Transactions of the Philological Society · PhilSoc The Philological Society

Transactions of the Philological Society Volume 00 (2023) 1–22 doi: 10.1111/1467-968X.12275

SOCIOLINGUISTIC TYPOLOGY MEETS HISTORICAL CORPUS LINGUISTICS

By George Walkden, Gemma Hunter McCarley, Raquel Montero, Molly Rolf, Sarah Einhaus and Henri Kauhanen
University of Konstanz

(Submitted: 28 April, 2023; Accepted: 30 September, 2023)

Abstract

This paper makes the case for using historical corpora to assess questions of sociolinguistic typology. A full account of any contact-induced change will need to establish WHAT the linguistic innovation in question was, WHO was in contact, WHERE and WHEN the contact took place and HOW the change happened, both at the individual level and at the population level. The historical corpus approach complements other methods by narrowing down the WHERE and the WHEN, allowing us to develop a clearer picture of how the change diffused. In support of our approach, we present three case studies of potential morphosyntactic simplification using quantitative evidence gleaned from historical corpora: the loss of number concord in the history of English, change in the null-subject system(s) of Latin American Spanish and reduction of the case system in the history of Balkan Slavic. All three cases allow us to test theoretical predictions and uncover new influencing factors in a way that would be impossible without fine-grained quantitative corpus research.

1. INTRODUCTION

In what has by now already become a classic work on language change, Trudgill (2011) articulates the hypothesis that different types of language contact situation may give rise to different types of change.[1] Drawing together two traditions of the literature on contact effects, he proposes that long-term contact scenarios that involve a high degree of child bilingualism are likely to lead to what he terms ADDITIVE COMPLEXIFICATION: linguistic features or properties will be transferred between the varieties in contact, in a way that may be additive rather than simply replacive. By contrast, short-term contact scenarios in which the bulk of the learning is done by adults rather than children are likely to lead to SIMPLIFICATION, at least of the varieties which are so acquired. This is the core of Trudgill’s theory of sociolinguistic typology.

This article is about how to assess Trudgill’s theory using the most powerful methodological tool that contemporary historical linguistics has at its disposal: the study of the historical record itself, in the form of textual corpora. The theory of sociolinguistic typology is, at its core, a theory about change; thus, any method that is able to assess historical change in progress ought to be ideally suited for testing the theory, provided we

[1] We would like to thank Fernanda Barrientos, Bozhil Hristov and Kaius Sinnemäki for discussions. The research reported in this paper was carried out as part of the project STARFISH; this project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no 851423).

have the information we need about the historical circumstances. Curiously, though, much of the work that has been done on sociolinguistic typology since the publication of Trudgill's book has left the historical record aside entirely, focussing on distributions of features across samples of present-day languages—which we term the typological approach (see section 2.1)—or on small groups of participants in artificial settings—which we term the experimental approach (see section 2.2).[2] In section 2, we present the historical corpus-based approach of the STARFISH project, and contrast it with both the typological and experimental approaches, weighing up the pros and cons of each.[3]

In support of our approach, we present three case studies of morphosyntactic phenomena for which the theory of sociolinguistic typology makes predictions, and assess them using historical corpus evidence. Section 3 deals with the loss of grammatical number concord in the history of English; section 4 deals with the development of null subject systems in Latin American Spanish; and section 5 deals with case, grammatical functions and their formal realisations in Balkan Slavic. These three case studies are not intended to be conclusive on their own, but rather to illustrate—taken together—how historical corpus evidence can be brought to bear on questions of sociolinguistic typology.

Given the quantitative richness of historical corpus evidence, a natural question to ask is whether we can go beyond simply characterising a given historical scenario as short-term or long-term after the fact, and instead make more precise inferences about the role of population dynamics in language change that can then be tested against the quantitative evidence we have at our disposal. Section 6 is devoted to this question; Section 7 then summarises and concludes the paper. Our aim is to demonstrate how an approach based on historical corpora can complement other methodologies by adding depth and nuance to our understanding of sociolinguistic typology.

2. METHODOLOGY

Trudgill (2011) provides a number of case studies in support of his approach. One such is Nubi, a language derived from Arabic via creolisation (Trudgill 2011: 44–5). Where Classical Arabic has an intricate templatic verbal morphology for ϕ-feature agreement, for example *ta-ktub-u* 'you write' vs. *ya-ktub-uw-na* 'they (m.) write', Nubi has the invariant form *gi-ˈkatifu* in all ϕ-feature combinations (Owens 2001: 349).[4] This is morphological simplification, in an intuitively clear sense. Trudgill's way of showing this, following Owens (2001), is basically the traditional method of historical-comparative linguistics: presenting a grammatical description of language stages before, and after, the change, and using these to draw conclusions about the nature of the change, constructing a historical narrative (see Mayr 2004: 32–3). The same is true of his treatment of simplification in the nominal morphology of English from Old English to the present day (Trudgill 2011, ch. 2).

In many cases, this tried-and-true method will yield clear results—but sociolinguistic typology can go beyond it. In this section, we discuss several methods for investigating the predictions of sociolinguistic typology, with a particular focus on our own, the historical corpus method. To do so, it is useful to list what we are trying to find out when we investigate a case of language contact to see whether it conforms to the predictions of Trudgill's account.

[2] The corpus-based studies of change in historical Mainland Scandinavian in Blaxter (2017) are an important exception, and a forerunner of our approach.

[3] The vexed question of what constitutes complexity in (morpho)syntax is one that lies beyond the scope of this paper. In setting out the relevant notion of complexity for our case studies, we follow Walkden & Breitbarth (2019a, 2019b), who argue that semantically uninterpretable features in syntax are L2-difficult; see that work for discussion and justification.

[4] *gi-* is a progressive prefix.

1. WHAT the linguistic variants in question were/are: an accurate characterisation of the linguistic realities on the ground is essential for all contact linguistics.
2. WHO is in contact: which people were involved, which languages did they know/use, and what sort of language users/acquirers were they (child, adolescent, adult)?
3. WHERE the contact took place: different geographical patterns can provide important circumstantial evidence for a historical narrative.
4. WHEN the contact took place: timescale (long-term vs. short-term) is also crucial for evaluating a scenario based on sociolinguistic typology.
5. HOW the contact took place at the INDIVIDUAL level: what is happening in individuals' minds/brains, that is what are the sociopsychological factors that play a role in the language contact situation? As McIntosh (1994: 137) puts it, 'what we mean by "languages in contact" is "users of language in contact" and to insist upon this is much more than a mere terminological quibble and has far from trivial consequences'.
6. HOW the contact took place at the POPULATION level: one swallow does not make a summer when it comes to language changes diffusing through a population, and the social and historical mechanisms leading to this diffusion are also central to our historical narratives.

The WHAT question is the starting point for all methods, as all methods presuppose a good linguistic characterisation of the phenomena in question: we cannot do without the methods of descriptive and theoretical linguistics. Beyond this, however, the different approaches we sketch have their strengths and weaknesses in answering different questions.

2.1. *The typological approach*

One way to scale up Trudgill's method is to use databases containing information about a large number of languages, and this has been done by, for example, Lupyan & Dale (2010), Bentz & Winter (2013), Sinnemäki & Di Garbo (2018), Koplenig (2019), Sinnemäki (2020), Kauhanen et al. (2023) and Shcherbakova et al. (2023). Typically, such studies are quantitative and correlational: they take data from databases of present-day languages such as WALS (Dryer & Haspelmath 2013) or Grambank (Skirgård et al. 2023), operationalise sociolinguistic information (such as population size or proportion of L2 speakers), and assess the relationship between these variables statistically. Sinnemäki & Di Garbo (2018), for instance, look at verbal inflectional synthesis in a dataset of 309 languages, finding that number of L1 speakers has a significant effect on degree of synthesis, with a borderline significant effect of proportion of L2 speakers as well.

In terms of the questions outlined at the beginning of this section, we can highlight that studies such as these usually have a good angle on WHO and WHERE: demographic information is crucial, and attempts are usually made to control for geographical proximity (as well as relatedness). By including variables to do with L1 and L2 acquisition, typological studies also allow us an angle on the HOW (INDIVIDUAL) question, albeit very indirectly. The WHEN aspect is usually left entirely out of consideration in such studies: as they are based on present-day datasets, in principle the causal historical events leading to the synchronically observable distributions could have happened either yesterday or ten thousand years ago. A further disadvantage is that such data sets are usually fairly coarse-grained in terms of the linguistic information they include, since they are based on descriptive grammars and code variables in categorical terms, without considering if there is variability within the population. On the other hand, the major advantage of the typological approach is that its broader empirical base can serve to relativise grand claims based on case studies from only a handful of languages where the causal factors at play may be underdetermined: In short, the typological approach has more statistical power at its disposal for disentangling signal from noise.

2.2. *The experimental approach*

Investigating the cognitive mechanisms that govern the outcome of language change is crucial since it is the individuals and their interactions within a society that ultimately shape language evolution. Various cognitive biases, such as the well-established correlation between age of acquisition and language proficiency, provide compelling evidence for non-target-like adult L2 learning, affecting both productive and receptive performance in multiple linguistic domains such as phonology, morphology and syntax (Trudgill 2011; Atkinson et al. 2018). Although the extent of these non-target-like acquisitions is highly variable among individuals and depends on various factors such as learning context and motivation, some linguistic features consistently pose challenges in adult language learning. Experimentally identifying these features and observing their behaviour in artificial contact scenarios allows for the integration of results obtained from typological and historical approaches and provides insights into the individual HOW of language change, ultimately explaining why certain properties are subject to change over others.

For instance, Berdicevskis & Semenuks (2022) demonstrated, via an iterated learning experiment modelling generational transmission, that morphological simplification and thus change correlates with the number of ‘short-time’ learners (i.e. learners with a reduced amount of linguistic input) involved in a transmission chain, with redundant agent marking on verbs being the most vulnerable feature. Likewise, an artificial language learning study by Atkinson et al. (2018) investigated similar effects in contact scenarios and revealed that speakers of a complex language variety tend to simplify when communicating with speakers of a simpler variety by regularising irregular verb forms. In another related experiment, Raviv et al. (2019) explored the effects of group size on language systematicity and found that not only did larger groups systematise their grammar to a greater extent overall, they were also faster in doing so than small groups, the latter sometimes not even systematising at all. These results illustrate that not only cognitive mechanisms in the individual can be identified in laboratory settings, but also group dynamics on a larger scale can be modelled and monitored to verify and refine hypotheses obtained from the historical and typological approaches.

2.3. *The historical corpus approach*

In contrast to the typological approach and the experimental approach, the historical corpus approach seeks to identify the forces at work WHEN and WHERE they apply, that is in the linguistic usage of people in relevant contact situations. A crucial ingredient is the variationist recognition that population-level patterns are not entirely random and that individuals’ usage can be non-categorical but still pattern meaningfully in probabilistic ways: this is what Weinreich et al. (1968) term ‘orderly heterogeneity’. Recognising orderly heterogeneity is the linguistic manifestation of what studies of cultural evolution call ‘population thinking’; see Roberts & Sneller (2020) for discussion of this link.

Once it is established WHAT the linguistic variants in question are, and which of these are (by hypothesis) simpler or more complex, the variationist historical corpus method can assess which variants are more prevalent in which places and at which times. Sociolinguistic typology makes straightforward predictions about what places and times will be affected more: in the case of simplification, the WHERE will be those places in which there are significant numbers of adult L2 acquirer-users, and the WHEN will be as soon as their grammar is realised in written usage of which we have records. The historical corpus approach assesses these predictions by comparing texts from different times and places and looking at the quantitative prevalence of different variants. For instance, if a particular simplificatory change in English is due to the influence of Norse L2 acquirers, we would expect to see it earliest and most prominently in Middle English or late Old English texts, and in texts from the north and east rather than from the south and west.

Investigating questions of sociolinguistic typology by means of historical corpora puts our approach firmly within the tradition of historical sociolinguistics (see Auer et al. 2015 and the papers in Conde-Silvestre & Hernández-Campoy 2012). In this tradition, the nature of the data used is of central importance—not only in terms of macrolevel factors such as geographical location, but also microlevel factors to do with the writers themselves and the situation: who were they, and who were they writing for? The historical record necessarily exhibits a strong bias towards the written output of the literate classes in formal settings, making it difficult to know whether generalisations made on this basis also hold for the rest of the population, for more oral language, and in informal settings. We also might expect to see effects of normative pressure dampening the textual frequency of innovations driven by adult language acquisition. These problems can be mitigated, though not solved, by drawing on a wide variety of documents produced by as wide a variety of language users in as wide a variety of situations as possible. In this paper, we draw on texts ranging from a Middle English horse treatise via Bulgarian saints' lives to Afro-Colombian poetry. These considerations become especially important in dealing with null subjects in section 4, where the level of orality of a text is found to play a major role.

In terms of scale, this approach occupies an intermediate position between the typological approach—maximally zoomed out, with whole language communities represented as single data points—and the experimental approach, with its focus on a small number of individuals at a single point in time. Like the other approaches, this one has its disadvantages. Most saliently, it is restricted to those situations for which we actually possess the right kind of texts, in sufficient breadth and number, to draw comparative conclusions. Most societies throughout most of human history have not been of this type: in particular, languages spoken in small societies of intimates—a situation that according to sociolinguistic typology is likely to lead to linguistic complexification—are unlikely to be documented in this way.[5] In general, the languages for which we have extensive written records are heavily skewed towards Europe, Asia, northern Africa and the language families that are represented there. Still, the rich information about WHERE, WHEN and HOW (POPULATION) that such studies can yield make them a useful complement to the typological and experimental approaches. As Walkden & Breitbarth (2019a: 187) put it, 'An issue as nuanced as the Trudgill conjecture ought to benefit from as many different lines of attack as possible, hopefully with convergent results'.

Convergent results converge for a reason: they point to an underlying reality, a mechanism from which the various observed phenomena, which initially may have seemed unrelated, flow. A powerful way of making sense of such underlying mechanisms is through the use of computational models. In section 6, we illustrate how a complex-systems account of population dynamics can shed light on contact-induced simplification by connecting insights from the other approaches into a single, consilient mechanistic model. This model takes its inspiration in the way it defines individual linguistic agents from the experimental approach, procures relevant demographic variables in the ways suggested by the typological approach, and compares model predictions against actual historical data gathered through the use of the historical corpus-linguistic approach. The result is an account of language contact dynamics which not only can be used to test already existing predictions but also suggests new (typically quantitatively sharpened) predictions for future research to put to the test.

[5] Wray & Grace (2007), following Thurston (1989), term this the ESOTERIC NICHE.

3. CASE STUDY: NUMBER

3.1. What: Syntagmatic redundancy

As our first case study, we will look at the development of syntagmatic redundancy, focussing in particular on plural concord in Early English. The notion of 'syntagmatic redundancy' goes back to the work by Trudgill (1977) where it was used to refer to cases in which the same meaning appeared in multiple exponents within the same clause or phrase. In Middle English, for example, the piece of information for which the referent is plural appeared encoded on the noun, but also on the adjective and quantifier, as in (1):

(1) a. *to alle men*
to all.PL man.PL
(*Aelred of Rievaulx's De Institutione Inclusarum*, 36.291)

b. *wit hote terys*
with hot.PL tear.PL
(*Aelred of Rievaulx's De Institutione Inclusarum*, 42.485)

This type of redundancy in language has been argued to be subject to loss in cases of language contact characterised by a high number of adult L2 speakers. Trudgill (2010), for example, argues that in the Norwegian dialect of Bergen predicative adjectives lost plural marking due to intense contact with Low German speakers (2). This is in line with work that suggests that adult learners have problems with redundancy due to its high memory cost (see for example Sagarra 2008).

(2) a. *vi er trøtt-∅* (Bergen Dialect)
we.PL are tired-PL

b. *vi er trøtt-**e*** (Other Dialects)
we.PL are tired-PL

In this respect, English was no exception. The Middle English period was characterised by a general loss of inflectional (redundant) morphology (see Curzan 2003, for the loss of gender agreement, and Allen 1997, for the loss of case). It has been argued that this loss was accelerated due to contact with Scandinavian speakers (Algeo & Pyles 1993: 127–9), but few studies offer a nuanced exploration of its precise impact, as it is often difficult to disentangle internal from external causes of change. The aim of this section is to show how the historical corpus approach can help us not only tackle the question of whether contact had an effect or not, but can also provide insights to more fine-grained questions such as:

1. WHAT type of redundancy is affected and to what degree: did all redundancy behave in the same way? Was plural morphology in quantifiers and adjectives affected to the same extent?

2. WHEN and WHERE: the contact with Scandinavian speakers lasted from roughly the 8th century until the end of the 11th. Which were the first texts that showed a change in the diachronic pattern, and how long did it take for these changes to spread to other dialectal areas?

3.2. What, where and when: Corpus study

We analysed 47 prose texts taken from the Penn–Helsinki Parsed Corpus of Middle English (Kroch et al. 2013), focussing on constructions consisting of a strong monosyllabic adjective/

quantifier followed by a plural noun (a total of 4,366 tokens were annotated).[6] The reason for looking at monosyllabic modifiers only is that by 1250 the plural -*e* ending was present only in monosyllabic strong adjectives and quantifiers, polysyllabic ones were uninflected (Baugh & Cable 2002: 146). Given that adjectives tend to contain more syllables than quantifiers do, we only compared elements with the same number of syllables, in order to prevent phonological processes from affecting the analysis.

How was syntagmatic redundancy affected? If one looks at the behaviour of both adjectives and quantifiers, there are some texts in which quantifiers show less agreement than strong adjectives (texts in dark in Figure 1), and another set of texts in which there is no difference between the two categories (texts in light).[7] That is, in some texts agreement in quantifiers is lost earlier than agreement with adjectives. Importantly, this difference does not depend on the date of composition of the manuscript, but rather on the dialectal region. The majority of texts which show a difference between quantifiers and adjectives come from areas where Scandinavian contact was intense. We calculated a linear regression model in R (R Core Team 2021) with the difference in agreement between adjectives and quantifiers as our dependent variable and region as independent variable (number of tokens per region: East Midlands = 2,063, North = 350, South = 875, West Midlands = 1,078). We used South as the reference level as this was the region with the least Scandinavian contact. The results showed a significant effect of region for the North ($\beta = 0.39$, SE $= 0.15$, $t = 2.6$, $p = 0.01$) and the East Midlands ($\beta = -0.2$, SE $= 0.10$, $t = -2.09$, $p = 0.04$), but not the West ($\beta = -0.13$, SE $= 0.13$, $t = -1.0$, $p = 0.3$).[8] This geographical distribution suggests that contact with Scandinavian speakers played a relevant role in the loss of inflectional morphology, but equally importantly, it shows that concord in quantifiers was more affected than concord with adjectives.

These results suggest that, contrary to the view that all 'redundancy stands for higher complexity because it defies the one-to-one mapping of form and function' (Audring 2014: 6), redundancy can be affected differently depending on syntactic category. Studies on L2 acquisition corroborate this point, showing that the inherent difficulty of redundancy is not homogenous across syntactic categories (for example, L2 learners of Spanish produce more agreement on determiners than on adjectives; see Hawkins 1998). These results, thus, highlight the importance of exploring the role of L2-difficulty when defining notions such as contact-induced simplification and complexification, and show how corpus studies can complement experimental approaches (see section 2.2).

When and where did the development take place? The first text in which quantifiers show less agreement than adjectives is the Northern text *Richard Rolle's Prose Treatises*, composed between 1250 and 1350 (the manuscript dates from 1420 to 1500). After this point in time, the pattern spread to other dialects, starting with the East Midlands text *Mandeville's Travels*, dating back to 1400, and reaching the south by 1450, in the text *A Late Middle English Treatise on Horses*.

[6] For data, statistical analyses, annotation guidelines and metadata on textual sources, see the section on 'Data and code availability'.

[7] There were three texts that show the opposite pattern (quantifiers had more agreement than adjectives), but for reasons of space we leave them out from the discussion.

[8] Regression residuals were checked using the DHARMa package (Hartig 2022); we found no evidence for significant deviations from uniformity.

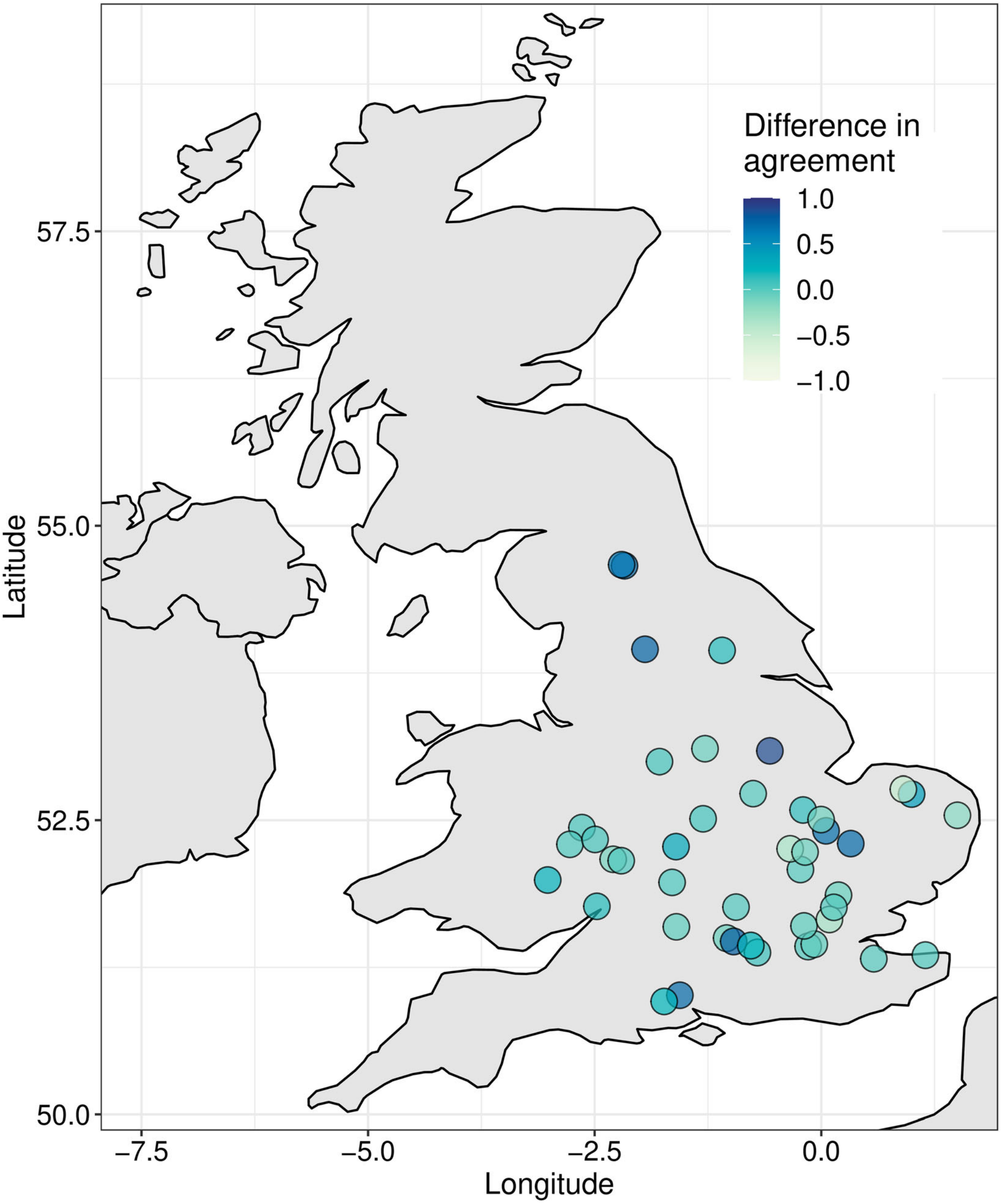


Figure 1. Difference in agreement between quantifier and adjective per text. Positive scores mean adjectives display more agreement than quantifiers; negative scores mean quantifiers have more agreement than adjectives (−1 and 1 are the extremes)

4. CASE STUDY: NULL SUBJECTS

4.1. *What: Null subjects*

Our second case study addresses the development of the null subject system in certain varieties of Latin American Spanish. Spanish has long been used alongside Italian as a poster child of consistent null subject languages (NSLs) (e.g. Rizzi 1982). That is, sentences like (3) that omit the subject pronoun are completely grammatical.

(3) *(Yo) hablo español todos los días.*
(I)* speak Spanish every day.

However, overt subject pronouns have been observed at higher rates in Spanish varieties spoken throughout Latin America (Toribio 2000; Travis 2005; Otheguy & Zentella 2007; Orozco & Guy 2008; Camacho 2013; Alfaraz 2014; Cerrón-Palomino 2018, among many others). Brazilian Portuguese (BP) has also seen an even higher rise in overtness, to the point that it has been recategorised as a partial NSL (PNSL) (Duarte 1993; Erker & Guy 2012). This development could be a change à la van Gelderen's (2011) subject cycle in which these varieties would be transitioning from an NSL to a non-NSL, just as French did. Latin American Spanish and BP are also linked in that both their speakers came into contact with enslaved Africans brought over during the colonial period.

4.2. *Who, when and where: The contact situation*

This contact scenario fits in with Trudgill's (2011) sociolinguistic typology as described in section 2. First, acquisition studies (e.g. Bini 1993; Pérez-Leroux & Glass 1999; Margaza & Bel 2006) suggest null subjects are L2-difficult, particularly for adults. If this is the case, then an increase in overt subjects would be an act of simplification brought on by loose-knit,[9] short-term, adult language learning. This describes exactly the colonial contact situation for enslaved Africans taken to the Caribbean and coastal South America. For many centuries, slave labour in the Caribbean was characterised by work on small- or medium-sized *haciendas* where they worked in smaller populations alongside white and *mestizo* (people of mixed European and Amerindian descent) labourers rather than the large-scale plantation system that would later come into effect in the nineteenth century (Sessarego 2015). While this first generation of slaves would have still needed to acquire a new language within a looser social network, their acquisition was smoother than that of the huge populations of slaves found in French and English colonies, explaining the comparative lack of Spanish creoles (cf. Mintz 1971; Megenney 1984; Chaudenson 1992). However, although the Spanish demographic situation was not as drastic, these African slaves would have still had to quickly learn Spanish as adult second language learners without a dense network to learn from, likely struggling with the L2-difficult null subject system. As a result, they would overuse overt subjects and their children would have nativised their system. Sessarego (2013) proposes precisely this as the origin of the Afro-Hispanic Languages of the Americas (AHLAs), which are the varieties spoken by the current descendants of the original enslaved Africans. Encouragingly, they show the exact kind of change this model predicts and most relevant to us, a high rate of overt subjects.

[9] The work started in and continued from Milroy (1980), Milroy & Milroy (1985) and Milroy (1992) establishes the key difference between dense and loose social networks. The former are characterised by communities where each member knows every other, leading to strong social ties that favour continuity and stability. The opposite holds for loose networks which in turn provide less incentive for continuity, enabling change.

The implications from this combination of patterns from acquisitional studies, current fieldwork and sociolinguistic typologies are very promising. However, the only way to truly verify this proposed origin of the increase in overt subject usage is to use historical corpus data. Specifically, we need to track whether a steady rise in overt pronouns following the point of contact between African and Spanish speakers can be found. The argument would be further strengthened if such an increase were higher in regions with larger populations of Afro-Hispanic speakers.

4.3. Methodology and corpus

Several methodological issues arise when both compiling and annotating a corpus for null subjects. First, historical written corpora pose a series of issues in and of themselves, inherently consisting of only written texts. This means they cannot reliably reflect oral speech which hinders the tracking of a highly oral change. That is not to say that all written texts are comparable. Indeed, balancing a corpus to make it as homogenous as possible is paramount. Traditionally, this means accounting for genre; however, genre is not necessarily enough for a diachronic corpus. For instance, a genre's level of orality (which we have measured quantitatively as an ORSCORE) can shift over time (Rosemeyer 2019), which has been found to have a positive correlation with overtness in our own corpus (Figure 2). We calculated a linear regression model in R (R Core Team 2021) with the proportion of overt pronominal subjects per text as the dependent variable and the degree of orality[10] of each text as the independent variable. Results showed a significant effect of orality ($\beta = 0.10$, $p < 0.0001$, $R^2 = 0.35 \pm 0.064$).[11]

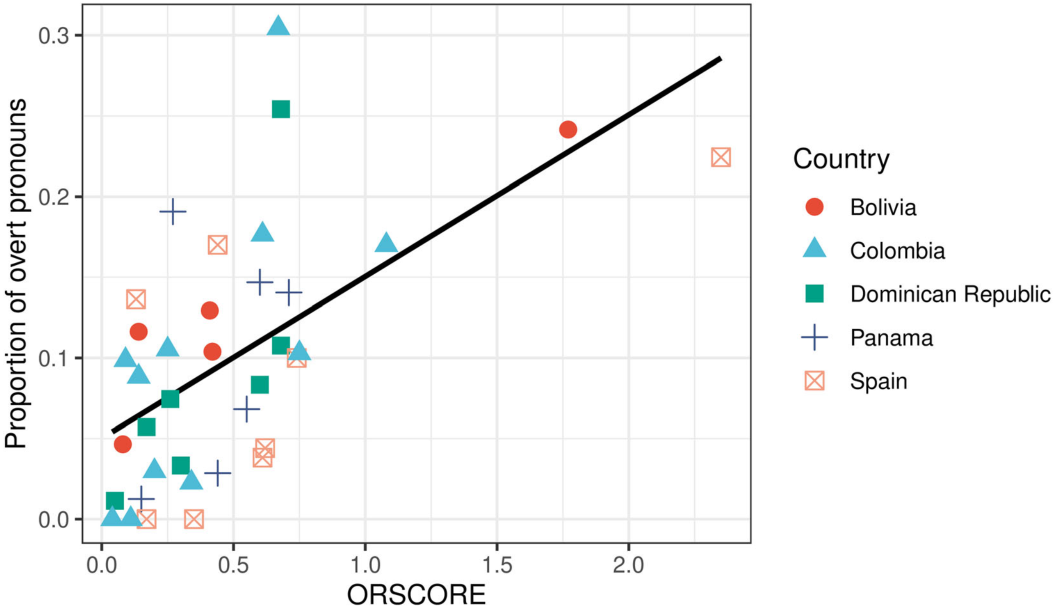


Figure 2. Orality vs. overtness

[10] The ORSCORE was calculated based on the adjusted frequencies of four of the five oral variables applied in Rosemeyer (2019): private verbs, the progressive, neuter demonstrative pronouns, and time and place adverbs. The total frequencies per text were added up, divided by word count, and then multiplied by 100 to create the final score.

[11] We again checked the regression residuals using DHARMa (Hartig 2022), finding no evidence for nonuniformity.

Additionally, the early enactors of this change, enslaved L2 learners and their subsequent generations, were significantly less likely to be literate in Spanish or have access to publishing. Thus, the extant texts were predominantly written in the standard.

The next hurdle is which corpus to use to investigate this development. Surprisingly, there is no syntactically parsed historical Spanish corpus comparable to those for English, French, Portuguese, etc. There especially is not a dialectal one. Although a parsed corpus would be helpful in compiling finite clauses already tagged for clause type and overt subjects, it admittedly would not mark null subjects, necessitating a degree of annotating by hand regardless. There are three major historical corpora: CDE (Davies 2002), CORDE/CDH (Real Academia Española 2013, n.d.), and CORDIAM (Academia Mexicana de la Lengua, n.d.). We can rule out the CDE's historical corpus immediately as it unfortunately is not tagged by region or country. CORDE/CDH and CORDIAM then have the same issue of not being able to be searched for finite verbs. The option that remains is to create your own corpus, pulling freely available texts from databases such as Cervantes Virtual (Biblioteca Virtual Miguel de Cervantes 1999), BDH (La Biblioteca Nacional de España 2008), and DLOC (Digital Library of the Caribbean 2004). Working from continuous samples comes with the convenience of more context for each clause which aids tagging for priming, topichood, switch reference, etc. This brings us to the next series of methodological choices: how to annotate null subjects.

The first choice one will need to make is if they are going to limit their scope by person. Choosing to just study first person can simplify matters a great deal as it eliminates potential person effects; for example, first and second person have been shown to often favour overt realisation (e.g. Shin & Erker 2015). Additionally, third person comes with its own set of problems as one will have to make decisions regarding expletives and nonreferential/impersonal constructions. Whether to count coordinated subjects is another preference that needs to be determined. These choices have significant effects on the frequency calculated, rendering comparison with other studies' counts tricky. There are many other linguistic variables that the previous literature has found to have an effect on pronoun realisation that one might also want to account for when annotating, including switch reference, focus, emphasis, topichood, information structure, priming, TAM morphology, clause type and syntactic/semantic verb class.

Our own corpus consists of over 50 texts from seven countries with attested AHLA speaking communities (Dominican Republic, Cuba, Panama, Colombia, Venezuela, Bolivia, and Peru), plus Spain as a control. To establish a baseline, the earliest texts are by peninsular Spanish speakers who settled in their respective countries in the sixteenth century. The corpus then continues through the nineteenth century and is split into two genres (literary and nonliterary). Texts were first chosen from a list of famous national authors from each country. Once those options were exhausted, we turned to the archives and libraries listed above to find the rest of the texts. The first text found that met the country, century and genre specifications for a cell in the corpus was selected. In addition to these main texts, the corpus was supplemented by texts found in Afro-Hispanic vernaculars. Each text was sampled for 2,000–3,000 word excerpts that were transcribed, tokenised and further annotated by hand. Only finite clauses were considered, impersonal/expletive constructions were tagged but set aside, and coordinated subjects were counted as one token.[12]

[12] Coordinated overt pronouns are so rare in the corpus that this choice really only has an effect on the frequency of null subjects.

4.4. *Preliminary results*

The first five countries of our corpus seem to show an increase in overt subjects over the last few centuries for every country apart from the Dominican Republic, as seen in Figure 3.[13]

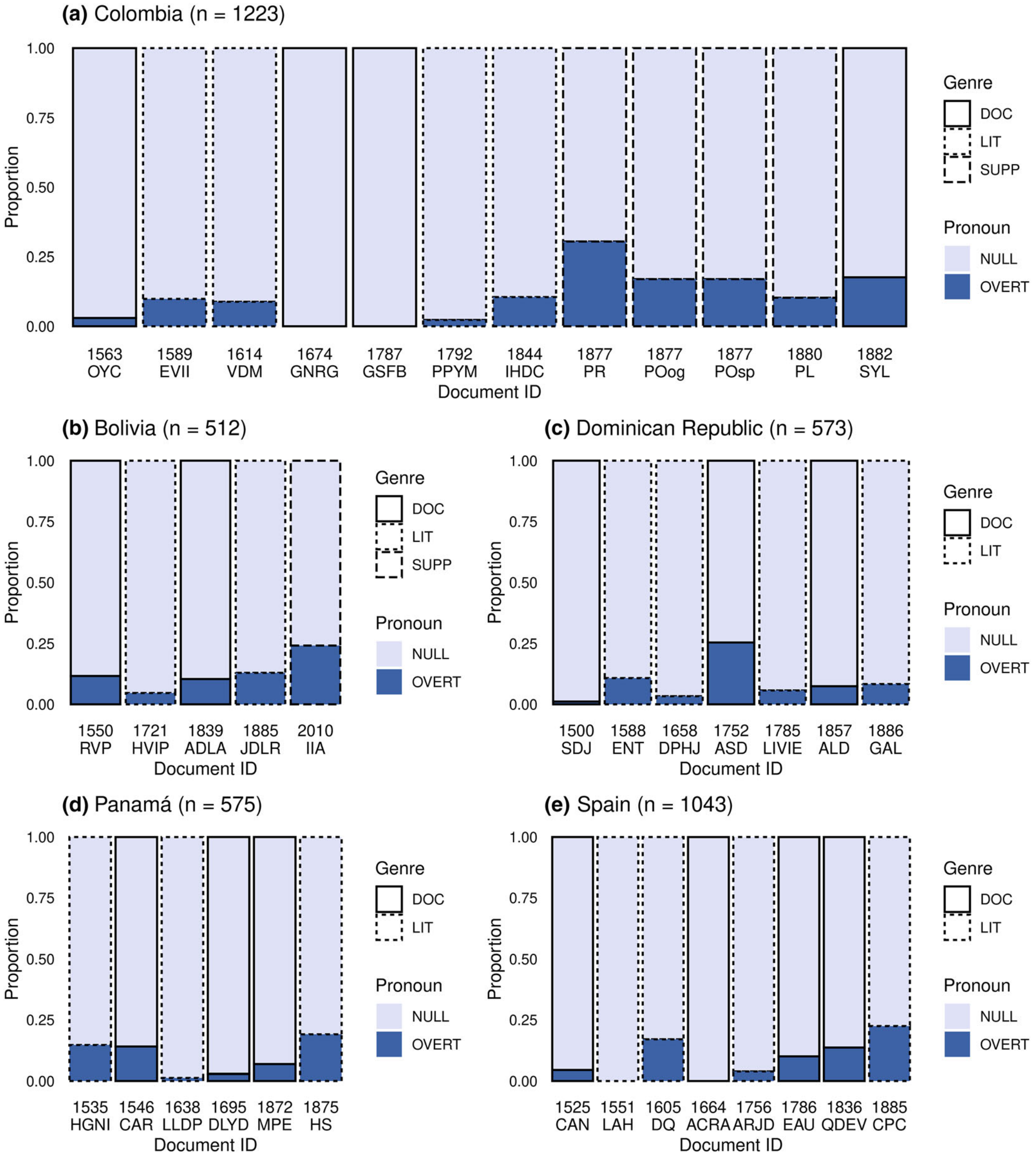


Figure 3. Pronoun realisation by region. DOC, document; LIT, literary; SUPP, supplement genre. For explanation of document IDs, see the section on 'Data and code availability'

[13] The document ID is an abbreviation of the title of a text (e.g. ENT = *Entremés*), and is given together with document year. The document genre is illustrated with a solid border, the literary genre with a dotted border. The bars with a dashed border are supplemental texts from Afro-Hispanic speakers. For full metadata on the text sources, including the meaning of the document ID abbreviations, see the section on 'Data and code availability'.

However, this trend also holds for Spain, which contradicts our sociolinguistic predictions. Since the orality effect discussed in section 4.3 may be at play in the bar charts, it is necessary to run a mixed-effects model to confirm any relationships across countries and centuries. A mixed model using glmer from the lme4 package (Bates et al. 2015) in R (R Core Team 2021) was run to find any such underlying patterns. The model had pronoun realisation (3,773 total tokens) as the binary-dependent variable; macroregion (Spain vs. Non-Spain), orality and year (both z-scored) as fixed effects; and document ID as a random effect. Country was originally included as a fixed effect but had too many levels and was replaced with macroregion. The interaction between year and orality was significant ($\beta = -0.44$, SE $= 0.21$, $p < 0.03$), where the negative coefficient means that for every year that passes, the effect of orality on overtness is lessened.[14] Thus, there is an underlying diachronic effect, but as we can see from Figure 3, it is a small effect. Macro-region was also found to be just significant ($\beta = 0.63$, SE $= 0.32$, $p < 0.05$), favouring overtness in every country but Spain. This result explains the apparent rise we saw for Spain in Figure 3 as just a product of orality: the later Spanish texts score higher in ORSCORE.

As was emphasised in section 2, a full account of historical situations such as this must ultimately also take into account the effect of population dynamics, and more specifically, any potential complex dependencies that may obtain between demographic factors, on the one hand, and psychological factors such as L2-difficulty, on the other. It is possible that the slow increase in overt pronoun realisation is because, irrespective of the difficulty faced by adult L2 learners, child L1 learners still have sufficient evidence in their linguistic input for favouring the null subject system over the competing overt subjects system. As we discuss in more detail in section 6, recent work in complex-systems modelling suggests that a critical threshold may exist which the proportion of L2 learners in a population must exceed for population-level simplification to occur. If this threshold is not met, then incipient simplificatory changes will ultimately fail to percolate through the entire population, that is the simplifications introduced by L2 learners will not be faithfully regularised by subsequent generations of L1 learners.

Despite the orality effect detected in section 4.3 rendering the corpus imperfectly balanced and potentially obscuring the trends we seek, its discovery yields important insights. Our corpus demonstrates that subject pronouns are more likely to be overtly realised in highly oral texts, which better contextualises the low frequencies we see in the preliminary data. Fortunately, a diachronic trend was found by the mixed model, providing information on WHEN the change took place. A regional distinction between Spain and the rest of the data was also found, establishing the WHERE of this change. We are hopeful that future work will be able to continue teasing apart these effects.

5. CASE STUDY: CASE

5.1. What: Case/adposition function variables

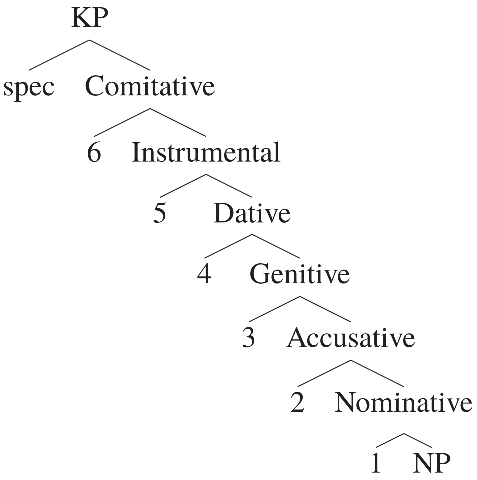
A widespread, well-known and well-documented diachronic pattern in the Indo-European language family is the loss of morphological case and subsequent rise of adpositions, which take over the cases' functional load. This can be seen in the Romance, Germanic, Baltic, Slavic, Indo-Aryan and Celtic families among others (Hewson & Bubenik 2006). It is these functions, which case markers and adpositions both have the ability to realise, that are focussed on here: grammatical (e.g. subject, direct object) and oblique/semantic (e.g. ablative, locative). We take Blake's (1994) Case Hierarchy as the theoretical basis for the makeup of

[14] We found no evidence for nonuniformity of residuals, using DHARMa (Hartig 2022).

case inventories and relationship between the cases (4), plus a Caha-esque KP (Caha 2009) for the syntactic representation of both cases and adpositions (5).

(4) nom > acc/erg > gen > dat > loc > abl/ins > other

(5)



In terms of L2-complexity, we propose two variables concerning the adpositional and case marking realisation of grammatical function. The first is adpositional vs. case marking expression in the manner just discussed: the adpositional expression of a function is the ‘simple’ value of this variable, and the case marking expression is the ‘complex’ value. This means here that the L2-difficulty in acquiring a case marker is higher than that of an adposition. The second variable concerns syncretism: a distinct case marker/adposition expressing a function is the ‘complex’ value, and a syncretic form is ‘simple’. This is based on the number of forms to be acquired: at this stage, we do not take into account fusional forms marking other grammatical categories as well as case, such as number or gender.[15]

We expect to see grammatical systems with higher proportions of adpositional expression and syncretic forms in languages with histories of high proportions of adult L2 speakers. In addition to this, one expects syncretisms to arise, which involve only contiguous regions of Blake’s hierarchy and the exploded KP, and a change from case marking realisation to adpositional realisation moving down the KP projection/down the hierarchy. This is broadly what we can already observe in Indo-European: remaining cases are those lower in the hierarchy.

5.2. The where *and the* who

The Balkan Sprachbund is well known for its long and intense contact situation and for the many shared syntactic ‘Balkanisms’ found in its languages which do not appear to have arisen through (internal) language shift. It contains languages with varying degrees of case loss. Lindstedt (2000, 2018) argues that the interplay of the sociolinguistic factors found in the Balkans predicts the simplification of the case systems, which happened most sharply in the Balkan Slavic languages (characterised here as Bulgarian and Macedonian and various

[15] There is clearly a lot more to be said about the various uses and ‘levels’ of case in terms of theory and L2-difficulty. We set aside here questions of structural Case where uninterpretable features come into play; see Walkden & Breitbarth (2019a: 200–2) for tentative discussion.

dialects). All seven inflectional cases from Indo-European except for the ablative have been lost in the standard languages (Friedman 2017). As Wahlström (2015) details, this case loss in Balkan Slavic took place mostly between the eleventh and sixteenth centuries.

Whereas Trudgill (2011) uses the Sprachbund as an example for his long-term complexification type, Lindstedt (2000, 2018) reports there is 'no reason' to think that there was much child bilingualism occurring; that it was predominantly adult men coming into contact with speakers of other languages. That is, the contact in Balkan Slavic is long-term, but among L2 adults. He also argues that Balkan Slavic had the highest rate of this adult L2 multilingualism due to prestige, actually enacting the simplification-type change.

The traditional historical-comparative method already shows the simplification of the case paradigms through case syncretism and loss of case forms in Balkan Slavic. The corpus method employed here is used to track the diachronic realisation of function, as opposed to (just) the rate of simplification in the case paradigm. This can therefore more accurately state in WHAT manner and to what extent simplification occurred and for WHO and WHERE, insofar as this extralinguistic information is available.

5.3. *Corpus and methodology*

The study uses the Diachronic corpus of prestandardised Balkan Slavic (Šimko 2021; Šimko & Escher 2021; Šimko et al. 2021), a collection of texts from the 1300s to 1928, with the bulk of the texts falling between 1700 and 1800. They all come from the Balkan Sprachbund area, mostly from modern-day Bulgaria. Some of the existing morphosyntactic tagging is used in data collection. These tagsets are the part of speech (PoS) tags and the syntactic Universal Dependency tags, noting the function of some of the nominals. Some examples of the latter tags are subject, object and direct object. Furthermore, more in-depth tagging of this functional nature was implemented, including on the adpositions. In order to be able to directly compare the function of adpositions and case marking, a more fine-grained distinction was made between the traditional case labels, which are often ambiguous and multifunctional. For instance, the associative function was distinguished from the comitative, and the dative split into the goal and recipient functions. These functions were then ordered in their relative positions in the KP based on their syncretic case and adposition forms, following Caha's original KP. Each instance of a nominal (noun, adjective and pronoun) plus adposition was tagged with one of these function tags based on the role played by the nominal phrase (or PP) in the larger phrase or clause.

The nature of the simplification predicted regarding case is as follows: It is predicted that at all times, the adpositional realisation proportion is higher going up the KP, and that the adpositional realisation proportions for each function can only increase over time.

5.4. *Preliminary results*

Figures 4 and 5 show preliminary results of some of the adpositional realisation proportions over 19 texts from the 1300s to 1860.

The instrumental function shows a clear increase in adpositional realisation, though many of the functions increase after c. 1500. The comitative and associative functions are consistently high in adposition realisation, and the direct object and subject functions consistently low, both of which are predicted. The synchronic and diachronic adposition proportions of the locative and ablative functions do not follow our predictions (i.e. they are often higher than the comitative and the instrumental particularly), indicating that, in some sense, they do not participate in the KP as the other functions do. Overall then, regarding the functions based on the *original* KP functions (comitative, associative, instrumental,

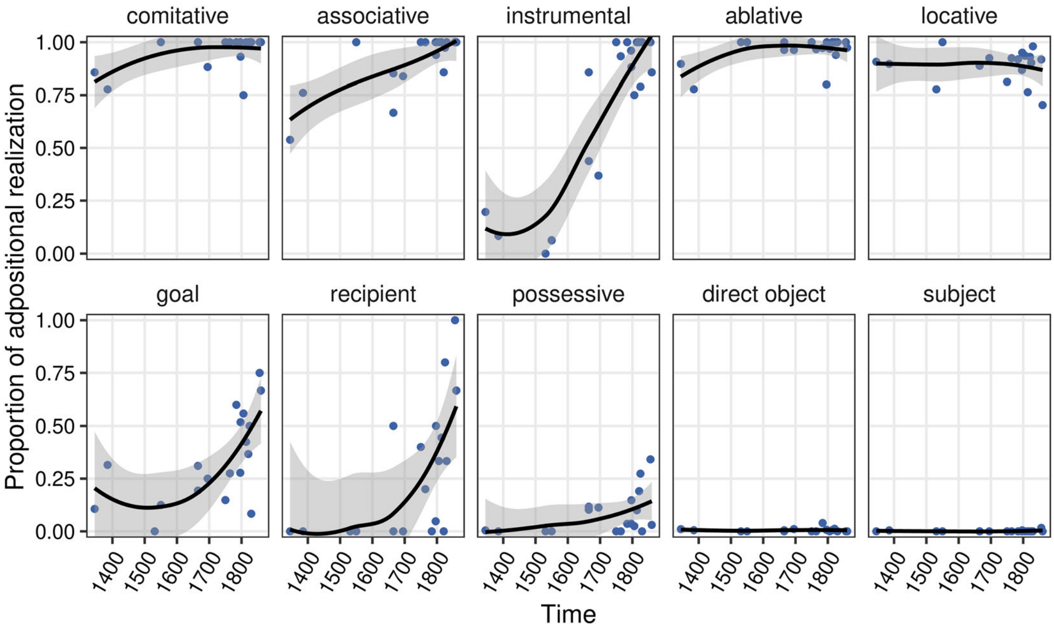


Figure 4. Diachronic proportions of adposition realisation. Each point represents a single text; curves give LOESS regressions together with confidence bands. Functions are ordered as in the KP, from top to bottom

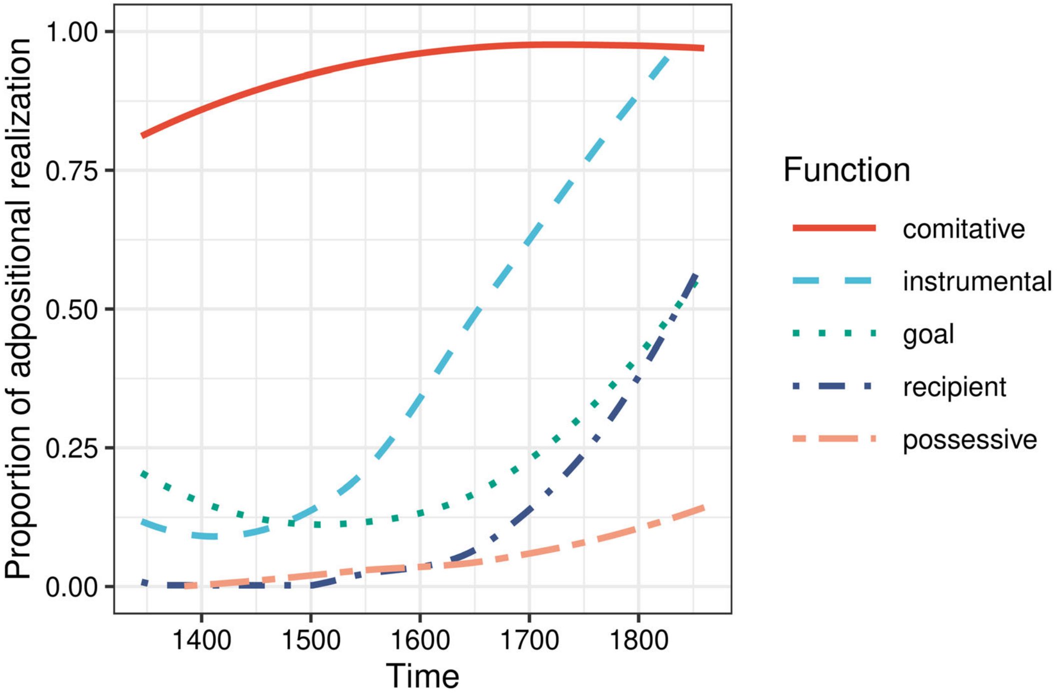


Figure 5. Diachronic proportions of adposition realisation for functions comitative, instrumental, goal, recipient and possessive (LOESS). Functions are ordered as in the KP, from top to bottom

benefactive, goal, recipient, possessive, direct object and subject), the corpus shows the predicted functional takeover of adpositions going down the KP.

We believe these preliminary data illustrate how corpus data can be used to track the already reported simplification regarding case paradigms and furthermore uncover the nature of such case loss by detailing the functional takeover by adpositions.

6. THE POPULATION PERSPECTIVE: HOW CONTACT-INDUCED CHANGES (DO NOT) PROPAGATE

We have discussed the relative merits of three kinds of approaches to sociolinguistic typology—typological, experimental and corpus-based—emphasising that these approaches are complementary, not competing. In closing, we will briefly discuss a fourth approach, which is complementary to these three and has the potential to unify them.

It is often difficult or impossible to scale empirical studies up to the population sizes and structures typical of real-life speech communities. A MECHANISTIC MODELLING approach can be used to achieve this, since far larger systems can be studied in computational models than is typically possible to do empirically. A mechanistic model of an empirical phenomenon views that phenomenon as a system of interacting elements; by making theoretically motivated assumptions about those elements and their interactions, the model then predicts a limited set of outcomes, which may be compared against empirically collected data (Lindsey 2001; Craver 2006; Baker et al. 2018).

To illustrate this, we will briefly discuss contact-induced simplification, leaving complexification processes for future study. Here, the crucial population-level question boils down to the following: just how many (adult) L2 learners must be present in a speech community for community-wide simplification to occur? Although typological studies have suggested that higher L2 speaker proportions increase the likelihood of simplification, and although Trudgill (2011: 57–8) has proposed that L2 speaker proportions on the order of 0.5 should lead to simplification, neither of these approaches can provide a conclusive answer to the question: the first is a statement solely about the direction of an effect, and the second is anecdotal.

In the theory of dynamical systems, a BIFURCATION refers to any major qualitative change to a system's behaviour prompted by minute quantitative variation in a control parameter. In our case, that qualitative change concerns whether simplification occurs or not, and the relevant control parameter is the proportion of L2 speakers. Let σ refer to that proportion (a number between 0 and 1). Is there a bifurcation point, a critical value σ_{crit} such that for actual proportions $\sigma > \sigma_{\text{crit}}$, simplification is predicted to occur, whereas with $\sigma < \sigma_{\text{crit}}$ full simplification is not expected? Kauhanen (2022) studied this with a model in which both L1 and L2 speakers are modelled as variational learners (Yang 2002), albeit the latter experience an additional L2-difficulty in acquiring one of the competing grammatical options. A bifurcation threshold was indeed found to exist in this model. The model thus replicates the central intuition about simplification from Trudgill's (2011) theory.

It is instructive to consider in a bit more detail what the ingredients are that enter the equation for the bifurcation threshold. Mathematically, the threshold is

$$(6) \qquad \sigma_{\text{crit}} = \frac{(\alpha - 1)(D + 1)}{\alpha D} = \left(1 - \frac{1}{\alpha}\right)\left(1 + \frac{1}{D}\right),$$

where D is a measure of the L2-difficulty suffered by one of the two competing grammatical options, and α is a measure of how much advantage the L2-difficult grammar has in L1 acquisition, relative to the competing option (see Kauhanen 2022 for details on how these quantities are calculated). Intuitively, we should expect higher values of D to lower the bifurcation threshold σ_{crit} and higher values of α to increase it—simplification should be more likely the more difficulty L2 learners incur in acquiring the L2-difficult option, but also less likely the more L1 learners favour the L2-difficult option. In (6), we show two algebraically equivalent forms of the bifurcation threshold; the form on the right makes it clear that α and D indeed have this effect on σ_{crit}.

Particularly, when α has a large value, so that its reciprocal $1/\alpha$ is very small, the predicted critical threshold σ_{crit} may be so high that it is not met in actual empirical situations. One of the case studies considered in Kauhanen (2022) was null subjects in Afro-Peruvian Spanish (on this variety, see Sessarego 2015). This case study is interesting in the light of our preliminary results concerning null subject systems in AHLAs (section 4). For null subjects in Afro-Peruvian Spanish, Kauhanen (2022) estimated the critical simplification threshold to lie at about $\sigma_{\mathrm{crit}} = 0.9$, precisely because of the high value of α in this case.[16] However, empirical estimates put the actual historically attested proportion of L2 learners in the relevant geographical regions between $\sigma = 0.2$ and $\sigma = 0.6$. In other words, the model predicts only partial simplification since the actual proportion of L2 learners never exceeded the critical threshold in the historical situation. Although more research is clearly needed (for a start, the model assumes a fully mixing population without any social network or stochastic effects), this is in line with the small degree of simplification—the slow growth in overt subjects—observed in our own corpus-based investigation (section 4).

The population perspective is important not only for the intrinsic interest of population-dynamical models; it also plays a crucial role in uniting the other approaches. Linguistic theory enters the picture in the form of an analysis of which variants are in competition, and why. The psychology of learning (particularly the specifics of L2 acquisition) enters the picture in terms of the learning algorithm assumed by the model and in terms of the L2-difficulty parameter. Demographics enter the picture by way of estimates of the relevant population proportions at the time of the change under study. Finally, corpus evidence enters the picture at least in two ways: firstly, by providing empirical estimates of the distributional advantages enjoyed by the competing grammars, and second, by providing a unique longitudinal description of the change process; this, crucially, allows us to ask questions such as whether the duration of change predicted by the mechanistic model aligns with the empirically observed duration.

7. CONCLUSION

Contact-induced changes in morphosyntax can be investigated using a variety of methods. Typological-correlational studies and experimental studies have been used with great success to investigate WHO is responsible for change, WHERE change takes place, and HOW mechanisms of cognition and interaction at the individual level produce different outcomes. In this paper, we have made the case for a third, complementary approach that utilises the historical textual record to its full potential: a sociolinguistically informed historical-corpus-based approach. By taking a fine-grained view of time, place and text type, this approach can shed light on the WHERE and WHEN of actual contact-induced changes, as well as HOW changes spread through a population.

[16] This high value essentially follows from the fact that a null subject grammar has far greater relative parsing advantage (cf. Yang 2000) than the competing overt subjects grammar: the latter scores a win with expletive subjects only, which are in the minority in production. Thus, L1 learners tend to favour null subjects, and consequently a relatively high frequency of L2-learner-induced simplifications is required for the overt subject system to completely take over. See Kauhanen (2022) and Simonenko et al. (2019) for details.

We have presented three case studies in support of the approach, focussing on different geographical areas (Britain, Latin America, the Balkans) and different linguistic features (number concord, null subjects, case). Crucial to all three studies is a precise linguistic characterisation of WHAT is changing, and an account of why this linguistic feature is suspected to be L2-difficult. The case study of number in Middle English in section 3 suggests a possible role for simplificatory contact with speakers of Scandinavian, with number concord on quantifiers lost more and earlier than number concord on adjectives. The case study of null subjects in Latin American Spanish in section 4 highlights effects of year and macroregion on overtness, which simultaneously supports the case of simplification and the predictions of the population-level model (section 6) once orality is accounted for. The case study of case in Balkan Slavic in section 5 shows that the predictions of case hierarchies are also borne out in the diachronic domain as morphological case is lost. Quantitatively rich datasets such as these can also serve as a testing ground for explanatory models of HOW L2 speaker proportions affect change at the population level, as section 6 makes clear.

Together, these case studies illustrate the added value of a historical corpus approach to the hypotheses of Trudgill's (2011) sociolinguistic typology, and to questions of historical language contact effects more broadly.

DATA AND CODE AVAILABILITY

All data and code necessary to replicate our results, as well as metadata on textual sources, can be obtained from https://github.com/erc-starfish/5wh or from https://doi.org/10.5281/zenodo.8392620.

Correspondence
George Walkden
University of Konstanz
Department of Linguistics & English Language
78457 Konstanz
Germany
Email: george.walkden@uni-konstanz.de

REFERENCES

ALFARAZ, GABRIELA G., 2014. 'Variation of overt and null subject pronouns in the Spanish of Santo Domingo', in Ana M. Carvalho, Rafael Orozco, & Naomi Lapidus Shin (eds.), *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*. Washington, DC: Georgetown University Press.

ALGEO, JOHN & THOMAS PYLES, 1993. *The Origins and Development of the English Language*. New York: Harcourt Brace Jovanovich.

ALLEN, CYNTHIA, 1997. 'Middle English case loss and the 'creolization' hypothesis', *English Language and Linguistics* 1 (1). 63–89.

ATKINSON, MARK, KENNY SMITH & SIMON KIRBY, 2018. 'Adult learning and language simplification', *Cognitive Science* 42(8). 2818–2854.

AUDRING, JENNY, 2014. 'Gender as a complex feature', *Language Sciences* 43. 5–17.

AUER, ANITA, CATHERINA PEERMANS, SIMON PICKL, GIJSBERT RUTTEN & RIK VOSTERS, 2015. 'Historical sociolinguistics: The field and its future', *Journal of Historical Sociolinguistics* 1(1). 1–12.

BAKER, RUTH E., JOSE-MARIA PEÑA, JAYARATNAM JAYAMOHAN & ANTOINE JÉRUSALEM, 2018. 'Mechanistic models versus machine learning, a fight worth fighting for the biological community?', *Biology Letters* 14(5). 20170660.

BATES, DOUGLAS, MARTIN MÄCHLER, BEN BOLKER & STEVE WALKER, 2015. 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* 67(1). 1–48.

BAUGH, ALBERT C. & THOMAS CABLE, 2002. *A History of the English Language*. London: Routledge.

BENTZ, CHRISTIAN & BODO WINTER, 2013. 'Languages with more second language learners tend to lose nominal case', *Language Dynamics and Change* 3(1). 1–27.

BERDICEVSKIS, ALEKSANDRS & ARTURS SEMENUKS, 2022. 'Imperfect language learning reduces morphological overspecification: Experimental evidence', *PLoS ONE* 17(1). e0262876.

Bini, Milena, 1993. ‘La adquisición del italiano: más allá de las propiedades sintácticas del parámetro pro-drop’, in Juana M. Liceras (ed.), *La lingüística y el análisis de los sistemas no nativos*. Ottawa: Dovehouse. 126–139.

Blake, Barry J., 1994. *Case*. Cambridge: Cambridge University Press.

Blaxter, Tam, 2017. *Speech in Space and Time: Contact, Change and Diffusion in Medieval Norway*. Cambridge: University of Cambridge dissertation.

Caha, Pavel, 2009. *The Nanosyntax of Case*. Tromsø: University of Tromsø dissertation.

Camacho, José A., 2013. *Null Subjects*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524407

Cerrón-Palomino, Álvaro, 2018. ‘Variable subject pronoun expression in Andean Spanish: A drift from the acrolect’, *Onomázein* 42(42). 53–73.

Chaudenson, Robert, 1992. *Des îles, des hommes, des langues*. Paris: L’Harmattan.

Conde-Silvestre, Juan Camilo & Juan M. Hernández-Campoy (eds.), 2012. *The Handbook of Historical Sociolinguistics*. Malden, MA & Oxford: Wiley-Blackwell.

Craver, Carl F., 2006. ‘When mechanistic models explain’, *Synthese* 153(3). 355–376.

Curzan, Anne, 2003. *Gender Shifts in the History of English*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486913

Dryer, Matthew S. & Martin Haspelmath, 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/

Duarte, Maria Eugênia, 1993. ‘Do pronome nulo ao pronome pleno: A trajetória do sujeito no português do Brasil’, in Ian Roberts & Mary A. Kato (eds.), *Português brasileiro: Uma viagem diacrônica*. Campinas: Ed. Da Unicamp. 107–128.

Erker, Daniel & Gregory Guy, 2012. ‘The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish’, *Language* 88(3). 526–557.

Friedman, Victor A., 2017. *Languages of the Balkans*. https://doi.org/10.1093/acrefore/9780199384655.013.348

Hartig, Florian, 2022. *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R package version 0.4.6. https://CRAN.R-project.org/package = DHARMa

Hawkins, Roger, 1998. *The Inaccessibility of Formal Features of Functional Categories in Second Language Acquisition*. Paper presented at the Pacific Second Language Research Forum, Tokyo.

Hewson, John & Vit Bubenik, 2006. *From Case to Adposition: The Development of Configurational Syntax in Indo-European Languages*. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.280

Kauhanen, Henri, 2022. ‘A bifurcation threshold for contact-induced language change’, *Glossa: A Journal of General Linguistics* 7(1). 1–32.

Kauhanen, Henri, Sarah Einhaus & George Walkden, 2023. ‘Language structure is influenced by the proportion of non-native speakers: A reply to Koplenig (2019)’, *Journal of Language Evolution* 8(1). 90–101.

Koplenig, Alexander, 2019. ‘Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers’, *Royal Society Open Science* 6(2). 181274.

Lindsey, J. K., 2001. *Nonlinear Models in Medical Statistics*. Oxford: Oxford University Press.

Lindstedt, Jouko, 2000. *Linguistic Balkanization: Contact-Induced Change by Mutual Reinforcement*. Leiden: Brill.

Lindstedt, Jouko, 2018. ‘Diachronic regularities explaining the tendency towards explicit analytic marking in Balkan syntax’, in Iliyana Krapova & Brian Joseph (eds.), *Balkan Syntax and (Universal) Principles of Grammar*. Berlin: de Gruyter. 70–84. https://doi.org/10.1515/9783110375930-005

Lupyan, Gary & Rick Dale, 2010. ‘Language structure is partly determined by social structure’, *PLoS ONE* 5(1). e8559.

Margaza, Panagiota & Aurora Bel, 2006. ‘Null subjects at the syntax–pragmatics interface: Evidence from Spanish interlanguage of Greek speakers’, in Mary Grantham O’Brien, Christine Shea, & John Archibald (eds.), *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA)*. Somerville, MA: Cascadilla Proceedings Project. 88–97.

Mayr, Ernst, 2004. *What Makes Biology Unique? Considerations on the Autonomy of a Scientific Discipline*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511617188

McIntosh, Angus, 1994. ‘Codes and cultures’, in Margaret Laing & Keith Williamson (eds.), *Speaking in our Tongues: Proceedings of a Colloquium on Medieval Dialectology and Related Disciplines*. Cambridge: D. S. Brewer. 135–137.

Megenney, W., 1984. ‘El habla bozal cubana ¿lenguaje criollo o adquisición imperfecta?’, *La Torre* 33. 109–139.

Milroy, James, 1992. *Linguistic Variation and Change*. Oxford: Blackwell.

Milroy, James & Lesley Milroy, 1985. ‘Linguistic change, social network and speaker innovation’, *Journal of Linguistics* 21(2). 339–384.

Milroy, Lesley, 1980. *Language and Social Networks*. Oxford: Blackwell.

Mintz, Sidney, 1971. ‘The socio-historical background to pidginization and creolization’, in Dell Hymes (ed.), *Pidginization and Creolization of Languages*. Cambridge: Cambridge University Press. 481–498.

Orozco, Rafael & Gregory Guy, 2008. ‘El uso variable de los pronombres sujetos: ¿qué pasa en la costa Caribe colombiana?’, in Maurice Westmoreland & Juan Antonio Thomas (eds.), *Selected Proceedings of the Fourth Workshop on Spanish Sociolinguistics*. Somerville, MA: Cascadilla Proceedings Project. 70–80.

Otheguy, Ricardo & Ana Celia Zentella, 2007. ‘Apuntes preliminares sobre el contacto lingüístico y dialectal en el uso pronominal del español en Nueva York’, in Kim Potowski & Richard Cameron (eds.), *Spanish in Contact: Policy, Social and Linguistic Inquiries*. Amsterdam: John Benjamins. 275–295. https://doi.org/10.1075/impact.22.20oth

Owens, Jonathan, 2001. 'Creole Arabic: The orphan of all orphans', *Anthropological Linguistics* 43. 348–378.

Pérez-Leroux, Ana T. & William R. Glass, 1999. 'Null anaphora in Spanish second language acquisition: Probabilistic versus generative approaches', *Second Language Research* 15(2). 220–249.

R Core Team, 2021. *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing Vienna. https://www.R-project.org/

Raviv, Limor, Antje Meyer & Shiri Lev-Ari, 2019. 'Larger communities create more systematic languages', *Proceedings of the Royal Society B: Biological Sciences* 286(1907). 1–9.

Rizzi, Luigi, 1982. *Issues in Italian Syntax*. Dordrecht: Foris. https://doi.org/10.1515/9783110883718

Roberts, Gareth & Betsy Sneller, 2020. 'Empirical foundations for an integrated study of language evolution', *Language Dynamics and Change* 10(2). 188–229.

Rosemeyer, Malte, 2019. 'Actual and apparent change in Brazilian Portuguese wh-interrogatives', *Language Variation and Change* 31(2). 165–191.

Sagarra, Nuria, 2008. 'Working memory and L2 processing of redundant grammatical forms', in Zhao Hong Han (ed.), *Understanding Second Language Process*. Clevedon: Multilingual Matters. 133–147. https://doi.org/10.21832/9781847690159-009

Sessarego, Sandro, 2013. 'Afro-Hispanic contact varieties as conventionalized advanced second languages', *Iberia* 5. 96–122.

Sessarego, Sandro, 2015. *Afro-Peruvian Spanish: Spanish Slavery and the Legacy of Spanish Creoles*. Amsterdam: Benjamins. https://doi.org/10.1075/cll.51

Shcherbakova, Olena, Susanne Maria Michaelis, Hannah J. Haynie, Sam Passmore, Volker Gast, Russell D. Gray, Simon J. Greenhill, Damián E. Blasi & Hedvig Skirgård, 2023. 'Societies of strangers do not speak less complex languages', *Science Advances* 9(33). eadf7704.

Shin, Naomi & Daniel Erker, 2015. 'The emergence of structured variability in morphosyntax: Childhood acquisition of Spanish subject pronouns', in Ana M. Carvalho, Rafael Orozco, & Naomi Lapidus Shin (eds.), *Subject Pronoun Expression in Spanish*. Washington, DC: Georgetown University Press. 169–190.

Simonenko, Alexandra, Benoit Crabbé & Sophie Prévost, 2019. 'Agreement syncretization and the loss of null subjects: Quantificational models for Medieval French', *Language Variation and Change* 31(3). 275–301.

Sinnemäki, Kaius, 2020. 'Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach', *Journal of Historical Sociolinguistics* 6(2). 20191010.

Sinnemäki, Kaius & Francesca Di Garbo, 2018. 'Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity', *Frontiers in Psychology* 9(1141). 1–22.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera, Nataliia Hübler Miranda, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Wu Henry, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray, 2023. 'Grambank reveals global patterns in the structural diversity of the world's languages', *Science Advances* 9(16). adg6175.

Thurston, William R., 1989. 'How exoteric languages build a lexicon: Esoterogeny in West New Britain', in Ray Harlow & Robert Hooper (eds.), *VICAL I: Papers in Oceanic Linguistics*. Auckland: Linguistic Society of New Zealand. 555–579.

Toribio, Almeida Jacqueline, 2000. 'Setting parametric limits on dialectal variation in Spanish', *Lingua* 110(5). 315–341.

Travis, Catherine E., 2005. 'Priming in subject expression in Colombian Spanish: The yo-yo effect', in Randall Gess & Edward J. Rubin (eds.), *Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers from the 34th Linguistic Symposium on Romance Languages (LSRL), Salt Lake City, March 2004*. Amsterdam and Philadelphia: John Benjamins. 329–349. https://doi.org/10.1075/cilt.272.20tra

Trudgill, Peter, 1977. 'Creolization in reverse: Reduction and simplification in the Albanian dialects of Greece', *Transactions of the Philological Society* 75(1). 32–50.

Trudgill, Peter, 2010. 'Contact and sociolinguistic typology', in Raymond Hickey (ed.), *The Handbook of Language Contact*. Chichester: Wiley. 299–319. https://doi.org/10.1002/9781444318159.ch15

Trudgill, Peter, 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

van Gelderen, Elly, 2011. *The Linguistic Cycle: Language Change and the Language Faculty*. Amsterdam: Benjamins. https://doi.org/10.1093/acprof:oso/9780199756056.001.0001

Wahlström, Max, 2015. *The Loss of Case Inflection in Bulgarian and Macedonian*. Helsinki: University of Helsinki dissertation.

Walkden, George & Anne Breitbarth, 2019a. 'Complexity as L2-difficulty: Implications for syntactic change', *Theoretical Linguistics* 45(3–4). 183–209.

Walkden, George & Anne Breitbarth, 2019b. 'Interpreting (un)interpretability', *Theoretical Linguistics* 45(3–4). 309–317.

Weinreich, Uriel, William Labov & Marvin I. Herzog, 1968. 'Empirical foundations for a theory of language change', in *Directions for Historical Linguistics: A Symposium*. Austin, TX: University of Texas Press. 95–195.

Wray, Alison & George W. Grace, 2007. 'The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form', *Lingua* 117(3). 543–578.

Yang, Charles D., 2000. 'Internal and external forces in language change', *Language Variation and Change* 12(3). 231–250.

Yang, Charles D., 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

PRIMARY SOURCES

Academia Mexicana de la Lengua, n.d. *Corpus Diacrónico y Diatópico del Español de América*. https://www.cordiam.org/

Biblioteca Virtual Miguel de Cervantes, 1999. *Virtual Library Miguel de Cervantes*. https://www.cervantesvirtual.com/

Davies, Mark, 2002. *Corpus del Español: Historical/Genres*. http://www.corpusdelespanol.org/hist-gen/

Digital Library of the Caribbean, 2004. *Digital Library of the Caribbean*. https://dloc.com/

Kroch, Anthony, Ann Taylor & Beatrice Santorini, 2013. *The Penn–Helsinki Parsed Corpus of Middle English (PPCME2)*. http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4

La Biblioteca Nacional de España, 2008. *Biblioteca Nacional de España—Biblioteca Digital Hispánica (BDH)*. http://bdh.bne.es/bnesearch/Inicio.do

Real Academia Española, 2013. *Corpus del Diccionario histórico de la lengua española (CDH)* [online]. https://apps.rae.es/CNDHE

Real Academia Española, n.d. *Banco de datos (CORDE)* [online]. http://www.rae.es

Šimko, Ivan, 2021. *Digital Editions of the Tale of Alexander the Elder*. Zürich: UZH Institute of Slavic Studies.

Šimko, Ivan & Anastasia Escher, 2021. *Digital Editions of the Life of St. Petka of Tarnovo*. Zürich: UZH Institute of Slavic Studies.

Šimko, Ivan, Polina Mihova, Olivier Winistörfer & Anastasia Escher, 2021. *Pop Punčov Sbornik—Digital Edition*. Zürich: UZH Institute of Slavic Studies.