Article

Matti Marttinen Larsson*

Modelling incipient probabilistic grammar change in real time: the grammaticalisation of possessive pronouns in European Spanish locative adverbial constructions

<https://doi.org/10.1515/cllt-2021-0030>

Received April 28, 2021; accepted October 6, 2021; published online October 25, 2021

Abstract: The present paper provides a methodological case study on how underlying incipient grammar change might be discerned even when frequencies of the incoming variant are apparently marginal and stable. Analysing the spread of tonic possessive pronouns in complements of locative adverbial constructions in European Spanish from a probabilistic perspective, more than 11,000 locative constructions from 1900 to 2004 were compiled, and probabilistic grammar change was operationalised as an interactive function between language-internal predictors and real time. The results reveal that numerous intralinguistic factors have been and are active in constraining the variation, with the innovation spreading significantly in spite of apparent stability in frequency. Crucially, the findings demonstrate that, even in a relatively standardised written language where the innovation has a considerably low frequency, the innovation grammaticalises along the same pathway as in colloquial vernaculars where the incoming variant is employed much more frequently.

Keywords: grammaticalisation; incipient change; morphosyntax; possessive pronouns; probabilistic grammar change

1 Introduction

The present paper adopts a probabilistic view on diachronic grammar change (Bresnan 2007; Szmrecsanyi 2016), and showcases how an incipient instance of morphosyntactic variation is undergoing a change in its linguistic conditioning

*Corresponding author: Matti Marttinen Larsson, Stockholm University, Stockholm, Sweden, E-mail: matti.marttinen@su.se. <https://orcid.org/0000-0002-6224-7872>

during a 100-year timespan in spite of seemingly stable and marginal frequencies of the grammaticalising item.¹ The phenomenon under study is the grammaticalisation of tonic (i.e. postpositive) possessive pronouns in locative adverbial phrases. In these adverbial constructions, traditional prepositional person-referential complements (e.g. *delante de nosotros* ‘in front of us’) alter with innovative tonic possessive complements (e.g. *delante nuestro* literally ‘in our front’), the latter form not being normatively acceptable. The possessive variant has spread through extension² from noun phrases where these two types of complements are interchangeable for the expression of possession (*la casa de nosotros*, literally ‘the house of us’ > *la casa nuestra* ‘our house’). However, the use of the tonic possessive pronoun in locative adverbial phrases is considered grammatically implausible since possessive pronouns are restricted to noun phrases where they entail traits of possession and person. The identification of the driving forces behind the functioning and diffusion of the innovative possessive variant in adverbial contexts remains to be scrutinised using diachronic data, which the present study provides. Firstly, a frequency account is provided which shows that the incoming possessive variant remains marginal throughout the entire studied timespan with minimal frequency fluctuations. Secondly, the paper illustrates the benefits of focussing on probabilistic approaches to the examination of incipient language change. Using this approach, what takes centre stage here are the probabilistic conditions underlying variant choice, that is, the reasons for which speakers choose a determined variant in discourse, rather than frequency competition between variants.

The analysis reveals that, during a time span of 100 years, the incoming variant represents only 6.9% of the compiled data and exhibits an overall stable and low frequency of use. Considering that this variable has apparently been subject to apparent social evaluation during its diffusion over registers, dialects and sociolects in Spain (see Section 3), it comes as no surprise to find that the

1 The initial part of this research was conducted during a research stay at the Quantitative Lexicography and Variational Linguistics research unit at KU Leuven. I thank Professor Benedikt Szmrecsanyi for valuable insights and feedback concerning regression modelling, both on a computational and theoretical level, as well as philosophical. At a later stage, different versions of this paper have been commented on by research seminar participants at Stockholm University and at the Sociolinguistics Lab at Berkeley, by students of Spanish grammar at the Humboldt University of Berlin, and by Livia Oushiro, Marie-Ève Bouchard, Carlota de Benito Moreno, Isaac Bleaman, Laura Álvarez López and Miriam Bouzouita, to whom I am thankful for their help in refining the work during its years of elaboration. Lastly, I am very grateful to the anonymous peer-reviewers who have generously devoted their time and effort to improving this paper. Naturally, all errors are solely attributed to myself.

2 The term ‘extension’ is used in Harris and Campbell’s (1995: 51) sense: “a mechanism which results in changes in the surface manifestation of a pattern and which does not involve immediate or intrinsic modification of underlying structure.”

innovation appears to be inhibited in written and relatively standardised textual sources as those that comprise the analysed corpora. Yet, what the present study shows is that, in spite of a general inhibition on the implementation of the innovation in written and rather formal registers, this restriction is limited to frequency and there appears to be no such inhibition in terms of the innovation's diffusion across linguistic contexts. As the findings highlight, not only are we able to discern that, in spite of apparent incipience, stability and marginality of the studied innovation, probabilistic constraints are operating throughout the studied time-span and we are able to identify how the alternation evolves over time with progressive actualisation³ of the innovation.

In all, the present study constitutes a methodological case study that shows how language-internal change, i.e. grammaticalisation of the innovation, might be identifiable even in instances of incipient and possibly inhibited change. This issue is of both theoretical as well as methodological importance since the study of incipient variation contributes to a fuller understanding of the genesis of variation and the full trajectory of language change.

The paper is organised as follows. Section 2 describes the probabilistic grammar approach and how this extends to the inquiry into diachronic variation and change. Section 3 outlines the phenomenon for the case study, which is an instance of morphosyntactic variation in European Spanish where tonic possessive pronouns are undergoing reanalysis and expanding to replace prepositional complements with personal pronouns in locative adverbial phrases. Earlier studies' indications of the possessive variant's diachronic spread are presented (Section 3), as well as its synchronic distribution (Section 3.1). In addition, the conditioning language-internal factors highlighted by earlier studies are scrutinised and several hypotheses concerning these factors' influences are proposed (Section 3.2). In Section 4, the data, the circumscription of the variable context and the annotation are described (Section 4.1). Section 4.2 elaborates on the employed statistical techniques. In Section 5, the results are outlined. Section 5.1 discusses the model selection process and provides the model characteristics. In the subsequent sections, the results from the mixed-effects model are reported in light of the influence of the grammatical person of the referent (Section 5.2), the grammatical number of the referent (Section 5.3) and the locative adverbs (Section 5.4). Lastly, a discussion is brought forward (Section 6) and the study's conclusions are outlined (Section 7).

³ Following De Smet (2012: 601) and the scholars cited by him, 'actualisation' is taken to refer to "the process that bears out the consequences of a prior reanalysis. Where reanalysis invisibly assigns a new syntactic representation to an existing surface sequence, actualization involves the emergence of overtly new syntactic behavior".

2 The probabilistic view on diachronic change

The present study is framed in the variation- and usage-centred probabilistic grammar framework (Bresnan 2007). This approach is concerned with language use and variation along with the probabilistic experiences that speakers have concerning linguistic variants in use. Drawing significant inspiration from Labovian variationist studies whose interest lie in scrutinising the how's and why's of variation, probabilistic grammar scholars narrow down these research endeavours with a focus on morphosyntactic variation; still, similarly to the sociolinguistic variationist enterprise, the key assumptions are that variation is subject to probabilistic constraints that influence variant alternation within and across varieties (see Grafmiller et al. 2018 for a comprehensive theoretical overview).

Crucially, the probabilistic grammar approach assumes that grammar constitutes the “cognitive organization of one's experience with language” (Bybee 2006: 711), meaning that variant alternation processes and their constraints are taken to be learned from speakers' exposure to and interaction with other speakers (Bresnan and Ford 2010; Bybee and Hopper 2001; Grafmiller et al. 2018: 1). In that sense, studies within probabilistic grammar adopt the premise that a) grammatical knowledge involves a probabilistic component that facilitates a predictive capacity to speakers and b) such grammatical knowledge is in large part experience-based, and the probabilistic conditioning shaping variation patterns evolves during speakers' lifetimes, adapting to their community input (Grafmiller et al. 2018: 2). Even though this approach is entirely compatible with that of sociolinguistic variationist studies within the Labovian school, probabilistic grammar focuses more specifically on the influence of intralinguistic factors in order to infer, on one hand, in what manner these influence variation and, on the other hand, what such conditioning reveals about speakers' grammatical knowledge (Grafmiller et al. 2018; Szmrecsanyi 2017: 693).

As Grafmiller et al. (2018: 3) outline, a notion that should hold true in light of the probabilistic grammar approach is the following: if speakers' behaviour is partly guided by universal cognitive processes, and speakers' behaviour forms the basis for community-level patterns of behaviour, we are able to predict that the effect of certain cognitive factors on syntactic variation between different (sub)varieties of a language should be relatively stable in terms of the direction of the effects of those factors. This prediction is based on experimental evidence that shows a convergence between linguistic conditioning of variant forms found in corpus data and experimental studies (Bresnan 2007). This key finding supports the notion of probabilistic analyses of corpus data being representative of speakers' knowledge of the language (Bresnan 2007; Bresnan and Ford 2010; Hilpert and Mair 2015; Lorenz 2012). Extending this notion to diachrony, this probabilistic view of grammar would allow

corpus-based research of morphosyntactic variation to model the manner and functioning of grammatical knowledge in real time (Hilpert and Mair 2015: 191). Crucially, it is assumed that “the general cognitive mechanisms that underlie speakers’ behaviour in the present have been the same in the past” (Hilpert and Mair 2015: 191). In that sense, the use of the probabilistic grammar approach in the present study aids us in documenting the way in which speakers’ expression of morpho-syntactic variation evolves over time.

In this paper, such probabilistic plasticity is discerned independently of frequency fluctuations; instead, by computing real time as an interaction effect with intralinguistic predictor variables, we are able to determine to what extent grammar change has occurred (Gries and Hilpert 2010; Hilpert and Mair 2015; Szmrecsanyi 2016: 165). Thus, following Szmrecsanyi (2016), the criterion for positing probabilistic grammar change is the identification of a significant interaction between language-internal constraints and real time. Operationalising grammar change as such, we scrutinise the different functions employed by each linguistic variant and, by extension, the evolution of probabilistic community grammars.

As a case study, the diachronic probabilistic grammar approach will allow us to model an instance of ongoing grammar change in European Spanish, namely the grammaticalisation of tonic possessive pronouns in locative adverbial constructions. Examining more than 11,000 observations from written corpora, the diachronic evolution that these adverbial constructions have undergone during 1900–2004 is scrutinised. In the following section, background concerning the linguistic variable is provided.

3 Morphosyntactic variation in Spanish locative constructions

The phenomenon under study is an instance of grammatical variation in Spanish locative adverbial constructions. This variation is exemplified below, where the locative adverbs *detrás* (‘behind’) and *delante* (‘in front [of]’) are followed by two variants that alternate with each other:

- 1) [...] Becerril hablaba *detrás* **mío**. Luego, Berbeitos me dijo una palabra larguísima, sentado *delante* **de mí** y yo perdía la noción de tiempo (CREA, *La media distancia*, Alejandro Gándara, 1984, Spain).
 ‘[...] Becerril spoke *behind* **me** (1.SG.POSS). Then, sitting *in front of* **me** (1.SG.PREP.PN.), Barbeitos told me something lengthy, and I forgot the concept of time.’

- 2) Uno de los chicos, en la alcoba *detrás de mí* comenzó a llorar. (CORDE, *La forja de un rebelde*, Arturo Barea, 1951, Spain).
'One of the boys, in the bedroom *behind me* (1.SG.PREP.PN), began to cry.'
- 3) Padre, si trata así a madre *delante mío*, me voy. No llores, mamá. (CORDE, *Una historia de pasión*, Miguel de Unamuno, 1917, Spain).
'Father, if you treat mother like that *in front of me* (1.SG.POSS), I'll leave.
Don't cry, mother.'

In this case of binary alternation, the adverbs are thus followed by either prepositional complements (e.g. *delante de mí*, 'in front of me'; *detrás de mí*, 'behind me') or possessive complements (e.g. *delante mío*, literally 'in front mine'; *detrás mío*, literally 'behind mine'). The latter possessive construction is an innovative variant that started to spread in the early 1900s in several varieties of Spanish (Marttinen Larsson and Álvarez López 2017), thus constituting a rather incipient phenomenon of variation. Until now, mainly synchronic studies (Eddington 2017; Hoff 2020; Marttinen Larsson and Bouzouita 2018; Salgado and Bouzouita 2017; Santana Marrero 2014) and diachronic descriptive univariate analyses (Marttinen Larsson and Álvarez López 2017⁴) have been brought forward to deal with this case of variation. As such, the different stages of grammaticalisation that the possessive variant has experienced remains to be determined through a diachronic examination of the variants' competition.

Even though there are some historical accounts regarding usage of the possessive innovative variant with certain locatives as early as the 15th century, it was not until the 20th century that noticeable competition between the two variants began (Marttinen Larsson and Álvarez López 2017; Octavio de Toledo y Huerta 2016: 217). By that time, the emergent use of the locative possessive constructions (e.g. *delante mío* [literally 'in front mine']) did not go unnoticed by normative grammarians and speakers, provoking some debate and prescriptive works in an attempt to restrain the innovation's advance. Reports concerning the innovation were harshly negative, being described as "a completely abnormal construction

⁴ Even though that study by Marttinen Larsson and Álvarez López (2017) was the first to document the diachronic diffusion of the innovative possessive variant in different Spanish-speaking regions, it nonetheless left unexplored the influence of several sources of variability that might have conditioned the course of actualisation of the phenomenon (e.g. grammatical person, grammatical number, type of locative, etc.). Furthermore, in contrast to this paper, the study by Marttinen Larsson and Álvarez López (2017) only scrutinised the spread of a subset of complements and locatives, and thus did not examine all the possible instances of the variable. Thus, the diachronic dimension of this linguistic variable needs to be revisited more thoroughly through a multivariate analysis, which the present study provides.

[that is] in expansion” (Carnicer 1967, March 9; my translation) and “an incorrec-tion that has been spreading rapidly during the last couple of years, and that, if God does not take care of it, will inundate all linguistic domains of Spanish, both metropolitan varieties as well as varieties overseas” (Llorente Maldonado 1986: 47–48; my translation), suggesting that the innovation historically constitutes a variable that is highly subjected to social valuation (Labov 2001: 196). As Marttinen Larsson and Bouzouita (under evaluation) point out, the possessive form still appears to be subject to certain stigmatisation, especially the more marked variants: for instance, the denominal locative locution *al lado* (‘by the side’) is comprised by a masculine noun (*lado*), and thus its combination with a feminine possessive (which is frequently found in the Southern parts of Spain, where there appears to be free variation between masculine and feminine pos-sessives in locative adverbial constructions; see Marttinen Larsson and Bou-zouita [under evaluation]) is deemed particularly incorrect. The following excerpt from Twitter exemplifies this, where a radio listener is directing himself to the hosts of Radio Seville in an attempt to correct the use of the denominal locative with a feminine possessive:

- 4) @javier_mrquez @valentingarcia2 @RadioSevilla Ahora es un buen momento para dejar de decir “*al lado mía*” “*al lado suya*”, etc saludos ‘@javier_mrquez @valentingarcia2 @RadioSevilla This might be a good time to stop saying “al lado.SG.MASC mía.1.SG.FEM.POSS” “al lado.SG.MASC suya.3.SG.FEM.POSS”, etc greetings’ (Marttinen Larsson and Bouzouita [under evaluation], Seville, August 13th 2018)

The following section outlines previous studies’ accounts of the use of the possessive variant.

3.1 Synchronic accounts

According to the Royal Spanish Academy (Real Academia Española [RAE]) and the Association of Academies of the Spanish Language (Asociación de Academias de la Lengua Española [ASALE]; RAE and ASALE 2009: 18.4ñ), the innovative possessive variant not only appears to have a high frequency of use in several varieties of Spanish, but it also seems to have diffused to a large number of strata and registers. European Spanish Twitter data examined by Marttinen Larsson and Bouzouita (2018) document the use of the possessive variant in 28.1% (62/221) of the locative constructions, and Hoff (2020) found an overall rate of the possessive variant of 47% (2,594/5,524) in Twitter data from Madrid. Santana Marrero (2014)

studied data from global Spanish-language Google News and found a large proportion of the possessive variant (29.78%, 187/628).⁵ Salgado and Bouzouita (2017: 775) report almost identical distributions in European Spanish oral corpora, with the innovation comprising 29.9% (87/291) of the compiled observations. Evidently, in synchronic data, the possessive variant is far from marginal with a documented usage in both formal linguistic production (such as news articles from the Spanish-speaking world; e.g. Santana Marrero 2014) and oral/colloquial language use (Hoff 2020; Marttinen Larsson and Bouzouita 2018; Salgado and Bouzouita 2017).

3.2 Conditioning intralinguistic factors

Synchronically, the most influential factors in conditioning the variation are the grammatical person and number of the referent and the type of locative adverb involved in the construction (Hoff 2020: 66; Marttinen Larsson and Bouzouita 2018; Salgado and Bouzouita 2017).

As concerns the grammatical person, studies indicate that the 1st person favours the possessive complement, whereas the 2nd person exhibits more variability and the 3rd person almost categorically disfavours the possessive (Hoff 2020: 66; Marttinen Larsson and Bouzouita 2018: 29–30; Salgado and Bouzouita 2017: 776). The existence of a discourse participation hierarchy (1st person > 2nd > 3rd; speech act participants > 3rd person) is well documented in typological studies and historical linguistics (Givón 1994; Manning 2003; Silverstein 1976; among others). Furthermore, in the case of Spanish possessives, these appear to a large extent to undergo reanalysis following this very order not only in adverbial phrases but also in noun phrases (Company Company 2017; Pato 2021) as well as verb phrases (Bouzouita and Casanova 2018; Casanova 2021) where innovative functions of possessive pronouns are firstly recruited to the 1st person and subsequently spread to the 2nd and 3rd person. The same pattern has been observed for other phenomena in Romance languages, for example in the case of subject pronoun expression in Brazilian Portuguese and Italian (de Oliveira 2000: 39; Detges 2006).⁶

⁵ Unfortunately, Santana Marrero (2014) does not provide the distributions of the prepositional versus the possessive variants for the regions represented in the dataset, but instead only provides aggregate distributions.

⁶ There are, of course, counterexamples of this direction in domains of pronominal hierarchies of Spanish. In that sense, the tendency does not appear to be entirely universal. Yet, at least as far as atonic and, most prominently, tonic possessives are concerned, the pattern 1P > 2P > 3P appears to be recurrent even if opposite cases (3P > 2P > 1P) do exist.

With regards to the grammatical number of the referent, various studies report that singular referents exhibit significantly higher probabilities of being expressed through possessive complements (e.g. *mío* ‘my’, *tuyo* ‘your’, etc.) than plural referents (e.g. *nuestro* ‘our’, *vuestro* plural ‘your’, etc.) (Hoff 2020: 66; Marttinen Larsson and Bouzouita 2018: 29–30). Similar patterns have been documented by other scholars; for example, typological studies of differential object marking (DOM) have found that the diffusion of animacy-driven DOM in Old Russian began in the singular and later spread to plurals (Witzlack-Makarevich and Seržant 2018: 7). Furthermore, Taylor (1991, 1996: 232) reports on a preference for singular possessors over plurals in corpus data of English noun phrases, suggesting that this finding relates quite naturally to cognitive motivations because “a single identifiable entity is likely to be better suited to a reference point function than a plurality of entities” (Taylor 1996: 232).

Lastly, concerning the influence of the locative, earlier studies found that there is significant variation in (dis)preferences of the possessive complements between the different locative adverbs. Univariate analyses dealing with European Spanish data have observed that *delante* (‘in front [of]’), *detrás* (‘behind’), *al lado* (‘next [to]’), *enfrente* (‘in front [of]’, ‘facing’) and *alrededor* (‘around’, ‘surrounding’) combine more frequently with possessive complements (Salgado and Bouzouita 2017; Santana Marrero 2014). While *delante*, *detrás* and *enfrente* are locatives that employ purely adverbial functions, *alrededor* and *al lado* have nominal bases and possess nominal syntactic characteristics (RAE and ASALE 2009), which would explain their more frequent combination with possessives (Marttinen Larsson and Bouzouita 2018: 8; Octavio de Toledo y Huerta 2016: 114, 217–219).

Even though recent studies (Hoff 2020: 67) have found an influence of other factors such as coordination, priming (weak, strong and anti-priming) and animacy, among others, these effects have mainly been marginal and with large standard errors (likely due to their overall infrequency in the data set). While the indications of these predictors’ effects are intriguing and warrant further research,⁷ especially in the case of animacy in 3rd person referents, the present study includes the factors that most recurrently have been found to influence the variation synchronically in order to determine these factors’ courses of actualisation in diachrony (namely grammatical person and number, and locative). These factors do indeed warrant further diachronic examination most urgently and exclusively. Nonetheless, it is encouraged that future research replicate the analysis brought

⁷ In addition, in analysing three varieties simultaneously (Spanish from Buenos Aires, Mexico City, and Madrid), Hoff (2020) does not compute these dialects as an interaction effect with the linguistic predictors. This procedure does not allow us to assess to what extent the constraint set-up differs between the studied speech communities.

forward by Hoff (2020) with other data sets. The following section discusses the data sources used as well as the coding of the different predictors included in the analysis.

4 Data and circumscription of variable context

The present study used data stemming from the RAE corpora *Corpus diacrónico del español* (CORDE) and *Corpus de referencia del español actual* (CREA).⁸ The 250-million word CORDE consists of written data dating from the first written documentation of Spanish until 1974, whereas the CREA corpus contains written and (to a much lesser extent) oral data from 1975 until 2004, consisting of 125 million words. The data were extracted through the corpora's interface. The corpora are almost exclusively comprised of relatively standardised written language, containing mainly publications of fictional works (prose and verse) and non-fiction (journalistic prose, scientific works, legal documents, religious and historical works, among others).⁹ Thus, colloquial language use is not particularly present in the corpora.¹⁰

Considering that noticeable competition between the grammatical variants started at the beginning of the 20th century (Marttinen Larsson and Álvarez López 2017), and before then the possessive construction was found in a latent state, the variable context was diachronically circumscribed to involve searches centred on 1900 and onwards.

Because the corpora are not syntactically tagged, each search string consisted of a specific adverb + complement, both the prepositional ones and possessives,

8 As previously noted by Marttinen Larsson and Bouzouita (2018: 2), the linguistic variable under study is rather infrequent in discourse and, as such, require large corpora for the data compilation. This is one of the reasons for which written data, instead of oral, has been used. Note also that Salgado and Bouzouita's (2017) paper analysed the linguistic variable in question by consulting 21 oral corpora from Spain. They obtained 291 occurrences of the variable out of which a mere 87 were of the possessive variant.

9 While the CORDE and CREA corpora offer metadata very broadly characterising the genre of the text (fiction vs. non-fiction, and so on; see Section 4), this variable has been set aside in the analysis. Indeed, the issue of register and genre variation is acutely relevant, but needs proper operationalisation and great scrutiny in order to be sensible. As Kabatek (2005: 172–173) puts it, categorisations such as those used by the CORDE and CREA are insufficient since they are so general that they within themselves exhibit significantly diverse forms of expression, hence intra-register variation.

10 The CREA corpus consists of 10% transcribed oral data. However, no occurrences of the linguistic variable were found in this part of the corpus. Hence, the compiled data all consists of written language.

accounting for all variant forms. The occurrences were manually extracted with the associated context (KWIC), consisting of 100 characters (including spaces), along with the available metadata (year, author, document title, country,¹¹ theme of document and place of publication).

4.1 Annotation

The obtained observations were coded for a series of predictors, which are outlined in the following paragraphs.

4.1.1 Complement

Binary dependent variable with two levels: ‘prepositional’ (e.g. *delante de mí*) versus ‘possessive’ (*delante mío*).

4.1.2 Grammatical person

The search strings included the possessives *mío* (1.SG.MASC), *mía* (1.SG.FEM), *tuyo* (2.SG.MASC), *tuya* (2.SG.FEM), *suyo* (3.SG/PL.MASC), *suya* (3.SG/PL.FEM), *nuestro* (1.PL.MASC), *nuestra* (1.PL.FEM), *vuestro* (2.PL.MASC) and *vuestra* (2.PL.FEM). The prepositional searches consisted of *de* + the respective prepositional pronouns, namely *mí* (1.SG), *ti* (2.SG), *usted* (2.SG), *él* (3.SG), *ella* (3.SG), *nosotros* (1.PL.MASC), *nosotras* (1.PL.FEM), *vosotros* (2.PL.MASC), *vosotras* (2.PL.FEM), *ustedes* (2.PL), *ellos* (3.PL.MASC) and *ellas* (3.PL.FEM).

This factor was coded as a categorical predictor with three levels – 1st person, 2nd person and 3rd person.

4.1.3 Grammatical number

This factor included singular versus plural referents. Because the 3rd person pronouns *suyo* and *suya* (3.SG/PL) are ambiguous in terms of number, all the occurrences of these possessives ($N = 449$) were manually inspected in order to discern the number of the referent.

¹¹ In this study, only Spain is of interest.

4.1.4 Locative adverb

The following locative adverbs and adverbial locutions were included in the search strings:

- *al lado* ('next [to]')
- *alrededor, rededor, en torno*¹² ('around', 'surrounding')
- *arriba, encima* ('above', 'on top [of]')
- *debajo, bajo, abajo* ('below', 'underneath')
- *delante, adelante* ('in front [of]')
- *detrás, atrás, tras* ('behind')
- *dentro, adentro* ('inside')
- *fuera, afuera* ('outside')
- *cerca* ('close [to]')
- *en medio* ('in the middle [of]', 'in the centre [of]')
- *enfrente, frente* ('in front [of]', 'facing')
- *lejos* ('far [from]')

In the final analysis, however, the locatives *abajo, adelante, adentro, afuera, arriba* and *atrás* were determined not to participate in the variable context because they had zero or extremely low counts (yielding in total only 16 occurrences). The locative *bajo*¹³ was also excluded from the analysis due to low frequency ($N = 14$).

4.1.5 Type of locative

The locatives outlined above were classified into two groups – nominals versus adverbials. The nominal group included locatives that are comprised by a noun and thus classified as nominals, namely *al lado* ('by the side [of]'), *alrededor* and *en torno* ('in the surroundings [of]'), (*al*) *frente*¹⁴ ('in (the) front of') and *en medio* ('in the centre/middle [of]'). The rest of the adverbs were classified as 'adverbials'. As discussed above, it is expected that nominal adverbs have an initially higher probability of combining with possessive complements than adverbials (see, e.g., Marttinen Larsson and Bouzouita 2018: 8; Octavio de Toledo y Huerta

¹² Note that some locatives, such as *en torno*, might be followed by two different types of prepositions, namely *de* and *a* (e.g. *en torno de mí* vs. *en torno a mí*), without denoting differences in significance. Both forms were included in the searches when applicable.

¹³ Actually, *bajo* as well as *tras* are prepositions. However, they also act as variants to the adverbs *debajo* and *detrás*, respectively, and were included in the search strings.

¹⁴ While *enfrente* is an adverbial locative, the variant (*al*) *frente* is an adverbial nominal locution that exhibits similar syntactic behavior to *al lado*.

2016: 114, 217–219), thus these might have diachronically opened up the way for adverbials to combine with possessives.

4.1.6 Time (period)

In an attempt to achieve the finest level of granularity, ‘year’ was first modelled as a continuous predictor (1900–2004) and then recoded into a centred factor. However, due to data scarcity for some years (especially for the incipient possessive variant), a slightly less fine-grained option was opted for in which ‘year’ was recoded into the continuous factor ‘period’, converting ‘year’ into chunks of 5 years (1900–1904, 1905–1909, etc.) which was then treated as a numerical variable. By doing this, the issue of data scarcity was alleviated while still maintaining a highly fine-grained diachronic continuous component in the analysis.

4.1.7 Author

In order to account for confounding effects of idiosyncratic variability, ‘author’ was included as a random effect in the analysis (see Tagliamonte and Baayen 2012). The compiled observations were produced by a total of 800 authors according to the corpus metadata (approximately 13.8 observations per author).

4.2 Data discussion and statistical treatment

A total of 11,019 occurrences were used for statistical analysis. Out of these, 759 observations (6.9%) were of the possessive variant. As Tagliamonte (2006: 82–84) highlights, incipient linguistic variables that exhibit minimal variation are not particularly viable for variationist analysis because of difficulties in reliable statistical modelling. However, this is mainly the case when employing techniques such as those offered by VARBRUL that only use fixed-effects regression analysis. Without the inclusion of relevant by-speaker and by-item random intercepts, unbalanced or skewed data will likely overstate the statistical influence of fixed effects. For the present research, a Generalised Linear Mixed-Effects Model (GLMEM) was fitted in a stepwise fashion, and it converged with the data despite the low frequencies of the innovative variant (759 out of 11,019). The model selection process is described in Section 5.1. Considering GLMEMs’ advantages over VARBRUL (see e.g. Gries and Hilpert [2010: 304–305] for an overview), and given that the statistical model converges, it is argued that it is

both statistically and theoretically warranted to model the studied variation despite certain data sparseness.

5 Results

In this section, the model-fitting procedure and the results from the multivariate analysis will be discussed. Firstly, though, a frequency-based distributional analysis is illustrated where the proportions of the competing variants are plotted against real time in order to provide a historical frequency account of the variation (Figure 1).

Generally, an increase in the frequency of an incoming variant is interpreted as a tell-tale sign of (incipient) grammaticalisation (Hopper and Traugott 2003; Krug 2000; Mair 2004). Nonetheless, as Figure 1 shows, the innovative possessive variant exhibits minimal fluctuations in terms of frequency along the diachronic axis. Instead, it remains rather constant throughout the studied timespan.

Contemplating the testified diffusion of the possessive variant (e.g. Carnicer 1967; Hoff 2020; Llorente Maldonado 1980; Marttinen Larsson and Bouzouita 2018;

Figure 1: Condensity plot of the distribution of complements in Spain 1900–2004.

Salgado and Bouzouita 2017; Santana Marrero 2014), the expected development would instead be an increased usage of the possessive variant as time advances. However, considering the type of data contained in the examined corpora, this frequency stability and marginality might not be too surprising: it may simply be the case that the variant under study is inhibited in standardised written language due to its social evaluation and stigmatisation, as described in Section 3. This overall marginal representation of the incoming variant, which remains low and stable throughout the studied time span, would thus reflect a general inhibition of the implementation of the innovation in standardised written language. In effect, as Section 3.2 outlines, the variant has a noticeable presence in colloquial language on, for instance, Twitter (Hoff 2020; Marttinen Larsson and Bouzouita 2018). Thus, there is an apparent contrast in terms of frequency between the two modalities.

Indeed, as Mair (2011) recognises, a crucial aspect hindering the diachronic analysis of incipient grammaticalisation phenomena from a frequency perspective is that the available sources for the investigation of such incipient variables are typically not very well-suited for the task. This is because incipient grammaticalisation generally does not occur firstly in relatively standardised written sources, but rather in spoken language. What is more, if the studied grammaticalising item has historically been subject to social evaluation, as is frequently the case in instances of sociolinguistic variation and change (Labov 2001: 28–29; Nevalainen and Palander-Collin 2011: 125), the incoming variant's diffusion in written formal registers is likely to be increasingly halted.

Relying on the frequency-based account provided in Figure 1, this might indeed appear to be the case in the studied data. The question remains, though, if this inhibition is limited to the implementation of the variant in terms of frequency, and if a probabilistic analysis of the variation might enable us to shed light onto underlying grammar change in the intralinguistic factors that constrain the variation. Indeed, in light of the probabilistic grammar approach, we should be able to discern *why* language users use the variants they use (Szmrecsanyi 2016: 154).

To sum up, examining the plotted distribution impressionistically, we would retreat to the interim conclusion that no grammar change appears to be going on in the studied data. However, in line with probabilistic approaches to grammar change, we have instead operationalised change not as frequency variability but as function variability – in that, grammar change can be posited if a given linguistic context probabilistically varies as a function of real time, as we shall see in the following analysis (Bresnan and Hay 2008; Gries and Hilpert 2010; Hinrichs and Szmrecsanyi 2007; Szmrecsanyi 2016).

5.1 Mixed-effects modelling

This part of the analysis is concerned with the probabilistic approach to diachronic grammar change. As stated in the previous section, a GLMEM was constructed through a stepwise model-selection process in order to build a model as simple but also as accurate as possible, with only significant factors included. The model was fitted iteratively with the coded predictors, and predictors that were not significant, either by themselves or in an interaction, were removed. The random factors that were included are ‘author’ (to account for idiosyncratic variation; see Tagliamonte and Baayen 2012) and a by-item (i.e. locative adverb) random slope for the effect of ‘period’ (to account for the fact that idiosyncrasies of lexical items might not stay stable across the studied century; see Baayen and Bates 2008; Westfall et al. 2014), and their statistical significance was assessed by means of a likelihood-ratio test between models including and excluding them (cf. Baayen 2008: 253).

The final minimal model has the following characteristics, and constitutes an overall close to perfect fit according to established criteria with a very high prediction accuracy:

- C index of concordance: 0.97
- Somer’s D_{xy} : 0.95
- Marginal GLMM² (variance explained by fixed effects): 24.9%
- Conditional GLMM² (variance explained by the entire model): 81%
- Classification accuracy: 96%

However, while these metrics serve as useful evaluative tools, their estimates are likely overinflated considering the class imbalance of the data. Indeed, when dealing with incipient linguistic variables such as the one examined in the present study, the sample of the traditional variant will be considerably larger than the sample of the incoming variant. For this reason, the classification metrics used for model evaluation will need to be able to account for this class imbalance. This is an issue that metrics such as prediction accuracy, C index and other metrics used for binary classification (like F1 scores) do not consider since they fail to adequately estimate the model’s classification capacity for the minority class, i.e. the incoming variant. This means that, while accuracy and F1 scores are sound options for evaluating models based on balanced datasets, they will provide unreliable and overly optimistic metrics for unbalanced data (Chicco and Jurman 2020: 11). A metric that circumvents this type of class imbalance is the Matthews correlation coefficient (MCC). This measurement considers not only the prediction accuracy of the majority class (like F1 and prediction accuracy do) but

also the minority, and is thus highly apt for classifying unbalanced data. It is based on the *phi* coefficient and yields a value between -1 and $+1$, with 0 indicating no relationship whatsoever (i.e. probability equal to coin flipping), -1 being equivalent to perfect misclassification and $+1$ to perfect classification (Chicco and Jurman 2020: 5). Thus, in difference to accuracy and F1 scores, a high MCC indicates that the prediction not only correctly classified a large proportion of the majority variant (i.e. the prepositional complement) but also a large proportion of the minority variant (i.e. the possessive complement), independently of the degree of class (im)balance (Chicco and Jurman 2020: 8). Using this metric, the obtained result is $MCC = 0.66^{15}$ which constitutes a strong positive relationship. In other words, while the accuracy and *C* index are overly optimistic, the MCC indicates that the model’s prediction is still strong and far more informative than the baseline or random classification.

The best-fitting model included the following factors:

Fixed effects:	Grammatical person	×	Period (continuous)
	Grammatical number ¹⁶		
	Type of locative	×	Period (continuous)
Random effects:	Locative		
	Author		

The full output of the GLMEM is given in Table 1 (in Appendix).¹⁷ In order to facilitate the interpretation of the output, the results will be visualised in the following sections.

5.2 The influence of the grammatical person on the referent

In Figure 2, the interaction between ‘grammatical person’ $\times$ ‘period’ is plotted using the *interactions* package in R (Long 2019). On the y-axis, the average predicted probabilities are shown for the possessive complement. ‘Grammatical person’ is illustrated in form of differently coloured regression lines, evolving as a function of real time.

¹⁵ True positives = 424, false positives = 98, true negatives = 10,162, false negatives = 335. The MCC calculation was done using the R package *mltools* (Gorman 2018).

¹⁶ The reason for which ‘grammatical number’ is not computed as an interaction effect is discussed in Section 5.3.

¹⁷ The dataset and coding scheme, and the R code used to obtain the final regression model, the plots and the model evaluation metrics are available on DOI 10.17605/OSF.IO/HX2BD.

Figure 2: Predicted interaction between ‘grammatical person’ × ‘period’ with 95% confidence intervals.

As Figure 2 shows, there is functional variability along the diachronic axis ($p = 0.007$; see Table 1 in Appendix for full results), and overall, during the vast majority of the studied timespan, the 1st person (blue solid line) has the highest probability of being expressed through a possessive complement. As revealed by the model found in Table 1 (Appendix), this interaction is very significant. The 2nd and 3rd person have a later onset and experience a probabilistic productivity increase after approximately 1925, and by around 1990 the 2nd person closes in on the 1st in terms of productivity, with the 3rd person also continuing to evolve rather steadily in a probabilistically possessive-favouring direction. In other words, it appears to be the case that the probabilistic increase of the use of possessive complements first found its grip with 1st person referents, and in the last couple of decades there has been a rapid spread to the 2nd contexts. Only subsequently, and moderately, there has been an extension to involve 3rd person referents.

Interestingly, the probabilistic patterning in the data accurately follows the discourse participation hierarchy (1st > 2nd > 3rd; Givón 1994; Silverstein 1976),

thus providing a blueprint of the increasing degree of discourse orientation¹⁸ (Narrog and Heine 2021: 92–93) during grammar change in progress. Synchronically, the results are consistent with findings reported by other scholars in that the 1st and 2nd person are most frequently found in the possessive form, more so than the 3rd person (Hoff 2020; Marttinen Larsson and Bouzouita 2018; Salgado and Bouzouita 2017). Nonetheless, what this examination has allowed us to conclude is the course of actualisation, which is most suitably determined by means of a diachronic examination.

5.3 The influence of the grammatical number on the referent

During the model selection process, ‘grammatical number’ was first computed as an interaction effect with ‘period’. However, ‘grammatical number’ was only significant as the main effect, i.e. it was constant throughout the timespan. More specifically, the regression model (Table 1) shows that singular referents significantly favour the possessive complement ($p < 0.0001$). The non-significance of the interaction indicates that the rate of implementation of the innovation is constant in all linguistic contexts. Considering this paper’s interest in the trajectory of grammaticalisation that the tonic possessives have undergone in different linguistic contexts, it is still informative, though, to visualise the effect as a function of real time in order to map out this (constant) spread of the innovation across linguistic contexts. For this reason, the predicted probabilities of the different grammatical numbers are plotted as a function of diachrony (Figure 3). It should be kept in mind though that said interaction is not significant. What Figure 3 shows is that, during the tonic possessive’s course of actualisation, singular contexts favour the possessive the most and have done so constantly throughout the studied timespan. Plural referents favour the possessive variant significantly less. What the graph appears to indicate, too, is that the effect of the grammatical number has in recent years neutralised almost entirely, which can be taken as a further indication of the increasing degree of grammaticalisation of the incoming variant.

18 This concept refers to that, during processes of grammaticalisation, “meanings tend to become increasingly oriented towards the speech act participants, that is, speaker and hearer, and towards organising speech of discourse itself” (Narrog and Heine 2021: 92–93). Such process subsumes three semantic changes, which involves i) an increasing orientation towards the speaker’s perspective (in this case, 1st person possessives), ii) an increasing orientation towards the hearer (2nd person possessives), and iii) an increasing orientation towards the discourse itself (3rd person possessives). Note that Narrog and Heine (2021: 325) do not claim that the order between these three tendencies is established yet, even though they suggest that discourse orientation is a later stage phenomenon.

Figure 3: Predicted probabilities of the possessive complement by 'grammatical number' $\times$ 'period' with 95% confidence intervals.

Considering these results, we are not able to determine that the grammatical number varies significantly as a function of real time; instead, similarly to the reports of typological studies and other scholars (see Section 3.3), the grammatical number is diachronically stable in conditioning the variation where singular referents probabilistically favour the possessive variant.

5.4 The influence of the type of locative adverb

In what follows, the influence of the type of locative on complement variation will be scrutinised.

As the predicted probabilities in Figure 4 suggest, the nominal group initially had a rather high probability of combining with possessive complements. In contrast, at the beginning of the 1900s, the adverbial group had a categorical probability of being expressed with prepositional complements. As expected,

there is a very slight increase in the probability of possessive complements with the adverbial group along the diachronic axis. In other words, it appears to be the case that the nominal-based locatives and their combination with possessives have enabled the use of possessives with adverbial-based locatives, leading to further actualisation. Nonetheless, and rather surprisingly, the probability of the nominal group to combine with possessive complements decreases dramatically. That is, as the probability of finding ‘adverbial base + possessive’ increases as a function of real time, the probability of finding ‘nominal base + possessive’ decreases, until a point in which the two groups synchronically appear to have a close to equal probability of combining with possessives (cf. 2000 and onwards in Figure 4’s x-axis). The reason behind this pattern might be that nominal locatives act as a type of booster that, over time, loses momentum and the constraint becomes less important. It is not entirely clear why there is a decreasing relevance of the constraint, but it may possibly be due to a growing stabilisation of the incoming variant across adverbial contexts. Another possibility is hyper-correction: as an increasing metalinguistic awareness is growing concerning the ‘incorrect’ status of ‘adverbial base + possessive’, this ripples (erroneously) onto nominal locative constructions with speakers perceiving all types of instances of ‘locative + possessive’ as grammatically incorrect.

Figure 4: Predicted interaction between type of locative × period with 95% confidence intervals.

6 Discussion

The present paper has provided a diachronic probabilistic grammar account of constraints on an incipient case of morphosyntactic variation in Spanish locative adverbial constructions, where prepositional complements vary with innovative possessive complements (e.g. *delante de mí* vs. *delante mío*). Departing from the premise that grammar change can be legitimately posited if probabilistic weights of conditioning language-internal factors alter as a function of real time (Bresnan and Hay 2008; Gries and Hilpert 2010; Szmrecsanyi 2016), the innovative possessive constructions' spread and functioning in European Spanish written corpora from 1900 and onwards has been scrutinised. Contrasting this approach with one that deals with mere frequency counts of competing variants, it was argued that a probabilistic account allows us to identify underlying abstract constraints in instances of incipient grammar change when frequencies fail to tell the full picture concerning the innovation's possible trajectory of grammaticalisation. Indeed, from a frequency perspective, incipient grammar change is generally defined as a stage of variation between alternating variants where the incoming form has a proportion of below 15% (Nevalainen and Raumolin-Brunberg 2017: 54–55). Therefore, due to an overall incipience, marginality and low degree of frequency of the incoming variant, instances of grammaticalisation processes with low-frequency innovations are oftentimes neglected in the study of grammar change, and are perceived as constituting problematic and difficult cases for empirical examination (Mair 2004: 127, 2011: 241). What the present study has shown, however, is that a focus on the underlying probabilistic constraints might aid us in determining the functioning of the respective variants in competition and how probabilistic community grammars evolve in real time, even when frequencies do not appear to reveal any such change.

Having characterised incipient grammar change in the above paragraph, it needs to be emphasised that the concept of 'low-frequency grammaticalisation' is rather ambiguous and requires clarification. On the one hand, it can be taken to refer to grammaticalisation processes where the overall occurrence of the variable is notably low, such as in the case of the infrequent English complex prepositions (e.g. *by dint of*, *on turnover of* and similar P + N + P structures) examined by Hoffmann (2004). In these cases, the question becomes how and to what extent such infrequent constructions manage to grammaticalise at all, given that frequency of use is generally assigned a pivotal role in grammaticalisation (see, for instance, the works by Bybee and colleagues). On the other hand, it can refer to the notion of a determined variable being present in a corpus, but the incoming variant having a low frequency of use since it constitutes an incipient change in the studied discursive context. This distinction is crucial for the argumentation of this

paper since I do not pretend to lay direct claims to the nature of frequency effects during processes of language variation and change. Instead, what is of concern here are the contexts in which the morphosyntactic variable is present in everyday language use, but where the incoming variant may constitute an incipient grammaticalising item in certain discursive contexts, leading it to have a low frequency of use in those particular types of discourses. This may, for instance, be the case in discursive contexts where the form entails a significant degree of sociolinguistic salience (Levon and Buchstaller 2015) such as relatively standardised written language (as opposed to colloquial vernaculars). In these circumstances, the mere low frequency of occurrence of an incoming variant states little about its frequency of use in other contexts and, by extension, its cognitive and probabilistic conditioning in grammar. Indeed, as Hoffmann (2004: 197) rightly asserts, “the frequency with which a particular linguistic feature is found in a corpus may, in fact, be quite different from the actual frequency with which an average language user is exposed to it in his or her daily language use”. Importantly, as sociolinguistic studies and corpus research on socially salient variables have demonstrated, language users are sensitive to negatively evaluated surface manifestations of forms, and are thus not as prone to employ such variants in standardised written language as they might be in other styles and contexts. In the context of the present argumentation, this means that incipient changes typically do not originate in written formal registers (Mair 2011), and are even less likely to do so if the incoming variant is subject to negative social evaluation (Labov 2001: 28–29; Nevalainen and Palander-Collin 2011: 125). This leads us to the burning question of why we are able to identify differences in probabilistic conditioning in spite of low and stable frequencies of the incoming variant. The present paper argues that this finding might be interpreted in light of Labov’s Interface Principle (Labov 1993) which states that “Members of the speech community evaluate the surface forms of language but not more abstract structural features. More specifically, social evaluation bears upon the allophones and lexical stems of the language, but not upon phonemic contrasts, rule ordering or the direction or order of variable constraints”. As the principle maintains, members of the community are insensitive to more abstract structural features and the probabilistic conditioning of morphosyntactic variants. It is because of this insensitivity to abstract constraints that we are able to identify changing probabilistic constraints (which reflect the evolving linguistic structure of the community grammar) as a process that is separate from socially informed surface frequencies in corpus data (cf. Levon and Buchstaller 2015: 322). Thus, the incoming variant firstly finds its grip in less standardised contexts where speakers are exposed to the incoming variant in everyday language use. As the variant grammaticalises, speakers’ abstract probabilistic grammars are continuously configured. It is this underlying, abstract

and evolving conditioning that we are able to identify, whereas the variant's low frequency of use reflects surface-level sociolinguistic monitoring (Labov et al. 2011; Levon and Buchstaller 2015).

Indeed, even if the innovative variant appears to be inhibited in relatively standardised written sources, this inhibition is only reflected in its frequency of use; from a probabilistic perspective, the results indicate that several language-internal constraints have been and are active in conditioning the studied variation in spite of apparent stability in terms of frequency. The innovative variant has spread from one linguistic context to another at a significant rate and the present study has confirmed what synchronically oriented investigations have identified in terms of language-internal conditioning factors and their effects (Hoff 2020; Marttinen Larsson and Bouzouita 2018; Salgado and Bouzouita 2017). In view of this convergence in linguistic conditioning, the results show that the diffusion of the innovative possessive constructions in standardised written Spanish has evolved along the same trajectory of grammaticalisation as it has in colloquial vernaculars. In that, the methodological approach adopted in the paper has illustrated that the same forces operate in low- and high-frequency contexts, and parallel grammaticalisation pathways are identifiable.

The present study has some limitations that need to be made explicit. Firstly, in spite of statistical significance, the effect sizes are mostly in the weak range due to the very incipience of the studied variant. Because of this, the results should be interpreted with a certain caution. As indicated, though, they are in line with evidence from synchronic studies, but future studies are urged to replicate the analysis using other diachronic data in order to verify the strength and direction of the examined predictors on the linguistic variable. Secondly, another limitation lies in the inability to account for the influence of extralinguistic factors other than social evaluation that might be at play in yielding the observed patterns. In effect, lacking from the present account is an examination of the effect of factors that, given the used data, are not operationalizable satisfactorily. The present study has, on the basis of a comparison between the results yielded from the analysed data and the results brought forward by earlier studies on the variant's use in colloquial language, subscribed to the idea that the examined linguistic variable is subject to register and genre variation. The question becomes to what extent standardised written language in itself exhibits variability in between registers. As an increasing body of research shows, register variation constitutes one of the most influential factors in conditioning the spread of a determining feature; however, given the complexity of operationalising 'register' and 'genre' sensibly, one needs to take into account contextual cues, social dynamics and textual and discursive

traditions that are, unfortunately, beyond the scope of this paper.¹⁹ Consider, though, that the random factor ‘author’ contained 800 levels. If we make the crude assumption that authors are more or less consistent in their genre of choice (that is, a fiction author mostly tends to write texts of that genre as opposed to, for instance, non-fictional legal or medical documents), one could argue that the random effect of ‘author’ acts as a proxy for the confounding effect of the genre in the regression analysis. It goes without saying that this is not to neglect that individuals and authors exhibit register and even genre variation, but the idiosyncrasies that this random predictor deals with might be viewed as a way of accounting for some impact that, potentially, the genre may have on the analysed variability. Nonetheless, these approximations are only tentative; future research is urged to consider a qualitative examination of the studied process of change in a more metadata-rich corpus using data from a large span of registers and genres in order to allow for a more detailed assessment of individual social and textual dynamics on this case of variation.

7 Conclusions

From a frequency perspective, we are evidently dealing with a case of linguistic change in progress that is incipient in relatively normative written language. In general, the odds are still in favour of the prepositional complement in terms of overall usage. Yet, what the current study has aimed to show are the abstract constraints that have fuelled the variation in the history of these two constructions’ competition and how these relate to the probabilistic diagnosis of grammar change. Considering the influence of language-internal factors on the studied phenomenon of ongoing language change, we can discern that we are dealing with a rather incipient phenomenon of variation that, nonetheless, is increasing in probabilistic productivity. This is evidenced by the way in which an increasing number of linguistic contexts are involving the possessive variant. The diachronic examination brought forward allows us to determine that 1st person and nominal adverbials act as the instigators of change, and only subsequently have language users started to generalise such usage to other linguistic contexts and at unequal

19 The open data published along with this paper (see Section 5.1) contains metadata concerning the textual theme and author of the respective data points. Using this dataset, future studies could be conducted as expansions of the present paper by examining the possible influence of some of the extralinguistic predictors discussed above. However, in the context of the present analysis, the modelling of textual, discursive and sociolinguistic predictors unfortunately remains outside the scope of the study.

rates. The effect of the grammatical number is influential but remains stable over time. Thus, in line with our operationalisation of probabilistic grammar change (cf. Szmrecsanyi 2016), we can discern that actual grammar change has taken place, albeit with weak effects which is not at odds given the incipience of the variation.

In spite of low and stable frequencies of the incoming variant in the studied corpora, incrementation in language use leads to the internal adjustments of the variable’s constraints (Labov 2007). This is evidenced by the way in which the increasing weakening of cognitive constraints as well as the widening scope of the innovation is transferred by each generation (Labov 1989, 1994: Ch. 14), independently of normative inhibition on use. In all, the present study has aimed at showcasing how even in cases of highly incipient innovations with apparent stability in diachrony, the probabilistic operationalisation of grammar variation might allow us to gain a better understanding of the full trajectory of language change and not only the intermediate stages that are most frequently subject to empirical scrutiny. In expanding the temporal scope to include the incipient tail of the curve, probabilistic analyses such as the one brought forward may be able to identify what factors operate at the genesis of variation and, potentially, approximate the modelling of the actuation of change (cf. Weinreich et al. 1968: 102).

Appendix

Table 1: Mixed-effects regression model.

Variable	Level	N	Estimate	S.E.	z	p-Value
Dependent variable: prepositional complement (N = 10,260) versus possessive complement (N = 759)						
(Intercept)			−8.53979	1.79367	−4.761	<0.0001***
Gram_Person	1st	3,050	Reference level			
	2nd	1,122	−2.23137	0.61841	−3.608	<0.001***
	3rd	6,847	−2.74642	0.39582	−6.939	<0.0001***
Gram_Number	PL	2,726	Reference level			
	SG	8,293	0.68063	0.14616	4.657	<0.0001***
Loc_Type	Adverbial	9,295	Reference level			
	Nominal	1,724	6.71872	2.52263	2.663	<0.008**
Period	Numerical	Mean	0.35224	0.16032	2.197	0.0280*
		1963				

Table 1: (continued)

Variable	Level	N	Estimate	S.E.	z	p-Value
Gram_Person × period	1st × period	N/A	Reference level			
	2nd × period	N/A	0.22218	0.08247	2.694	0.00706**
	3rd × period	N/A	0.15937	0.05124	3.110	0.00187**
Loc_Type × period	Adverbial × period	N/A	Reference level			
	Nominal × period	N/A	−0.45681	0.19280	−2.369	0.01782*
Random effects:						
Author	800 levels					
Locative	15 levels					
Model evaluation	AIC: 2,862.9					
	BIC: 2,957.9					
	C index: 0.97					
	MCC: 0.66					

References

Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, Harald & Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Bouzouita, Miriam & Vanessa Casanova. 2018. *¿Vos gustás mío? Yo gusto tuyo*. The diatopic distribution of the use of possessive verbal complements in Latin American Spanish. Paper presented at the Conference possessive constructions in Romance (POSSROM). University of Gent.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin: Mouton de Gruyter.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245–259.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.

Bybee, Joan & Paul Hopper. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Carnicer, Ramón. (1967, March 9th). *Falsos posesivos*. La Vanguardia Española 56.

Casanova, Vanessa. 2021. El uso del complemento posesivo verbal por el complemento de régimen preposicional en español actual. In Miriam Bouzouita & Matti Marttinen Larsson (eds.), *Possessive constructions in Romance*. [Special issue]. *Moderna språk*, vol. 114(3), 264–301.

- Chicco, Davide & Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(6). 1–13.
- Company Company, Concepción. 2017. El posesivo átono con artículo definido y con artículo indefinido. Similitudes y diferencias. In Concepción Company Company & Norohella Huerta Flores (eds.), *La posesión en la lengua española*, 133–175. Madrid: Consejo Superior de Investigaciones Científicas.
- De Smet, Hendrik. 2012. The course of actualization. *Language* 88(3). 601–633.
- Detges, Ulrich. 2006. From speaker to subject. The obligatorization of the Old French subject pronouns. In Hanne Leth Andersen, Merete Birkelund & Maj-Britt Mosegaard Hansen (eds.), *La Linguistique au Coeur. Velance verbal, grammaticalisation et corpus. Mélanges offerts à Lene Schøsler à l'occasion de son 60e anniversaire*, 75–103. Odense: University Press of Southern Denmark.
- Givón, Talmy. 1994. *Voice and inversion*. Amsterdam/Philadelphia: John Benjamins.
- Gorman, Ben. 2018. mltools: Machine learning tools. R package version 0.3.5. Available at: <https://cran.r-project.org/web/packages/mltools/mltools.pdf>.
- Grafmiller, Jason, Benedikt Szmrecsanyi, Melanie Röthlisberger & Benedikt Heller. 2018. General introduction: A comparative perspective on probabilistic variation in grammar. *Glossa: A Journal of General Linguistics* 3(1). 94.
- Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language & Linguistics* 14(3). 293–320.
- Harris, Alice C. & Lyle Campbell. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.
- Hilpert, Martin & Christian Mair. 2015. Grammatical change. In Biber Douglas & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics*, 180–200. Cambridge: Cambridge University Press.
- Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language & Linguistics* 11(3). 437–474.
- Hoff, Mark. 2020. *Cerca mío/a or cerca de mí?* A variationist analysis of Spanish locative + possessive on Twitter. *Studies in Hispanic and Lusophone Linguistics* 13(1). 51–78.
- Hoffmann, Sebastian. 2004. Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In Hans Lundquist & Christian Mair (eds.), *Corpus approaches to grammaticalization in English*, 171–210. Amsterdam: John Benjamins.
- Hopper, Paul J. & Elisabeth Closs Traugott. 2003. *Grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.
- Kabatek, Johannes. 2005. Tradiciones discursivas y cambio lingüístico. *Lexis* 29(2). 151–177.
- Krug, Manfred G. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: Mouton de Gruyter.
- Labov, William. 1989. The child as linguistic historian. *Language Variation and Change* 1(1). 85–97.
- Labov, William. 1993. The unobservability of structure and its linguistic consequences. Paper presented at the 22nd new ways of analyzing variation (NWAV) conference. University of Ottawa.
- Labov, William. 1994. *Principles of language change: Internal factors*. Oxford: Blackwell.

- Labov, William. 2001. *Principles of language change: Social factors*. Oxford: Blackwell.
- Labov, William. 2007. Transmission and diffusion. *Language* 83(2). 344–387.
- Labov, William, Sharon Ash, Maya Ravindranath, Tracey Weldon, Maciej Baranowski & Naomi Nagy. 2011. Properties of the sociolinguistic monitor. *Journal of Sociolinguistics* 15(4). 431–463.
- Levon, Erez & Isabelle Buchstaller. 2015. Perception, cognition, and linguistic structure: The effect of linguistic modularity and cognitive style on sociolinguistic processing. *Language Variation and Change* 27. 319–348.
- Llorente Maldonado, Antonio de Guevara. 1980. Consideraciones sobre el español actual. *Anuario de letras* 18. 5–61.
- Llorente Maldonado, Antonio de Guevara. 1986. *El lenguaje estándar español y sus variantes*. Salamanca: Universidad de Salamanca.
- Long, Jacob A. 2019. Interactions: Comprehensive, user-friendly toolkit for probing interactions. R package version 1.1.0. Available at: <https://cran.r-project.org/web/packages/interactions/interactions.pdf>.
- Lorenz, David. 2012. *Semi-modal constructions in English: Emancipation through frequency*. Freiburg: University of Freiburg PhD thesis.
- Mair, Christian. 2004. Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In Hans Lundquist & Christian Mair (eds.), *Corpus approaches to grammaticalization in English*, 121–150. Amsterdam: John Benjamins.
- Mair, Christian. 2011. Grammaticalization and corpus linguistics. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of grammaticalization*, 239–250. Oxford: Oxford University Press.
- Manning, Christopher D. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 289–341. Cambridge, MA: MIT Press.
- Marttinen Larsson, Matti & Laura Álvarez López. 2017. ‘Delante suyo’ vs. ‘Delante de él’: El uso de las locuciones adverbiales locativas desde una perspectiva diacrónica y diatópica. *Signo y Seña* 31(1). 85–104.
- Marttinen Larsson, Matti & Miriam Bouzouita. 2018. *Encima de mí* vs. *encima mío*: un análisis variacionista de las construcciones adverbiales locativas con complementos preposicionales y posesivos en Twitter. *Moderna Språk* 112(1). 1–39.
- Marttinen Larsson, Matti & Miriam Bouzouita. Feminine morphology in possessive complements of adverbial constructions in Andalusian varieties, under evaluation.
- Santana Marrero, Juana. 2014. La estructura adverbio + posesivo en medios de comunicación digitales. *Español actual* 101. 7–30.
- Narrog, Heiko & Bernd Heine. 2021. *Grammaticalization*. Oxford: Oxford University Press.
- Nevalainen, Terttu & Minna Palander-Collin. 2011. Grammaticalization and sociolinguistics. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of grammaticalization*, 118–129. Oxford: Oxford University Press.
- Nevalainen, Terttu & Helena Raumolin-Brunberg. 2017. *Historical sociolinguistics: Language change in Tudor and Stuart England*. Abingdon: Routledge.
- Octavio de Toledo y Huerta, Álvaro. 2016. *Los relacionantes locativos en la historia del español*. Berlin: De Gruyter Mouton.
- de Oliveira, Marilza. 2000. The pronominal subject in Italian and Brazilian Portuguese. In Mary Aizawa Kato & Esmeralda Vailati Negrão (eds.), *Brazilian Portuguese and the null subject parameter*, 37–53. Frankfurt/Madrid: Vervuert Iberoamericana.

- Pato, Enrique. 2021. Posesivos pleonásticos, redundancia y énfasis. De nuevo sobre la construcción *una mi amiga* en las variedades mexicano-centroamericanas. In Miriam Bouzouita & Matti Marttinen Larsson (eds.), *Possessive constructions in Romance*. [Special issue]. *Moderna språk*, vol. 114(3), 141–160.
- Real Academia Española/Asociación de Academias de la Lengua Española. 2009. *Nueva gramática de la lengua española*, vol. I–II. Madrid: Santillana.
- Real Academia Española/Asociación de Academias de la Lengua Española. Banco de datos (CORDE) [online] Corpus diacrónico del español. Available at: <http://www.rae.es>.
- Real Academia Española/Asociación de Academias de la Lengua Española. Banco de datos (CREA) [online] Corpus de referencia del español actual. Available at: <http://www.rae.es>.
- Salgado, Hugo & Miriam Bouzouita. 2017. El uso de las construcciones adverbiales locativas con pronombre posesivo en el español peninsular: un primer acercamiento diatópico. *Zeitschrift für Romanische Philologie* 133(3). 766–794.
- Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Robert M. W. Dixon (ed.), *Grammatical categories in Australian languages*, 112–171. Atlantic Highlands: Humanities Press.
- Szmrecsanyi, Benedikt. 2016. About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory* 12(1). 153–171.
- Szmrecsanyi, Benedikt. 2017. Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. *The Canadian Journal of Linguistics/La revue Canadienne de linguistique* 62(4). 685–701.
- Tagliamonte, Sali. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
- Tagliamonte, Sali & Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
- Taylor, John R. 1991. Possessive genitives in English: A discourse perspective. *South African Journal of Linguistics* 9. 59–63.
- Taylor, John R. 1996. *Possessives in English: An exploration in cognitive grammar*. Oxford: Clarendon Press.
- Weinreich, Uriel, William Labov & Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics: A symposium*, 95–195. Austin: University of Texas Press.
- Westfall, Jacob, David A. Kenny & Charles M. Judd. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General* 143(5). 2020–2045.
- Witzlack-Makarevich, Alena & Ilja A. Seržant. 2018. Differential argument marking: Patterns of variation. In Alena Witzlack-Makarevich & Ilja A. Seržant (eds.), *Diachrony of differential argument marking*, 1–40. Berlin: Language Science Press.