Supplementary Materials for: 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

A Supplementary Materials

A.1 A List of Terms Used for Identification of Translations in Posts

translated, translation, translating, transliteration, verb, sentence, translates, translate, noun, translations, informal, phrase, translator, context, grammatical, formal, sentences, meaning, speaker, expression, pronounced, literally, native, lines, lyrics, spanish, languages, grammar

A.2 Annotation Guidelines for Identification of Code-switched Posts

Sample annotation instructions for English-Romanian:

Please only pick this task if you are a Bilingual English-Romanian speaker. Please read the instructions and examples carefully.)

In this task, we exploreCode-Switching (sometimes called code-mixing); a phenomenon where a person switches from one language to another in mid-conversation. You will be given posts that someone wrote, and your job is to specify whether there is any code-switching in the post.

Code-switching occurs whenever there is an alternation between two languages within an utterance, with the following exceptions:

1. Quotes do not count as code-switching. Quotes are text from other people, websites, books, songs, etc. They are often, but not always, marked by quotation marks. Thisdoes not exclude common phrases and idioms in a given language, such as in my humble opinion, bigger sh to fry orspeak of the devil, which do count as code-switching when in a different language.

2. Translations do not count as code-switching. Authors sometimes repeat something they said in one language in the other language, but this is not code-switching.

3. Use of named entities does not count as code-switching. A named entity is a person, location, organization, company, product, etc., with a proper name, such as Facebook, Goldman Sachs, Trump, Canada, Silicon Valley, GoPro.

Note: code-switching may consist of any number of words (as few as one word), and may occur within a single sentence, between sentences, or across paragraphs.

We pre-processed our data to the best of our abilities to identify posts with more than one language. We also replaced as many named entities and quotes as possible, marking their location with NE and Quote, however, we could not remove all of them.

If a post contains at least one example of valid code-switching (a switch in language that is not a quote, translation, or named entity) then we call it a code-switch post and you should answer ‘yes’. If there are no switches in language, or if all the switches are one of the exceptions above, then it is not a code-switch post and you should answer ‘no’. In (the very rare) case a post is very unclear, and you are unable to tell if it’s code-switch, answer ‘maybe’.

Please carefully study all 7 examples below to learn what exactly we consider code-switching.

Examples:

1. Is it possible to make enough money and to go to University in Romania? I have seen that university costs for business course costs 1980-2200 euro per year, without taking the course in Romanian Language.(1980 - 2700 per year). Nu am gandit sa raman in Romania, dar sa studiez acolo. este possibil sa gasesc de lucur care plateste deajuns?

Correct answer: ”Yes”. The first 2 sentences are in English and the last one is in Romanian.

2. Meanwhile, aia care fura castiga alegerile si ciclul se repeta.

Correct answer: ”Yes”. The word meanwhile is English and everything else is Romanian.

3. adica stirile si discutiile negative despre tigani, unguri, nemti, vasluieni, bucuresteni, politicieni, etc sa fie complet interzise si sa ramana doar posturile culturale, alea despre vreme, alea cu intrebari si alea in care OP descrie cum ii place sa suga pula. Censorship is telling a man he can’t have a steak just because a baby can’t chew it.

Correct answer: ”No”. The English part of this post is a quote by a person (Mark Twain) , please look out for quotes by specific people, websites or books.

4. E foarte usor sa dai exemple selective care sa iti sustina punctul de vedere. Uite cum arata restrictia privind fumatul in statul New York: smoking is banned statewide in all enclosed workplaces in New York, including all bars and restaurants and construction sites.[311] The law exempts (1) private homes and automobiles, (2) hotel/motel rooms, (3) retail tobacco businesses, (4) private clubs, (5) cigar bars (A cigar bar that makes 10 percent of its gross income from the on-site sale of tobacco products and the rental of on-site humidors, not including vending machines sales are exempt from the ban), (6) outdoor areas of restaurants and bars, and (7) enclosed rooms in restaurants, bars, convention halls, etc., when hosting private functions organized for the promotion and sampling of tobacco products

Correct answer: ”No”. The English part of this post is a quote from a website, not the authors own words.

5. E adevarat ca sunt acum parte din ”establishment”, dar prin ”new money” ma refeream la corporatiile care inca sunt conduse de fondatorii originali

Correct answer: ”Yes”. The words ”establishment” and ”new money” are English, and everything else is Romanian. The quotes represent emphasis rather than directly quoting someone.

6. Richard Turner, unul dintre cei mai renumii magicieni cu cri din toate timpurile, uimete publicul din ntreaga lume cu legendara sa ndemnare. Ceea ce audiena lui s-ar putea s nu realizeze, este c este complet orb

I can’t think right now for another ”sleight of hand” translation, so I used ”ndemnare” (skill, ability) for the translation.

Correct answer: ”No”. The author is clearly translating something and not code-switching.

7. Bineinteles, Sillicon Valley sustine politica imperialista a SUA

Correct answer: ”No”. Silicon Valley is a place and is considered a named entity.