different language sets, and uploading the by-products to the Linguistic Linked Open Data cloud. The case studies involve, on the one hand, the corpus-derived CEFR-graded resources compiled for a number of languages in the CEFRLex project and, on the other hand, the lexicographic datasets for Dutch of K Dictionaries and the Dutch Language Institute. The challenges we discuss include (a) evaluating the words in CEFR lists and linking their appropriate senses in dictionaries, (b) ensuring the additional lexicographic components exclude words from higher CEFR levels, and (c) creating CEFR lists for languages that don't have them yet and linking them to lexicographic data. Additionally, we seek diverse partners with language learning expertise, lexicographic resources, and linked data know-how, to join in a future project.

Keywords: language learning; proficiency levels; CEFR lists; lexicographic resources, LLOD cloud

2. Towards an infrastructure for the semi-automatic development of corpus-based language exercises. (13:40～14:10, Video presentation)

・ Katrien Depuydt (Dutch Language Institute)

・ Jesse de Does (Dutch Language Institute)

発表概要（Abstract)

We report on the development of an environment for the semi-automatic creation of corpus-based language exercises using linguistically annotated corpora and lexical resources, both dedicated CEFR-lexical[1] and general-purpose lexical resources[2]. The aim is to reduce manual work, to produce more natural exercises “in context”, and to allow for more efficient creation of exercise material adjusted to specific text types, e.g. “easy” material[3] or sports-related texts.

We have created a working environment, which combines the BlackLab[4] corpus exploration environment and the lexical resources in a Vue.js[5]-based frontend. A language learning specialist selects vocabulary and choses a type of exercise (cloze, scrambling, …), corresponding to a parametrized corpus query suitable for BlackLab. The exercises constructed automatically from the query results can be previewed, pruned and adjusted by the specialist.

The environment has been used to create exercises for the prototype of a mobile app developed in cooperation by Oefenen.nl[6] and Game Architect[7].

References

[1] notably NT2Lex, https://cental.uclouvain.be/cefrlex/nt2lex/

[2] Especially the morphosyntactic lexicon GiGaNT-Molex,

https://taalmaterialen.ivdnt.org/download/gigant-molex2-0/

[3] In our case: the Start!-krant, https://www.eenvoudigcommuniceren.nl/kranten/start-krant

[4] https://inl.github.io/BlackLab/

[5] https://vuejs.org/

[6] https://oefenen.nl/

[7] https://gamearchitect.eu

3. Methodological issues regarding a corpus-based analysis for the development of the CEFR Grammar Profiles for Estonian (14:10〜14:40, Online presentation)

・Jelena Kallas (Institute of the Estonian Language)

発表概要（Abstract）

In the talk, we will introduce the Estonian Profile （the Estonian Vocabulary Profile and the Estonian Grammar Profile), which is designed to support the CEFR illustrative descriptors scales of linguistic competence with language-specific description. We use corpus research techniques as our principal method for investigating how learners acquire grammar. Different types of corpora were used for the development and validation of the Estonian Profile, including the L2 Estonian coursebook corpora Estonian as a Second Language Coursebook Content Corpus 2017, Estonian as a Second Language School Coursebook Content Corpus 2021, and the L2 Estonian learner language corpora Estonian learners' language corpus EMMA and the TLU interlanguage corpus. For the analysis, we used the Corpus Query Systems KORP, EMMA and Sketch Engine. We will demonstrate the main functionalities of CQSs used for grammar profile development and propose possible future research to facilitate a corpus-based analysis for the development of the CEFR Grammar Profiles.

Keywords: learners' corpora; textbook corpora; CEFR proficiency levels; lexicographic resources, Estonian

4. Q & A's (14:40～14:55)

15:00～16:00 Part 2 招待講演（Plenary Talk）

DDL 研究会担当(Organized by the JAECS SIG on DDL)

Speaker: Dr. Luciana Forti (University for Foreigners of Perugia, Italy)

Title: Exploring the affordances of CEFR-based learner corpora in Data-driven learning

講演概要（Abstract）

Data-driven learning (DDL) practices are known for being predominantly based on L1 corpora. According to the review by Boulton & Vyatkina (2021), in fact, only 4% of the total number of DDL studies published over 30 years (n = 489) rely on learner corpora instead. Yet, claims on the potentially beneficial applications of learner corpora in DDL have been made in many learner corpus research (LCR) papers, deeming learner corpora as rich repertoires of linguistic input for the second language learner to explore.

This presentation will discuss not only the benefits of using learner corpora in DDL, but also present the specific pedagogical advantages of learner corpora based on CEFR-categorised texts. Special attention will be devoted to some of the pedagogical applications of the CELI corpus (Spina et al., 2022), a corpus of learner Italian texts produced under language certification exam conditions, evenly divided into proficiency levels B1, B2, C1 and C2.

References

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions. *Language Learning and Technology, 25*(3), 66–89.

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., & Zanda, F. (2022). Il Corpus CELI: Una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue, 14*(1), 116–138.