CodeSwitch-Reddit: Exploration of Written Multilingual Discourse in Online Discussion Forums*

Ella Rabinovich **Masih Sultani** **Suzanne Stevenson**
Dept. of Computer Science, University of Toronto, Canada
{ella,masih,suzanne}@cs.toronto.edu

Multilingual communities adopt various communicative strategies that navigate among multiple languages. One of the most notable of such strategies is *code-switching* (CS) – when a bilingual mixes two or more languages within a discourse, or even within a single utterance.

The sociolinguistic underpinnings of code-switching as an *oral* conversational strategy have been investigated extensively for many decades. By contrast, the analysis of *written* code-switching has only recently enjoyed a surge of interest, and remains seriously under-studied. Written text often differs greatly from conversation in its levels of both spontaneity and formality, and findings thus far have differed in their conclusions regarding the extent to which various genres of written text reflect the same communicative functions of CS as observed in oral conversation.

The growing popularity of social media and online discussion platforms poses both opportunities and new research questions regarding written code-switching. Global online forums, in which English is a lingua franca, not only draw on but create wide-reaching multilingual communities. The resulting communications lead to a wealth of data that potentially includes a large amount of code-switching across multiple language pairs. Moreover, communication on discussion platforms often resembles a hybrid between speech and more formal writing. These differing characteristics lead to new research questions regarding the extent to which findings from oral CS carry over to these online interactions.

Research is only just beginning to grapple with these issues. Computational work on code-switching in online venues has largely focused on the practical challenges that multiple interleaved languages pose to the application of standard NLP tools, rather than on the communicative purposes of CS. More broadly, computational investigation of the sociolinguistic aspects of written CS is dominated by studies conducted with a limited number of language-pairs and/or authors, thereby constraining the nature of questions that can be addressed with this data. Our work here seeks to address these gaps in the study of code-switching in online interactions. We begin by introducing the *CodeSwitch-Reddit* corpus: a novel, large, and diverse dataset of written code-switched productions, carefully curated from topical threads of multiple (including understudied) bilingual communities on the Reddit discussion platform. The corpus comprises over 135K CS messages by over 20K unique authors, spanning five language-pairs, with average post length of 75 tokens.

The uniform nature of our data (written communication from a single online discussion platform), as well as its ample size, pose novel opportunities for large-scale empirical investigation of research questions on code-switching – questions that have thus far been mainly addressed in the context of oral language. As a first study, here we explore fundamental questions about both the content and style of code-switched posts, as well as about the English proficiency level of authors who frequently code-switch.

The contribution of this work is twofold: First, we construct a novel code-switching corpus, whose size, number of language pairs, and diversity of content (consisting of posts of unrestricted length in a variety of topics) make it a desirable testbed for a range of research questions on CS in online discussion forums. Second, we demonstrate the usefulness of this dataset through an empirical investigation that sheds new light on postulated universals of CS – involving linguistic proficiency, style, and content – when inspected through the lens of online communication.

* Accepted for publication at EMNLP2019

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, page 446
Hong Kong, Nov 4, 2019.