1-1-2007

Baby-Steps Towards Building a Spanglish Language Model

Juan C. Franco
University of Texas at El Paso, jcfranco@utep.edu

Thamar Solorio
tsolorio@utep.edu

Comments:
Technical Report UTEP-CS-07-05

Baby-Steps towards Building a Spanglish Language Model

Juan C. Franco and Thamar Solorio

University of Texas at El Paso
El Paso, TX, 79912
{jcfranco, tsolorio}@utep.edu

Abstract. Spanglish is the simultaneous use, or alternating of both, traditional Spanish and English within the same conversational event. This interlanguage is commonly used in U.S. populations with large percentages of Spanish speakers. Despite the popularity of this dialect, and the wide spread of automated voice systems, currently there are no spoken dialog applications that can process Spanglish. In this paper we present the first attempt towards creating a Spanglish language model.

1 What is Spanglish?

Spanglish has existed for a long time, but has not been formally recognized as a language, nor has it been classified as a particular linguistic phenomenon. This interlanguage is more of a continuum of the mix between English and Spanish. From a linguistic point of view, it is difficult to decide what to consider Spanglish. It is debatable whether to consider Spanglish as an interlanguage, a pidgin, or a creole language. An interlanguage is a language that is often spoken between linguistic borders [1]; Spanglish does not fit this category, as it is also spoken in areas where no such borders exist, New York City being an example of this. A pidgin is a communication system created when people communicate despite their lack of knowledge in the other language [1]; this might explain its origin, but it certainly does not apply to its use, as most of the Spanglish speakers are bilingual. A creole language originates when a community adopts a pidgin as their primary source for communication [1]; a fragment of Spanglish speakers fall under this category since they cannot use traditional English or Spanish because of lack of proper training, but this cannot be generalized to all the Spanglish speakers, a large percentage of Spanglish speakers are bilingual who can express themselves in either of the traditional languages.

The origins of Spanglish in the U.S. are attributed, to a large extent, to socio-historical circumstances. The Mexican-American war, which according to history, started with the annexation of Texas to the U.S., resulted in Mexico ceding the territories of California and New Mexico to the U.S. in the mid eighteen hundreds. For many years Spanish speakers were going back and forth across these regions maintaining contact with English speakers. Many years later, the U.S. experienced a considerable immigration from Spanish speaking countries like Mexico, Cuba, Venezuela, Colombia and even Spain. In recent years, the flow of immigrants from Spanish speaking countries has not ceased to occur. In addition to this, the constant contact among the border cities between the U.S. and Mexico certainly has had influence on the proliferation of Spanglish.

In this paper we report results from building a Language Model (LM) with a small Spanglish corpus we collected. To the best of our knowledge, we are the first attempting to build a LM for Spanglish. Such LM is one of the first steps towards advancing the state-of-the-art regarding the automated processing of interlanguages, an achievement that will open the road for exploring interesting research avenues and applications. A good example is the possibility for building an automated speech recognizer for spoken dialog systems capable of processing requests from Spanglish speakers. We present here evaluation results of the language model, and although they show the language model to be weak, the results are promising. We will continue working on gathering more data to improve the corpus. However, the corpus already represents a valuable asset for deeper analysis of bilingualism. It will allow a statistical analysis that can support a formal characterization of Spanglish. The next section describes some of the most salient features of Spanglish.

2 Linguistic Features of Spanglish

In the linguistic, sociolinguistic, psychology, and psycholinguistic literature, bilingualism and the inherent phenomena it exhibits has been studied for nearly a century [7, 8, 11–13, 16, 20]. Despite the numerous previous studies of linguistic characteristics of bilingualism, there is not a clear consensus on the use of concepts related to the language alternation patterns in bilingual speakers. The alternation of languages within a sentence is known as code-mixing, but it has also been refereed as intrasentential code-switching, and intrasentential alternation [1, 10, 18]. Alternation across sentence boundaries is known as intersentential code-switching, or just code-switching. Yet there is another alternation mode defined as borrowing, which consists on adopting words, or idiomatic expressions, of a foreign language, usually modifying the original word, or expression, to suit the grammar or morphology of the receiving language [19].

In this paper we present Ardila's classification of Spanglish characteristics into two groups: shallow and deep phenomena. From his definition, shallow phenomena encompass code-mixing and code-switching; these are the linguistic features of Spanglish that can be easily spotted by humans. In contrast, deep phenomena includes, among other things, the transformation of Spanish to approximate English; the transformations can be so subtle that they are harder to detect, even for speakers of traditional Spanish, and include false cognates, also known as false friends. For our research purpose we are interested mostly in shallow phenomena of Spanglish, thus, the following subsections are focused on this type of features. The interested reader can find more information regarding the deep phenomena in [1].

2.1 Code-switching

Code-switching is defined as the change of language from one sentence to the following, or when starting a new topic. That is, the speaker starts an utterance in a given language, then switches to the other, and continues his/her utterance in the other language. What is very interesting about code-switching is that most speakers don't even notice when they are changing tongues [6]. It is likely that this is due to the speaker being more focused on expressing an idea, and in the process of formulating an accurate expression of that idea they make use of the known vocabulary in the two languages.

Toribio states that, " . . . code-switched forms are context-bound, practiced by bilinguals for bilinguals" [21]. This might explain why most English speakers are unaware of the existence of Spanglish, which in turn explains partially why Spanglish has received little to no attention by linguistic researchers. A common misconception of code-switching is the belief that it is just random mixtures of languages, when in fact, " . . . it is rule-governed and systematic, demonstrating the operation of underlying grammatical restrictions" [21]. Spanglish speakers, however, don't receive instruction on how to code-switch, they just use it. Toribio published a study where she defines syntactic rules governing Spanglish; these rules however, have not been validated by a statistical analysis. The lack of a good quality Spanglish corpus makes difficult to perform such a study.

2.2 Borrowing

Borrowing refers to the situation in which a sentence is composed by all words, but one, from the same language. The borrowed word can be one that is commonly used in the other language, thus it is retrieved first. Also, the word might be borrowed from the other language because there is no equivalence of meaning in the first language. There are other language alternations that can also be considered as borrowing, these are explained somewhere else [1, 19].

2.3 Code-mixing

In contrast to code-switching, when the change of language occurs at the end of sentences, or topics, in code-mixing the change of language is present within the same sentence. That is, a sentence might begin in one language, and then switch and end in the other [1].

2.4 Examples of Shallow Phenomena

Table 1 presents examples illustrating each of the linguistic features described above. Now that we described the salient linguistic features of Spanglish, and presented the motivation behind our work, we give a brief introduction to language models. Then, we will continue this paper with the description of the data collected, and the results of testing the LM developed. Examples of the linguistic phenomena described above is shown in Table 1.

Table 1. Examples of some of the linguistic phenomena present in Spanglish

Linguistic Phenomenon	**Example**
code-switching	*Le dejé un mensaje en la contestadora.* She called me back and. . . (I left a message on the answering machine. She called me back and. . .)
borrowing	*Vámonos al* mall. (Lets go to the mall)
code-mixing	I need to tell her *que no voy a poder ir* I need to tell her that I won't be able to make it

3 Language Models

Language models are statistical models of word sequences. LMs can assign probabilities to sequences of words. The way LMs estimate the probability of a word sequence W with length n is by looking back in history to the previous words. More specifically, the probability of a word sequence W, denoted as $p(W)$ can be approximated as follows:

$$p(W) = \prod_{i=1}^{n} p(w_i|w_1, \ldots, w_{i-1}) \tag{1}$$

Since it would be difficult to find a corpus from which we can reliably estimate all the terms $p(W)$, we approximate them using a shorter history. Then Equation 1 becomes:

$$p(W) = \prod_{i=1}^{n} p(w_i|w_{i-2}, w_{i-1}) \tag{2}$$

for a trigram model. Since we are using a history of $n-1$ words, this is also called an n-gram language model. LMs need a corpus appropriate for the target task in order to estimate, as accurately as possible, the probability of observing sequences of words. But for any corpus of finite size, there will always be unobserved events from which the language model will assign a zero probability. In such cases we can use smoothing, or discounting techniques, to assign a non-zero probability to these events. The language model we build in this work uses the Good-Turing discounting method [9]. To determine the quality of a LM we can use two measures from information theory that are commonly used for speech recognition and many other NLP tasks: entropy and perplexity. In this context, entropy measures how well an n-gram grammar matches a corpus. This measure is defined in equation 3.

$$H(x) = -\sum_{x \in X} p(x) log_2 p(x) \quad (3)$$

where x is a random variable over the set of words, X, in the vocabulary of the model, and $p(x)$ refers to the probability of observing word x. Perplexity is an estimate of the branching factor of the recognition task, it measures the complexity of a text source from the point of view of the recognizer [14]. The perplexity of a given LM is computed as 2^H, where H is the entropy measure described above. Both perplexity and entropy are estimated over a separate test text. LMs have been used successfully in many NLP problems. Some examples are speech recognition, hand-writing recognition, text classification, augmentative communication for the disabled, and spelling error detection [3, 15].

4 Data Collection

A Spanglish conversation was the basis for building the corpora. The conversation was recorded between three staff members of a southwestern university of the United States. The volunteers were recorded during their lunch break and were instructed to ignore as much as possible the fact that they were being recorded. It is clear that at the beginning of the recording session the subjects were very self conscious, but after a few minutes they ignored the recorder and started talking spontaneously. The three speakers come from a highly bilingual background. Two of the speakers were raised in Mexico, and they moved to the U.S. in their early adulthood. The third speaker was born and raised in the U.S. but started learning Spanish when she moved to a border city as a teenager.

This recording session has around 40 minutes of continuous speech. The conversation ranges over four topics and shows the casual interaction of Spanglish speakers. The vocabulary of the transcription has a total of 1,516 different word forms. This transcription and the audio file are freely available for research purposes[1].

One of the major problems faced during this project is that the current corpus is far too small, thus it has a very limited vocabulary. It also contains incomplete sentences, due to overlapping and stuttering segments of speakers within the source audio file. All of these factors prevent the the corpus from being ideal for training. A language model will need a corpus size in the order of several thousands, or even millions, of words to estimate more accurately the probability of the n-grams. However, this is just a first approximation, as more data become available we can retrain the language model and achieve better results.

[1] By contacting the authors

5 Tools of the Trade

We describe in this section the software tools we used, first in the task of transcribing the Spanglish conversation, and then in the development of the language model. For the transcription task we used Transcriber, a free distribution software program that facilitates the manual annotation of speech. This program features a user-friendly graphical user interface (GUI) for segmenting speech recordings and is ideal for transcribing long speech files. More information and access to download the tool can be found in [2]. For the LM we used the CMU-Cambridge Statistical Language Modeling (SLM) toolkit [5]. This toolkit is a practical component for the creation and evaluation of language models; among the functionality provided with this tool are the generation of word frequency lists and vocabularies, word n-gram counts, vocabulary counts, n-gram-related statistics, various back-off n-gram language models, out-of-vocabulary (OOV) rate, n-gram-hit ratios, distribution of back-off cases, annotation of test data with language scores, and perplexity and entropy calculation [5].

An additional program that was used is the Universal Text Imitator (UTI) [22] , which is a program that serves as an all-purpose text generator targeted to generate sentences approximating the style and content of any given document. UTI works in conjunction with the Charniak parser to build a probabilistic context-free grammar, then it generates sentences by traversing the grammar. This tool can also by used in combination with the SML toolkit.

6 Test Phase and Results

There were two tests phases for this project. One involved using the SML toolkit to generate a language model and evaluating it, and the other one consisted of having UTI build a grammar and generate random sentences with such model. The results of both experiments are discussed below.

6.1 SML Test

We divided the transcription file into a training file and testing file, 85% of the original transcription was used for training and the remaining 15% was used for testing. Then, we input to the SML toolkit the training file and we generated several n-gram models. After the language models were trained, we used the test file to measure the entropy and perplexity of the different models. We performed different tests by varying the frequency threshold of the vocabulary. Table 2 presents the results of these experiments.

We can see that the best results were obtained with the 2-gram and 3-gram models, both showing very similar results. This was not surprising as we know that 3-gram models are still the state-of-the-art on speech recognizers [14]. For a vocabulary domain, such as the one from the conversation, these numbers are not bad. However, they represent an optimistic view of what would be achieved in real situations, where the speakers, and the topics, would be different from the ones in this conversation.

Table 2. Perplexity (P) and entropy (E) for the random sentence generation with different n-grams, where three types of vocabulary files were created; one with the top 3000 words (-top 3000), and with words repeated at least 5 (-gt 5) and 2 times (-gt 2).

2-gram model	**-top 3000**	**-gt 5**	**-gt 2**	**3-gram model**	**-top 3000**	**-gt 5**	**-gt 2**
P	99.16	49.96	66.40	**P**	100.17	49.40	66.48
E	6.63	5.64	6.05	**E**	6.65	5.63	6.05

4-gram model	**-top 3000**	**-gt 5**	**-gt 2**	**5-gram model**	**-top 3000**	**-gt 5**	**-gt 2**
P	102.66	49.79	69.19	**P**	104.01	50.95	71.73
E	6.68	5.64	6.11	**E**	6.67	5.67	6.16

6.2 UTI Test

The language model generated by the SML toolkit and a training text, were used by the UTI to generate the probabilistic context-free grammar. Then, a total of 220 sentences were produced for the experiments. For analyzing the resulting sentences generated by UTI, we categorize them into coherent sentences, semi-coherent and incoherent sentences. We grouped as coherent the sentences generated that "make sense", that is, sentences that we believed could have been uttered by humans. Semi-coherent phrases are those that sound weird, or are not likely to be used by humans, but they have the syntax of either of the traditional languages. A semi-coherent phrase can be turned into a coherent phrase by substituting a word in the sentence to another word with the same part of speech. Lastly, incoherent phrases are those that don't resemble human-like sentences. Table 3 presents a summary of the results on these experiments.

Table 3. Results of generating sentences using the UTI. Column label **S** stands for Spanish, column **E** for English, and column labeled **Spg** for Spanglish. Coherent sentences are grammatical sentences that sound human-like; semi-coherent sentences are grammatical sentences that do not make sense, but by a simple replacement of words with the same part-of-speech they can be turned into coherent sentences; incoherent sentences are ungrammatical sentences that are not likely to be uttered by humans

N-gram	**Coherent**				**Semi-Coherent**				**Incoherent**
Model	**S**	**E**	**Spg**	**Total**	**S**	**E**	**Spg**	**Total**	
2-gram	1.81%	4.09%	1.81%	7.71%	0.90%	2.27%	3.63%	6.80%	85.45%
3-gram	2.27%	5.00%	4.54%	11.81%	0.90%	3.18%	4.54%	8.62%	79.54%
4-gram	0.90%	5.00%	1.81%	7.71%	0.45%	3.18%	1.36%	4.99%	86.81%
5-gram	0.45%	5.00%	1.36%	6.81%	0.90%	1.36%	1.36%	3.62%	89.54%

For the random sentence evaluation, the bigram and trigram models worked better than the other models that were used. Although the sentences generated by the UTI software were obtained with the help of a context-free grammar file, it is evident from Table 3 that the model that produced the majority of Spanglish phrases was the 3-gram model.

Table 4. Examples of the random sentences generated by the UTI

Coherent sentences:

- *Como muy convinced.* (Like very convinced.)
- *I dije, you know.* (I said, you know.)
- *Y the first girl.* (And the first girl.)
- *So, they know your entire body through thirty- de esos.* (So, they know your entire body through thirty of those.)
- *En the parade.* (At the parade.)

Semi-coherent sentences in Spanglish:

- *I dije you was little.* (I told you was little.)
- *I call a las cinco, and I had to go y to sign him.* (I call at five and I had to go and to sign him.)
- *Le dije, it didn't volunteer we to a notary about a moment.* (I told him/her, it didn't volunteer we to a notary about a moment.)

Incoherent sentences:

- *Tonto no iban gonna grabando.* (Fool they were not gonna record.)
- *Remind upstairs the gusta que.* (Remind upstairs the like that.)
- *Digo, you called fue dando antibiotics, you, and, also, to the.* (I mean, you called was giving antibiotics, you, and, also, to the.)

Table 4 shows some of the sentences generated by the UTI software. We show only Spanglish-like sentences, although the grammar also generated several sentences in English and a few of them in Spanish. Spanish phrases were generated only with the bigram model. It is observed that within the corpus, the use of English dominated, possibly explaining the lack of Spanish output sentences.

7 Final Remarks

The term Spanglish has existed for several decades now, but the negative connotation associated with it in the past, or as Nash wrote, slightly derogatory label [17], has changed in recent years; the ever-increasing number of Spanglish speakers, as well as the raise in sensitivity, and understanding of bilingualism, has contributed to the fact that newer generations do not consider the word as a derogative one, but simply as the best label so far to describe the very interesting phenomenon of the long interaction between English and Spanish.

According to projections from the U.S. Census Bureau by 2050, one out of every four people in the U.S. will be Hispanic [4]. Currently, the Hispanic population of the U.S. is the largest minority, and is a powerful consumer base that is being neglected by automated voice systems unable to handle their speaking preferences. Hispanics have contributed towards an increase in the amount of call traffic where automated voice systems are used. Unfortunately these calls end up being transferred to the human operator after several failed attempts from the system to parse the utterances of the frustrated caller.

Extending automated voice systems with Spanglish LM and acoustic models is a must for large companies searching to increase their Hispanic market. This will be a relevant advance for state-of-the-art ASR systems, and the experience of achieving this goal for Spanglish can shed light into the advancement of the automated recognition of other interlanguages. It is also important to remark that once we have an ASR system for Spanglish, we can then focus on applications of higher level NLP tasks including intelligent tutors for second language learners, summarization and topic segmentation for security, and writing assisting tools for Spanglish speakers, to name a few. Our research effort is the first step towards opening this research road.

8 Current and Future Work

There is a large list of exciting research paths that arise as a result of this work. We provide here a short description of what is currently under way, and things we want to explore in the near future.

– In order to come up with a more reliable language model, more Spanglish resources are needed. Future audio conversation recordings are currently being planned. In addition, we are currently looking at ways to gather written Spanglish samples, like e-mails, chat forums, or blogs.
– We are also working in a statistical analysis of the structure of Spanglish. Our goal is to develop a parser for Spanglish. This will allow to focus on other higher level NLP tasks dealing with Spanglish.
– We are also planning to experiment with the prediction of code-switching points by using prosodic and syntactic features. If we can predict when a change of language is very likely, then it would be possible to divide the Spanglish utterances into fragments belonging to either of the traditional languages, which in turn can be processed by existing tools.

9 Acknowledgements

We thank the National Science Foundation for supporting this work under grant 0080940, and the data transcription under award IIS-0415150. We would also like to thank Olac Fuentes and the reviewers of this paper for their useful comments.

References

1. Alfredo Ardila. Spanglish: An anglicized spanish dialect. *Hispanic Journal of Behavioral Sciences*, 27(1):60–81, 2005.
2. Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2), 2001. Software downloaded from `http://trans.sourceforge.net/en/download.php`.

3. Chris Brockett, William B. Dolan, and Michael Gamon. Correcting esl errors using phrasal smt techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, July 2006. Association for Computational Linguistics.
4. U.S. Census Bureau. U.s. interim projections by age, sex, race, and hispanic origin, 2004. Retrieved August 30, 2006 from `http://www.census.gov/ipc/www/usinterimproj/`.
5. Philip R. Clarkson and Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings ESCA Eurospeech 1997*, 1997.
6. Elena M. de Jongh. Interpreting in Miami's federal courts: Code-switching and Spanglish. *Hispania*, 73(1):274–78, March 1990.
7. S. Ervin and C. Osgood. Second language learning and bilingualism. *Journal of abnormal and social phsychology, supplement 49*, pages 139–146, 1954.
8. Aurelio M. Espinosa. Speech mixture in New Mexico: the influence of English language on New Mexican Spanish. *H. Stevens and H. Bolton, eds., The Pacific Ocean in history*, pages 408–428, 1917.
9. I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:16–264, 1953.
10. François Grosjean. *Life with Two Languages: An Introduction to Bilingualism.* Harvard University Press, 1982.
11. John J. Gumperz. Linguistic and social interaction in two communities. In John J. Gumperz, editor, *Language in social groups*, pages 151–176, Stanford, 1964. Stanford University Press.
12. John J. Gumperz. Bilingualism, bidialectism and classroom interaction. In *Language in social groups*, pages 311–339, Stanford, 1971. Stanford University Press.
13. John J. Gumperz and Eduardo Hernandez-Chavez. *Cognitive aspects of bilingual communication.* Oxford university Press, London, 1971.
14. Frederick Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, 1998.
15. Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing.* Prentice Hall, 2000.
16. John M. Lipski. Code-switching and the problem of bilingual competence. In M. Paradis, editor, *Aspects of bilingualism*, pages 250–264, Columbia, SC, 1978. Hornbeam.
17. Rose Nash. Spanglish: Language contact in Puerto Rico. *American Speech*, 45(3/4):223–233, 1970.
18. Shana Poplack. Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics*, 18(7/8):581–618, 1980.
19. Shana Poplack, David Sankoff, and Chris Miller. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104, 1988.
20. David Sankoff. *Social aspects of multilingualism in New Guinea.* Ph.D. thesis, McGill University, 1968.
21. Almeida Jacquline Toribio. Spanish/english speech practices: Bringing chaos to order. *International Journal of Bilingual Education and Bilingualism*, 7(2–3):133–155, 2004.
22. Sam Wintermute. The universal text imitator, October 2006. Software downloaded from `http://www-personal.umich.edu/~swinterm/nlpproj/`.