Using quantitative measures to investigate the relative roles of languages participating in code-switched utterances

Cathy Lonngren-Sampaio
University of Hertfordshire
C.Lonngren-Sampaio@Herts.ac.uk

1 Introduction

Jake and Myers-Scotton define Code-switching (CS) as 'language use that consists of material from two or more language varieties at any level from the discourse to the clause.'(2009:207). They propose that in 'Classic' CS there is always asymmetry between the two (or more) languages participating in CS clauses. According to the Asymmetry Principle (ibid:209) the abstract morphosyntactic frame of the bilingual clause largely, or entirely, comes from one of the languages, named the Matrix Language (ML) while the other participating language is called the Embedded Language (EL) and typically contributes content morphemes.

The aim of this study is to use quantitative methods to investigate this Asymmetry Principle in a corpus of child bilingual language and answer the following research questions: 1) Do word frequency measures provide evidence for the ML/EL asymmetry?; 2) Can correlations be found between vocabulary diversity scores and the ML/EL?; 3) What can mean word and utterance scores contribute to the investigation of the Asymmetry Principle in bilingual corpora?; and 4) Can these quantitative measures combined provide a useful method for determining the participatory roles of the languages of code-switched discourse?

2 Methodology

Twenty-five hours of recordings of naturalistic interactions between two bilingual Brazilian/English siblings (JAM, 3;6 and MEG, 5;10) and other family members, were transcribed and coded using the CHAT (Codes for the Human Analysis of Transcription) transcription system developed by MacWhinney and colleagues (MacWhinney 2012a). The resulting corpus, named the LOBILL (Lonngren BILingual Language) Corpus, then received more specific coding: language codes to differentiate English and Portuguese material were inserted throughout and a specially developed postcode was used to code bilingual utterances (see example below). Addressee information for each utterance was also included.

*MEG: <<mas a agua>[@pt]>[//] <the water is very very cold>[@en] ? [+ pe]
%add: MOT

Quantitative analyses were carried out using the CLAN (Computerized Language Analysis) software (MacWhinney 2012b), in particular the commands KWAL (which outputs specified utterances), FREQ (which outputs frequency word lists), VOCD (which outputs vocabulary diversity scores) and WDLEN (which outputs mean word and utterance lengths). An example command line is shown below:

kwal @ +t%add +t*JAM +s"PAI" +u +d | vocd +r6 +s"[+ *]"

Consisting of two parts (separated by the upright slash), first KWAL selects all utterances JAM addresses to PAI (his father). Then VOCD performs an analysis on only the CS utterances. By subsequently adding the strings -s"<@pt>" and -s"<@en>", VOCD then performs analyses on only the English material in CS utterances and then on only the Portuguese material in CS utterances.

By systematically substituting the speaker codes (JAM, MEG, MOT PAI), the addressee codes (PAI, MOT, MEG, JAM) and the commands (freq, vocd +r6, wdlen) 36 analyses were performed for each speaker, making a total of 144 analyses.

The results were examined and triangulated with other quantitative and qualitative analyses previously performed on the LOBILL Corpus. Patterns were observed and correlations were made between the four different types of measures (word frequency, vocabulary diversity, mean word length and mean utterance length) and the roles of the two languages in code-switched utterances.

3 Results

The first set of results, obtained from the frequency analyses (performed by FREQ), provided a measure of the proportion of English and Portuguese words used by the siblings and their parents when code-switching with each other. The following correlations were found: a high proportion of words in one language indicated that that particular language was acting as the Matrix Language; a low proportion of words indicated a role more akin to the Embedded Language.

Secondly, the vocabulary diversity analyses (performed by VOCD) provided separate diversity (D) scores for each language in code-switched utterances. Low D scores were found to correlate with the Matrix Language whereas high D scores were representative of the Embedded Language.

What is postulated in both cases (frequency proportions and diversity scores) is that the greater the relative difference between the values for each language, the more asymmetrical their participatory roles appear to be. This in turn means that where the relative difference in values is less disparate we would expect more equal participation of both languages in code-switched utterances.

The third and forth sets of values, which resulted from the WDLEN analyses, measured mean word and mean utterance lengths. A low mean word length was found to correlate with the Matrix Language while a high mean word length correlated with the Embedded Language. In terms of utterance length, the reverse correlation was found: a low mean utterance length provided evidence that the language was acting as the Embedded Language whereas a high mean utterance length was found to be representative of the Matrix Language. As for the first and second sets of results, the evidence suggests that the greater the comparative difference in values between the two languages (this time in terms of means) the more asymmetrical their participatory roles become.

4 Conclusion

This study set out to examine whether quantitative measures could contribute to the investigation of the Asymmetry Principle, a common feature of code-switched discourse. The results of the analyses performed with the CLAN commands KWAL, FREQ, VOCD and WDLEN, led to the formulation of correlations between the four types of quantitative values and the participatory roles (Matrix and Embedded) of the languages involved. These correlations have in turn led to the development and propos o be us h values - switch

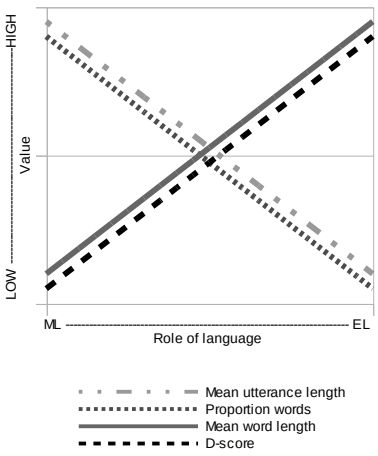

Figure 1. Schema for the interpretation of FREQ, VOCD and WDLEN values when used to investigate the relative roles of the languages participating in code-switched discourse.

If the methodology used in this study were replicated on other code-switching data sets, it is believed that the proposed schema (Figure 1) would allow for cross-linguistic comparison: the original results from the LOBILL Corpus could be compared with the results from other bilingual corpora. Such comparisons would have the potential to shed more light on the contribution that the use of quantitative measures could make when investigating the Asymmetry Principle in code-switched discourse.

References

MacWhinney, B. 2012a. *The CHILDES Project, Tools for Analyzing Talk – Electronic Edition.*. Carnegie Mellon University. Available online at: http://childes.psy.cmu.edu/manuals/chat/pdf .

MacWhinney, B. 2012b. *The CHILDES Project, Tools for Analyzing Talk – Electronic Edition.*. Carnegie Mellon University. Available online at: http://childes.psy.cmu.edu/manuals/clan/pdf .

Myers-Scotton, C and Jake, J.L. 2009. "A universal model of code-switching and bilingual language processing and production". In B. Bullock and J. Toribio (eds.) *The Cambridge Handbook of Linguistic Code-switching*. Cambridge: Cambridge University Press.