Testing models of diffusion of morphosyntactic innovations in Twitter data

Deepthi Gopal[1], Tamsin Blaxter[1], David Willis[3]
and Adrian Leemann[2]
[1]University of Cambridge / [2]University of Bern / [3]University of Oxford

Established models of the spatial diffusion of linguistic innovations vary in their relationship to population density. Differences in prediction between the gravity models (Trudgill 1974), in which probability of diffusion is sensitive to settlement size, and the traditional wave models can be challenging to test due to the difficulty of large-scale and finely-grained geographical sampling. This paper tests the suitability of data derived from Twitter in establishing diffusion patterns. Using two case studies from British English – variation in the realisation of ditransitives, and preposition drop with *go* – we propose that the correlation between (local) population density and linguistic similarity to geographical neighbours can be used as a measure of hierarchical patterning for an individual innovation.

Keywords: dialectology, syntactic variation, computational sociolinguistics, British English, dative alternation

1. Introduction

How does the presence of large centres of population affect the spatial distribution of linguistic innovations? In most existing conceptions of language change, innovations originate in focal areas of high importance, before diffusing outward through space. The established models of the nature of this diffusion vary in their sensitivity to the heterogeneity of population density, and consequently in their predictions of the observed relationship between settlement size and grammatical variation. The essential distinction is between models that incorporate no such relationship and predict no empirical correlation, and those that assign weight to the distribution of the population and predict that large settlements measurably differ from smaller ones.

https://doi.org/10.1075/silv.27.11bla

In traditional wave models (Schmidt 1872; Bloomfield 1933), change begins at a specific point in space and spreads evenly outward, with no formal dependence on any non-geographic quantity ('contagion diffusion'). The ultimate consequence of this dynamic is a tendency for variants to be distributed in contiguous, internally homogeneous regions. In the gravity models (Trudgill 1974; drawing on the work of geographers: Hägerstrand 1952; Haggett 1965; Olsson 1965), the probability of change is still distance-dependent, but is additionally determined by the relative population of the areas involved: novel forms may jump from high-population city to high-population city ('hierarchical diffusion'), bypassing spatially intermediate but lower-population points and giving rise to discontinuous patterns in the synchronic geographical distribution of variants. Bailey et al. (1993) and Wikle and Bailey (1997) observe an apparent inversion in the direction of this hierarchical patterning (contra-hierarchical patterning) with larger cities lagging behind lower-population areas in the adoption of a variant. A re-ordering of this type retains, however, the underlying notion that population size has a measurable effect on the distribution of variants at any single instant. In these classes of model (see also Labov's (2001) cascade), the sharp gradient in population density that defines the edge of a city slows diffusion between the city and the surrounding area, making these likely locations of isoglosses; at the same, an assumed higher rate of long-distance social connections between high-population areas – potentially simply a consequence of the fact that such areas contain a large proportion of the total population – facilitates diffusion between them (Burridge 2018).

The problem with which this paper is concerned is the reliable detection and quantification of patterns of diffusion, given the distribution of a variable across space, but in the absence of the time dimension. Individual models of spatial diffusion, although not necessarily mutually exclusive, differ in their predictions as to the effect of urbanisation on the distribution of a variable at any single time-point. In order to evaluate such predictions, we ideally need datasets that are well-dispersed along multiple dimensions – distributed across a spatially contiguous area, and spanning the largest possible range of population densities and states of urbanisation. The extension of these problems to large geographical scales then requires quantities of data that cannot easily be produced via traditional dialectological data collection, which typically samples very few informants per site. Furthermore, these issues are not necessarily amenable to the methods of variationist sociolinguistics, in which relatively small numbers of distinct geographical locations are typically considered. We argue here that one possible source of appropriately rich data lies in the use of large-scale social media. Such datasets provide us with access to the real-time language use of millions of individuals, and in particular allow us to resample multiple different geographical regions at similar levels of granularity. To this end, we establish (Section 2) a localised corpus of Twitter data in British

English, covering the time period between October 2017 and May 2019. This corpus consists of all tweets posted by a set of users, for whom profile metadata and keyword analysis (Section 2.1.1) allow localisation to within a civil parish (or smaller administrative unit) within Britain and Ireland, representing more than a million unique individuals in total.

Using these data, we investigate (Section 3) the current distribution of two variables in British English: the extensively studied variability in the ditransitive construction, or the dative alternation (recent work: Siewierska and Hollmann 2007; Bresnan and Ford 2010; Wolk et al. 2013; Yáñez-Bouza and Denison 2015, among others), and the more recent case of the innovation of preposition drop with *go* and certain noun phrases (*go __ (the) pub, go __ school*; Myler 2013; Biggs 2015). These variables differ substantially in their historical context, in their geographical extent (and their attendant presence or absence in individual high-population areas), and in the structure of the space of variants. For these variables, we test a potential measure of 'hierarchicality', or the detectable presence of a recent process of gravity-like diffusion, by examining the correlation between local population density and linguistic similarity to geographical nearest neighbours; we find that the ditransitive variants and preposition drop differ under this metric. The probabilistic distribution of variants of ditransitives is sensitive to geography, showing clear regional boundaries that recall those in the *Survey of English Dialects*, but generally insensitive to the size of localities. In contrast, the distribution of preposition drop in *go*-phrases shows a stronger direct dependence on population density. We relate these differences to the history and current status of each variable under consideration, and argue overall that the characterisation of diffusion processes is incomplete without a rigorous model of their operation.

2. Methodology and corpus construction

The essential property introduced above (and expanded upon in Section 4 below) is the presence or absence of a relationship to settlement size, and therefore to local population density at each point in a geographical sample. The predictions of the gravity and wave models coincide in the limit of perfectly uniform population density, and we therefore ideally require a sufficiently non-uniform distribution of settlement size in the region of sampling. In order to test the validity of any individual model, therefore, we need our dataset to have two properties: (i) enough underlying variation in the true population density must exist for different models of diffusion to make statistically distinguishable predictions, and (ii) our respondents must be densely distributed over the range of different settlement sizes.

If settlement size and population density have a truly flat distribution, failing property (i) above, then irrespective of the depth of experimental sampling, differences in the underlying mechanism of diffusion are unlikely to result in statistically significant differences in the resultant distribution of linguistic variation. In order to maximise the probability that our tests of diffusion give rise to meaningful results, we need to consider the distribution of population density itself in the geographical region of interest. The optimal test case spans a range of orders of magnitude with uneven distribution, so that a high-population locality always has low-population geographic neighbours with which a comparison can be made (the number of highly dense areas is much smaller than the number of less dense areas). We note (Figure 1) that a dataset of British English inherently draws from an underlyingly non-linear spatial distribution of population densities, and satisfies condition (i); this would not be the case for a country like the Netherlands, where the range of settlement sizes is much narrower.

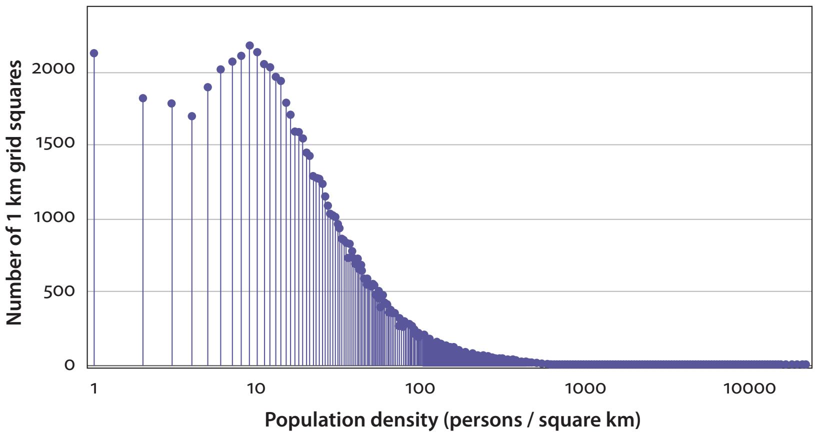

Figure 1. Population density in Great Britain over a 1km grid, based on the 2011 Census and the 2015 Land Cover Map (Reis et al. 2017)

One solution to the second problem of sampling – the requirement that our dataset must represent as much of the population density curve as possible – is the use of large-scale corpora. In principle, social-media platforms have unprecedented scale and wide coverage of the population, therefore allowing access to much larger and more broadly-distributed volumes of information than might be acquired using traditional methods. This should then facilitate the investigation of large-scale problems of spatial variation. The use of Twitter-based corpora is, concordantly, on the rise in dialectological research (Russ 2012; Bamman, Eisenstein and Schnoebelen 2014; Doyle 2014; Eisenstein et al. 2014; Gonçalves and Sánchez 2014; Jones 2015; Huang et al. 2016; Grieve, Nini and Gou 2017; Grieve et al. 2019; among others).

Although a number of works exist that investigate the distribution of morphosyntactic variables in Twitter corpora (Haddican and Johnson 2012; Doyle 2014; Stevenson 2016; Strelluf 2019), the use of Twitter data remains less established for this purpose than for the investigation of lexical variation, and we intend this discussion to advance this area. Much of the existing work using Twitter corpora also deals with very widely spoken world languages (e.g. American English; world Spanish) across geographical regions that are many orders of magnitude larger than Britain and Ireland, on which we focus here. There is comparatively little existing work even on British English (although see Stevenson 2016; Shoemark et al. 2017; Grieve et al. 2019). The further major goals of the following discussion are to extend the use of Twitter corpora to the investigation of morphosyntactic variables, and to replicate the results of traditional (dialectological and sociolinguistic) methodologies at various scales of geography.

2.1 Corpus structure

Due to the limitations of both the Twitter search API[1] and data storage, posts made by UK-based users cannot be singled out exhaustively. A cap applies to the use of the free Twitter streaming API – any individual search query can return no more than 1% of all data. Downloading all posts tagged in English generates a very large volume of irrelevant (largely American) material, which would render analysis prohibitively difficult, and would likely lose a large fraction of British and Irish English posts to the data cap. In order to extract an appropriate set of posts, therefore, our final corpus of British English has two components. Initially, we collected via the Twitter 'streaming' API all posts made between October 2017 and May 2019, and geolocated within the ranges 49.8°N to 61°N and −11°E to 2°E (covering the UK and Ireland). This set was cleaned to remove all retweets not containing original content, all posts whose language was identified as non-English either by Twitter's own language-detection or by the *Chromium Compact Language Detector 2 (CLD2)* library, and all formulaic automated posts made by applications such as Youtube or Foursquare. This amounted to 104,657,500 posts from 1,734,260 unique individuals.

Geo-location on Twitter is opt-in, and, as such, the set of users with geolocation metadata available (and who appear in a search based on a bounding box) is subject to geographical and demographic biases away from both the overall set of Twitter users and the general population, for example towards younger users and towards urban areas (Malik et al. 2015; Pavalanathan and Eisenstein 2015; see Section 2.1.1).

1. 'Application Programming Interface'.

In an attempt to partially mitigate this effect, the original data were supplemented with a further set of 49,857,358 posts from 826,653 unique individuals who were not themselves present in the geolocated corpus, but whose usernames were mentioned by the users in that corpus *and* who could be localised to Britain or Ireland by the procedure outlined in Section 2.1.1. Finally, tweets were annotated and aggregated for syntactic context, as outlined for the case studies in Section 3.

2.1.1 *Localisation*

Essentially, all previous research using Twitter corpora for the purpose of spatial dialectology uses Twitter's own geolocation metadata. As noted above, these metadata are only available for a small fraction of the total set of users (1–2%; Eisenstein 2018: 369), for which the degree of representativeness of the overall population is unknown. Pavalanathan and Eisenstein (2015) identify biases in age and gender associated with geotagged data for American English, and Hecht and Stephens (2014) measure bias towards urban areas. In dialectological research, a further issue arises: ideally, we want to associate individuals with the location of their language acquisition, rather than with the set of coordinates at which any particular tweet is written.

With this in mind, we developed a keyword-based strategy for assigning locations to individual users. We give only a brief overview here, however, a more technical discussion of this procedure, with specific reference to Welsh data, has already appeared in print (Willis 2020). The index of place-name 'keywords' corresponded to the set of possible places, at the smallest geographical scale, with which a user could be associated; this combined the Index of Place Names for Great Britain (Office For National Statistics 2016) with the Ordnance Survey Ireland (2016) gazetteer of Irish townland names. Rows corresponding to uninhabited places and overwhelmingly common dictionary words were removed. Commonly attested spelling variants were added, along with abbreviations (with varying degrees of official status; matching 'Hull' to Kingston-upon-Hull, and 'Cdiff' to Cardiff).

Across the total set of users, we considered two Twitter metadata fields, along with the set of all posts produced by each user within our corpus: the user-provided 'location' and 'bio' fields, both of which allow free text entry. In principle, these fields are intended to contain a description of the user's (current) location and a short personal biography of the user respectively, although the validity and relevance of this user-provided information is very variable. While the majority of users (82.9%) provided some information in the location field in particular, this was not necessarily useful, as we could derive little from very large-scale locations such as 'UK', or entirely non-geographic locations such as 'Hell' or 'The kitchen'. High-quality, small-scale matches to individual place-names appeared in the 'location' field for

676,361 users (39%), and in the ‘bio’ field for 222,354 (13%). For each individual, we extracted all matches to our set of place name keywords that appeared in each metadata field; we then extracted a number of small-scale place mentions appearing in each user’s tweets.

The result of this procedure was an ordered set of usernames paired with all candidate small-scale places, or candidate ‘localisations’, for which at least one match was found: this list of place name matches was assigned an initial score defined on a provisional basis (10 for the presence of a match in the location field; 5 for a match in the user description; 1 for each instance appearing in the user’s post text). These scores were then weighted by all references in the user’s metadata to larger-scale hierarchical geographical areas (‘Lancashire’ or ‘North Wales’ increased the score of all geographically subordinate candidate points), and by the presence of high-frequency local demonyms which were associated with both point localities (‘Mackem’ for Sunderland) and with broader regions (‘Yorkshireman’ corresponded to all candidate points in Yorkshire). Scores were incremented if mentions of small-scale place names appeared in a (manually defined and non-exhaustive) list of constructions that suggested direct relationships thereto (*X born and bred; I’m from X originally*), and lowered in the opposite case (*from X living in Y*). Remaining ties were broken by choosing the smaller of two locations in any ties involving major metropolitan areas (e.g. formulations such as ‘London/Penrith’ were mapped to Penrith), and choosing the largest possible location in the case of ambiguous name (e.g. Newcastle-upon-Tyne rather than Newcastle-under-Lyme for ‘Newcastle’, if no other disambiguating information was available). This then resulted in a set of 1,033,058 users (59.6%), for whom a best-guess low-level localisation within Britain and Ireland was produced.

3. Mapping the distribution of morphosyntactic variants

With the localised corpus established, datasets were extracted corresponding to two different cases of morphosyntactic variation in British English. Section 3.1 discusses variation in ditransitives, and Section 3.2 discusses the distribution of preposition drop with *go*. These case studies differ in geographical distribution, in the number of competing variants, and in the presumed rate at which change is taking place at the present day. We establish an overview of these cases in this section, and follow with a discussion of the evidence that they provide for diffusion processes in Section 4.

3.1 Dative alternation revisited

In most varieties of English, the realisation of ditransitive verbs frequently involves competition between two semantically synonymous constructions, which together constitute the 'dative alternation': a prepositional dative with the Goal following the Theme and marked by a preposition, and a double object construction with Goal preceding Theme. A third variant, in which the Theme precedes the Goal but the Goal is not marked by a preposition, has been associated in the literature largely with the north-west and the Midlands of England. Language-internally, it is associated with clauses in which both objects are pronominal (Gast 2007; Siewierska and Hollmann 2007; Haddican 2010; Gerwin 2013; Biggs 2016). Examples are given in (1) below.

(1) British English ditransitives with pronominal objects
  a. *Give it me*. Theme-Goal ditransitive
  b. *Give me it*. Goal-Theme double object construction
  c. *Give it to me*. Theme-Goal prepositional dative

Although the more general variation in ditransitive constructions is not restricted to pronominal forms and may as well appear with full nominal referents, the dataset in this discussion is restricted to cases of the form V it to Pronoun/V Pronoun it/V it Pronoun, for which there exists some evidence of a distinct historical development (Gast 2007; Gerwin 2013; Yáñez-Bouza and Denison 2015).

In order to construct a relevant dataset, all expressions matching the prepositional dative (*X it to Y*, where Y was any pronoun – including abbreviated or modified spellings found frequently on Twitter) were extracted; all *X* identified conclusively as non-verbs by manual inspection were then removed, and the remaining list was used to extract from the corpus all expressions matching *X Y it* and *X it Y*. This produced a set of 27,757 tweets by 23,530 users; of these, 18,065 tweets (14,769 users) contained the prepositional dative, 3,703 (3,346 users) the Theme-Goal ditransitive, and 5,989 (5,415 users) the Goal-Theme double-object construction. Figures 2, 3 and 4 map the spatial distribution of the resultant data; users localised to the same point (corresponding essentially to a single centre of population) are pooled to give an overall value at that point, and k-nearest-neighbour kernel density estimation is then applied (with k = number of localities)[2] to smooth fluctuations in the data.

2. This is a heuristic applied due to the complexity of bandwidth selection in kernel density estimation.

How closely do our results align with existing evidence for the distribution of ditransitive variation? The literature on the syntactic conditioning and diachrony of this alternation is extensive, but a consensus picture of the geo-spatial patterns involved is less forthcoming. Individual studies have established the availability of particular variants in specific dialect regions, but the most up-to-date account of the detailed geography of 'present-day' ditransitive variation essentially remains the *Survey of English Dialects* (SED; Orton 1962; Orton, Sanderson and Widdowson 1978: map S1; and for the most recent maps, Upton and Widdowson 2006), despite the now considerable age of the data (collected between 1950 and 1961) and the unrepresentativeness of the mostly rural elderly speakers surveyed. The distribution in the SED samples only a single informant at each location, and therefore not necessarily a complete picture of the proportional availability of each option. This assigns the prepositional dative, which is the minority variant in the SED, to a large contiguous region in the southwest of England and to the vicinity of London, the Goal-Theme double object construction to a corridor along the eastern half of England, extending across to Cumbria in the far northwest, and the Theme-Goal construction to the western Midlands and the traditionally defined northwest of England surrounding Manchester and Lancashire.

More generally, despite the often substantial distance in apparent time, the SED distribution is concordant with the results of more recent work that evaluates the prevalence of variants over coarsely grained dialect regions in both historical and present-day corpora (Siewierska and Hollmann 2007; Gerwin 2013; Szmrecsanyi 2013; Yáñez-Bouza and Denison 2015). On a more geographically fine-grained level, Stevenson (2016) tests a sub-case of the current one (considering the past tenses of *give* and *send* only) over a Twitter-based corpus, and finds fairly robust agreement with the distribution in the SED. This is also the case for our dataset; the most apparent deviation from traditional surveys lies in the much higher incidence of the prepositional dative. While clear regional patterns corresponding to preferences for particular variants are apparent, there is no locality in our dataset for which the rate of occurrence of the prepositional dative is lower than 15%, and it is clearly the majority variant overall. One interpretation of this is that the relationship between Twitter data and spoken data is not entirely straightforward, and that 'overuse' of the prepositional dative may represent the use of a perceived prescriptive written standard. The geographical patterning of the remaining variants in our data is however in line with traditional predictions. We can distinguish Scotland, Cumbria, and the northeast of England – preferring the Goal-Theme double object construction – from the northwest of England, south and west Yorkshire, and the West Midlands, where higher rates of the Theme-Goal construction are observed. The spatial distribution of the relevant data is shown in Figures 2, 3 and 4.

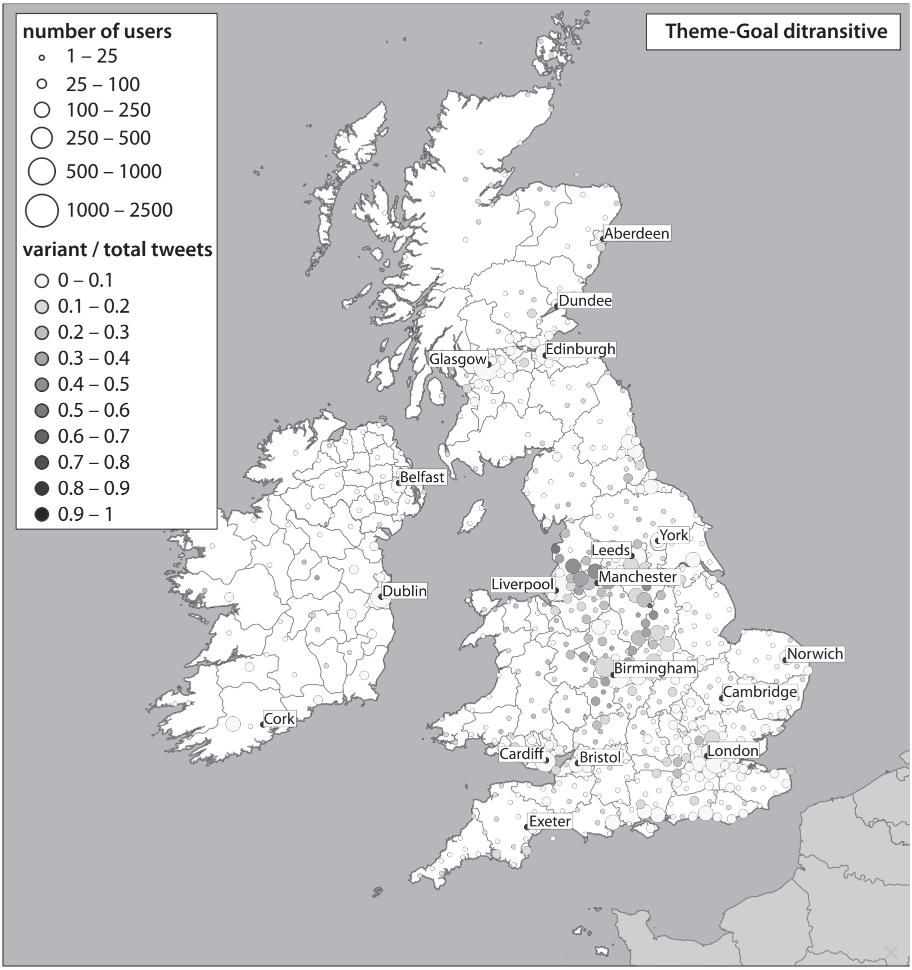

Figure 2. Incidence of the Theme-Goal ditransitive (*give it me*) with two pronominal objects, as a fraction of all ditransitive contexts for 23,530 UK and Ireland Twitter users with k-nearest-neighbour smoothing ($k = \sqrt{(\text{number of localities})} = 24.7$)

One consequent interpretation of our dataset here is that, although all three major variants tend to appear in competition across the entire area, the regions defined by transition from one highly-preferred variant to another have been relatively stable over time – isogloss movement since the period of the SED has been comparatively small-scale. On broad geographic scales, there is then little evidence for any identifiable pattern of diffusion in this data, and we do not expect any measure of hierarchicality to show significant results for this case. If patterns of gravity diffusion and of hierarchical structure emerge from the spread of innovations through

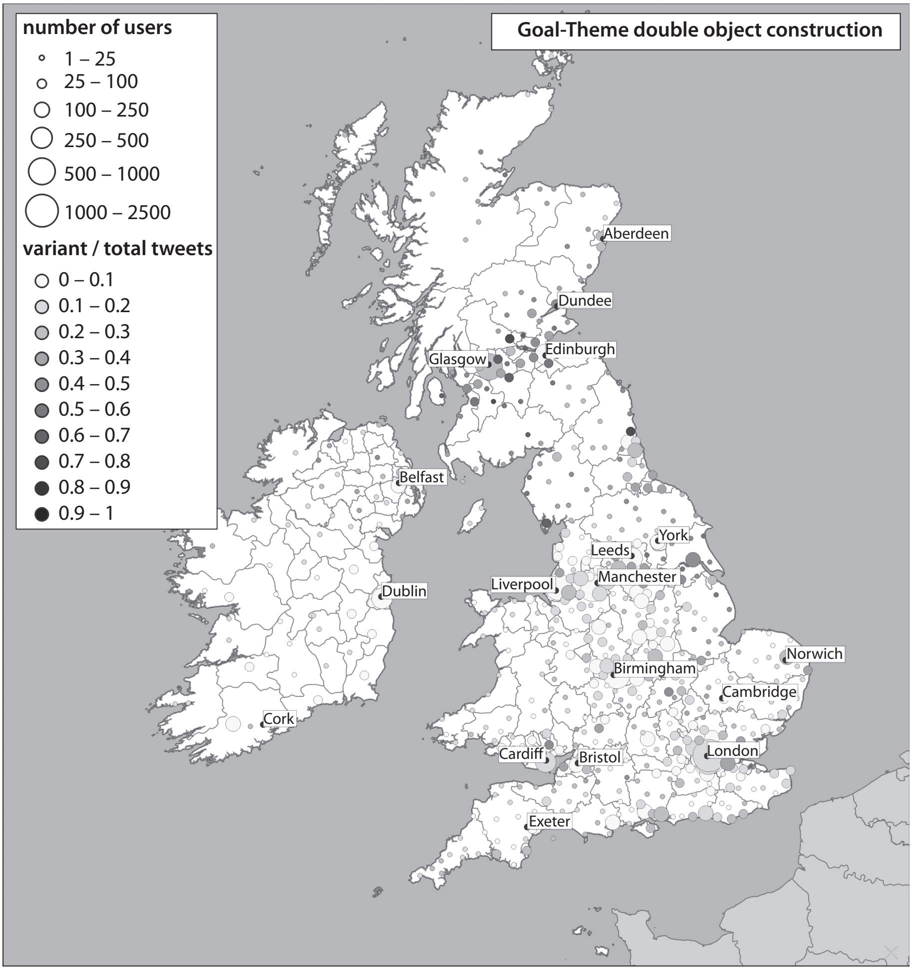

Figure 3. Incidence of the Goal-Theme ditransitive (*give me it*) with two pronominal objects, as a fraction of all ditransitive contexts for 23,530 UK and Ireland Twitter users with *k*-nearest-neighbour smoothing ($k = \sqrt{(\text{number of localities})} = 24.7$)

both local and long-distance social networks, then this observation is expected, given the long-standing nature of the variation in English ditransitives. The case in this subsection may then constitute a control case for the statistical detection of patterns of diffusion. We establish in the following subsection (Section 3.2) a contrasting case, which instead involves the recent spread of a single variant through major urban centres.

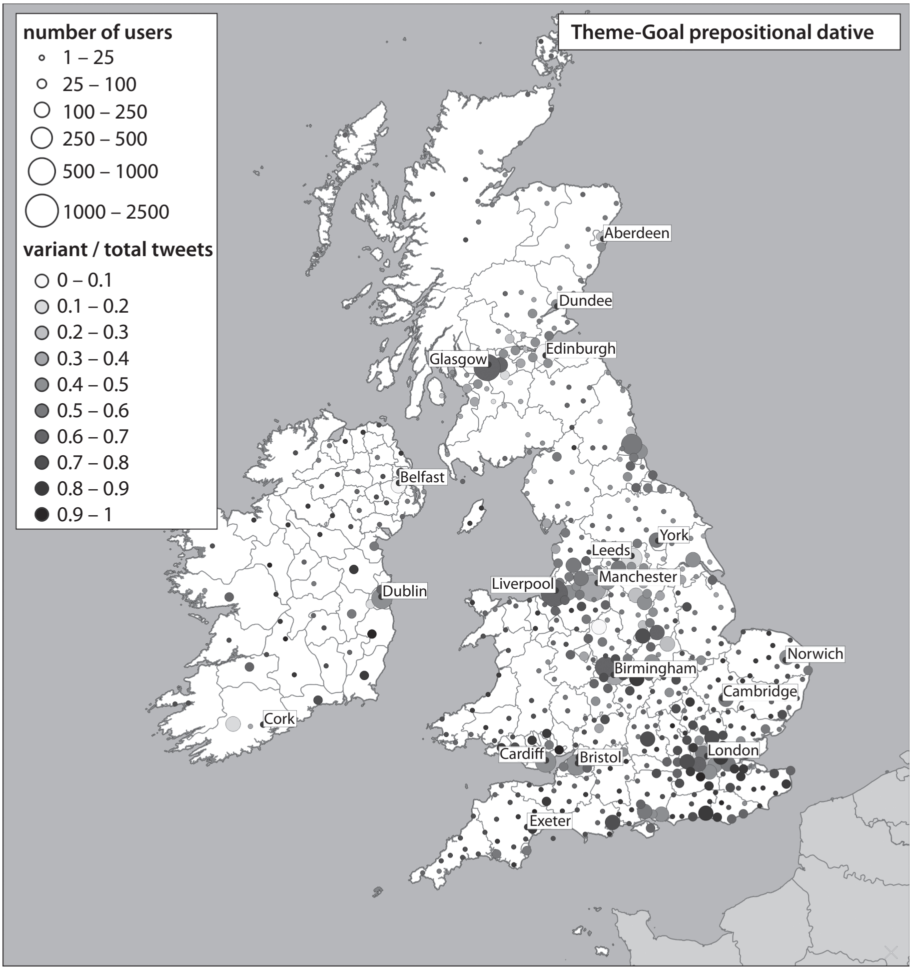

Figure 4. Incidence of the Theme-Goal prepositional dative (*give it to me*) with two pronominal objects, as a fraction of all ditransitive contexts for 23,530 UK and Ireland Twitter users with k-nearest-neighbour smoothing ($k = \sqrt{}$(number of localities) = 24.7)

3.2 Preposition drop

In Section 3.1, we established that the Twitter dataset has the capacity to replicate results arrived at by traditional means for a single variable. The variation in British English ditransitives thus provides a test case for our measurements of diffusion type in the following sense: the history and current distribution of the variants suggest that we should predict little trace of a single diffusion process, and therefore a negative result from any test of hierarchicality. As a contrast, this section presents a case of recent, rapid change in the geographical distribution and rate of use of an innovative variant.

A number of varieties of British English show optional non-realisation of the preposition *to* following certain, typically directional verbs, as in (2).

(2) Preposition-deletion across varieties of British English
Manchester: *She went _ the pub.* (Haddican 2010: 2430)
Liverpool: *She said we'd go _ the pub, and _ the pub we went.* (Biggs 2014: 19)
London: *We went _ pub last night.* (Hall 2019: 2)

This has been reported for several varieties of northwest England (Haddican 2010; Myler 2013; Biggs 2014, 2015), and for London and the surrounding area (Bailey 2018; Hall 2019), with differences in the corresponding set of environments, as well as in the presence or absence of the definite article. Speakers may be unaware that these constructions are not grammatical in varieties other than their own (Biggs 2014). Based on these factors, this variable seems to be a good candidate for a relatively recent and potentially ongoing change.

Previous work on preposition drop has been based on relatively small and discrete samples of speakers of largely urban varieties, and as such there is no pre-existing picture of the full geographical extent of this variation. The literature does establish syntactic differences between the preposition drop in different dialects, which cannot be dealt with fully here. Relevant to our dataset is the observation (Bailey 2018; Hall 2019) that preposition drop in the southeast of England is subject to the semantic restriction that the noun phrase, which must be interpreted as the directional Goal, must denote a familiar or anaphoric location or an institution. In the northwestern varieties described by Myler (2013) and Biggs (2014), the range of verbs permitted (including at least *come, go, run, drive, nip, jog* in Lancashire, and wider still in Liverpool) is much broader than in the southeastern ones, in which the verb must be *come* or *go*; the determiner may optionally be present in the northwest, but is obligatorily dropped in the southeast.

For the purpose of this analysis, we wish to establish a fairly general overview, and as such discussion is restricted to those contexts where preposition drop is possible in all previously described varieties; that is, we consider only collocations involving the verb *go*, subject to 'familiarity' restrictions on the goal noun. We also collapse the variation in the presence or absence of the determiner. In order to construct the dataset, all sequences were extracted that involved *go* and a small set of frequent, semantically appropriate noun phrases: *go* (*to*) *Amsterdam, go* (*to*) *Asda, go* (*to*) (*the*) *chicken shop, go* (*to*) *college, go* (*to*) *jail, go* (*to*) *London, go* (*to*) *Manchester, go* (*to*) *Nando's, go* (*to*) *Paris, go* (*to*) *prison, go* (*to*) (*the*) *pub, go* (*to*) *school, go* (*to*) *Tesco*. This yielded a total of 34,615 tweets produced by 26,618 individuals. The resulting spatial distribution appears in Figure 5. As in the previous case, users localised to the same point are pooled to give an overall frequency of use at that point, and *k*-nearest-neighbour kernel density estimation is applied to reduce the visual impact of random fluctuations in the data.

The geographic extent of preposition drop turns out to be superset of the regions identified in the existing literature. The presence of preposition drop is strongly associated with both the northwest of England (Liverpool, Manchester), and with London and the surrounding areas, which corresponds to previous descriptions. To this the large urban areas of the West Midlands can be added. Doing

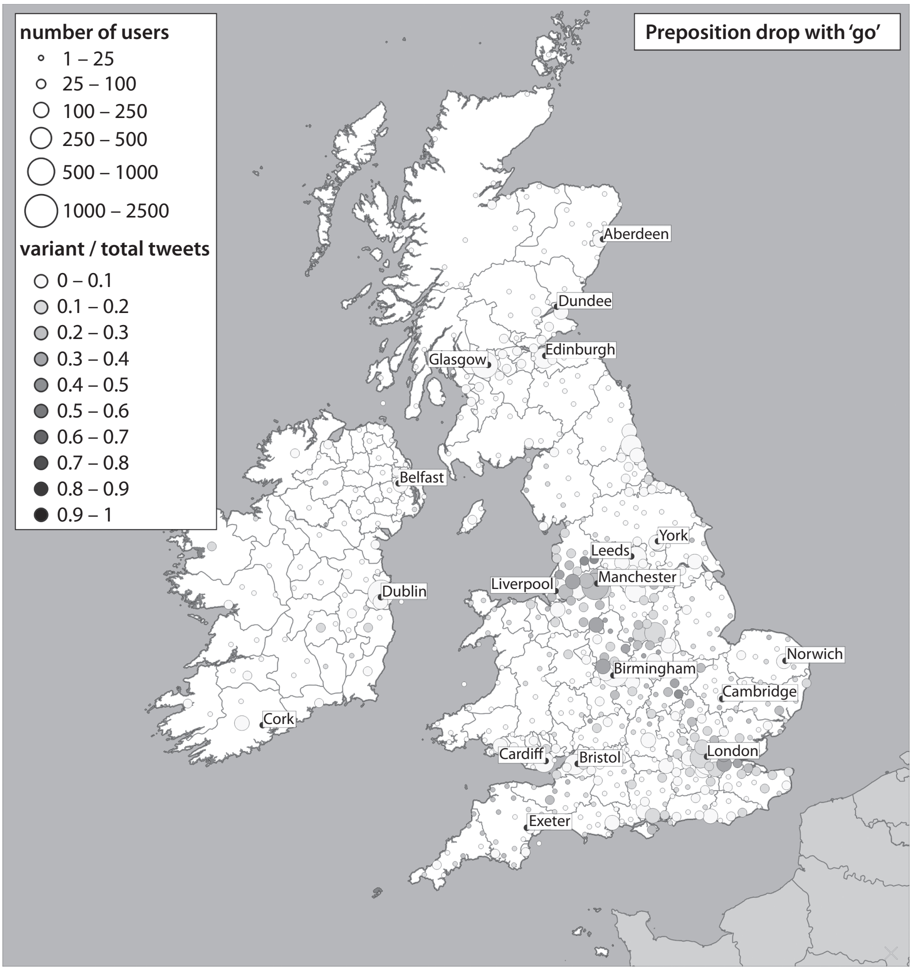

Figure 5. Overview of the occurrence of preposition drop in *go* _ (*the*) *N* for 26,618 UK and Ireland Twitter users, with *k*-nearest-neighbour smoothing ($k = \sqrt{(\text{number of localities})} = 25.8$)

so forms a contiguous corridor corresponding to highly densely-populated and well-connected areas, excluding several geographically neighbouring regions of comparatively low population density (East Anglia, the south coast of England, and most of Wales). The current distribution of preposition drop therefore constitutes a candidate case for true hierarchical diffusion.

4. Approaches to the quantification of diffusion

We now consider the problem of the correlates of diffusion, which is the focus of the remainder of this article. In Section 2, we set out the essential structure of the individual datasets we considered, and mapped the distributions of each possible variant in Figures 2–5. These distributions are entirely synchronic; although we might in principle reconstruct apparent-time information from the content of the Twitter corpus (Nguyen et al. 2014), this is beyond the scope of this paper, and we restrict ourselves here to the inference of diffusion processes from the stationary geographic distribution of individual features.

Given a dataset that takes the form of a single snapshot of the state of the population, our task is then to quantify the extent to which it resembles either the output of a gravity-like process or a wave-like one. This requires both that we define a standard for comparison, and that we establish the expected properties of the output in either idealised theoretical case. As such, we begin by briefly recapitulating the predictions and prerequisites of the existing models of spatial diffusion.

In a wave model, change diffuses evenly outward from the point of origin. From a formal point of view, this reduces to an inverse relationship between the rate and probability of change and distance, such that the influence of any locality on any other is dependent only on the distance between them. In Trudgill's (1974) formulation of a gravity dynamic, a relationship to the relative population of localities is introduced; the likelihood that a pair of locations interact remains inversely related to the distance between them, but must also incorporate their relative size. At any given instant in the progression of a diffusion process, the influence of any one locality on any other must in either model be determined by the relevant set of parameters. Adapting Trudgill (1974), we can then establish the influence of one population centre on another in the form of (3).

(3) The influence of centre *i* on centre *j* (partially adapted from Trudgill 1974).

$$I_{ij} = \Big(\underbrace{\frac{1}{d_{ij}^2}}_{\textsc{Wave}} \quad or \quad \underbrace{\frac{n_i n_j}{d_{ij}^2}}_{\textsc{Gravity}} \Big) \times \frac{n_i}{n_i + n_j}$$

where d_{ij} is the distance from i to j, and n_i is the population at i.

This implementation of influence requires translation into quantities with which we can engage more directly. While we reserve a serious treatment of this issue for future work, we provide here a brief elaboration in order to motivate our measurements of hierarchicality. For the sake of abstraction, if it is assumed that the diffusion process in question manipulates a binary variable, then the interaction in (3) can be treated as the probability p, normalised as in (4), that any individual in j will 'flip' to the new feature value due to the influence of i.

(4) Normalised probability of change due to i at j

$$p_{ij} = \frac{I_{ij}}{\sum_{x=1\ldots N} I_{xj}}$$

The expected number of individuals whose feature value at j changes due to i is, in formal terms, given by the binomial distribution with parameters p_{ij} and n_j; that is, the result of flipping n_j biased coins (representing each individual at j) with probability p_{ij}, which has the simple expected value $n_j\, p_{ij}$. The final expected value at j is then the sum over these expected values scaled by the proportion of individuals at i with the innovative feature. If the number of individuals with the new feature value is k_j at location j, then at each point in time the size of the innovation-using population can be estimated as in (5).[3]

(5) Expected size k of the innovation-using population at locality j at time t + 1at time t + 1

$$k_j\,(t+1) = k_j\,(t) + \sum_{i \neq j} n_j p_{ij} \frac{k_i}{n_i}$$

Equation (5) allows us to produce idealised simulations of the output of a hierarchical or a wave process, which will be considered in Section 4.1.1. Any such implementation carries several cautions. First, the delineation of an individual centre of population is not necessarily straightforward or uniquely determined – some implications of this observation are considered in Section 4.2. Second, the definition of either model as a process operating over individual localities is an abstraction away from the individual-level dynamics that are conjectured to underlie mechanisms of diffusion. One potential desideratum for measures of diffusion is that they must be extensible to cases where interacting individuals are considered to be the locus of change, rather than interacting localities.

3. As noted in the introduction to Section 2, a necessary consequence of the structure of either process is their sensitivity to the various initial conditions on the set of localities and associated populations over which they apply. In the limiting case of uniform population density, in which all localities have similar populations ($n_i \cong n_j$ for all i, j), the factor $n_i n_j$ is essentially constant, and the gravity process is not distinguishable from the wave process.

4.1 Measurement

Consider the simplified examples in Figures 6 and 7. These are the output of iterating the processes in (3)–(5) over regularly spaced points, with the innovative variant present only at the highest-population point (0,0) at time $t = 0$. The relationship between the highest-population points, at (0,0) and (2,0), and their lower-population neighbourhoods is not constant across models. At a single timepoint during the operation of a hierarchical process, high-population points have higher rates of use of the new variant than lower-population points, and the nearest geographical neighbours of a high-population point P are less similar in innovation prevalence to P than they would be in the result of a wave process.

This generalisation thus offers a heuristic justification for a potential metric of degree of hierarchicality. Given a measure of the similarity between localities that is defined over the distribution of relevant variants at each point, we can evaluate the similarity of a single locality to its nearest geographical neighbours. In the intermediate stages of the operation of a hierarchical diffusion process, this local neighbourhood similarity is expected to have an inverse relationship with population. An alternative, more intuitive measurement of diffusion type is the relationship between population and fraction of the innovative variant. Under certain initial conditions, our proposal avoids the need to pre-specify the innovative variant, and is more clearly extensible to cases of change involving large numbers of competing forms. These initial conditions are that the number of large settlements or contiguous high-density areas must be small relative to the number of

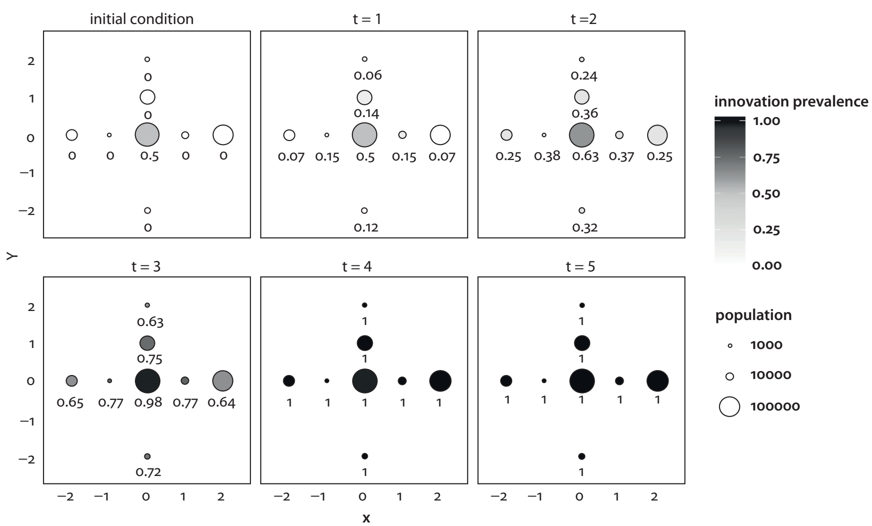

Figure 6. Illustrations of the output of a wave process over gridded points

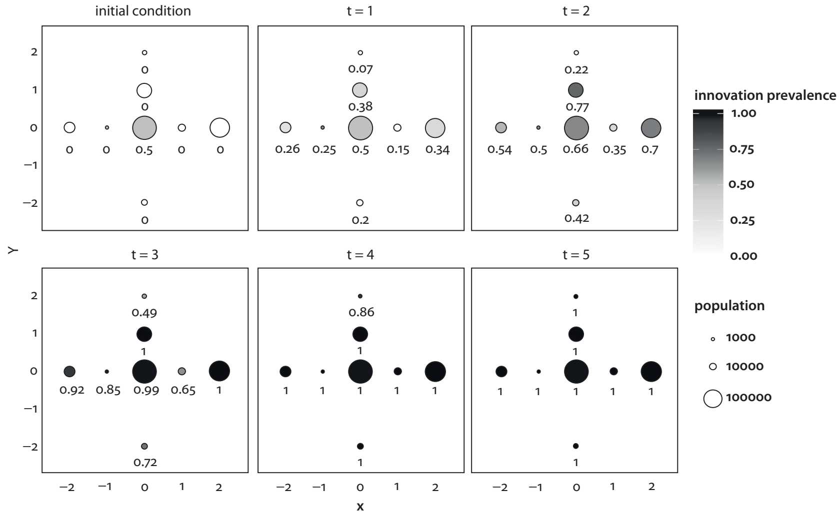

Figure 7. Illustrations of the output of a hierarchical process over gridded points

low-density settlements, such that large settlements only have small neighbours. As noted in Section 2, this is likely also a prerequisite for hierarchical diffusion to be at all observable.

If large settlements are outnumbered by smaller ones, as is the case for most real-world scenarios, then the neighbours of any large settlement are lower-population, and as such, likely to be behind their high-population neighbours in the spread of innovation. At the same time, the neighbours of smaller settlements are on average similarly sized, and, as such, should pattern together with one another. A relationship of this type is not expected to hold in the absence of hierarchical diffusion; if there is no underlying relationship between a locality's population and its position in the course of change, then no relationship between nearest-neighbourhood similarity and population is expected to hold.

With these observations in mind, it remains to define the measure of nearest-neighbourhood similarity and its domain of evaluation. There is a sampling problem related to the use of discrete individual locations in real-world datasets, especially in the Twitter dataset under consideration: location boundaries are arbitrary, since the definitions of single point locations can differ in their overall level of granularity, and variation in whether data are associated with a generic point ('London') or a specific one (an individual London borough) can interfere with the identification of the true local population density. In order to mitigate this, we take the domain of evaluation of the similarity metric to be individual grid cells rather than localities, binning the continuous geographical space of each dataset into various sizes of an $n \times n$ grid. A more realistic idea of local population density is obtained by normalising data into a grid over the whole map.

We define the similarity between two cells in a standard way, as the *cosine similarity*: that is, the inner product of vectors representing the relative proportions of variants in each cell. For two cells A and B, with normalised variant proportions (a_1, a_2, a_3) and (b_1, b_2, b_3), the cosine similarity is $a_1b_1 + a_2b_2 + a_3b_3$: this is 1 if the cells are identical, and 0 if the cells never overlap at all. For each cell, we compute this quantity for each of its four 'von Neumann'-neighbours (i.e. with which it shares an edge), and average over all four for a measure of the similarity between the cell and its local neighbourhood.

4.1.1 *Simulated data*

One illustration of the appropriateness of this measure is its performance over 'realistic' simulated data; that is, over simulated data with similar distributional characteristics to the datasets considered in Section 3, but for which the process of diffusion is simulated as in (5). The outcome of a measure of hierarchicality should then be straightforwardly predictable for such data, giving statistically distinguishable results for simulated processes of gravity and wave diffusion. Note, however, that this does not constitute a formal proof of validity.

Figure 8 shows the output of five iterations of the processes in (3)–(5), over a set of points with associated populations generated by sampling 50,000 randomly chosen users from the corpus and their locations. In both cases, the initial frequency of the 'new variant' was set to 0 at every point outside the largest point, and to 0.5 at the largest point. We then evaluate nearest-neighbourhood similarity and consider its relationship to logarithmically scaled population in Figure 9. For the simulated

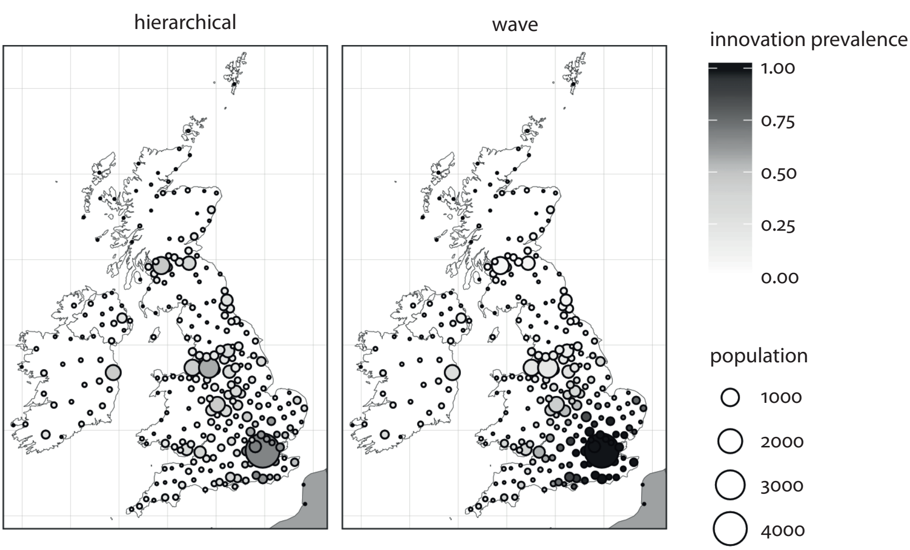

Figure 8. Simulated output of hierarchical and wave processes

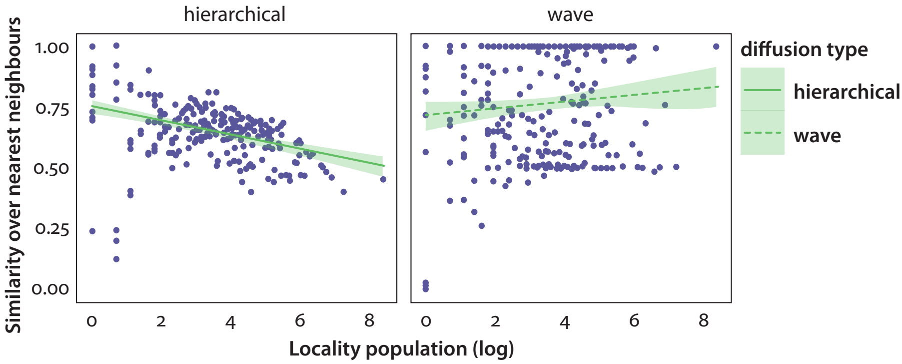

Figure 9. Averaged similarity over von Neumann-neighbourhood against log-scaled population, for the simulated data in Figure 7, over a 20 × 20 grid, with best-case linear fit

process of gravity or hierarchical diffusion, nearest-neighbourhood similarity has a statistically-significant and an essentially linear relationship (with negative slope) to the logarithm of the local population (Pearson's $r = -0.3253$, $p < 0.00001$). In the simulated wave case, as expected, no statistically meaningful relationship holds (Pearson's $r = 0.0630$, $p = 0.3583$).

4.2 Evaluating real data

At this point, we have seen that a case of true hierarchical diffusion will show a statistically meaningful inverse relationship between the (log-scaled) population in a cell and the similarity in distribution of variants between that cell and its local neighbourhood. We can turn now to evaluate this claim on real-world, large-scale datasets. Section 3 previously presented two cases of variation for which there are grounds to expect measurably different responses.

Recall the variation in the ditransitive construction from Section 3.1. The geographical distribution essentially acts as the lower bound of the overall reliability of our social media corpus, in that it has a fairly substantial resemblance to the findings of more traditional methodologies. At the same time, the hypothesis is that this variation has been subject to relatively little active spatial diffusion in its recent history, and as such we do not expect to see any significant hierarchicality.

The relationship between neighbourhood similarity and the log-scaled population of each grid cell is plotted in Figure 10 below, and the relevant tests of correlation are given in Table 1. We note, with respect to real-world data, that the choice of grid size – which we have not yet considered – has a substantial effect on the outcome. If the spatial region of interest is divided into a small number of very large grid cells, we risk collapsing together high-density localities and surrounding

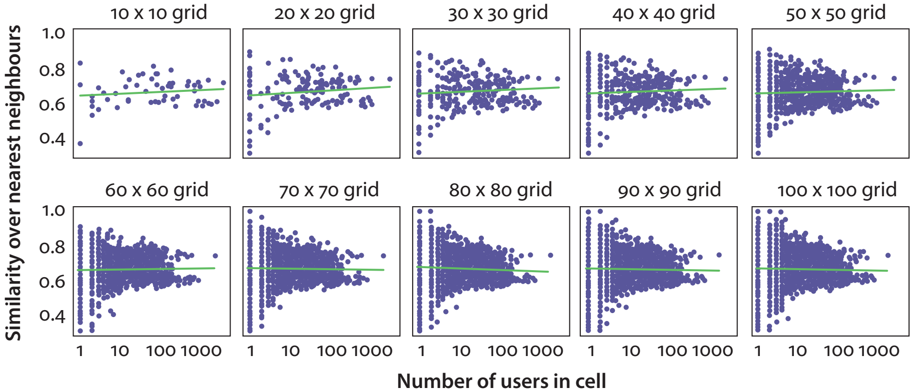

Figure 10. Averaged similarity over the von Neumann neighbourhood plotted against log-scaled cell population, for the ditransitive variation shown in Figures 2 and 3

sparser regions, while the use of a very fine grid would overfill the lowest-population regions of the plot. In order to remain sensitive to this measurement effect, multiple potential grid sizes must be considered.

Irrespective of the choice of grid over which to evaluate neighbourhood similarity, we find no evidence here for a statistically significant relationship between the population of an individual cell and its similarity to its nearest neighbours. All correlation coefficients in Table 1 are small, with p-values never approaching significance. Similarity to the local neighbourhood appears essentially independent of population for this dataset, and rather normally distributed. This is as expected if no hierarchical diffusion has taken place.

Consider now the less historically established case of preposition drop that was set out in Section 3.2. Unlike ditransitive variation, preposition drop appears to be a relatively recent phenomenon. The fact of its relative recency suggests that we may well expect traces of a recent mechanism of spatial diffusion, and the current geographical distribution that we observe in Figure 5 suggests that areas of high population density are more advanced in the spread of the change than closely

Table 1. Pearson's r and p-value for neighbourhood similarity against log-scaled cell population, for the ditransitive data

Grid size	Pearson's r	p-value	Grid size	Pearson's r	p-value
10 × 10	−0.0371	0.7839	60 × 60	0.0113	0.7621
20 × 20	0.0120	0.8781	70 × 70	−0.0005	0.9877
30 × 30	0.0121	0.8377	80 × 80	−0.0021	0.9450
40 × 40	0.0093	0.8481	90 × 90	0.0043	0.8823
50 × 50	0.0090	0.8290	100 × 100	0.0055	0.8398

neighbouring but less densely populated regions. The relationship between neighbourhood similarity and log-scaled population is shown graphically for this dataset in Figure 11, and the corresponding tests of correlation are shown in Table 2.

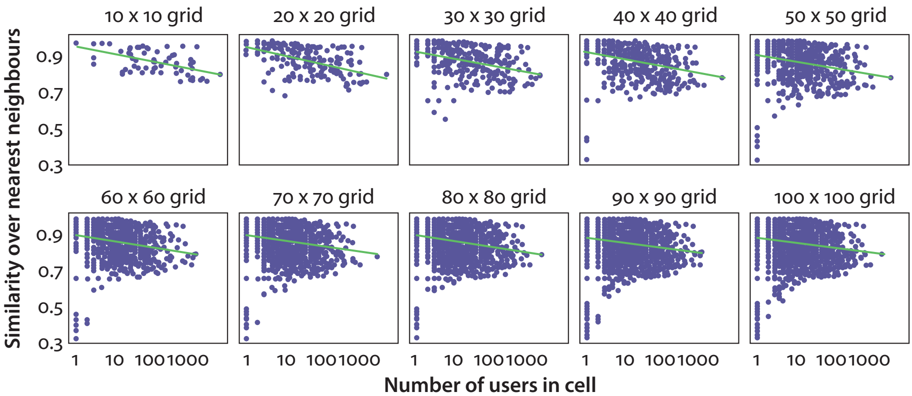

Figure 11. Averaged similarity over the von Neumann neighbourhood plotted against log-scaled locality population, for the preposition-drop data in Figure 5

There is an immediately apparent and statistically robust difference between the measure in this case and the measure as evaluated in the previous dataset. The preposition-drop data show a significant inverse correlation between the population of a cell and its similarity to its nearest neighbours, irrespective of the total number of grid cells chosen. As the population of a cell increases, the probability that it is dissimilar to its immediate neighbourhood increases, and, as such, high-density areas stand out from their low-density surroundings. This is the characteristic appearance of hierarchicality. The impressionistic claim that preposition drop has undergone a recent, underlyingly gravity-type spatial process of diffusion therefore seems to be substantiated.

Table 2. Pearson's r and p-value for neighbourhood similarity against log-scaled locality population, for the preposition-drop data

Grid size	Pearson's r	p-value	Grid size	Pearson's r	p-value
10×10	−0.3553	0.0040	60×60	−0.1165	0.0019
20×20	−0.2256	0.0025	70×70	−0.1224	0.0004
30×30	−0.1905	0.0008	80×80	−0.0901	0.0054
40×40	−0.1697	0.0003	90×90	−0.0808	0.0083
50×50	−0.1328	0.0013	100×100	−0.0791	0.0077

5. Conclusions

This paper has tested the utility of social-media data in producing geographically finely grained descriptions of the distribution of individual variables, and in characterising the spatial diffusion process by which a particular distribution arose. We have seen in the case of variation with ditransitives that social-media data are sufficient to reproduce the large-scale geographic trends seen in datasets established by more traditional means. For preposition drop, it was possible to use data of this type to establish the existence of variation in varieties other than those for which such constructions have previously been documented. The existence of a statistically reliable and measurable inverse relationship between point population density and similarity to the surrounding neighbourhood was used as an indicator of recent hierarchical diffusion; this relationship is present in simulated and real hierarchical data, and absent in other cases.

There are several possible directions for future work in this line. Computational simulation of the output of an idealised spatial diffusion process is suggested, but underexplored in this work, and more sophisticated simulated data would allow comparison of more varied quantitative measures of diffusion. One essential component of a more complete understanding of spatial diffusion is the time dimension; the inference of age-related metadata for individual Twitter users, while a difficult problem (Nguyen et al. 2014), offers potential for the application of geographically rich social-media datasets to apparent-time analysis.

Funding

This work was funded by the Economic and Social Research Council, grant ES/P00752X 'Investigating the diffusion of morphosyntactic innovations using social media'.

References

Bailey, Guy et al. 1993. Some patterns of linguistic diffusion. *Language Variation and Change* 5(3). 359–390. https://doi.org/10.1017/S095439450000154X

Bailey, Laura R. 2018. Some characteristics of Southeast English preposition dropping. *Iberia: An International Journal of Theoretical Linguistics* 10. 48–70.

Bamman, David, Jacob Eisenstein and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135–160. https://doi.org/10.1111/josl.12080

Biggs, Alison. 2014. Passive variation in the dialects of Northwest British English. Paper presented at the *3rd Conference of the International Society for the Linguistics of English (ISLE)*, University of Zürich, 24–27 August. https://www.isle-linguistics.org/assets/content/documents/hogg/Biggs--passive_variation--2014.pdf. (4 February 2020)

Biggs, Alison. 2015. A new case for A-movement in Northwest British English. In Ulrike Steindl et al. (eds.), *Proceedings of the 32nd West Coast Conference on Formal Linguistics (WCCFL 32)*, 218–227. Somerville, MA: Cascadilla Proceedings Project.

Biggs, Alison. 2016. Locating variation in the dative alternation. *Linguistic Variation* 16(2). 151–182. https://doi.org/10.1075/lv.16.2.01big

Bloomfield, Leonard. 1933. *Language*. New York, NY: Holt & Company.

Bresnan, Joan W. and Marilyn Ford. 2010. Predicting syntax. Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213. https://doi.org/10.1353/lan.0.0189

Burridge, James. 2018. Unifying models of dialect spread and extinction using surface tension dynamics. *Royal Society Open Science* 5(1). https://doi.org/10.1098/rsos.171446

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In Shuly Wintner, Sharon Goldwater and Stefan Riezler (eds.), *Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics*, 98–106. Gothenburg: Association for Computational Linguistics. https://www.aclweb.org/anthology/E14-1011.pdf. https://doi.org/10.3115/v1/E14-1011

Eisenstein, Jacob et al. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11). https://doi.org/10.1371/journal.pone.0113114

Eisenstein, Jacob. 2018. Identifying regional dialects in on-line social media. In Charles Boberg, John Nerbonne and Dominic Watt (eds.), *The handbook of dialectology*, 368–383. Hoboken, NJ: Wiley–Blackwell.

Gast, Volker. 2007. I gave it him – on the motivation of the 'alternative double object construction' in varieties of British English. *Functions of Language* 14(1). 31–56. https://doi.org/10.1075/fol.14.1.04gas

Gerwin, Johanna. 2013. Give it me!: Pronominal ditransitives in English dialects. *English Language and Linguistics* 17(3). 445–463. https://doi.org/10.1017/S1360674313000117

Gonçalves, Bruno and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PLoS ONE* 9(11). https://doi.org/10.1371/journal.pone.0112074

Grieve, Jack et al. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2. Article 11. https://doi.org/10.3389/frai.2019.00011

Grieve, Jack, Andrea Nini and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21(1). 99–127. https://doi.org/10.1017/S1360674316000113

Haddican, William. 2010. Theme-goal ditransitives and theme passivisation in British English dialects. *Lingua* 120(10). 2424–2443. https://doi.org/10.1016/j.lingua.2009.11.003

Haddican, William and Daniel E. Johnson. 2012. Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics* 18(2). 31–40.

Hägerstrand, Torsten. 1952. *The propagation of innovation in waves* (Lund Studies in Geography, Ser. B, Human Geography, 4). Lund: Royal University of Lund, Department of Geography.

Haggett, Peter. 1965. *Locational analysis in human geography*. London: Arnold.

Hall, David. 2019. P D drop and pseuo-incorporation in London English. In Maggie Baird and Jonathan Pesetsky (eds.), *NELS 49: Proceedings of the Forty-Ninth Annual Meeting of the North East Linguistic Society: Volume 2*, 85–89. Amherst, MA: GLSA.

Hecht, Brent and Monica Stephens. 2014. A tale of cities: Urban biases in volunteered geographic information. In Eytan Adar and Paul Resnick (eds.), *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 197–205. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Huang, Yuan et al. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59. 244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Jones, Taylor. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech* 90(4). 403–440. https://doi.org/10.1215/00031283-3442117

Labov, William. 2001. *Principles of linguistic change, vol. 2: Social factors* (Language in Society 29). Malden, MA: Wiley–Blackwell.

Malik, Momin et al. 2015. Population bias in geotagged tweets. In Derek Ruths and Jürgen Pfeffer (eds.), *Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662. (10 January 2020.)

Myler, Neil. 2013. On coming the pub in the North West of England: Accusative unaccusatives, dependent case, and preposition incorporation. *Journal of Comparative Germanic Linguistics* 16(2/3). 189–207. https://doi.org/10.1007/s10828-013-9055-1

Nguyen, Dong et al. 2014. Why gender and age prediction from tweets is hard. Lessons from a crowdsourcing experiment. In Junichi Tsujii and Jan Hajić (eds.), *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, 1950–1961. Dublin: Dublin City University and Association for Computational Linguistics.

Office For National Statistics, Geography Division. 2016. Index of place names in Great Britain (July 2016). https://www.ons.gov.uk/methodology/geography/geographicalproducts/otherproducts/indexofplacenamesipn

Olsson, Gunnar. 1965. *Distance and human interaction. A review and bibliography* (*Bibliography Series 2*). Philadelphia, PA: Regional Science Research Institute.

Ordnance Survey Ireland. 2016. Townlands – OSi national placenames gazetteer. https://data-osi.opendata.arcgis.com/datasets/townlands-osi-national-placenames-gazetteer. (10 January, 2020)

Orton, Harold. 1962. *Survey of English Dialects (A): Introduction*. Leeds: Arnold.

Orton, Harold, Stewart Sanderson and John D. A. Widdowson. 1978. *The linguistic atlas of England*. London: Croom Helm.

Pavalanathan, Umashanthi and Jacob Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. In Lluís Màrquez, Chris Callison-Burch, Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2138–2148. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1256. https://www.aclweb.org/anthology/D15-1256

Reis, Stefan. et al. 2017. UK gridded population 2011 based on Census 2011 and Land Cover Map 2015. NERC Environmental Information Data Centre. (Dataset). https://doi.org/10.5285/0995e94d-6d42-40c1-8ed4-5090d82471e1

Russ, Brice. 2012. Examining large-scale regional variation through online geotagged corpora. Paper presented at the *Annual Meeting of the American Dialect Society*, Portland. http://www.briceruss.com/ADStalk.pdf

Schmidt, Johannes. 1872. *Die verwantschaftsverhältnisse der indoger manischen sprachen*. Weimar: Böhlau.

Shoemark, Philippa et al. 2017. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In Mirella Lapata, Phil Blunsom and Alexander Koller (eds.), *Proceedings of the 15th conference of the European chapter of the association for computational linguistics, vol. 1, long papers*. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1116

Siewierska, Anna and Willem B. Hollmann. 2007. Ditransitive clauses in English with special reference to Lancashire dialect. In Mike Hannay and Gerard J. Steen (eds.), *Structural-functional studies in English grammar. In honour of Lachlan Mackenzie, vol. 83: Studies in Language Companion Series*, 83–102. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.83.06sie

Stevenson, Jonathan. 2016. Dialect in digitally mediated written interaction: A survey of the geohistorical distribution of the ditransitive in British English using Twitter. Master's Thesis. York: University of York.

Strelluf, Christopher. 2019. Anymore, it's on Twitter. Positive anymore, American regional dialects, and polarity licensing in tweets. *American Speech* 94(3). 313–351. https://doi.org/10.1215/00031283-7587883

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects. A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2). 215–246. https://doi.org/10.1017/S0047404500004358

Upton, Clive and John D. A. Widdowson. 1996. *An atlas of English dialects: Region and dialect*. Oxford: Oxford University Press.

Wikle, Thomas and Guy Bailey. 1997. The spatial diffusion of linguistic features in Oklahoma. *Proceedings of the Oklahoma Academy of Science* 77. 1–15.

Willis, David. 2020. Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa: a journal of general linguistics* 5(1). 103. https://doi.org/10.5334/gjgl.1073

Wolk, Christoph et al. 2013. Dative and genitive variability in Late Modern English. Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419. https://doi.org/10.1075/dia.30.3.04wol

Yáñez-Bouza, Nuria and David Denison. 2015. Which comes first in the double object construction? *English Language and Linguistics* 19(2). 247–268. https://doi.org/10.1017/S136067431500012X