ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SYLLABLE-FINAL /S/ LENITION IN THE LDC'S CALLHOME SPANISH CORPUS

Michelle A. Fox

Department of Linguistics
University of Pennsylvania
Philadelphia, Pennsylvania 19104, USA
minnick@unagi.cis.upenn.edu

ABSTRACT

This paper describes a data corpus which is being made available through the Linguistic Data Consortium (LDC) that codes lenition of syllable-final /s/ in Latin American Spanish in the LDC's CallHome Spanish corpus. This lenition is a process whereby the /s/ may be aspirated (pronounced [h]) or deleted altogether. Since syllable-final /s/ is frequent in Spanish, lenition has a great effect on overall pronunciation. While previous data collected on syllable-final /s/ lenition has been dialect-specific, the CallHome Spanish corpus contains speech from many speakers of different dialects and provides the type of speech data needed to adequately model the phenomenon.

Data analysis indicates that over 40% of instances of syllable-final /s/ are deleted. The deletion rate is extremely speaker- and dialect-dependent and also depends on linguistic factors such as phonetic environment and grammatical status of the /s/, as well as the identity of the individual word.

1. INTRODUCTION

It is a well-known fact that syllable-final /s/ (hereafter -/s/) is subject to lenition in many Latin American Spanish dialects. Lenition of -/s/ is a variable phonological process in which an -/s/ may be aspirated (pronounced [h]) or deleted altogether (Ø). Lenition of -/s/ has been widely studied by sociolinguists, who have identified various linguistic and extralinguistic factors that favor the process (cf. for example, [1, 2, 3, 4]).

The most important extralinguistic factor influencing a speaker's use of -/s/ lenition is dialect. Countries of the Caribbean are well-known for /s/-aspiration and deletion: Cuba, Dominican Republic, Puerto Rico, Venezuela, Honduras, and Nicaragua. Lenition also occurs in many other regions of Latin America: El Salvador, Panama, Chile, Paraguay, Uruguay, much of Argentina, part of Bolivia, and small parts of Peru, Ecuador, and Colombia [3]. It is important to note that different regions in the same country may display very different characteristics, particularly between geographically separated areas such as the coast and mountainous regions. Other extralinguistic factors that affect the overall rate of lenition are social factors (lower social classes tend to have more lenition [5, 6]) and speech formality (less lenition in more careful speech [5]).

While a particular speaker may have an overall tendency to aspirate or delete -/s/ in their speech, several linguistic factors have been identified that play a role in determining whether this variable phonological process applies in a particular instance. For word-final -/s/, phonetic environment is important, with a following consonant favoring lenition much more than a following vowel, and a following unstressed syllable favoring lenition over a following stressed syllable [2, 4]. It has also been noted that for word-final -/s/, the more syllables in the word, the higher the probability of lenition [5].

In addition to phonetic influences, the grammatical status of a word-final /s/ also affects the probability of lenition. In Spanish, a word-final /s/ may be simply part of the lexical item and therefore not contribute to its meaning (*entonces*, after) or it can be inflectional, either marking plurality on nouns, adjectives, determiners and quantifiers or disambiguating the 2nd person singular from the 3rd person singular of verbs (*tú estás*, you are, vs. *él está*, he is). *Lexical* -/s/ is less likely to be aspirated or deleted than inflectional /s/ [1]. When several words with a plural -/s/ form a noun phrase, the first word of the noun phrase is most likely to retain the /s/ [2, 5].

Although sociolinguistic studies have identified these extralinguistic and linguistic factors, such studies have concentrated on small numbers of speakers of a single dialect, often one of the Caribbean dialects. However, in order to adequately understand and model the phenomenon, data is needed that will allow comparison and analysis across many dialects. The goal of this project was to generate this linguistic data and to make it available.

2. SPEECH DATA TO BE CODED

The speech data used as the basis for this syllable-final /s/ corpus is from the CallHome Spanish corpus published by the Linguistic Data Consortium (LDC), which contains telephone conversations between native speakers of Spanish. This corpus is especially well-suited to the task of studying variation in -/s/ lenition because it contains informal speech by a large number of speakers from many different dialects. General information regarding each of the speakers, including dialect, is identified, so that dialectal studies can be performed with the data.

Each of the telephone calls in the CallHome Spanish corpus was transcribed orthographically, with no pronunciation information, so instances of underlying -/s/ were easily identified by searching through the transcriptions using the pronunciations given in the LDC Spanish Lexicon [7]. Although syllabification is not given in the lexicon, all

instances of word-internal /s/ followed by a consonant are syllable-final [5]. All occurrences of word-final /s/ were also included, even though when immediately followed by a vowel, it may be re-syllabified in fast speech. In addition, in Spanish, surface /z/ is actually an underlying /s/ [5], so all syllable-final instances of /z/ in the lexicon were treated as /s/.

Although the CallHome Spanish corpus includes nearly ideal data for the study of -/s/ lenition across dialects, the fact that it is telephone quality speech makes the coding significantly more challenging. The phoneme /s/ is characterized by high frequency sound energy, mostly above 4kHz, while telephone speech does not contain frequencies above 4kHz. Because of this, spectrograms were of little help in identifying whether an -/s/ was retained, and the spectrograms were not used during the coding.

3. CODING PROCEDURE

Once a list of all words containing -/s/ was made, a large amount of redundancy was added to the list in order to measure the repeatability of coding. Since the task is a difficult one, it was important to see whether each coder consistently coded tokens, and to see whether the two coders used the same criteria. The list of tokens was then randomized to prevent the coders from being affected by listening to multiple tokens by the same speaker, either (1) by expecting the speaker to retain or delete an -/s/, and hearing what they expected, or (2) by adjusting the coding criteria to the speaker (e.g. if a particular speaker pronounced /s/ very strongly in most cases, a weaker /s/ might be mis-coded as a deletion). In a further attempt to retain constant criteria, samples of -/s/ pronounced as [s], [h], and Ø were presented to the coders at regular intervals during the coding process.

Two students at the University of Pennsylvania performed the coding. The first coder is a female native speaker of English who is proficient in Spanish and a linguistics student. The second coder is a male bilingual speaker of English and Puerto Rican Spanish not familiar with linguistics. Both were familiar with the -/s/ lenition phenomenon before beginning the project.

For each token of -/s/ to be coded, the coder was shown the orthographic transcription of the entire sentence, along with an indication of which -/s/ to code. An automatic alignment of the speech files was used to determine the approximate start and end times of the given word; from this alignment a window of speech starting 20ms before the hypothesized beginning of the word and ending 20ms after the end of the word was played. The coder was able to replay the speech and to change the window of speech as needed.

The coding categories available were:

- ***s:*** the /s/ was retained;
- ***z***: the /s/ was retained and voiced;
- ***h:*** the /s/ was retained, but only as aspiration;
- *Ø*: the /s/ was not deleted;
- *R*: the recording was distorted and so analysis could not be made
- *f*: the following segment was also /s/, so the -/s/ in question could not be categorized
- *t*: the entire syllable was truncated
- *T*: the original transcript was incorrect and there was no word with a syllable-final /s/

As noted above, the coding task was a difficult one. The coders were instructed to make a selection for each occurrence of -/s/ unless the recording quality was poor. However, when the coders felt uncertain about a classification, they were able to indicate that the classification had low confidence. The results of this confidence rating are not reported here, although they are available with the -/s/ data corpus.

4. RESULTS

A total of 24,473 different instances of -/s/ from the training and development test files of the CallHome Spanish corpus were coded. 4,727 of these tokens were coded twice, and 843 of these were coded three times. In this corpus of -/s/ data, each coding includes references to the word's location in the speech and transcript files and information about the phonetic environment and the speaker. This data is being made available through the LDC; see www.ldc.upenn.edu/fox/ch_span/ for more information.

4.1. Overall Deletion/Aspiration

Table 1 shows the compiled data with the percentage of tokens receiving each of the classifications. The category "conflicting coding" indicates that a token was classified once as [s] and once as Ø, and so could not be classified as one or the other. It is somewhat surprising that the category [h] was selected so infrequently (less than 1% of tokens) given that -/s/ aspiration is widely cited in the literature. This may be due to the fact that the telephone-quality recording makes differentiation of [s] and [h] more difficult.

In the rest of this paper, the classifications [s], [z], and [h] are considered to be *s-retention*, while Ø is considered to be *s-deletion*. The tokens with all other classifications have been removed from consideration. After making this modification, the overall deletion rate is 42.7%.

4.2. Repeatability of Coding

Table 2 shows the distribution of codes for the tokens that were coded more than once. Both of the coders had high percentages of repeatability considering the difficulty of the task (88% for Coder 1 and 86% for Coder 2). However, the agreement across speakers was significantly lower, 76%, with the majority of disagreement being labeled [s] by Coder 1 and Ø by Coder 2.

Classification	Num tokens	Percentage
[s]	11114	45.4
[z]	1242	5.1
[h]	232	0.9
Ø	9371	38.3
Syllable truncated	333	1.4
Followed by /s/	639	2.6
Poor Recording	600	2.5
Conflicting coding	482	2.0
Transcript wrong	460	1.9

Table 1. Classification of all tokens of -/s/

This is mostly due to a difference in coding criteria by the two coders; Coder 1 only 40% of all tokens as deletion while Coder 2's labeled 53% of tokens as deletion.

4.3. Effect of Speaker and Dialect

Since -/s/ lenition is dialect- and speaker- dependent, the overall deletion rate for the entire -/s/ corpus does not necessarily represent the results for any actual speaker or dialect. It is therefore necessary to look at the data by these factors. Figure 1 shows the percentage deletion for each of the 172 speakers with more than 50 tokens of -/s/, sorted by amount deletion. The values range from 5% to 90%. Although the literature speaks of dialects with high -/s/ lenition and others dialects without, this data indicates that most speakers fall somewhere in the middle of these two extremes and there are no sharp divisions between speakers of different dialects.

Figure 2 shows the percent deletion for each dialect, separated by location in the word. As expected, most of the countries reported to have high lenition rates [3] do have high deletion in this data. The data for Honduras is an exception, since the data here shows a small mount of deletion, but this may be due to the small number of speakers (only 2). Furthermore, even the countries which are not reported to have -/s/ lenition do show a fare amount of deletion, especially word-finally.

		Coder 1		Coder 2	
		[s]	Ø	[s]	Ø
Coder 1	[s]	1755	383	808	367
	Ø		1049	96	659
Coder 2	[s]			282	98
	Ø				313

Table 2. Distribution of codes for tokens coded more than once

4.4. Linguistic Factors

Phonetic environment. The rate of -/s/ deletion varies greatly with respect to the following segment. Before a vowel, 30.0% of word-final -/s/ deleted; 45.4% before a pause; and 55.1% before a consonant. Table 2 shows the word-final data for each of the dialects having a sufficient number of tokens. Different dialects operate under different constraints. For example, the Venezuelan data does not show great variation by context, while the Mexican, Peruvian, and Argentinean data have much lower deletion rates before a vowel than before other contexts. In addition, the Mexican and Colombian data have less deletion before voiceless stops /p/, /t/, and /k/ in comparison to the other dialects.

Grammatical status of /s/. As described in the introduction, the lenition rate has been linked to the grammatical status of a

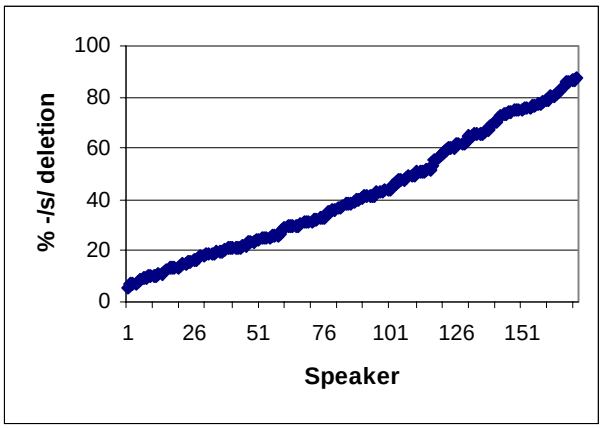

Figure 1. Percentage deletion for each of the speakers

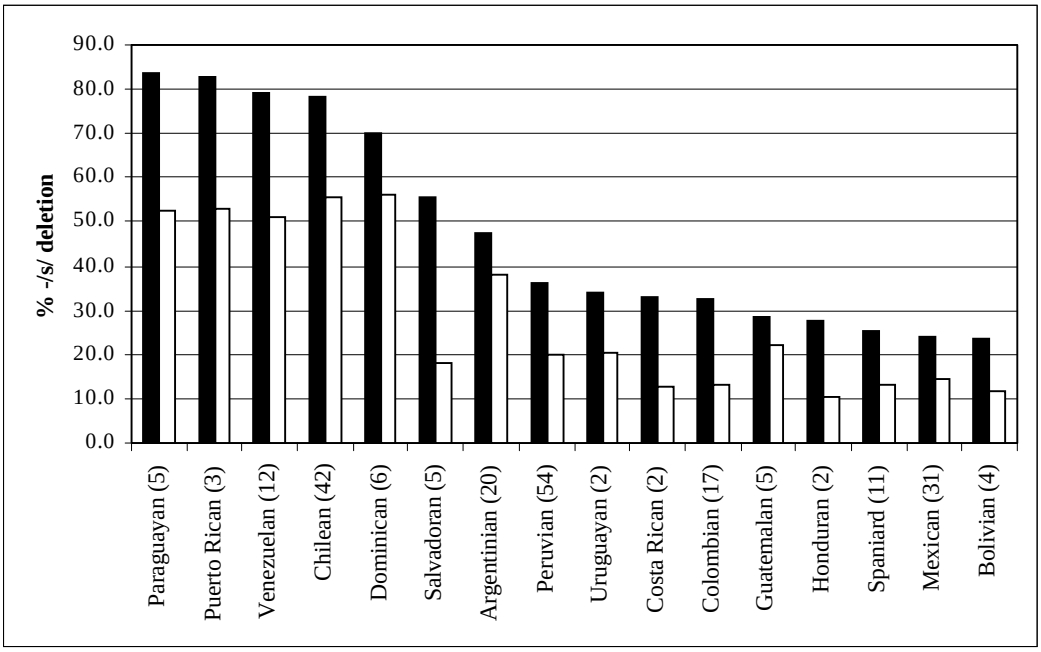

Figure 2. Percent -/s/ deletion for word-final /s/ (black) and word-internal syllable-final /s/ (white) for each dialect. The number in parentheses indicates the number of speakers from each dialect.

	vowel	pause	p/t/k	b/d/g	m/n	l
Mexican	9.1	23.0	15.4	32.8	46.9	64.9
Peruvian	13.5	30.7	42.8	52.8	56.8	64.0
Argentinean	13.5	40.0	69.0	64.6	76.8	72.4
Colombian	20.7	32.1	19.4	42.2	52.7	80.0
Salvadoran	34.8	43.6	51.1	86.2	88.2	89.5
Dominican	51.4	59.1	87.5	77.4	93.8	84.6
Chilean	77.2	78.8	81.0	72.7	81.0	87.1
Venezuelan	78.5	78.7	77.3	83.8	77.6	84.1

Table 3. Percent deletion of word-final -/s/ by dialect and following phoneme

Position in NP	**% Ø**
First/only	**50.2**
2nd	**54.5**
after -[s]	38.7
after -Ø	69.5
3rd	**61.4**
after -[s], -[s]	32.4
after -Ø,-[s]	50.0
after -[s],-Ø	61.8
after -Ø,-Ø	74.8

Table 4. Percent deletion for word-final /s/ in noun phrases by location of word within the noun phrase

word-final -/s/. In this syllable-final /s/ corpus, there was a small difference in deletion rates according to the grammatical status of the -/s/: plural /s/ deleted 52.1%; lexical /s/ 46.0%; 2nd person singular of verb 42.5%. (Word-internal -/s/ deleted only 30.5% of the time). Poplack [2] also found that within a plurally-marked noun phrase, the 2nd or 3rd words have different deletion rates depending on whether the preceding -/s/ deleted. The data in Table 4 is consistent with this; the first instance of plural /s/ deletes the least, and the second and third instances are affected by the realizations of the preceding plural markers.

Individual words. While the linguistic constraints of phonetic environment and grammatical status of -/s/ have an effect on the variable application of lenition, some individual words, especially ones that occur frequently, may act differently. For example, consider the words *estoy* (39% deletion), *esta* (28%), and *este* (20%). The word-internal /s/ is in the same phonetic context in all three frequently occurring words, but they have significantly different deletion rates. This indicates that at least some speakers may have alternate lexical specifications for such words.

5. CONCLUSION

The preliminary data analysis reported in this paper is mostly consistent with what has been reported by previous studies of -/s/ lenition. Because the data includes speech by many speakers of different dialects, this -/s/ corpus can be valuable in the study of phonological variation both for linguists and in speech technology.

Plans for future studies using this data include a more detailed analysis of the type of factors described in this paper. Previous studies have assumed that all linguistic and extralinguistic factors are independent of each other, but the data in Table 3 show that this is not the case. Further analysis of the data may find other areas where the factors are not independent, which may allow more insights that are not visible without finer-grained analysis.

It has been claimed that the information-bearing status of an -/s/ affects its lenition [2, 8]. Since complete transcripts of the conversations are available in the CallHome Spanish corpus, it is possible to perform syntactic analysis to determine whether the deletion of -/s/ is dependent on whether ambiguities result from its deletion. It has also been suggested that even when an -/s/ is deleted, some acoustic cues remain to allow listeners to determine that the speaker intended the -/s/. Such cues may be the preceding vowel quality or the preceding vowel and/or following consonant length. Acoustic analysis of the speech files can be made to determine whether such cues exist (and if so, subject to what constraints), and if so, how they interact with the lenition of the -/s/ itself.

6. ACKNOWLEDGEMENTS

The Linguistic Data Consortium provided funding for this project. I am greatly indebted to my advisor Mark Liberman for his assistance with this project. Thanks to Susan Garrett for input in developing the coding scheme, determining the criteria, and for many helpful suggestions for improving the data collection.

7. REFERENCES

1. Poplack, S. The notion of the plural in Puerto Rican Spanish: Competing constraints on /s/ deletion. In *Locating Language in Time and Space*. W. Labov (ed.). Academic Press, New York, 55-68, 1979.

2. Poplack, S. Mortal phonemes as plural morphemes. In *Variation Omnibus*. D. Sankoff & H. Cedergren (eds.). Linguistic Research, Edmonton, Alberta, 59-72, 1981.

3. Canfield, D. L. *Spanish pronunciation in the Americas.* University of Chicago Press, Chicago, 1981.

4. Reynolds, W. *Variation and phonological theory*. PhD. Dissertation, University of Pennsylvania, 1994.

5. Barrutia, R. and Terrell, T. *Fonetica y fonologia españolas*. John Wiley & Sons, New York, 1982.

6. Fontanella de Weinberg, M. Aspectos sociolingüísticos del uso de -s en el español bonaerense. *Orbis* 23:85-98, 1974.

7. Garrett, S., Morton, T. and McLemore, C. *LDC Spanish Lexicon*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 1997.

8. Hochberg, J. /s/ Deletion and pronoun usage in Puerto Rican Spanish. In *Diversity and Diachrony*. D. Sankoff (ed.). Benjamins, Amsterdam, 199-210, 1986.