Classification and Analysis of Neologisms Produced by Learners of Spanish: Effects of Proficiency and Task

Shira Wein
Georgetown University
sw1158@georgetown.edu

Abstract

The Spanish Learner Language Oral Corpora (SPLLOC) of transcribed conversations between investigators and language learners contains a set of neologism tags. In this work, the utterances tagged as neologisms are broken down into three categories: true neologisms, loanwords, and errors. This work examines the relationships between neologism, loanword, and error production and both language learner level and conversation task.

The results of this study suggest that loanwords and errors are produced most frequently by language learners with moderate experience, while neologisms are produced most frequently by native speakers. This study also indicates that tasks that require descriptions of images draw more neologism, loanword and error production. We ultimately present a unique analysis of the implications of neologism, loanword, and error production useful for further work in second language acquisition research, as well as for language educators.

1 Introduction

Neologisms, as opposed to nonce-words, are new words, phonemes, or locutions appearing in the language, that have been accepted by a speech community (Picone, 1996). Nonce-words or nonce-formations are speech units produced by one specific author that have not yet been accepted by a speech community (Stekauer, 2002). Bauer suggests the use of a dictionary and a large corpus to detect neologisms while excluding nonce-words (Bauer, 2001). A study by Luz Rello and Eduardo Basterrechea explores the properties of neologisms with respect to linguistic creativity, and concludes that more than 50 % of Spanish verbs identified in the dataset do not appear in the largest Spanish dictionary (Rello, 2010). Loanwords are words that borrow from a language other than the target language.

The Spanish Learner Language Oral Corpora (SPLLOC) consists of transcribed conversations between investigators and native English speakers learning Spanish. Each conversation is focused around a specific task, such that the investigator asks questions about that topic. The tasks that are tested in SPLLOC 1 are four discussion tasks (Loch Ness, Discussion, Modern Times, and Clitics) and one image description task (Photo). The conversations take place with language learners at four levels: native speakers, Undergraduate students, students in Year 13, and students in Year 9.

SPLLOC contains a set of neologism tags. We propose that the 362 words tagged as neologisms in SPLLOC 1 in fact encompass a range of **coinages**, not all of which are neologisms. In this work, we break down these coinages into three categories: 28 true neologisms, 119 loanwords, and 215 errors. We hypothesize that the three categories of coinages will have different frequencies amongst various learner levels, such that speakers with high proficiency would produce the most neologisms and loanwords, whereas speakers with low proficiency would produce the most errors.

2 Categorization Technique

Transformations of words in other languages were categorized as **loanwords**. These loanwords included borrowings from English, Portuguese, and French. Some examples of the loanwords produced include: *pictura* from English picture, *decremento* from English decrement, and *paquigente* from English packaging.

Errors are malformations of Spanish words due to a production issue, such as incorrect gender ending, pluralization, or tense formation. Examples of errors include: *periodisto* which incorrectly places a masculine ending on Spanish *periodista*, meaning journalist, or *cuatros* which incorrectly pluralizes a cardinal number *cuatro*.

The remaining words, which were neither loanwords nor errors, were categorized as **neologisms**. Two neologisms from the corpus are *previstas* in place of Spanish *preestrenos*, meaning previews, and *chiquititos* as an extension of the Spanish *chicos*.

In this work, we collect data on the production frequency of, and differences between, each of these three categories of coinages. We then investigate the relationship between each category of coinage and two key variables: speaker's proficiency level and conversational task.

3 Results

Level	Neologisms	Loanwords	Errors	Conversations	Words	Words Per Conversation
Native	**7**	1	3	40	159313	3982.8
Undergraduate	5	27	63	91	448240	**4925.7**
Year13	8	72	115	70	245331	3504.7
Year9	3	23	35	60	162563	2709.4
All Levels	23	123	216	261	1015447	3890.6

Table 1: Key statistics for each language level, including total number of neologisms, loanwords, and errors produced, how many interviews take place with speakers at that level (conversations), total number of words, and number of words per conversation.

Level	Neologisms per Conv.	Loanwords Per Conv.	Errors Per Conv.
Native	**0.175**	0.025	0.075
Undergraduate	0.055	0.297	0.692
Year13	0.114	**1.03**	**1.64**
Year9	0.05	0.383	0.583
All Levels	0.088	0.471	0.828

Table 2: Average number of neologisms, loanwords, and errors produced per conversation by each learner level.

Task	Neologisms Per Conv.	Loanwords Per Conv.	Errors Per Conv.
LochNess	0.093	0.44	0.493
Discussion	0.0769	0.462	0.654
ModernTimes	0.00	0.36	0.6
Clitics	0.00	0.117	0.483
Non-Image	**0.0483**	**0.328**	**0.527**
Photo	**0.187**	**0.823**	**1.573**
All Tasks	0.0881	0.471	0.828

Table 3: Neologisms, loanwords, and errors produced per conversation for each task. Non-Image is the rate of neologisms, loanwords, and errors produced per conversation over all of the non-image based tasks: LochNess, Discussion, ModernTimes, and Clitics.

As seen in Table 1, Undergraduate students had the longest conversations as measured by words per conversation (text), followed by native speakers, year 13 students, and lastly by year 9 students. Using Pearson's correlation test, the p-value for correlation between number of neologisms produced and number of words in the conversation is 1.608e-05. This is a very strong correlation between neologism production and length of conversation.

As evidenced by the rates of coinage production per conversion shown in Table 2, while neologisms are produced most frequently by native speakers, loanwords and errors are produced most frequently by language learners with low to moderate proficiency.

The fact that loanword and error production both have similar frequency patterns of loanword and error production, such that Year 13 students have the highest rates, followed by similar rates of undergraduate and Year 9 students, and lastly followed by a much smaller rate of native speakers, suggests that loanwords and errors are more closely related than loanwords and neologisms. While loanwords could be an indication of mastery of both Spanish and the borrowed language, such as use of Spanglish by native Spanish speakers, in this corpus loanwords are more similar to errors. The loanwords that appear in SPLLOC 1 are borrowed words from another language transformed to sound more like English, which is supported by the examples of loanwords that appear in Section 2. This suggests that the appearance of loanwords in a learners corpus is an indication of low mastery of the target language, as is the case with error production, rather than mastery of the target language, as is the case with neologism production.

Neologism production is a clear indication of mastery of the language. The high rates of neologism production by native speakers is unsurprising because creative generation of language requires high mastery of the language.

Year 13 students producing a higher rate of loanwords and errors than Year 9 students also suggests that there is a certain degree of linguistic risk taking required to produce more loanwords and errors. Year 9 students may be making fewer errors and producing fewer loanwords out of an abundance of linguistic caution and concern over accurately producing Spanish words. Once students develop a degree of mastery over the language, this type of error and loanword production is no longer necessary or prevalent, as is the case for undergraduate students and native speakers.

Additionally, Table 3 shows that tasks requiring descriptions of images elicit a higher rate of coinage production per conversation than non-image based tasks. The Photo task requires the student to describe what is happening in an image presented to them. This task is unique amongst the set of tasks in that there is a finite set of objects in the image, so it is expected that the student use specific language to describe the objects and actions seen in the image. This disparity in rate of neologism, loanword, and error production between the photo description task and the four other tasks signals that the neologisms, loanwords and errors being produced are often substitutions for specific Spanish words. The need for specific words draws out more neologism and nonce-words, as evidenced by the high rates of neologism, loanwords, and errors production for the Photo task.

4 Conclusion

This work explores the relationships between neologism, loanword, and error production and conversational task as well as learner level. Our results indicate that production of loanwords by language learners may be illustrative of low mastery of the target language, similarly to error production. This motivates future research investigating the relationship between speaker's confidence in the language and the production of errors and loanwords.

We also illuminate a difference in neologism production between the discussion tasks and the photo description task. This disparity indicates that loanwords, errors, and even neologisms are produced in substitution for specific, sometimes unknown Spanish words, as the photo task requires specific words to describe the objects and actions that appear in the photo. These findings suggest that language educators should use photo description tasks to test a student's vocabulary and discussion tasks to test general fluency.

Our analysis of the implications of neologism, loanword, and error production is useful for language educators as well as for future work in second language acquisition research. This work is relevant to natural language understanding, specifically analyses of learner language, because it demonstrates that the types of coinages that appear in learner data differs from the language of native speakers. As a result, we suggest that language technologies targeted at non-native speakers should recognize and support differences in coinages. Future work should examine whether these patterns hold in other languages; specifically, whether loanword production is consistently a mark of low mastery of the language.

References

Laurie Bauer. 2001. Morphological Productivity. *Cambridge Studies in Linguistics* 95:39, 158-159.

Özkan Kılıç. 2014. Using Corpus Statistics to Evaluate Nonce Words. *Pristine Perspectives on Logic, Language, and Computation. ESSLLI 2013, ESSLLI 2012.*

Luke Plonsky. 2013. STUDY QUALITY IN SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research. *Studies in Second Language Acquisition* 35.4:655-687

Michael Picone. 1996. Anglicisms, Neologisms, and Dynamic French. *John Benjamins B.V.* 3

Laura de Vaan, Robert Schreuder and R. Harald Baayen. 2007. Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon* 32(1):1-24

Pavol Štekauer. 2002. On the Theory of Neologisms and Nonce-formations. *Australian Journal of Linguistics* 22:1, 97-112.

Luz Rello and Eduardo Basterrechea. 2010. Automatic conjugation and identification of regular and irregular verb neologisms in Spanish. *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity, Association for Computational Linguistics.* 1-5.

Karine Megerdoomian and Ali Hadjarian. 2010. Mining and Classification of Neologisms in Persian Blogs. *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity, Association for Computational Linguistics.* 6-13.

Quirin Würschinger, Mohammad Fazleh Elahi, Desislava Zhekova, and Hans-Jörg Schmid 2016. Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The case of rapefugee, rapeugee, and rapugee. *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task.* 35-43.

John P. McCrae. 2019. Identification of Adjective-Noun Neologisms using Pretrained Language Models. *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Association for Computational Linguistics.* 135-141.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. 2018. Toward Better Loanword Identification in Uyghur Using Cross-lingual Word Embeddings. *Proceedings of the 27th International Conference on Computational Linguistics.* 3027–3037.

Paul Cook and Suzanne Stevenson. 2010. Automatically Identifying the Source Words of Lexical Blends in English. *Computational Linguistics 36:1.*

Jack C. Richards. 1974. Error Analysis: Perspectives on Second Language Acquisition.