Construction of the corpus of everyday Japanese conversation : An interim report

journal or publication title	Proceedings of the LREC 2018 Special Speech Sessions
page range	29-29
year	2018-05-09
URL	http://doi.org/10.15084/00001913

Construction of the Corpus of Everyday Japanese Conversation: An Interim Report

Hanae Koiso[†], Yasuharu Den[‡,†], Yuriko Iseki[†], Wakako Kashino[†], Yoshiko Kawabata[†], Ken'ya Nishikawa[†], Yayoi Tanaka[†], Yasuyuki Usuda[†]

[†]National Institute for Japanese Language and Linguistics
10–2 Midori-cho, Tachikawa, Tokyo 190–8561, Japan
koiso, iseki, waka, kawabata, nishikawa, yayoi, usuda@ninjal.ac.jp

[‡]Graduate School of Humanities, Chiba University
1–33 Yayoicho, Inage-ku, Chiba 263–8522, Japan
den@chiba-u.jp

Abstract

In 2016, we launched a new corpus project in which we are building a large-scale corpus of everyday Japanese conversation in a balanced manner, aiming at exploring characteristics of conversations in contemporary Japanese through multiple approaches. The corpus targets various kinds of naturally occurring conversations in daily situations, such as conversations during dinner with the family at home, meetings with colleagues at work, and conversations while driving. In this paper, we first introduce an overview of the corpus, including corpus size, conversation variations, recording methods, structure of the corpus, and annotations to be included in the corpus. Next, we report on the current stage of the development of the corpus and legal and ethical issues discussed so far. Then we present some results of the preliminary evaluation of the data being collected. We focus on whether or not the 94 hours of conversations collected so far vary in a balanced manner by reference to the survey results of everyday conversational behavior that we conducted previously to build an empirical foundation for the corpus design. We will publish the whole corpus in 2022, consisting of more than 200 hours of recordings.

Keywords: Corpus of everyday Japanese conversation, corpus design, legal and ethical issues, corpus evaluation

The content of this talk is identical to that of the paper 469 of the LREC main conference. The full PDF is available in the proceedings of the main conference, pp. 4259–4264.