The role of metadata in the Balanced Corpus of Contemporary Written Japanese in the analysis of register

ホドシチェク・ボル*, 山元啓史**
Bor HODOŠČEK*, Keiji Yamamoto**
*国立国語研究所,
**東京工業大学／カリフォルニア大学サンディエゴ校

Summary: This study proposes to evaluate the discriminatory power of the metadata contained within the Balanced Corpus of Contemporary Written Japanese (BCCWJ) for the modeling of linguistic variation (register). The available metadata is analyzed into several categories thought to influence register (NDC category hierarchy, gender, topic, media, etc.), which are then used to partition the documents within the corpus along different category groupings. The resulting similarity scores between the linguistic features of the category groupings reveal the relationships between--as well as the constraints and gaps within--the metadata, which is essential information for the reliable measurement of differences in register.