Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora

Itsuko FUJIMURA (Nagoya University)
Shoju CHIBA (Reitaku University)
Mieko OHSO (Nagoya University)

In this paper, we describe the lexical and grammatical characteristics of spoken Japanese by comparing the lexical profile of a conversation corpus with a 100 million word-sized balanced corpus of written Japanese.

As the basic material for our research, we use NUCC, the Nagoya University Conversation Corpus, which was released in 2009 and is now available for research purposes from the site (URL: http://dbms.ninjal.ac.jp/nuc/index.php?mode=viewnuc) free of charge. It contains 129 natural uncontrolled conversations, each between 30-60 minutes long. 198 participants, who are all native speakers of Japanese of various ages and from diverse academic backgrounds, took part in the recordings held from 2001 to 2003. The total recording time amounts to 100 hours. The data was then carefully transcribed and morphologically analyzed. The corpus contains about 1.5 million morphemes, which shows that the corpus is the largest corpora of spontaneous spoken Japanese. As a caveat, since there are more female participants (161) than male (37), and many of the participants are graduate students majoring in linguistic subjects, the data taken from this corpus may reflect this lack of balance of the participants. In this study, we carefully exclude the disparity resulting from this population bias.

For the current quantitative analysis, we devised a tool which compares two corpora quantitatively, called Lexical Profiling System. With this tool we compare NUCC with a 100 million-word sized large corpus of written Japanese (BCCWJ). BCCWJ, the Balanced Corpus of Contemporary Written Japanese, includes about 170,000 samples of written texts, which is classified into carefully designed subcorpora (genres), namely books, newspapers, magazines, whitepaper texts, Internet texts, Diet minutes, among others. It was compiled from 2006 to 2010, and is to be published in October 2011. We see BCCWJ as a good sample of written Japanese, because the corpus contains samples from many genres, each of which is large enough for the current purposes of this research. It also utilizes unique sampling strategies so that the corpus represents the most recent status of contemporary written Japanese (Maekawa 2007). We can evaluate with BCCWJ, for example, the pervasiveness of a morpheme or a morpheme sequence in the light of the distribution in different genres (frequency per subcorpus) or among the files (occurrences per file).

The Lexical Profiling System is designed to compare corpora of different size, genre, or even an individual part of a corpus with the whole of the same corpus. For that purpose, the data to be compared are morphologically analyzed, and the frequency of lemmas, word forms, bi- and trigrams is counted and stored in a database. The tool then computes the frequencies of these units using different statistical measures such as LLR (Log-likelihood ratio, Dunning 1993), Dice coefficient, MI-score (Church and Hanks 1990), among others.

In this study, we claim with numerous examples that the systematic comparison of lemma, word form, n-gram etc., can show not only lexical, but also grammatical structures peculiar to spoken Japanese (cf. Blanche-Benveniste 1997 for spoken French). Our study reveals that many aspects characteristics of spoken Japanese are only observable through lexical profiling techniques, because mere frequency information cannot show the deviation it actually implies. Our outcome thus shows that full quantitative comparison of the target corpus and the sample (hence large and balanced) corpus is needed to assess the lexical and grammatical features of the target corpus.

Then we argue that the linguistic features of NUCC indicate that spoken Japanese shows a "fragmented" nature in essence, which is opposed to the structure of "elaboration" peculiar to written Japanese. Our findings thus show that the lexical and grammatical structure of spoken/written Japanese is much like Korean (Fragmented vs. Elaborated Structure), rather than like English (Involved vs. Informational Production, cf. Biber 1995). Although this result may reflect the fact that the grammatical structures of Japanese and Korean resemble each other, it is clear that Japanese conversations also show the feature of Involved Production as well. For example, Japanese interactional particles (like "ne", "yo") or the Japanese nodding-word "uN" appear mainly in spoken discourse, while not in written discourse.

References

Biber, Douglas (1995). Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge: Cambridge University Press.

Blanche-Benveniste, Claire (1997) Approches de la langue parlée en français. Paris :Ophrys.

Church, Kenneth W. and Hanks, Patrick (1990). “Word association norms, mutual information, and lexicography.” Computational Linguistics 16(1), 22-29.

Dunning, Ted (1993). “Accurate Methods for the Statistics of Surprise and Coincidence.” Computational Linguistics 19(1), 61-74.

Maekawa, Kikuo (2007). “Design of a balanced corpus of contemporary written Japanese.” Proceedings of Symposium on Large-Scale Knowledge Resources (LKR2007), 1-3. March 2007, Tokyo, Japan. Pp. 55-58.