Automated translation between languages has been a hot topic since the advent of artificial intelligence (i.e. the classic example of “The vodka is good but the meat is rotten” for “The spirit is willing but the flesh is weak” in an early English-Russian-English translation scheme) and has become even more important as natural language processing moves to the forefront of the tech domain. Even outside of academia, machine translation pervades society in systems such as Google Translate (text-to-text and even picture-to-text translation) and Duolingo (an automated program for language learning). Oftentimes, both industrially available (i.e. Google Translate) and (mainly) academic systems focus on literal word- or sentence-level translations of either short texts (GT) or formal documents. Literal translation works well in many cases, but it is highly insufficient to get at the meanings (and then the translations) of language devices such as colloquialisms and irony. For my final project, I would like to focus on colloquialisms because they are such basic but crucial parts of any language which are essential for learning and communicating in that language. While colloquial language has surely been extensively studied in linguistic (but nontechnical) settings, and a few studies have focused on translating colloquial english, as far as I can tell, there is relatively little work done for processing colloquialisms in any language.

I propose to research a system that can translate colloquial English into a variety of languages, with Spanish and French to start. I have chosen these languages because they are the ones I am most familiar with. With inspiration taken from Barzegar (2008), ideally, I would like to build three corpora as training data. One will contain (sections of) transcripts of movies released in English, one will contain the Spanish translation of those transcripts, and the final will contain the French translation. There is no guarantee that these translations will be correct, but without specific and fluent (technical and/or cultural) knowledge of other languages, they should be decent foundations.

One simple framework for identifying colloquialisms from these transcripts (or other translation data) might be as follows:

1) Match and process the transcripts word-to-word.
2) When a word $w_i$ from an English transcript does not directly translate to its counterpart in another transcript, this may signal the beginning of colloquial language that cannot (or should not) be literally translated.
   2a) A word may be a colloquialism (i.e. “Sweet!” in English, meaning “awesome” or “cool”, and not Spanish “dulce” (“Esa galleta es dulce” = “That cookie is sweet”) if it ends or constitutes a sentence, so that the boundaries of the potential colloquialisms do not (inappropriately) cross sentences.
3) Incrementally build the potential colloquialism from $w_i$ to $w_n$ where at $w_{n+1}$, the transcripts match up directly once again.

3a) $w_{n+1}$ should not be a stop word (i.e. ‘it’s raining cats and dogs’ should not stop at ‘it is’ or ‘it’ or ‘it’s raining cats and dogs’).

3b) However, stop words or other very common words may be included in the colloquialism before it has been detected (i.e. “It’s raining cats and dogs” is detected at “raining”), so when processing a word $w_i$, it is useful to keep track of the prior $p$ words for some $p$ such that prior($w_i$) = {$w_{i-p}..., w_{i-1}$} so that they may be included in the potential colloquialism.

4) Once a colloquialism has been identified (in the English transcript and for its corresponding Spanish and French translations), add it to a dictionary of colloquialisms and their translations for further identification in test cases.

This system would be part of, or provide for, a larger machine translation system, so that colloquialisms are more easily identified and more sufficiently translated in, say, Google Translate. Additionally, it should be possible to add other training data to the model (i.e. book translations, user input) so that it can further learn how to translate colloquial language. And finally, translating from English to Spanish or French should also work in reverse, since not only are English colloquialisms detected in the training data, but their Spanish/French counterparts as well. The system could then build on itself by, for example, finding Spanish-English and Spanish-French translations and identifying colloquialisms across languages (i.e. “There’s nothing that really means ‘it’s raining cats and dogs’ in Spanish and no one would use a phrase like that, but this [Spanish phrase] is close” versus $a_{English} > b_{Spanish} > a_{English}$ because $b_{Spanish}$ is also a common colloquialism).

There are several issues with my setup. First, I need to find transcripts for movies in all three languages. Second, again, the transcripts may not provide accurate translations of colloquial English. Third, my system would need (access to) another two dictionaries from English to Spanish and English to French in order to tell if two words translate directly for (2) above. Finally, in the above system, any introduced/inferred colloquialisms must be stored separately. There is no way to connect them so that they might build compositionally to help identify current ones or new ones. This would seem very inefficient, but intuitively, I believe that in terms of lexical storage and access, while some processing is or may be compositional (“walked” = “walk” + “-ed/PAST”, “I walked” = “walk” + “-ed” + “1SG”), it is necessary to mentally separate colloquialisms, so this framework is not problematic in terms of human capacity for language. Overall, regardless of these issues, I still hope to refine the setup so that I can eventually work towards to overall goal to translation colloquial language, regardless of how I have it currently framed.