Is It Still a Village? Tracing Grammaticalization with Word Embeddings

Joseph Larson[1], Patrícia Amaral[1]

[1]Department of Spanish and Portuguese, Indiana University, Bloomington Indiana, USA

Abstract

Computational studies of language change tend to focus on predicting lexical semantic change that reflects cultural and societal changes. In this paper we focus instead on the syntactic and semantic shift from lexical to grammatical (grammaticalization), and we choose an understudied variety of Spanish. This paper investigates the grammaticalization of the noun *caleta* 'cove, village' to a degree expression (an intensifier) meaning 'a lot', as part of the system of degree words in Chilean Spanish. We use word embeddings trained on a corpus of tweets to show the ongoing syntactic and semantic change of *caleta*. Our distributional analysis also reveals how high degree is expressed in this variety of Spanish, showing the potential of these methods to explore lesser-known linguistic subsystems. Our study unveils degree expressions not previously studied in contemporary colloquial Chilean Spanish and also provides further evidence for an existing typology of degree modifiers across languages.

Keywords

grammaticalization, degree, quantifiers, historical linguistics, Chilean Spanish, word embeddings

1. Introduction

Studies of language change using distributional methods have shown the potential of word embeddings to trace syntactic and semantic change over time [1, 2, a.o.]. However, such research tends to focus on predicting changes that affect sets of lexical items shifting from one semantic domain to another, which typically reflects cultural and societal changes. Fewer studies have explored both semantic and morphosyntactic change (but see Fonteyn et al. 3). In this paper, we focus on the semantic and syntactic shift from lexical to grammatical, known as grammaticalization [4, 5], and the stages of this process. Specifically, we study the creation of degree expressions.

Traditionally, degree expressions have been associated with adjectives, considered the prototypical gradable category. However, degree modification is also compatible with nouns and verbs, which shows that gradability cuts across syntactic categories [6, 7, 8]. As a word becomes a degree expression over time, it typically expands its distribution along different categories: e.g. it first combines with nouns before co-occurring with verbs and adjectives. Hence, the grammaticalization of degree expressions provides insight into the semantics of degree and patterns in the distribution of degree words [9, 10]. This paper examines an understudied variety, Chilean Spanish, and uses word embeddings to investigate the emerging system of degree words to which one grammaticalized word shifts. We investigate the grammaticalization of *caleta* in Chilean Spanish, from a noun denoting 'cove, hiding place (where merchandise can be stored)', 'village', as in ex. (1), to a quantifier and degree adverb 'much, a lot', as in (2), where *caleta* modifies the verb and denotes high degree.

(1) Esta experiencia la realizamos en
this experience CL.FEM.SG.ACC do.PST.1PL in
Zapallar, en la caleta de pescadores
Zapallar in the caleta of fishermen
"We did this experience in Zapallar, in the fishermen's cove"

(2) me gustó caleta
CL.1SG.DAT like.PST.3SG caleta
"I liked it a lot."

We use word embeddings to examine to what extent the grammatialization of *caleta* has developed while also shedding light on the system of degree modifiers in Chilean Spanish. We ask, (i) how far along has *caleta* grammaticalized in Chilean Spanish, and (ii) what types of evidence do word embeddings provide of different stages of grammaticalization of degree words?

2. Previous Work

Linguists have provided analyses of the gradual process by which lexical items acquire grammatical functions: for example, in this diachronic change, nouns lose their categorial properties like occurring after a determiner or being pluralized. The grammaticalization of nouns into degree adverbs (e.g. the development from *lot* 'a set of objects' to *a lot* 'much') is well attested cross-linguistically:

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy.

joelarso@iu.edu (J. Larson); pamaral@iu.edu (P. Amaral)
https://github.com/joelarso/ (J. Larson);
https://sites.google.com/site/patriciamatosamaral/home (P. Amaral)
0000-0001-6651-0319 (P. Amaral)

other examples are French adverb *beaucoup* from *un beau coup* 'a good strike' and English *a bit* from 'a bite, a portion that fits in the mouth' [11, 12, 13, 14, 15].

This research has shown that a typical structure in which nouns occur - modification by a prepositional phrase, as in *a lot [$_{PP}$ of chairs], a mountain [$_{PP}$ of books]* - provides a starting point for quantity and degree interpretations. This structure undergoes subsequent syntactic reanalysis, where the head noun (e.g. *lot*) loses nominal properties and *a lot of* becomes an adverb modifying the second noun. The development of so-called binominal structures Det N_1 of N_2, which may or may not further evolve to a fully adverbial category, plays a crucial role in the grammaticalization of degree words. In our study, we also include the structure *(Det) caleta of N*, hence we investigate the distribution of *caleta de*.

As argued by 8, degree words across languages show a systematic behavior in terms of classes of words they can modify. These well-attested patterns correspond to types along a continuum of word classes defined by their syntactic-semantic properties. For example, since French *trop* 'too much' can modify all word classes, within this typology it is considered to be a Type C modifier. On the other hand, English *very* can only modify gradable adjectives ("very kind" is possible, while expressions like "*I traveled very" or "*very water" are not grammatical), therefore *very* is classified as a Type A modifier. For a complete summary of the continuum of word classes and typology, see Figure 1. As words develop into one type, they are predicted to modify words in the order along the continuum; for instance, if a word co-occurs with words of category V, it is expected to co-occur with words of category IV before it appears with words of category III. [1] As we investigate whether *caleta* has grammaticalized into a degree word, we will examine its stage of development with respect to Doetjes' continuum.

While some computational studies of grammaticalization have adopted case-driven approaches similar to ours [16, 17, 18], we also investigate how a distributional analysis of *caleta* can provide insight on the set of degree expressions currently used in colloquial Chilean Spanish. In other words, we aim to examine not just the grammaticalization of *caleta* but also how this word fits in the system of degree words in Chilean Spanish and in types of degree expressions across languages.

[1] Doetjes differentiates between 'gradable' and 'eventive' adjectives and verbs by whether or not the modifier is targeting the degree or is quantifying over events. The example she gives is from Dutch: *Jan is veel ziek* 'Jan is sick a lot' vs. *Jan is erg ziek* 'Jan is very sick.' In the former, *veel* as a quantifier targets eventive adjectives, thus it can only modify the quantity of sick events. In the latter, *erg* expresses the degree of sickness, i.e. the severity of his illness.

Category	**Word Class**					
I	gradable adjectives	*Type A* *very*^{E}	*Type B* *erg*^{D} *očen*^{R}	*Type C* *trop*^{F} *muito*^{P} *molto*^{I}		
IIa	gradable nominal predicates	*Type D* *beaucoup*^{F} *a lot*^{E}				
IIb	gradable verbs					
III	eventive verbs eventive adjectives comparatives		*Type E* *veel*^{D} *mnogo*^{R}			
IV	mass nouns				*Type F* *a mountain*^{E}	
V	plural nouns					*Type G* *many*^{E}

Figure 1: Typology of degree expressions according to their distribution along a continuum of word classes. Table adapted with modifications from [8, 138]. Superscripts indicate language: R for Russian, D for Dutch, F for French, E for English, P for Portuguese, and I for Italian.

3. Methodology

3.1. Corpus Creation

To ensure we had a good representation of colloquial Chilean Spanish, we created a subcorpus from an already existing corpus of online data [19]. The already existing corpus contained roughly 19GB of data, from diverse sources, including news, tweets, online reviews and other miscellaneous web content. We chose to create a subcorpus just from tweets to reduce the computational load for our later experiments and since we only wanted informal instances of language; *caleta* typically only occurs in less formal registers. The resulting subcorpus of $27,306,582$ tweets consisted of exactly $342,979,307$ tokens. The time span of these tweets is from 2010 to 2020.

3.2. Preprocessing

We first normalized the text in the corpus: we removed case, punctuation, diacritics, URLs, hashtags, and any repeated letters. For this last step, we only allowed double letters where they occur within normative Spanish orthography (i.e. $< r >, < c >, < l >$), elsewhere only single letters were allowed. Then we input the corpus into a plain text file separated by newlines. The resulting file was then lemmatized using SpaCy's Spanish lemma-

tizer [20].[2]

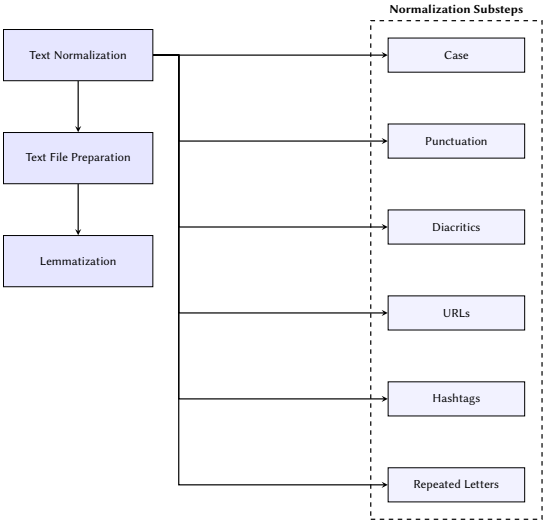

Figure 2: Preprocessing steps.

3.3. Model Selection

To represent the distributional patterns of words in our corpus, we decided to use static word embeddings over contextualized word embeddings. Non-contextualized embeddings allow us to compare our target word with other words in Chilean Spanish to examine the current stage of grammaticalization of *caleta* as determined by its closeness to different subsystems in the language.

The algorithm we use is Skip-Gram with Negative Sampling (SGNS) implemented in word2vec [21] to extract embeddings, based on previous research that showed good results for studies of semantic change [22, a.o.]. For this reason, we do not consider it necessary to use a more computationally expensive operation (e.g. dynamic word embeddings). We trained each model for five epochs, a minimum token count of 10 and the skip-gram algorithm. Initially, we experimented with several hyperparameters: the window size, the minimal word count and the vector size. The only hyperparameter that proved to be significant was the window size (see next section for more details). The resulting model used a vector length of 100 and a minimal word count of 10. To verify the validity of the model, we used analogy tests targetting gender-based morphological and semantic relations (see Table 1 for specifics). We performed the tests on both models we used for the embeddings (see following section for details). For both models, the analogy tests returned the expected word, except for the last pair with $w = 1$: where *perra* 'dog (female)' was expected, the most similar word embedding was for *quiltra* 'mutt (female)'.

Relationship	Word Pair 1		Word Pair 2		Accuracy
	Word A	Word B	Word A	Word B	
Age-based	*Hombre* 'Man'	*Mujer* 'Woman'	*Niño* 'Boy'	*Niña* 'Girl'	1.0
Familial	*Padre* 'Father'	*Madre* 'Mother'	*Hijo* 'Son'	*Hija* 'Daughter'	1.0
Feline	*Niño* 'Boy'	*Gato* 'Cat (male)'	*Niña* 'Girl'	*Gata* 'Cat (female)'	1.0
Canine	*Niño* 'Boy'	*Perro* 'Dog (male)'	*Niña* 'Girl'	*Perra* 'Dog (female)'	0.5

Table 1
The four analogy tests used to validate Word2Vec model. The equation used was $WB_2 = WA_1 - WA_2 + WB_1$.

3.4. Window Size

As mentioned in the previous section, the only hyperparameter we adjusted for the model was the window size. We extracted models for $w = [1, 10]$.[3] Although other authors have shown that small window sizes often produce noisy and unstable embeddings [23], for this project we expected small window sizes to be appropriate. Our hypothesis was that in our case, lower window sizes would capture the grammaticalized meaning of *caleta*, since the scope of grammatical words like quantifiers lies within its immediate neighbors, whereas higher window sizes show neighbors within the same semantic field (therefore its lexical use). However, since we use a corpus of tweets, window size is fairly limited by the genre itself (a possible limitation we address later).

4. Results

4.1. *Caleta*

Here we display only the results of the experiments with a small ($w = 1$) and a large ($w = 10$) window size.[4]This allows us to compare the information obtained by manipulating this parameter. In Figure 3, the word embeddings show both neighbors of the lexical noun and neighbors

[2]As an anonymous reviewer noted, our preprocessing might have worked better if we had normalized the text and lemmatized in one step. This is something we will consider for future experiments.

[3]As a reviewer suggested, we experimented with other window sizes e.g. $w = 2$. While we do not show the results for this window size, we note that there was not a signficiant difference for this window size and $w = 1$ for *caleta de*, but there was for *caleta*. For $w = 2$, *caleta* had almost no neighbors that were quantifiers. The other neighbors were *ene*, *caleta de* and then mostly toponyms, similar to the t-SNE's we show here for both strings with $w = 10$. This demonstrates that instances of just *caleta* within our corpus are more lexical uses, whereas *caleta de* demonstrates more grammaticalized uses.

[4]To generate the t-SNE graphs for both *caleta* and *caleta de*, we used the PCA (Principal Component Analysis) method since our data points were dense vectors, and we used a perplexity of 10.

of the degree word. Nearest neighbors of the noun are toponyms (i.e. names of villages) and other nouns with related meanings (e.g. *playa* 'beach' and *muelle* 'wharf'). As for the neighbors of the degree word, we find degree expressions, both adverbs and quantifiers like *mucho* and *ene*, both meaning 'a lot'. *Caleta de* also appears among the neighbors (please see subsequent section for these results).

The co-occurrence of neighbors of both meanings shows that *caleta* has partially grammaticalized; it still retains its lexical use as a noun. These findings provide evidence for a situation of layering [24], i.e. the synchronic co-existence of older and more recent functions of a form in a language.

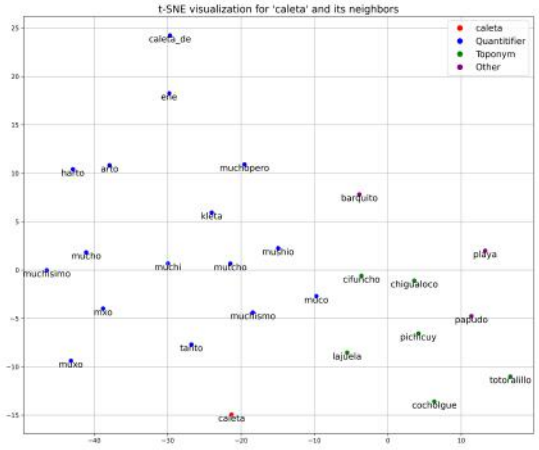

Figure 3: TSNE representation of *caleta* and its top 25 neighbors. Embeddings were created with a window size of 1. Blue corresponds to words that are quantifiers, green corresponds to toponyms (i.e. names of villages), and purple corresponds to semantically related nouns.

If we now use a larger window size, the results are different, with more neighbors associated with the lexical item. In Figure 4 we find the plural noun (*caletas*); as mentioned in historical analyses, the ability to be pluralized is a syntactic property of nouns. This attests to the persistence of some nominal categorial properties of *caleta*. We also find the noun *pescadores* 'fishermen', as the noun *caleta* typically refers to a village of fishermen and hence the nouns often co-occur (in *caleta de pescadores*), and related nouns like *muelle* 'pier' and *poza* 'puddle'.

4.2. *Caleta de*

We analyzed the results of *caleta de* separately from those of *caleta* since the former is the vestige of a binominal quantifier preceding the grammaticalization of the latter. Figure 5 and Figure 6 show the TSNE representations

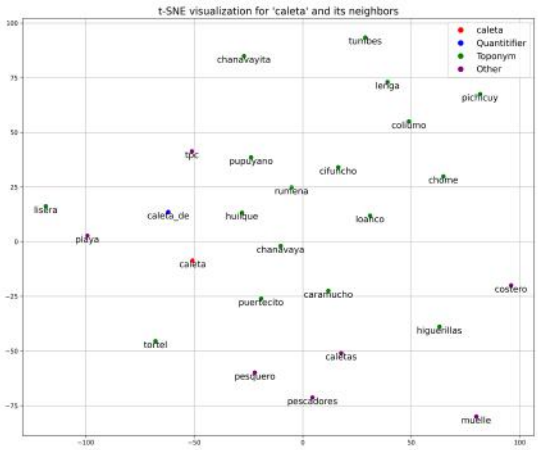

Figure 4: TSNE representation of *caleta* and its top 25 neighbors. Embeddings were created with a window size of 10. Blue corresponds to words that are quantifiers, green corresponds to toponyms (i.e. examples of villages), and purple corresponds to semantically related nouns.

of the nearest neighbors of *caleta de*. For the smaller window size, we see other quantifiers like *ene* (more in the next section), *caleta*, etc. The majority of neighbors here are quantifiers in their orthographical variants found in tweets (e.g. *mucho*, *mxo*, *nucho*, etc). Two other words that form part of binominal quantifiers are also present, *monton* and *montones*, both meaning 'pile' and 'piles', but which have grammaticalized in the same fashion as *caleta* to denote a large quantity (*un montón de N* 'a lot of N'). In this window size, only one proper noun is present, *Chorromil*, the name of a village. Lastly, we find other quantifiers, like *cualquiers* and *cualesquiers*, both orthographical variations of *cualquier*, 'whichever', and *puras*, a determiner in Chilean Spanish.

In the larger window size, we see *caleta* as its nearest neighbor. Other quantifiers like *mucho*, *ene*, *harto*, etc. are present, but they are much further away than semantically related nouns like *pescadores* 'fishermen', *artesanales* 'craftsmen', *reinetas*, a plural noun denoting a variety of white fish, as well as toponyms that are names of *caletas*. These results show once more how important the hyperparameter of window size is in capturing distributional properties of relatively newly grammaticalized words in a language.

In the following, we provide further analysis of the nearest neighbors of *caleta* and *caleta de*.

4.3. *Ene*

We decided to display the top 10 neighbors for the word *ene*, since *ene* always appeared as a top neighbor for *caleta* and *caleta de*. *Ene* comes from the Spanish pronunciation

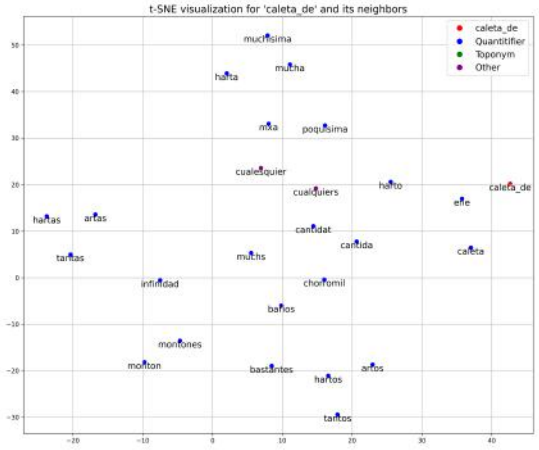

Figure 5: TSNE representation of *caleta de* and its top 25 neighbors. Embeddings were created with a window size of 1. Blue corresponds to words related to quantity, green corresponds to toponyms (i.e. examples of *caletas*), and purple corresponds to syntactically and semantically-related words.

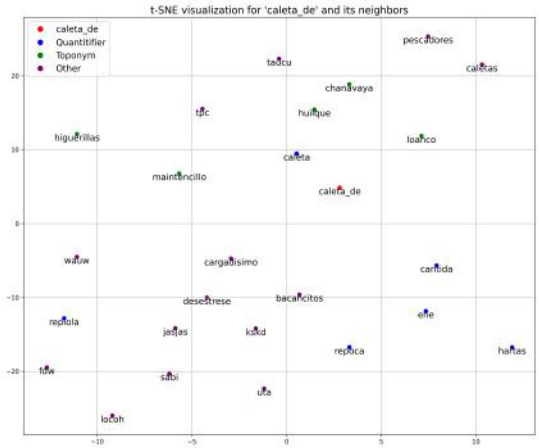

Figure 6: TSNE representation of *caleta de* and its top 25 neighbors. Embeddings were created with a window size of 10. Blue corresponds to words related to quantity, green corresponds to toponyms (i.e. examples of *caletas*), and purple corresponds to syntactically and semantically-related words.

of the grapheme $< n >$ and is used in Mathematics to denote an unspecified integer. Over time, in this variety of Spanish *ene* has grammaticalized like *caleta* to denote a large quantity and high degree. Our results show that *ene* is another example of a grammaticalized degree word, albeit in a different stage of grammaticalization. To the best of our knowledge, this has not been observed or studied. Example (3) shows a lexical use of *ene*, taken from the Dictionary of the Spanish Real Academy [25], since no such example could be found in our corpus. Example (4) shows the degree adverb (here, modifying a verb), i.e. the grammaticalized item. Lastly, example (5) shows *ene* in combination with *ctm*, a commonly used abbreviation of the phrase *concha (de) tu madre* (literally 'your mother's pussy'), which is used as a vulgar intensifier similar to *fucking* in English.

(3) El fenómeno se repite ene
The phenomenon CL.REFL repeat.PRS.3SG *n*
veces.
times
"The phenomenon is repeated *n* times."

(4) me gustó ene
CL.1SG.DAT like.PST.3SG ene
"I liked it a lot."

(5) me gustó ene ctm
CL.1SG.DAT like.PST.3SG ene ctm
"I fucking liked it a lot."

Table 2 and 3 show the closest neighbors for *ene* in our corpus. For both window sizes, none of the neighbors are semantically related to Mathematics, which would be expected if *ene* still retained some of its original lexical meaning. For the smaller window size, all of the neighbors are degree words meaning 'much' (including the noun *cantidad* which can appear in a binominal structure *cantidad de N* 'a large quantity of N'). For the larger window size, half of the neighbors are quantifiers. We also see the expressive *puxis* (an orthographical variation of *pucha*, meaning 'darn'), spellings of laughter and the vulgar term *autodelicioso*. This is evidence for what has been previously described in the literature that degree modifiers, as highly volatile units of language, are subject to rapid change and become expressives [26].

Rank	Word	Score
1	***caleta de*** 'a lot of'	0.78
2	***cantitat*** (*cantidad*, orthographical variation, 'quantity')	0.67
3	***harto*** 'a lot'	0.66
4	***caleta*** 'a lot' or 'village'	0.66
5	***kleta*** 'caleta' (orthographical variation)	0.65
6	***arto*** 'harto' (orthographical variation)	0.64
7	***mucho*** 'a lot'	0.64
8	***tanto*** 'so much'	0.63
9	***mxo*** 'mucho' (orthographical variation)	0.62
10	***muchopero*** (*mucho pero* as one word, 'a lot but...')	0.61

Table 2
Ranked words with their scores (cosine) for *ene* for $w = 1$

4.4. Other Quantifiers

Lastly, we show word embeddings of other degree words, in this case 'stable' quantifiers in Chilean Spanish: *harto* 'a lot', *mucho* 'a lot', *tanto* 'so many.' It is worth mentioning that unlike *caleta*, *caleta de* and *ene* (which syntacti-

Rank	Word	Score
1	***kleta*** (orthographical variation of *caleta*)	0.71
2	***caleta de*** 'a lot of'	0.68
3	***cantitat*** (*cantidad*, orthographical variation, 'quantity')	0.67
4	*graziash* (*gracias*, orthographical variation, 'thanks')	0.66
5	*jsjsjd* 'laughter'	0.66
6	***harto*** 'a lot'	0.66
7	*puxis* (orthographic variation of *pucha*, 'darn')	0.66
8	*autodelicioso* (lit. 'self-delicious', term used for masturbation)	0.64
10	***muchosaño*** (*muchos años* as one word, 'many years')	0.63

Table 3
Ranked words with their scores (cosine) for *ene* for $w = 10$. Bold words correspond to quantifiers.

cally can be considered degree adverbs), these quantifiers inflect for gender and number when modifying a noun. The purpose of using the lemmatizer was to control for this, but as the results show, some inflected tokens of these quantifiers were not properly lemmatized.

Tables 4, 5, 6, 7, 8 and 9 show the nearest neighbors for *harto*, *mucho* and *tanto* at the two window sizes. For *harto*, we see that the majority of its neighbors are other quantifiers for both window sizes, as well as orthographical variations (e.g. *harrto*, *arto*) and inflected versions of the lexeme, like the feminine form *harta*. Likewise, *tanto* as its neighbors for the smaller window size shows mostly orthographical variations (e.g. *tsnto*, *tabto*), while for the larger window size we can see similar results to *ene*, where nouns like 'laughter' are amongst the neighbors. For *mucho*, we can see mostly orthographical variants for the smaller window size (e.g. *muxo*, *muxho*) and for the larger window size we see less orthographical variations and more of other quantifiers, even its antonym *poco*, which also occurs with intensifying affixes: *re-poco* and *poc-azo* 'very little'.

Rank	Word (Gloss)	Score
1	***arto*** 'harto' (orthographical variation)	0.94
2	***mucho*** 'a lot'	0.84
3	***bastante*** 'quite'	0.78
4	***harrto*** 'harto' (orthographical variation)	0.74
5	***mxo*** 'mucho' (orthographical variation)	0.72
6	***muchisimo*** 'mucho' (superlative)	0.71
7	***muxo*** 'mucho' (orthographical variation)	0.69
8	***mutcho*** 'mucho' (orthographical variation)	0.68
9	***mucjo*** 'mucho' (orthographical variation)	0.67
10	***nucho*** 'mucho' (orthographical variation)	0.66

Table 4
Ranked words with their scores (cosine) for *harto* for $w = 1$. Bold words correspond to quantifiers.

5. Discussion

Our word embedding results for *caleta* show that nowadays the word is used to express high degree. In addition,

Rank	Word (Gloss)	Score
1	***arto*** 'harto' (orthographical variation)	0.81
2	***mucho*** 'a lot'	0.72
3	*sosi* (*eso sí*, abbreviation, 'though')	0.69
4	***bastante*** 'quite'	0.68
5	***harta*** 'a lot'	0.68
6	***ene*** 'a lot'	0.66
7	*pucha* 'darn'	0.63
8	***haarto*** 'harto' (orthographical variation)	0.63
9	***repoco*** 'poco' (intensifier)	0.63
10	***pocazo*** 'poco' (augmentative)	0.61

Table 5
Ranked words with their scores (cosine) for *harto* for $w = 10$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***tsnto***'tanto' (orthographical variation)	0.76
2	***demasia*** (*demasiado*, phonetic variation, 'too much'	0.70
3	***tantotanto*** 'tanto' (repeated)	0.69
4	***mucho*** 'a lot'	0.69
5	***tantoy*** (*tanto y* as one word, 'so much and')	0.69
6	***tabto*** 'tanto' (orthographical variation)	0.68
7	***tantisimo*** 'tanto' (superlative)	0.67
8	***tnto*** 'tanto' (orthographical variation)	0.64
9	***tanro*** 'tanto' (orthographical variation)	0.64
10	***mutcho*** 'mucho' (orthographical variation)	0.64

Table 6
Ranked words with their scores (cosine) for *tanto* for $w = 1$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***mucho*** 'a lot'	0.71
2	***tsnto*** 'tanto' (orthographical variation)	0.65
3	***tantotanto*** 'tanto' (repeated)	0.63
4	**tantisimo** 'tanto' (superlative)	0.60
5	***simuchas*** (*sí muchas* as one word, 'yes a lot')	0.60
6	*jskdkd* 'laughter'	0.60
7	*jajajajajajaun* 'laughter'	0.60
8	***muchogracias*** (muchas gracias as one word, 'thanks a lot')	0.59
9	*tisin* (*tí sin* as one word, 'you (prepositional), without)	0.58
10	*pueso* (portmanteau of *pues eso*, 'exactly')	0.58

Table 7
Ranked words with their scores (cosine) for *tanto* for $w = 10$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***muchisimo*** 'mucho' (superlative)	0.91
2	***mxo*** 'mucho' (orthographical variation)	0.88
3	***harto*** 'a lot'	0.82
4	***muxo*** 'mucho' (orthographical variation)	0.81
5	***mucjo*** 'mucho' (orthographical variation)	0.80
6	***muchi*** 'mucho' (diminutive)	0.77
7	***muho*** 'mucho' (orthographical variation)	0.77
8	***muxho*** 'mucho' (orthographical variation)	0.77
9	***arto*** 'harto' (orthographical variation)	0.76
10	***nucho*** 'mucho' (orthographical variation)	0.75

Table 8
Ranked words with their scores (cosine) for *mucho* for $w = 1$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***muchisimo*** 'mucho' (superlative)	0.79
2	***harto*** 'a lot'	0.74
3	***tanto*** 'so much'	0.71
4	***poco*** 'a little'	0.67
5	***muchoy*** (*mucho y* as one word, 'a lot and'	0.65
6	***muccho***'mucho' (orthographical variation)	0.65
7	***bastante*** 'quite'	0.65
8	***muchopero*** (*mucho pero* as one word, 'a lot but')	0.64
9	*aunpero* (*aún pero* as one word, 'still but')	0.63
10	***muchisisisismo*** 'mucho' (repeated superlative)	0.61

Table 9
Ranked words with their scores (cosine) for *mucho* for $w = 10$. Bold words correspond to quantifiers.

in our results both the lexical noun and the degree modifier are present. The choice of hyperparameters, specifically window size, has important consequences: a small window size yields nearest neighbors for both forms, while a larger window size results in more neighbors of the lexical noun. We hypothesize that this is due to the fact that as a degree word, *caleta* is a modifier, and occurs in close adjacency to the modified word. Hence, a small window captures this distribution. On the other hand, as a lexical noun *caleta* is less syntactically constrained, with more positional freedom and semantic content.

While cosine similarity scores give us insight into a changing word's distribution, they alone do not tell us about its syntactic properties in detail. To better understand *caleta*'s current status as a degree modifier, we performed a *post-hoc* analysis of the top 20 collocates of *caleta* and *caleta de*. We looked specifically at the top tokens that immediately precede and proceed the two strings in our unlemmatized corpus. We were interested in the kinds of words that *caleta* and *caleta de* have come to modify, in accordance to Doetjes's typology of degree modifiers (see Section 2).

Our analysis shows that *caleta* has evolved extensively beyond its original lexical usage, wherein it was only compatible with count nouns that were semantically related e.g. *pescadores* 'fishermen' *camarones* 'shrimp (plural)', headed by the preposition *de*. The structure *caleta de* is now compatible with count nouns beyond the semantic domain of a fishing village: *años* 'years', *veces* 'times/instances' (see (6)), as well as mass nouns e.g. *plata* 'money (informal), *tiempo* 'time' (see (7)). It can also modify comparatives e.g. *mejor* 'better', *peor* 'worse' (see (9)); eventive verbs e.g. *dormir* 'to sleep', *reír* 'to laugh' (see (8)); gradable verbs *gustar* 'to like', *querer* 'to want' (see (2); and finally gradable nominal predicates[5] e.g. *hambre* 'hunger', *pena*, 'sorrow', as in (10).

(6) Hace caleta de años
make.PRS.3SG caleta of years
"Many years ago"

(7) es caleta de plata
be.PRS.3SG caleta of money
"it's a lot of money."

(8) Yo igual reí caleta.
1SG.NOM same laugh.PST.1SG caleta
"I laughed a lot, anyway."

(9) hay que cuidarse caleta mejor...
be.EXIST.PRS.3SG that care.INF.REF caleta better
"one has to take care of themselves much better."

(10) Hace caleta de frío.
make.PRS.3SG caleta of coldness
"It's really cold."

There were no cases of *caleta* modifying either eventive adjectives or gradable adjectives within our corpus. This, according to Doetjes's classification, indicates that *caleta* has evolved into a type D degree modifier. Figure 7 shows *caleta*'s position in this typology, in comparison to the other degree expressions in Chilean Spanish that we have discussed in this paper. Our results align with claims in the literature that Type C and D are the most common in the Romance languages [8]. Lastly, within our results, *caleta* has no nearest neighbors with Type A modifiers (e.g. *muy* 'very'), which combine exclusively with gradable adjectives. This is not surprising since Type A modifiers have no overlap in word classes with Type D modifiers; their distributions are disjoint. This highlights how embeddings capture syntactic properties of words, as opposed to just similarity of meaning.

Our study has two main findings, which answer the research questions above. First, we have shown that *caleta* is undergoing grammaticalization: both the older and the new meaning are captured by the word embeddings. Importantly, we see a difference in the results depending on the window size, when compared to other degree words which are grammatical items and not undergoing change, like *mucho* and *harto*. In the latter case, window size does not significantly impact the neighbors. Additionally, our *post-hoc* analysis provided insight on the properties of *caleta* as a degree word.

Second, our word embeddings have allowed us to reveal the inventory of degree words in colloquial Chilean Spanish, including a word that to date had never been investigated, *ene*. These words denote high degree (intensifiers), words that are known to change rapidly due to social and expressive pressure [26]. Since *caleta* and *ene* are not normative forms, they are left out of tradi-

[5] Gradable nominal predicates, in Doetjes's definition, are nouns which are generally the objects of light verb expressions. The examples she gives are from French e.g. *Elle a très soif* 'She is very thirsty.' In Spanish, such light verb constructions also exist, so we consider cases like *tener sed* 'to be thirsty (lit. to have thirst)' to also be examples of nominal predicates.

<table>
<tr><th>Category</th><th>Word Class</th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>I</td><td>gradable adjectives</td><td>Type A</td><td rowspan="2">Type B</td><td rowspan="5">Type C
harto
bastante
demasiado</td><td rowspan="4"></td><td rowspan="5"></td></tr>
<tr><td>IIa</td><td>gradable nominal predicates</td><td>Type D</td></tr>
<tr><td>IIb</td><td>gradable verbs</td><td>caleta</td></tr>
<tr><td rowspan="3">III</td><td>eventive verbs
eventive adjectives
comparatives</td><td>ene
mucho
tanto</td><td rowspan="3">Type E</td></tr>
<tr><td>IV</td><td>mass nouns</td><td rowspan="2">Type F
un
montón
cantidad
montones</td></tr>
<tr><td>V</td><td>plural nouns</td><td>Type G
vario</td></tr>
</table>

Figure 7: Degree words found in our results and their corresponding types according to Doetjes' model; modified table from [8, 138]

tional studies. This entails that we may miss instances of change possibly of interest to current linguistic theory. Hence, word embeddings can be a tool to study lesser-known subsystems of a language and capture ongoing changes in synchrony.

6. Conclusion

Our study contributes to studies of language change by analyzing intensifiers in colloquial Chilean Spanish (an understudied variety) from the past twenty years. We do not yet have data from multiple temporal slices to demonstrate direct evidence of changes in grammatical behavior. For this reason, we infer grammaticalization from synchronic distributional patterns. Nevertheless, we reveal an ongoing change that had not been previously studied. Using spontaneous speech from tweets, we gained access to informal speech where speakers communicate in an unedited way, which has allowed us to study the use of older and more recent degree expressions. In the future, we plan on expanding the time span of the data, depending on the availability of more text reflecting spontaneous speech in this variety of Spanish.

We have shown that static word embeddings provide evidence for this change and can reveal meaning relations not previously studied. Moreover, we show that different choices of hyperparameters have an effect on which meaning of the word undergoing change (the lexical vs. the grammatical) is represented. Nevertheless, comparing our results with dynamic embeddings in the future could prove interesting.

Some limitations of our study are due to the genre itself. One such limitation is the difficulty with lemmatization: as we have mentioned, these are tweets, so we find strings that do not conform to normative orthography (for example, typos, abbreviations etc), therefore the lemmatizer has difficulty with detecting words of the same lexeme. In addition, Twitter users tend to adopt orthographical forms that reflect pronunciation and sometimes are intended to be expressive, like repeating vowels in a word to express a very high degree. Furthermore, using a corpus of tweets means that the character limit has an impact on the possible window sizes. To obviate this problem, further studies on *caleta* could use longer texts that have the same register as tweets, e.g. blog posts.

Lastly, the only hyperparmeter we significantly experimented with were the window size and the minimal word count. More hyperparameter fine tuning (e.g. adjustment of negative sampling and vector size) could potentially yield more robust results.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

[1] W. L. Hamilton, J. Leskovec, D. Jurafsky, Cultural shift or linguistic drift? comparing two computational measures of semantic change, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2116–2121. URL: https://aclanthology.org/D16-1229/. doi:10.18653/v1/D16-1229.

[2] A. Kutuzov, L. Øvrelid, T. Szymanski, E. Velldal, Diachronic word embeddings and semantic shifts: a survey, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1384–1397. URL: https://aclanthology.org/C18-1117/.

[3] L. Fonteyn, E. Manjavacas, S. Budts, Exploring morphosyntactic varation and change with distributional semantic models, Journal of Historical Syntax 6 (2022) 1–41.

[4] A. Meillet, L' évolution des formes grammaticales, Scientia 12 (1912) 130–148.

[5] P. J. Hopper, E. C. Traugott, Grammaticalization, Cambridge Textbooks in Linguistics, 2 ed., Cambridge University Press, 2003.

[6] D. Bolinger, Degree Words, De Gruyter Mouton, Berlin, Boston, 1972. URL: https://doi.org/10.1515/9783110877786. doi:doi:10.1515/9783110877786.

[7] A. Neeleman, H. Van de Koot, J. Doetjes, Degree expressions, The Linguistic Review 21 (2004) 1–66. doi:doi:10.1515/tlir.2004.001.

[8] J. Doetjes, Adjectives and Degree Modification, in: L. McNally, C. Kennedy (Eds.), Adjectives and Adverbs: Syntax, Semantics, and Discourse, Oxford University Press, 2008, pp. 123–155. doi:10.1093/oso/9780199211616.003.0006.

[9] P. Amaral, When Something Becomes a Bit, Diachronica 33 (2016) 151–186. doi:10.1075/dia.33.2.01ama.

[10] Y. Luo, D. Jurafsky, B. Levin, From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers, in: N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu (Eds.), Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1–13. URL: https://aclanthology.org/W19-4701/. doi:10.18653/v1/W19-4701.

[11] A. Abeillé, O. Bonami, D. Godard, J. Tseng, The Syntax of French de-N' Phrases, Proceedings of the International Conference on Head-Driven Phrase Structure Grammar (2004) 6–26. doi:10.21248/hpsg.2004.1.

[12] C. Marchello-Nizia, Grammaticalisation et changement linguistique, De Boeck, 2006.

[13] K. Verveckken, Towards a Constructional Account of High and Low Frequency binominal Quantifiers in Spanish, Cognitive Linguistics 23 (2012). doi:10.1515/cog-2012-0013.

[14] E. Traugott, Grammaticalization, Constructions and the Incremental Development of Language: Suggestions from the Development of Degree Modifiers in English, Variation, Selection, Development: Probing the Evolutionary Model of Language Change (2008) 219–250.

[15] P. Amaral, Bocado: Scalar Semantics and Polarity Sensitivity, Zeitschrift für romanische Philologie 136 (2020) 1114–1136.

[16] L. Fonteyn, E. Manjavacas, Adjusting scope: a computational approach to case-driven research on semantic change, in: Proceedings of the Workshop on Computational Humanities Research (CHR 2021), volume 2898 of *CEUR Workshop Proceedings*, 2021, pp. 280–298. URL: http://ceur-ws.org/Vol-2989/long_paper26.pdf.

[17] P. Amaral, H. Hu, S. Kübler, Tracing semantic change with distributional methods: The contexts of algo, Diachronica 40 (2023) 153–194.

[18] R. Nagata, Y. Kawasaki, N. Otani, H. Takamura, A Computational Approach to Quantifying Grammaticization of English Deverbal Prepositions, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 211–220. URL: https://aclanthology.org/2024.lrec-main.19.

[19] J. Ortiz-Fuentes, Chilean Spanish Corpus, 2023. URL: https://huggingface.co/datasets/jorgeortizfuentes/chilean-spanish-corpus. doi:10.57967/hf/3181.

[20] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, The Journal of Open Source Software 5 (2020) 2914. doi:10.5281/zenodo.1212303.

[21] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at ICLR 2013 (2013) 1–12.

[22] H. Hu, P. Amaral, S. Kübler, Word embeddings and semantic shifts in historical spanish: Methodological considerations, Digital Scholarship in the Humanities 37 (2022) 441–461.

[23] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: K. Toutanova, H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 302–308. URL: https://aclanthology.org/P14-2050/. doi:10.3115/v1/P14-2050.

[24] P. Hopper, On some principles of grammaticization, in: Approaches to Grammaticalization, Benjamins, 1991, pp. 17–35.

[25] Real Academia Española, Diccionario de la lengua española, 2025. URL: <https://dle.rae.es>[6/1/2025].

[26] R. Ito, S. Tagliamonte, Well weird, right dodgy, very strange, really cool: Layering and recycling in english intensifiers, Language in Society 32 (2003) 257–279. doi:10.1017/S0047404503322055.