The effect of the speaker's sex and age on the relative frequencies of nouns, pronouns and verbs

When a quantitative feature of a language is utilized for a typological study, a sociolinguist would wonder what kind of language sample is utilized there. The question is not trivial, as empirical studies teach us that the age and the sex of the speaker could be a critical factor determining their verbal behavior. How can we be so sure that they do not matter in looking into the variation of the relative frequencies of nouns, pronouns and verbs?

This paper examines the effect of those two factors, the age and the sex of the speaker, using the data from a longitudinal language survey which traced people's linguistic behavior for 55 years. The survey, Okazaki Survey on Honorifics, is a questionnaire-based survey on honorific use and its consciousness in Okazaki City, Japan, conducted by National Institute for Japanese Language and Linguistics. The survey has been conducted in 1953, 1972 and 2008. Among the samples taken in the surveys here we use a panel sample that traced 19 speakers (11 males and 8 females) from 1953 (when they were 16 to 34 year old), to 2008 (71 to 89 years old).

For the analysis, we use the speakers' responses to eleven questions asking what they would say in a given situation (e.g., complaining for a short change to a shop keeper). Note that the same questions were used in the three surveys, enabling a proper control of the possible syntactic forms in the responses. The transcriptions were tagged for parts-of-speech by MeCab, a part-of-speech and morphological analyzer, using a UniDic dictionary (Den et al. 2007).

The result shows that: (1) Even though the amount of speech shows a sizable increase from 1953 to 2008, the ratios of nouns and pronouns to verbs (calculated by the formula *(Nouns + Pronouns)/(Nouns+ Pronouns+ Verbs)* following Seifart (2011) remain the same for both sexes (around 0.54 for males and 0.48 for females); (2) Males show higher values than females for all three surveys; and (3) Females tend to show larger fluctuations. A use of a different morphological analyzer, ChaSen, to the same dataset yielded different values but did not change these findings in any significant manner.

Such a robust nature of the relative frequencies of the three parts of speech in a language suggests that they are deeply rooted in the grammar of the language, which, once acquired in the adolescence, would stay the same to the rest of the life. This then would mean that one could safely take a sample for the quantitative study of the parts of speech from a discourse at whichever stage of a person, as long as the issue is their *relative* frequencies. The sex of the speaker, however, should be fixed, as speakers of different sexes would show different ratios in a discourse of a similar nature.