Language Structure, Variation and Change

The Case of Old Spanish Syntax

Ian E. Mackenzie

Language Structure, Variation and Change

"In his latest book, Mackenzie marries grammatical theory and quantitative methods to make significant advances in the analysis of a range of syntactic phenomena in Old Spanish. Alongside meticulous corpus-based empirical work, he critically engages with recent concepts in the theory of quantitative historical linguistics to great effect. This book is highly recommended to anyone interested in the history of the Spanish language and indeed in language change more broadly."

—Richard Waltereit, Professor of Romance Linguistics, *Humboldt University Berlin, Germany*

Ian E. Mackenzie

Language Structure, Variation and Change

The Case of Old Spanish Syntax

palgrave
macmillan

Ian E. Mackenzie
School of Modern Languages
Newcastle University
Newcastle Upon Tyne, UK

ISBN 978-3-030-10566-2 ISBN 978-3-030-10567-9 (eBook)
https://doi.org/10.1007/978-3-030-10567-9

Library of Congress Control Number: 2019931381

Cover illustration: © Granger Historical Picture Archive / Alamy Stock Photo

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

The work leading to the publication of this book was supported by a generous Research Fellowship (2017–2018) from the Leverhulme Trust.

Contents

List of Figures

List of Tables

1

Preliminary Concepts: Old Spanish, How to Measure the Speed of Change and the Structure of the Corpus

1.1 Old Spanish

1.1.1 Deconstructing the Name

The term 'Old Spanish' is widely used in the syntactic literature and indeed the linguistic literature more generally, as a counterpoint both to 'modern Spanish' and also to names for other varieties of Old Romance, such as 'Old Portuguese', 'Old French' and 'Old Catalan'. The label is a useful one and it occurs frequently in this book, together with the term 'medieval Spanish', with which it can be regarded as being extensionally equivalent. It should nevertheless be noted that for most of the period of its existence, what we now think of as Old Spanish would not actually have been referred to as *español* (or any ancestral form of this word). Indeed, as late as the thirteenth century, usage does not point to the existence of any agreed name for the language. For example, in the prose manuscripts of the Alfonsine corpus, designations range from plain *romanz* or *romançe* through to *lenguage castellano*, *romanz de Castiella* and even the modern-looking *castellano*.[1] To the extent, then, that there was any generalized awareness on the part of speakers that they spoke a

I. E. Mackenzie, *Language Structure, Variation and Change*,
https://doi.org/10.1007/978-3-030-10567-9_1

particular language, the relevant linguistic identity was conceptualized in terms of a Castilian way of speaking or writing. Referring to the language of the medieval kingdom of Castile as 'Spanish' thus involves the retrospective application of a modern nomenclature.

This approach should not necessarily be regarded as being ahistorical, however. For modern diachronic syntactic theory (see e.g. Roberts 2007) envisages language change as being realized in specific areas of the grammar rather than in the language as a whole. Moreover, these localized changes, as externally manifested, are long-term processes, evolving as gradual curves (Kroch 1989) rather than in discrete stages. At any given time, therefore, the different components of a language's grammar will be at different evolutionary moments, implying, logically, that no particular temporal slice in a language's overall history has any preferential claim to a specific identity of its own. Thus the use of a single term for the entire, seamless continuum is actually well motivated, and for that particular role, it makes sense to employ the familiar, modern name. Dissociated in this way from any linkage to a specific historical period, the word 'Spanish' comes to refer not just to the language of the post-medieval Spanish state, but also to all of its previous incarnations. The secondary label 'Old Spanish' should accordingly be seen as a meronym, designating no more than a part or segment of the diachronic whole.

This perspective immediately invites the question of what temporal boundaries should be assumed to define the relevant segment of the continuum. Approaching this issue in the first instance with a comparison, it can be noted that what linguists have in mind when they talk about Old Spanish begins rather later than the language variety that falls under the label 'Old English'. For while the term 'Old Spanish' is not generally used to refer to linguistic data which significantly predate the onset of the High Middle Ages, Old English is identified as the language of the Anglo-Saxons, who settled in Britain from the mid-fifth century. Indeed, despite the formal similarity of their names, Old Spanish and Old English occupy different positions within their respective genealogies. In Spanish terms, the nearest counterpart to Old English would be something like late spoken Iberian Latin, i.e. what Wright (1994) refers to as early Ibero-Romance. In practice, Middle English is a better match for Old Spanish,

even if, as is implied by the discussion below, the notional beginning of the latter predates that of the former by about two centuries.[2]

As with other Romance varieties, the Spanish case is complicated by the relatively late appearance of a customized way of writing the language. In its recognizable written guise, Old Spanish effectively dates from the thirteenth century, texts prior to this period tending to be Latinate in appearance.[3] At the spoken level, how far back in time what we might be prepared to think of as Old Spanish goes is anyone's guess. Spanish philological tradition takes its cue from Menéndez Pidal's seminal work *Orígenes del español* (1926), a detailed reconstruction of the early spoken language based on documents from the tenth and eleventh centuries.[4] Linguistically speaking, there is no particular reason to identify those centuries as marking the emergence of a new linguistic entity. However, there is a fairly widespread assumption that they represent, in some way, the *época de orígenes* 'origins period' of the Spanish language. The more fundamental point is that there is a disjuncture between Old Spanish as manifested with full clarity by the bespoke writing system that came on stream in the thirteenth century and Old Spanish as an older but largely presumptive linguistic variety, revealed to us through the prism of a written code devised originally for the speech of many centuries earlier.

1.1.2 Syntactic Continuity

While the term 'Old Spanish' in principle covers both of the linguistic manifestations just highlighted, the second one can be referred to more specifically as **pre-literary Spanish**, where 'literary' alludes to writing in general rather than to literature specifically. This latter form of Spanish is itself presented to us in a wide variety of guises, close approximations to Latin grammar and spelling characterizing one polar extreme and innovative experimental forms the other. In between we find texts which, to varying degrees, mix Latin words and case endings with syntactic structures, vocabulary and spelling patterns that clearly belong to Old Spanish. Menéndez Pidal's *Orígenes del español* references both the experimental glosses associated with the San Millán and Santo Domingo de Silos

monasteries and a variety of documents embodying the hybrid text type. The latter, it has to be said, point to the existence of a spoken language whose syntax in most major areas is very similar to what we find in the linguistically more transparent manuscripts of the thirteenth century. This is hardly surprising, given that in the majority of cases Menéndez Pidal's documents pre-date those of the early literary period by no more than two hundred years, a timeframe which realistically does not allow for any dramatic transformation of the grammar.

The similarity in syntax between the presumptive spoken language of the *época de orígenes* and the well-attested Spanish of the thirteenth century is illustrated in the extract in (1) below, which is taken from a document originating in the Palencia area and written, according to Menéndez Pidal, in 1097.[5]

(1) Et si ego mici mortem ante uobis uenerit, si de mea ereditate comodo et demeo ganato, aueatis uos jlas duas partes, et jla tertja, siue de ereditate comodo et deganato, jntre promea anima asancti Zoili. Et siuobis uiro meo aut germanis meis jla tertja quesieritis recolere, aprecient jla quantum ualere, et date precio pro jlas duas partes, et jla tertia lexola por amor de Dios; (Archivo Histórico Nacional de Madrid, San Zoil de Carrión *P*–7; Menéndez Pidal 1926: 35)
'And if death comes to me before you, of my land as with my cattle, you shall have two parts, and the third, be it land or cattle, goes for my soul to San Zoil. And if you my husband or my brothers want to recover the third, its value should be established, then pay a price for two thirds of it, but the remainder I give for free.'

If one looks past the Latin interference, such as the anachronistic dative forms *mici* (Classical Latin: *mihi*) and *uobis* or outmoded spellings like *aut*, *pro* and *comodo* (Classical Latin: *quomodo*), the language in the extract should seem very familiar to anyone versed in the grammar of the post-1200 period. Rather obviously, the consistent use of *jla* and *jlas* to introduce noun phrases points to a fully operational definite article and the linear sequences *aueatis . . . jlas duas partes* 'shall have (the) two parts' and *date precio* 'give a price' indicate a predominantly VO word order. In addition, *aueatis uos* 'you shall have' appears to be an instance of the common medieval VS(O) pattern that is discussed here in Chap. 2, while *jla tertja quesieritis recolere* 'the third [you] want to recover' can be analysed

as discourse-driven constituent fronting, a wide-ranging and emblematic feature of Old Spanish which forms the principle subject matter of Chap. 2. As regards weak pronoun placement, the enclitic structures *aprecient jla* 'they/one should weigh it up' and *jla tertia lexola* 'the third, I leave it' are fully consistent with the Tobler–Mussafia system, which is commonly discussed in relation to thirteenth and fourteenth century Spanish (see Sect. 3.2.2 in this book). More specifically, *aprecient jla*, which immediately follows an adverbial clause, illustrates the inhibiting effect of a clausal boundary on proclisis, while *jla tertia lexola* involves Clitic Left Dislocation, which is known to correlate with pronominal enclisis in Tobler–Mussafia languages (see Benincà 2006; also Sect. 2.3.1.2 [latter stages] in this book). Finally, if *ualere* is assumed to be an early form of finite *valiere* 'it is worth', the *jla quantum ualere* component of the structure *aprecient jla quantum ualere* is no more than a pronominal counterpart to thirteenth century interrogative formulations of the kind discussed in Sect. 5.3.1.2, i.e. formulations like *la caça qual es* as it occurs in the example below (= (41) from Chap. 5):

(2) E despues daquesto quando la sacaren a caça conuiene que uea el caçador antes **la caça qual es**. ca mas uale que la ecchen ante a la mayor caça que a la menor; (*Libro de las animalias que cazan*, fol. 62v)
'And after this, when it is taken hunting, the huntsman should first of all check what is being hunted, for it is best to set [the bird] after larger game before smaller animals.'

As these brief comments on the language of (1) indicate, when texts from the pre-literary period exhibit a syntactic pattern that is also attested in the well-codified Old Spanish of the literary period, one can be fairly confident that we are dealing with the same phenomenon in both cases. In the converse situation, i.e. when an "expected" Spanish feature is missing in a Latinate text, the analysis is not necessarily as straightforward. In this type of case, it may in practice be impossible to say with certainty whether the scribe is merely applying a Latin rule, learned as part of his training, or whether the structure in question was actually subject to a degree of variation. An instance of this problem arises in connection with negation patterns in the Latinate section (dispositions I to CIX) of the *Fuero de Madrid*. This is a particularly interesting case, as the relevant part

of the *Fuero*'s manuscript dates from 1202, making it roughly contemporaneous with the emergence of literary Spanish. Now Camus Bergareche (1986) posits that preverbal negative quantifiers (other than *nunca*) were uniformly accompanied in medieval Spanish by the negative morpheme *non* until about the middle of the fifteenth century. Accordingly, with respect to examples like (3) and (4) below, he suggests (ibid. p. 113) that 'probablemente la oración no deba considerarse propiamente castellana sino latina'[6] and hence that the examples should not be treated as exceptions to the proposed general rule.

(3) Nullus respondeat sine rancuroso (*Fuero de Madrid*, XXXII)
'Nobody should be charged if no complaint is made.'

(4) [. . .] nichil pectet (*Fuero de Madrid*, LXV)
'. . . he pays nothing.'

On the face of it, this analysis seems to be the obvious one to advance. A prominent feature of Latin is that its negative quantifiers *nullus* 'no/nobody', *nihil* 'nothing', *nemo* 'nobody' etc. behaved exactly like their equivalents in English; that is to say, they could not construe with the clausal negator *non* (except to express a genuine double negation, amounting to an affirmative assertion), regardless of whether they were preverbal or postverbal. Examples (3) and (4) thus appear to reflect no more than the scribe's disciplined adherence to a salient principle of Latin syntax. However, looking at the *Fuero de Madrid* overall, we find that rather than being categorically excluded, the co-occurrence of *non* with Latin negative quantifiers is actually variable. Thus in addition to examples like (3) and (4), we find other examples, such as (5) and (6) below, where the negation-related syntax is consistent with the known patterns of Old Spanish (see Sect. 6.3 in this book):

(5) **Nullus non** pignoret qui uenerit cum mercadura. Todo el omne qui ad Madrid uenerit in requa, & alguna cosa adduxerit ad Madrit, **nullus homo non** pendret ei, et qui lo pendraret, pectet II morabetinos a los iurados del rei et tornet la pendra sene fiadura. (*Fuero de Madrid*, LXIV)

'Nobody from outside may bring goods to pawn. In respect of any man who comes to Madrid with a mule caravan, bringing goods to Madrid, no man may take collateral from him, and anyone who does must pay two maravedis to the king's officers and the loan is invalidated.'

(6) Toto homine de Madrid qui messare aut firieret uel mataret pastor aut bacherizo [. . .] & pignos noluerit dare cum bonas testemunas, **non pectet nullam calumpnia**, nisi calumpnia regi (*Fuero de Madrid*, XXII)
'Any man from Madrid who assaults, wounds or kills a shepherd or cowhand and fails to provide character witnesses pays no fine, except to the king'

Interestingly, the use of *non* is categorical in the specific case illustrated by (6), i.e. when the negative quantifier is postverbal. In other words, the locus of the variation is the preverbal field, which is exactly the context in which the use of *no*(*n*) with negative quantifiers has been diachronically unstable, both in Spanish and Romance more generally (see Sect. 6.4).

In light of the overall pattern of *non*-usage, then, one cannot be certain that the *non*-less syntax of examples like (3) and (4) is genuinely the product of Latin interference or whether, in actual fact, it should not be taken as evidence of early variation in this particular area of the spoken grammar. Such variation would not be entirely unexpected, given that overtly Old Spanish examples like (7) below, from a mid-thirteenth century manuscript, suggest that the *non*-less structure which came into its own in the fifteenth century may well have been available, if not widely used, at a significantly earlier time (in this connection, see Fig. 6.1 in Chap. 6).

(7) E **ninguno deue** tomar della otra cosa; fueras aquello quel es otorgado por el derecho de santa eglesia. (*Libro de las leyes*, fol. 90r)
'And nobody should take from it anything other than that to which they are entitled by the law of the holy church.'

As the discussion indicates, we are not in this this type of case talking about the wholesale absence of an expected Old Spanish feature, but its partial absence. The problem, therefore, is largely a methodological one;

namely, are the examples without the expected feature simply cases in which the scribe has remembered the relevant Latin rule, or might they be thought of as early tokens of a pattern that is independently known to have emerged? A case with potentially wider-ranging implications relates to the use of the definite article with prenominal possessives, as in *el so nieto* 'his grandson', a structure robustly attested from the thirteenth century onwards and, it is argued here in Chaps. 4 and 5, just one manifestation of a more general syntactic paradigm. As (8) below shows, the structure is attested in the famous eleventh century San Millán glosses:

(8) Facanos Deus omnipotes tal serbitjo fere que denante **ela sua face** gaudioso segamos. (*Códice Emilianense 60*, fol. 70r; cited Menéndez Pidal 1926: 8)
'May almighty God make us perform such service that before his face we are joyful.'

However, Batllori (2010: 420) finds that the structure is 'prácticamente ausente' in texts from the pre-literary period and Menéndez Pidal himself does not appear to have regarded it as belonging to the linguistic landscape of the *época de orígenes*, given that he makes no mention of it in his survey of the early possessive (ibid. § 67). One could infer from this that examples like (8) are not representative in this regard and that the use of the definite article with prenominal possessives reflects a type of syntax that did not acquire a detectable quantitative footprint until the latter stages of the High Middle Ages. That is of course perfectly possible. However, the quantitative data adduced here in Chap. 4 point to a rate of occurrence in the mid-thirteenth century of close to 20%, a level of productivity which remained approximately stable for the next two hundred years. This is suggestive of a structure that was already quite well embedded at the outset of the literary period. Viewed in that light, its apparently negligible presence in the (predominantly Latinate) documentation from before that time is surprising, particularly given its attestation in the early Romance gloss shown in (8).

It is not implausible, perhaps, to link this state of affairs to the relatively small number of 'possessive + noun' combinations that in practice

occur in the documents of the pre-literary period. In a text type embodying the discourse of property transfers and municipal charters to the exclusion of almost everything else, possessive noun phrases such as *sua parte* 'his/her part', *suo termino* 'its limit' *suos molinos* 'its mills', *suas sernas* 'its arable land' etc. may well have become frozen in their article-less form at an early date, thereby developing an unusually high degree of resistance to the pressure of the spoken language. And it is of course worth remembering that the 'definite article + possessive + noun' structure was not a majority option even in the thirteenth century (or at any time afterwards). Presumably, therefore, any impulse felt by the scribes of the pre-literary period to prefix the possessive determiner with the definite article would not have been a strong one. This particular combination of circumstances might conceivably have prevented a structure that one could reasonably expect to have been available in speech, if only at "background" levels of frequency, from appearing in the textual record.

Notwithstanding this latter puzzle, one can be fairly confident that the syntax of pre-literary Spanish, certainly from the *época de orígenes* onwards, was fundamentally similar to the syntax which is revealed in the prose manuscripts of the thirteenth century, which marks the beginning of the time window with which this book is primarily concerned. The right-hand boundary of this window, assuming a left-to-right visualization, is located in the early seventeenth century, a time that in conventional terms falls outside the Old Spanish period, which is usually assumed to come to an end in the fifteenth century (cf. Wanner 1991: 349). However, the majority of the changes which notionally represent the transition from Old Spanish to Modern Spanish are not actually complete by the end of the Middle Ages. Indeed, even the relatively late upper time boundary chosen here cuts off some of the changes before their diachronic curve has fully arrived at its saturation point, although the direction of travel is by that time usually pretty clear. This only reinforces the point implicitly made in Sect. 1.1.1, and explicitly made by Wright (1999), that linguistic periodization is ultimately self-defeating.

1.2 Measuring the Speed of Change

1.2.1 Constant Change

One of the most significant findings in diachronic quantitative linguistics in recent decades is that syntactic change, as externally manifested, tends to evolve at the same speed across all of the affected contexts. That this was the case had long been entertained as a possibility, but Kroch (1989) was the first linguist to demonstrate this empirically. His statement of what is now called the **Constant Rate Effect** is given in (9) below:

Constant Rate Effect (CRE)

(9) When one grammatical option replaces another with which it is in competition across a set of linguistic contexts, the rate of replacement, properly measured, is the same in all of them. (Kroch 1989: 200)

Quite apart from being an interesting phenomenon in its own right, the CRE is a very useful tool for analysing groups of changes in a language. Searching for a CRE in any given case can often help to determine whether two or more changes are entirely separate events or whether, in contrast, they are manifestations of a single, more general change. Resolving this latter type of question is in turn fundamental to attaining a more structural or systemic perspective on a language's evolution. The ability to measure the rate or speed of syntactic change is, therefore, an important component of the diachronic quantitative linguist's toolkit.

Quantitative change can obviously involve either an increase or a decrease, the first case being known as **growth** and the second as **decay**. Because linguists usually think in terms of one variant replacing another, discussions of syntactic change are typically about growth (in the frequency of the "incoming" structure). Clearly, though, structures may decline in frequency and in that case we need to measure the rate of decay. Growth and decay are measured in exactly the same way, but the interpretation of decay rates requires an additional consideration, whose discussion I will postpone until Sect. 1.2.3. As regards growth, Fig. 1.1

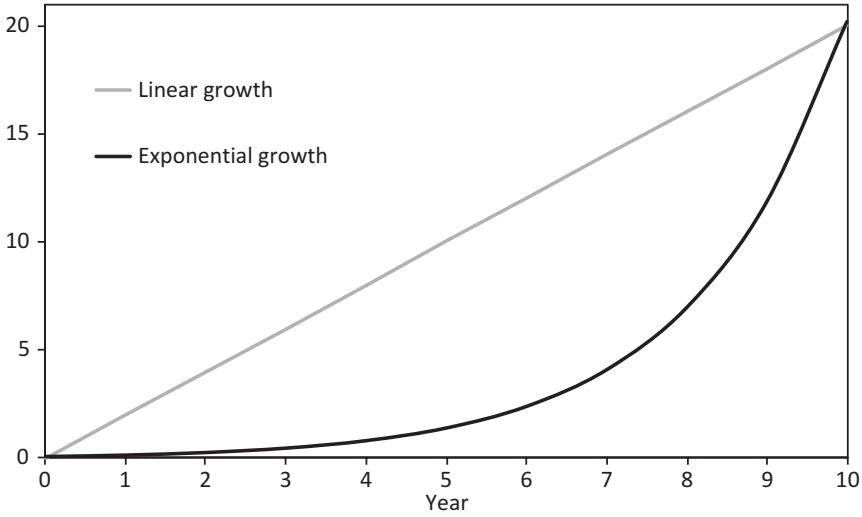

Fig. 1.1 Linear growth and exponential growth

illustrates what are probably the two commonest kinds, viz. linear growth and exponential growth.

Linear growth is additive, in the sense that for each unit increase in the independent variable, a specific number of units is added to the value of the dependent variable. Thus the grey line in the figure represents a situation in which every year the value of the quantity shown on the vertical axis increases by two units: in year 1, the value is 2; in year 2, it is 4; in year 3, it is 6 etc. This type of growth, when represented in a graph or chart, corresponds to a straight line. In contrast, exponential growth is proportional, in the sense that each new value of the independent variable is proportional to the previous one. In Fig. 1.1 the exponential curve corresponds to a situation in which, each year, the value of the dependent variable increases by 70% with respect to the previous year. Expressing this in terms of a ratio, we can say that the value of the dependent variable each year is 1.7 times higher than the previous year's value. As can be seen from the figure, this growth model delivers an approximately J-shaped curve.

Now in both of the cases just considered, the growth rate (the rate of increase) can be said to be constant, in the sense that it can be characterized

in terms of a single numerical value. In respect of the grey straight line we can say that the growth rate is two units per year and as regards the black J-curve we can say that the growth rate is 70% per year (equating to a yearly growth ratio of 1.7). If syntactic changes were either linear or exponential, we could simply use one of these two measures to characterize the evolution of any given structure. However, as is well-known, syntactic changes (and linguistic changes generally) tend to describe an S-curve when their evolution is plotted on a graph. By definition, a sigmoidal curve cannot express a constant rate of change. In order, then, to bring syntactic change within the scope of the model implied by Kroch's Constant Rate Effect, we need to conceptualize it in a different way from that implied by the usual graphical visualizations.

1.2.2 Logistic Growth and the Odds Ratio

The problem can be illustrated using data from Ellegård's classic 1953 study of the quantitative evolution of *do*-support in English, the process whereby simple verb–subject inversion (e.g. *Knowest thou the ordinances of heaven?* [*Job* 38:33]) in questions and negations was replaced by the modern periphrastic structure involving *do*. Figure 1.2 shows the curve corresponding to the advance of *do*-support in the context defined by Ellegård as including affirmative transitive questions that are either of the yes/no variety or involve an adverbial *wh*-word such as *when* or *where*.

As can be seen, both from the unbroken line corresponding to Ellegård's data and the dashed trendline, the growth in the percentage rate or (absolute) probability of *do*-support is neither linear nor properly exponential. Rather, it has an S-shaped trajectory, which diachronic linguists now usually identify with the **logistic curve** originally used by Pierre Verhulst to model population growth (see Altmann et al. 1983; Kroch 1989; Yang 2000; Kauhanen and Walkden 2017). Logistic growth is characterized by the existence of a natural terminus or **saturation point**, this being, in Fig. 1.2, the maximum percentage value of 100%. From an exponential perspective, the growth rate declines gradually over time, causing the curve to evolve from a convex shape to a concave one, the inflection point or change in curvature being located at the curve's midpoint (around

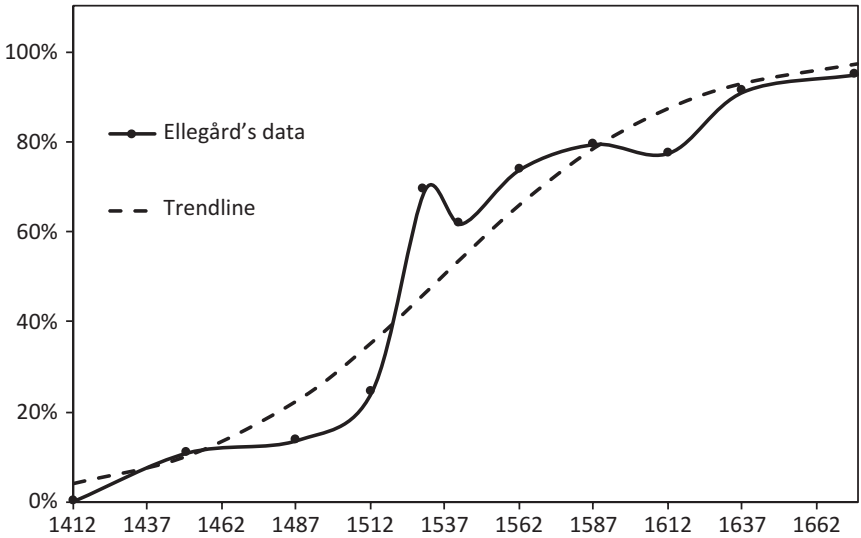

Fig. 1.2 Advance of *do*-support in affirmative transitive yes/no and adverbial questions (1400–1700)

1540 in the present case). In linear terms, the growth rate increases up to the inflection point and then starts to decline. Viewed from either perspective, the growth rate in the absolute probability is constantly changing. In other words, if we use the percentage scale to quantify the evolving frequency of the incoming structure – as is completely normal – there is no single numerical value that will capture the rate of change.

To resolve this problem, the absolute probability values need to be converted into the equivalent **relative probability** values, also known as **odds**. The odds are a measure of probability, but unlike absolute probability, which is calculated with reference to all possible outcomes, odds are calculated in relation to the likelihood of the relevant event *not* happening. For example, abstracting away from such matters as quality of squad and managerial skill, the absolute probability of any given team winning the English Premier League is 1/20 or 0.05, whereas the *odds* of this event happening are 1/19 (or 0.053 to three decimal places), which in betting parlance would be expressed as 'nineteen to one against'.[7] Absolute probability can be converted into odds using the equation below:

$$Odds = \frac{probability}{1 - probability}$$

A useful property of odds is that they have no upper limit: they approach infinity as absolute probability approaches 1. Better still, logistic growth in the absolute probability equates to exponential growth in the odds.[8] Accordingly, if we re-express Fig. 1.2 using the odds scale, we end up with the exponential pattern of evolution shown in Fig. 1.3.

Viewed in this way, the advance of *do*-support in the relevant context does have a constant rate. Specifically, it can be estimated by logistic regression analysis that the odds of *do*-support in this context increase by 28.9%, or a factor of 1.289, each decade. The factor increase of 1.289 is also known as an **odds ratio** (OR), a concept which is of fundamental importance in many areas of statistical analysis.

It should be noted that, rather than the OR itself, Kroch (1989) and his followers use the natural logarithm of the odds ratio, known as the

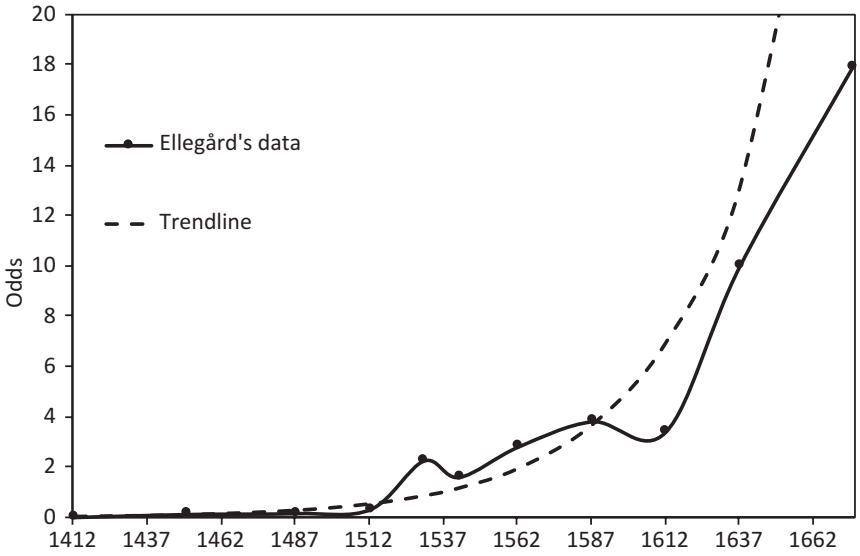

Fig. 1.3 Evolving odds of *do*-support in affirmative transitive yes/no and adverbial questions (1400–1700)

slope, as their measure of the growth rate. Here, for example, the slope would be ln(1.289) = 0.254, where 'ln' means 'natural logarithm of'.[9] While the slope is referenced in one or two places in this book, in general the more intuitive measure of the odds ratio is used to capture rates of change. This does not mean that synchronic rates of usage need to be expressed as odds, in the manner of Fig. 1.3, rather than as percentages (Fig. 1.2). The concept of the odds is only needed to enable a value to be assigned to the rate of change; for general purposes, including diachronic visualizations, we can use the familiar percentage scale. In reality, odds and percentages are just different ways of presenting the same underlying data, and it is to the latter that the logistic regression procedure is applied.[10]

1.2.3 Quantitative Decay and Failed Changes

If an odds ratio in excess of 1 implies growth, an OR of less than 1 indicates decay (and an OR of exactly 1 implies no change at all). For example, if the decadal OR for structure *X* in context *Y* is 0.74, this means that the odds of *X* in context *Y* decay at the rate of 26% per decade, where 26% is simply 1–0.74, expressed as a percentage. One should be careful, however, not to fall into the trap of assuming that the downward curve delivered by an OR of 0.74 is the mirror image of the upward curve delivered by an OR of 1.26 (the latter implying growth of 26% per unit of time). This would be analogous to thinking that a 50% increase in an investment of £100 would be exactly reversed by a subsequent decrease of 50%. Given that the value of the investment after the increase is £150, a 50% decrease would leave the investor with only £75 pounds. For the investment to revert to its initial value, £50 must be lost from its post-increase value, implying a decline of one third or 33.3%. As a measure of decay, this latter figure equates to a "growth" ratio of 0.667, a value which is simply the reciprocal of the growth ratio of 1.5 corresponding to the 50% increase; that is to say, 0.667 = 1/1.5 (to three decimal places). Exactly the same principle holds for decay in language change. For example, if we have an upward curve with an OR of 1.25, implying growth of 25% per time unit, the equivalent downward curve must have an OR of 1/1.25 = 0.8, which is equivalent to a decay rate of 20% per time unit.

Being able to determine whether a decaying OR matches a previous upward-trending OR is useful for evaluating up-and-down quantitative events, called **failed changes** by Postma (2010). In the context of unemphatic declarative sentences, for example, *do*-support initially increased in frequency (cf. examples like *Me thinke I doe heare a good manerly Begger at the doore*, cited by Kroch 1989: 229) only to decline later and eventually disappear. Figure 1.4, again based on Ellegård's data, illustrates the bell curve which events of this type characteristically describe.

How to analyse this type of phenomenon is far from settled at present. However, one obvious question that can be asked is whether the growth rate represented by the left-hand flank of the bell and the decay rate corresponding to the right-hand flank are (approximately) equal or not. Symmetry in this regard might imply that the curve overall represents a single, 'inherently failing' change, as Postma (2010, 2017) has proposed. Conversely, asymmetry might suggest that the rise and decline correspond to separate changes, implying that the failure is 'accidental' rather than predetermined. Either way, assuming that change broadly corresponds to

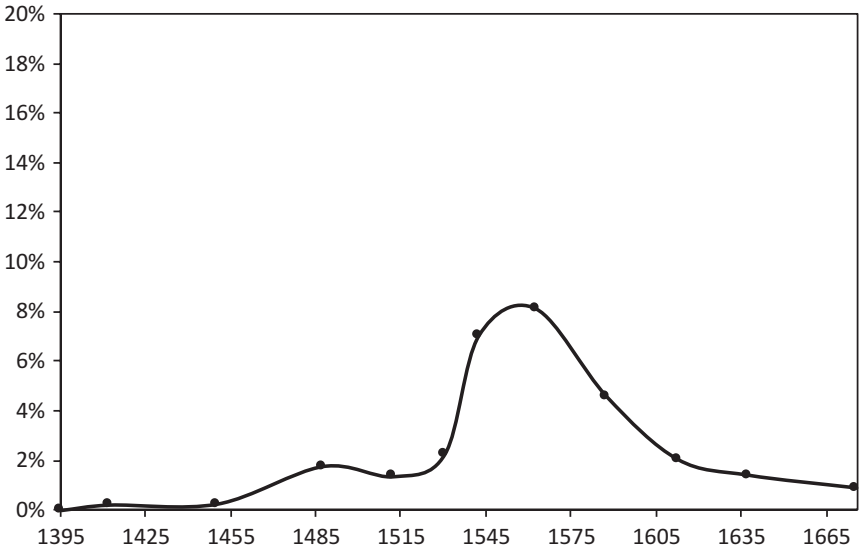

Fig. 1.4 Rise and fall of *do*-support in unemphatic affirmative declarative sentences (1390–1700)

exponential growth or decay in the odds, the OR provides a useful way of quantifying the speed of change on each side of the peak.

1.3 Overview of the Corpus

The main corpus for the present study is constructed from Castilian prose texts covering the period 1250–1609, with an overall size of 14,340,000 words. Tokens of the structures being investigated are aggregated at twenty-year intervals, meaning that there are eighteen data points, the first covering the period 1250–1269 (identified as 1260) and the final one covering the period 1590–1609 (identified as 1600). This fine-grained segmentation of the data enables relatively precise quantitative curves to be constructed, which in turn assists with hypotheses as to how the syntactic changes were related to one another. Manuscript texts have been assigned to data points on the basis of their date of copy, using the bio-bibliographical database PhiloBiblon, hosted at http://bancroft.berkeley.edu/philobiblon/. In a number of cases, a given manuscript is assigned to two or more adjacent data points. In such cases the relevant data are aggregated on a weighted basis, e.g. 50% to each data point if assigned to two data points or 33.33% if assigned to three. The post-medieval printed texts are assigned to data points on the basis of their initial publication date.

The texts in the first thirteen data points (1260–1500) have all been surveyed in their manuscript or incunabulum form, using the electronic transcriptions provided by the following outputs of the Hispanic Seminary of Medieval Studies (full details of which are given in the bibliography):

1. *Prose works of Alfonso X el sabio.* Digital library of Old Spanish texts. (Gago Jover 2011a)
2. *Spanish legal texts.* Digital library of Old Spanish texts. (Gago Jover 2011b)
3. *Spanish medical texts.* Digital library of Old Spanish texts. (Gago Jover 2011c)

4. *Spanish chronicle texts.* Digital library of Old Spanish texts. (Gago Jover 2011d)
5. *Electronic texts and concordances of the Madison Corpus of early Spanish manuscripts and printings* (O'Neill 1999)
6. *Textos y concordancias electrónicos de documentos castellanos de Alfonso X* (Herrera et al. 1999)

The specific texts drawn from the above sources that are included in the corpus are listed in Appendix 2, with manuscript details given in parentheses after each entry.

The majority of the texts in the last five data points (1520–1600) have been surveyed using the electronic editions provided by the *Biblioteca virtual Miguel de Cervantes* (hosted by the University of Alicante at http://www.cervantesvirtual.com/). These editions range from direct transcriptions of the relevant princeps edition to machine-readable versions of modern scholarly editions. In selecting texts for the corpus from this source, care has been taken to ensure that any editorial intervention is limited at most to updating the orthography. As with the medieval texts referred to above, the specific early modern texts included in the corpus are listed in Appendix 2, together with details of the printed editions on which the relevant electronic editions are based.

Notes

1. According to Wright (2013: 31), the word *romanz*/*romançe* and its cognates were originally applied to (non-Latinate) ways of writing rather than to actual speech. Presumably, language in a more general sense would have been referred to using terms like *lenguage*, *lengua* or *fabla* 'speech'.
2. In contrast, in the French domain, the 'Middle' instantiation of the language is relatively modern, being essentially the variety of French which post-dates the Old French period but is not yet the Classical French associated with the likes of Racine, Molière and Vaugelas.
3. The *Poema de mio Cid*, the first major literary work written in Old Spanish, is often presented as being a twelfth century text. If it is, it is likely to be a late twelfth century text. The sole surviving manuscript is actually from the fourteenth century, but folio 74r. states that 'Per abbat

le escriuio enel mes de mayo En era de mill & .C.C xL.v. annos' ('Abbot Per wrote it down in the month of May, era of Caesar 1245 [i.e. 1207 AD]'). Modulo its poetic style, the syntax of the *Poema de mio Cid* is not appreciably different from what one finds in prose texts from the second half of the thirteenth century.

4. It should be noted that the majority of the manuscripts transcribed in the *Orígenes del español* are not technically from the Castile area. Assuming that Spanish is the Romance variety associated with Castile, it is not certain that early manuscripts from Aragon, say, can be regarded as embodying an embryonic form of Spanish. Pountain (2001: 19) makes an analogous point with respect to the San Millán glosses, which are from the Rioja area. On the other hand, Wright (2013: 32) advocates the exercise of latitude when applying the label 'Spanish' to the language of that period.
5. See (51) and (52) in Sect. 2.4.1 for additional illustrations.
6. 'Probably the clause should not be seen as being a genuinely Spanish one but rather a Latin one.'
7. It should be noted that the odds quoted by a professional bookmaker are, first and foremost, a conditional financial commitment rather than a disinterested estimate of relative probability.
8. Growth in the odds, in the statistical sense, equates in principle to a 'shortening' of the odds in everyday usage. For example, if the odds shorten from three-to-one to two-to-one, they actually increase from 1/3 or 0.333 (to three decimal places) to 1/2 or 0.5.
9. The decadal value given in the text equates to a slope of 2.54 on a century basis. Alert readers will note that this differs from Kroch's own (century-based) estimate of the slope for the advance of *do*-support in this context, which is 3.62 (Kroch 1989: 225). The reason for the discrepancy is that Kroch's estimate is based on just the first seven of Ellegård's data points, rather than all eleven as here.
10. All the logistic regression estimates reported in this book employ maximum likelihood estimation, which is standardly available in statistical and mathematical software packages.

References

Altmann, Gabriel, Haro Buttlar, Walter Rott, and Udo Strauss. 1983. A law of change in language. In *Historical linguistics*, ed. B. Brainerd, 104–115. Bochum: Studienverlag Dr. N. Brockmeyer.