

---

**RESEARCH**

# Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh

David Willis

University of Oxford, Jesus College, Oxford, GB

david.willis@ling-phil.ox.ac.uk

---

Data gathered from social media have been used extensively to examine lexical dialect variation in widely used languages such as English and Spanish, but their use to date in morphosyntax and for lesser-used languages has been more limited. This paper tests the usefulness of using data derived from Twitter to address traditional questions in dialect syntax and sociolinguistics. It uses two cases studies from Welsh – the form of the second-person singular pronoun in various syntactic contexts, and the availability of auxiliary deletion – to assess whether datasets based on Twitter data can successfully replicate and enhance results derived by traditional means. The results of the case studies coincide to a large extent with distributions established in existing studies, even ones using entirely different methods, such as dialect questionnaires or acceptability judgment tests. Twitter data also show considerable success in establishing implicational hierarchies and conditioning factors comparable to those typical of the field. Where the results differ from existing studies, the differences may be due to the younger demographics of Twitter users, or to differences in the quantity of data provided by different methodologies. The results produce patterns closer to spoken data than to written data, giving us reasonable confidence in such data as a relatively good proxy for spoken usage of large numbers of language users.

---

**Keywords:** Dialectology; Syntactic variation; Computational sociolinguistics; Welsh; Personal pronouns; Auxiliary deletion

---

## 1 Introduction

While data from social-media platforms such as Twitter and Facebook have been used by linguists to investigate lexical variation (Russ 2012, Gonçalves & Sánchez 2014 etc.) and change (Grieve, Nini & Guo 2016; 2018), use of such material for morphosyntax has been relatively limited to date. This paper aims to demonstrate the successful application of Twitter data to investigate morphosyntax, to examine ways of dealing with methodological problems, and to test the extent to which it is possible to replicate results produced by traditional methods of investigating geospatial variation in morphosyntax (dialect surveys and spoken corpora) using social-media data.

Use of social-media data to examine dialect variation has grown substantially in recent years. Early work typically limited itself in several ways: first, by restricting itself to a corpus of tweets for which users enabled the automatic inclusion of GPS location metadata with their tweet; and, above all, by focusing on variables that can easily be extracted via searches for particular strings of characters, typically lexical variables. The pioneering work of Gonçalves & Sánchez (2014), for instance, used a corpus of 50 million GPS-localized

tweets in Spanish to map lexical variation across the Spanish-speaking world. Other studies in this tradition include Scheffler et al. (2014), Gonçalves & Sánchez (2016), Huang et al. (2016), Donoso & Sánchez (2017), Eisenstein (2017), Shoemark, Kirby & Goldwater (2017) and Grieve et al. (2019).

This work successfully demonstrates broad-brush variation; for instance, Gonçalves & Sanchez (2014) show that the Spanish word for ‘swimming pool’ is *alberca* in Mexico, *pileta* in Argentina and Paraguay, and *piscina* everywhere else. While promising, it has not been fully established that these methods can identify variation within regions at the local level. Furthermore, these studies do not address central questions in current work on language variation and change, much of which focuses on understanding how phonological and morphosyntactic innovations arise (actuate) and diffuse through space. To be fully integrated into mainstream work in language variation and change, social-media data need to be applied to core theoretical questions. For instance, typical variables in quantitative work show that the frequency of variants (whether innovative or in stable variation) is conditioned by both linguistic and external factors. External factors (gender, age, social status, network) may be difficult to establish in Twitter data, while the lexical variables chosen are often too simple for linguistic factors (phonological environment, clause type etc.) to be relevant. Abitbol et al. (2018: 1126) thus highlight the need for future work both to go beyond the study of lexical variation in English and to introduce subtler external factors such as social class into computational sociolinguistics.

Current work has begun to address some of these challenges. Eisenstein (2015) extends Twitter-based linguistic research to phonology, showing that the frequency of phonetically based unconventional spellings is sensitive to phonological context in a way that is typical for phonological variables in sociolinguistic studies. For instance, orthographic deletion of < t d > in tweets in words such as *jus(t)* or *ol(d)* is inhibited by a following vowel in the same way that phonological deletion of /t d/ (coronal stop deletion) is inhibited by a following vowel in speech. Nevertheless, the existing literature on this variable demonstrates complex hierarchies of conditioning by both preceding and following phonological context (Fasold 1972: 38–115; Tagliamonte & Temple 2005; Hazen 2011 etc.) alongside, in some studies, morphological factors (Guy & Boyd 1990). In this research context, the role of a following vowel is a relatively minor issue. Ideally, we would be able to address the same range of conditioning environments as found in sociolinguistically oriented studies.

Other studies have also demonstrated phonological variation using social-media data. Van Halteren, Van Hout & Roumans (2018) present data to suggest that the geospatial distribution of phonological variation in Limburg Dutch in Twitter data mirrors traditional dialect findings. Jones (2015) demonstrates previously undocumented geographically patterned phonological variation within tweets in African American Vernacular English (AAVE).

The question of observing ongoing change is addressed by Grieve, Nini & Guo (2016; 2018), who use Twitter to look at real-time diffusion of lexical innovations in American English over a period of several years.

A number of works have also begun to look at morphosyntactic variation. Haddican & Johnson (2012) used a Twitter corpus to test for differences between and within US and British English in the frequency of discontinuous orders with particle verbs (*put the lights out* vs. *put out the lights*). Doyle (2014) showed that Twitter data could broadly establish the distribution of double modal *might could* across the southern United States, although the results still do not provide good resolution at local level, with questionable pockets of double modals appearing in major cities. Furthermore, conditioning factors could not be considered and the variable could not be defined in the standard way as the relative

frequency of one variant compared to other possible variants (cf. Van Halteren, Van Hout & Roumans 2018: 140), since other variants, such as *might have been able to* were not collected. Stevenson (2016) used Twitter to establish the geospatial distribution of variation in the syntax of ditransitive verbs (specifically the past tense of *give* and *send*) with pronominal objects in British English (*gave it me/me it/it to me* etc.), and showed that these patterns closely match those of the *Survey of English Dialects* (Upton & Widdowson 1996: 52). Claes (2017) used Twitter data to show that plural agreement in Spanish existential constructions is conditioned by tense, negation and the semantics of the associate noun phrase, and that this finding is replicated in a corpus of traditional spoken Spanish and in previous studies of Caribbean Spanish. Ljubešić, Miličević Petrović & Samardžić (2018) plot the distribution for 16 variables, including 7 morphosyntactic ones, across the Slavic languages of the western Balkans using Twitter data. While they focus mainly on examining whether dialect differences match current political boundaries, they also successfully demonstrate the viability of using social-media data to investigate both morphological and syntactic variables. Strelluf (2019) paints a picture of the geospatial distribution of positive *anymore* in American English tweets that is in line with established distributions, and analyses those patterns in terms of language-internal factors whose relevance for the phenomenon is well established, namely negative-polarity licensing contexts and clausal position.

Despite these welcome contributions, the total volume of work on morphosyntax remains limited, and only a few attempts (notably Claes 2017 and Strelluf 2019) have been made to incorporate language-internal conditioning factors into investigations. Again, with a handful of exceptions, work also remains heavily focused on English and Spanish, languages where the volume of tweets available is vast, and little attention has been paid to what techniques are appropriate for languages with a lesser presence in social media. Finally, while consensus isoglosses have been successfully replicated in a number of cases, apart from Van Halteren, Van Hout & Roumans's (2018) work on phonological variation within Limburg, this has mostly been done at the macro rather than the micro level. Identification of micro-level patterns, typical of existing dialect atlases, is naturally a more challenging task and success therefore more difficult to demonstrate.

This paper addresses some of these issues by considering two cases of morphosyntactic dialect variation in Welsh, a language with a relatively small presence in social media.

The first case study concerns the dialect distribution of the Welsh strong second-person singular pronoun *chdi*. This occurs in northwestern Welsh dialects in various syntactic environments (after a non-inflecting preposition, in fronted focus position, as subject of an auxiliary etc.). The exact set of possible syntactic environments for its use varies from dialect to dialect according to an implicational hierarchy. Traditional studies of the dialect distribution of *chdi* show a concentric-ring pattern, with a core dialect in which *chdi* is permitted in the largest set of environments, and successively more distant dialects allowing it in fewer and fewer of them. This pattern results from historical wave-like distribution of *chdi* in more and more contexts via contagious diffusion from a central core (Bailey 1973: 65–109; see also Britain 2013 [2002] on models of diffusion more generally). The research question here is therefore to what extent both the overall pattern and the details of the linguistic factors that give rise to the implicational hierarchy can be established from Twitter data.

The second case study involves deletion of finite forms of auxiliary 'be' before subject pronouns in spoken Welsh verb-initial word order. In addition to the question of the geospatial distribution of this feature, and whether it can be accurately established using Twitter data, this variable allows us to test a second question, namely the extent to which social-media data provide a good proxy for spoken data. Auxiliary deletion is known to

occur at exceptionally high rates in spoken Welsh, but it is not a feature of the formal written language. If Twitter data are a good proxy for spoken data, we would expect to find rates closer to those found in spoken corpora than in written corpora.

This paper begins (section 2) by considering the data-collection procedure and methodological issues involved in Twitter research for a small language. Sections 3 and 4 set out the two case studies in turn, beginning in each case by describing the linguistic variable in question and the existing state of knowledge established using traditional methods (the reference distribution), before setting out and mapping the Twitter data for these variables in comparison. In both cases, the Twitter results turn out to be broadly consistent with the reference distribution and with other findings of existing studies. The reasons for mismatches in individual points of detail are discussed both in the case studies and in the conclusion (section 5).

## 2 Methods and data collection

Using Twitter to work with Welsh presents somewhat different issues from working with a major world language. For major world languages, the volume of data is such that it may be possible to discard a very large proportion (even 99%) of it, and still have a useful body of evidence. Welsh is regularly used in social media: Kevin Scannell ([indigenoustweets.com](http://indigenoustweets.com)) reports over 14,000 Twitter users as tweeting in Welsh with some 5.7 million tweets having been composed in Welsh, and the number has likely risen significantly since these figures were last updated in 2014. While substantial enough to be the basis for research, this corpus is by no means so large that we can afford to disregard large quantities of useful material from it. For this reason, it is not feasible to limit oneself to tweets by users who have enabled automatic GPS geotagging of their tweets on their mobile phones, as many studies have done. Only a small proportion of tweets (1–2%, Eisenstein 2017: 369) have such metadata. Users who enable this geotagging on their phones are also likely to be more urban and technologically minded than average, exacerbating existing biases in Twitter data (for a measure of the bias against rural users in GPS-localized Twitter data in the United States, see Hecht & Stephens 2014). Furthermore, relying on GPS-localized tweets alone would reduce the number of distinct users and hence the independence of the dataset, potentially introducing overreliance on idiosyncrasies of particular individuals in some part of the data. It was therefore decided to use all available tweets containing the relevant linguistic variables during a period of observation and, for mapping, to develop a strategy for assigning geographic locations to tweets based on other information provided by Twitter users.

As linguists mapping dialect variation, we are interested not in the location of a user when they are composing a tweet, nor even particularly where they live at the time they are tweeting, but rather where they acquired their language. We would also ideally like to know other demographic information associated with the user, such as their gender, age, social class and occupation, and, in the context of a lesser-used language such as Welsh, aspects of their language background, including the means by which they acquired Welsh (in the home, at school, as an adult learner etc.) and perhaps even the extent to which they participate in Welsh-language culture. These, and others, are all demographic factors that would be taken into consideration in a well-designed dialectological or sociolinguistic study of a given linguistic variable. Unfortunately, none of them is straightforwardly available to us for Twitter users. However, the sheer volume of easily available data is highly attractive, if these limitations can be overcome: a study of variation in Welsh with 14,000 informants is vastly beyond what could normally be achieved.

A corpus of Twitter data containing second-person singular pronouns was collected over a total of 52 days during 2016 (15–25 May, 9–17 June, 2–11 July, 30 November–10



December, 12–22 December). Data were collected using the Chorus Project Tweetcatcher application (Brooker, Barnett & Cribbin 2016). In addition to the base forms *ti*, *di* and *chdi*, various abbreviated and combined spellings in common use on Twitter were also included: *bochdi*, *boti* (i.e. *bod chdi* and *bod ti* ‘that you are’); *chdin*, *tin*, *din*, *chdn*, *tn*, *dn* (i.e. *chdi’n* and *ti’n* ‘you’ + progressive particle); *genti* (i.e. *gen ti* ‘with you’); *ichdi*, *iti* (i.e. *i chdi* and *i ti* ‘to you’); *ochdi*, *oti* (i.e. *oeddet ti* ‘you were’); *sachdi*, *sati* (i.e. *buasa chdi* and *buaset ti* ‘you would be’); *tisho*, *tisio*, *tishe*, *tisie* (i.e. *ti eisiau* ‘you want’) (plus the in fact non-existent equivalents with *chdi* in the cases where no form with *chdi* is given in this list). It is possible, indeed likely, that this does not exhaust the possible orthographic variation in genuine instances of these pronouns, but the effect of such omissions is likely to be small.

Although the corpus contained only tweets marked by the Twitter language-identification algorithm as being in Welsh, a considerable number were in other languages (mostly Kurdish, Bahasa Indonesia, Tagalog or Italian). These were removed manually. Also excluded were tweets containing only quotations (Bible verses, poetry), proper names, and obvious spam. None of the tweets appeared to have been produced by automated bots. Retweets and tweets from national-level institutional accounts (government organizations, broadcasters) were removed, along with resources for learners and tweets from users who identified themselves as second-language (L2) learners in their user description. The result was a dataset of 6,664 tweets in Welsh containing second-person singular pronouns from 2,932 distinct users. Although, in principle, such Twitter searches provide only a sample of tweets, manual checking of the data suggested that all or nearly all tweets coded as Welsh by Twitter’s language-identification algorithm and matching the query terms had been returned.

Tweets from accounts of local institutions were retained as potentially reflecting local usage. The choice to retain local but not wider institutional tweets inevitably influences the extent to which spoken forms are found, since even local institutions are likely to be more formal in the linguistic preferences than personal users. Self-identified L2 speakers were removed, but it is likely that others were present, hence an unknowable number remain in the dataset.

From subsequent analysis, it became clear that a proportion of the tweets collected were conventional social interactions frequently performed in Wales in Welsh even by non-Welsh-speakers (thanking people, wishing people a happy birthday and wishing people a happy Christmas). These are included in the dataset, but treated separately, and their effect will be considered separately in the analysis below.

Information about users’ geographic origin is crucial to any study of dialect variation. Tweets include various metadata that are potentially of use for this task, most obviously the location and description fields of the metadata provided by users themselves. In the current study, the location field of the metadata was left blank in 26.1% of tweets. While the majority of data (73.9%) thus provides some kind of user-provided information, this was not always useful: many users provided their location simply as the “UK” (35 tweets, 0.5%), “Wales” (175 tweets, 2.6%) or its Welsh-language equivalent “Cymru” (427 tweets, 6.4%) or similar. A few gave non-geographic locations of the type “In my kitchen”. Even when more specific locations were provided, they were not always particularly useful: a description such as “North Wales” or “Gogledd Cymru” (257 tweets, 3.9%) is of limited direct practical use for linguistic geography and such information was also disregarded in producing a localization. However, in many cases (3,062 tweets, 45.9%), the user location field did provide a specific city, town, village or small region sufficient to associate the tweet with a specific location. Where a user mentioned two locations, it was assumed that they had grown up in the smaller one and moved to the larger: formulations such as “Llanrwst/Cardiff” seemed mostly to be used by students to indicate their home town

and university town/city. Such users were therefore treated as coming from the smaller location.

Where the user location field proved inadequate, the user description field was used. This is a free text field where users can provide any information they like about themselves, and, while some (1181 tweets, 17.7%) left this field blank, most users wrote something here, including interests, hobbies, political views, parenthood, age, and either current location or the place they grew up in or identified with. Any information here was added to that obtained from the user location field, providing a new or more fine-grained localization for a further 304 tweets (4.6%).

Of course, this procedure is no guarantee that users are mapped to the places where they grew up, but it is the best approximation that can be made. Furthermore, this procedure can be applied to many more tweets than those for which geotags (GPS location data) are available and, in any case, is more likely to yield users' actual places of upbringing. In total, a usable localization was thus established for 3,366 tweets (50.5% of the dataset). While it is inevitable that this procedure leads to some tweets being assigned localizations that do not accurately reflect where the users acquired their language, it was anticipated that such errors would be relatively insignificant in the overall picture. If the geospatial distribution of features that emerges from the study matches or enhances what we find in established work, that expectation will have been borne out.

Relatively few users (336 tweets, 5.0%) gave direct information about their age. Users who did provide such information were almost always in their twenties or late teens. This age distribution is very similar to that found by Sloan et al. (2015) for a corpus of English-language tweets where age was identified directly from information in the user's description field. This is clearly much younger than the general population, although it is also clear from manual inspection of the data that Twitter users in their thirties and older are simply much less likely to state their age directly in their user description. The overall age profile of Twitter users in the UK is younger than the general population. IPSOS Connect (2017: 18) estimate 25% of users to be under the age of 25 (compared to 15% of the general population) and 46% under 35 (compared to 32% in the general population) (Great Britain only). Chaffey (2019) estimates that, as of October–December 2014, 28% of Twitter users were aged 16–24 and over 30% in the 25–24 age group (entire UK). If we assume a similar profile in our data, it is clear that many users in their thirties and above are present in the data, but do not identify themselves as such via their user description. The limited range of ages identifiable within the data makes it difficult to say anything easily about apparent-time variation within the data without further analysis (e.g. by assuming that, collectively, the group identifying themselves as parents is on average older than the group identifying themselves as students). Pending further investigation of how to extract information on age, gender, social class etc. in social-media-based linguistic research, the age dimension will not be considered further in this study.

Finally, tweets were annotated for the form of the pronoun and for syntactic context as explained for the individual case studies below. Since both case studies involve second person singular pronouns in some form or other, the same base dataset is used in both cases. This dataset is provided in anonymized form in the Supplementary Files linked to this article.

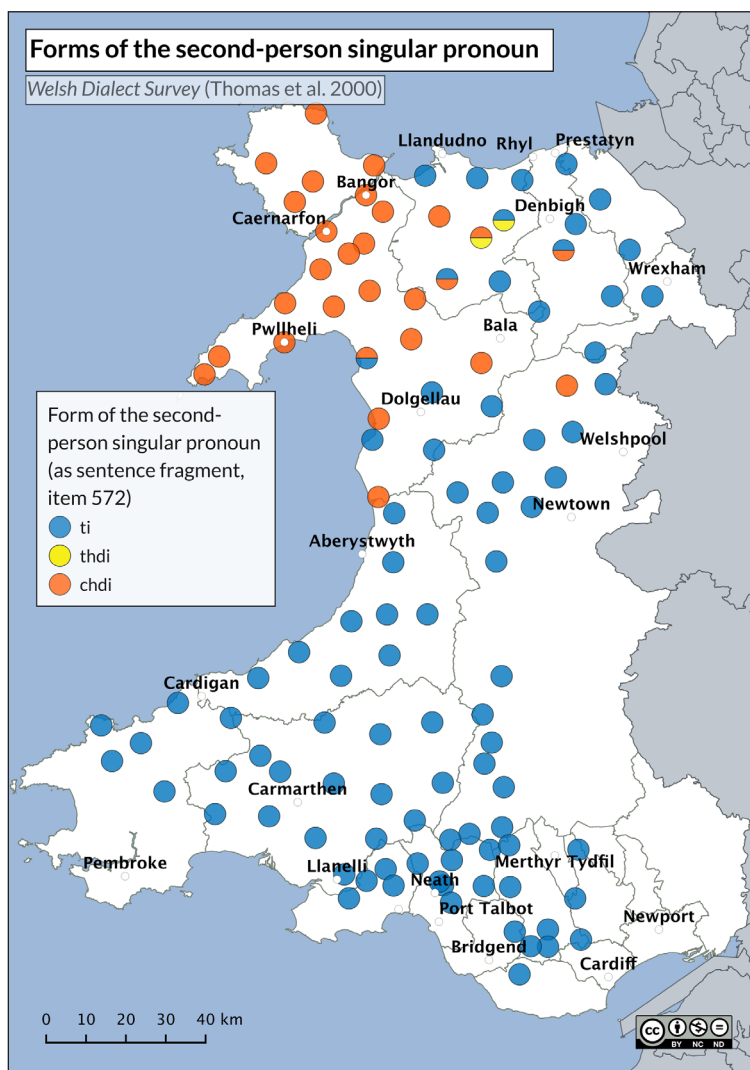
### 3 The form of the second-person singular pronoun

#### 3.1 The variable

In most varieties of Welsh, the second-person singular (informal) pronoun is *ti*. Another form, *di*, occurs in certain syntactic environments, such as the postverbal subject of certain auxiliaries that end in a vowel or as a possessor in a possessed noun phrases. In

northwestern varieties of Welsh, an alternate form of the pronoun has arisen, namely *chdi*. Historically, this is the result of phonological reduction and dissimilation of the Middle Welsh strong second-person singular pronoun *tydi*. After certain prepositions and conjunctions, the consonant alternation known as aspirate mutation was (and still is) triggered (e.g. *tŷ* ‘house’ but *â thŷ* ‘with a house’, with /t/ > /θ/). This applied regularly to *tydi* in the relevant syntactic environments, hence Early Modern Welsh *â thydi* ‘with you’. Syncope of the unstressed schwa led to *â thdi* (normally spelled *â’th di* in Early Modern Welsh). In northwestern and central northern varieties of Welsh, *thdi* was reanalysed as a new pronoun and spread beyond syntactic environments where aspirate mutation was triggered. Finally, a dissimilation occurred in some northwestern variety of nineteenth-century Welsh, so that *thdi* /θdi/ became *chdi* /χdi/, and this new form eventually supplanted the earlier form entirely (Willis 2017).

The dialect distribution of *chdi* is relatively well established. It was the subject of one question asked for the *Welsh Dialect Survey* (Thomas et al. 2000: 555), most of whose informants were born in the 1920s; see Figure 1.<sup>1</sup> This shows *chdi* to be solidly attested



**Figure 1:** Distribution of the forms of the second-person singular pronoun (based on data from Thomas et al. 2000: 555).

<sup>1</sup> From the contextual data given by Thomas et al. (2000: 555), it appears that informants were asked for the form of the pronoun used in sentence-initial focus position or as an isolated sentence fragment in coordination (*fi a t(h)i* ‘me and you’). The patterns that emerge in Figure 1 are consistent with this.

in the whole of the northwest, with some scattered attestation in adjacent areas of the northeast; it is not present in the south.

The syntactic distribution of *chdi* is quite complicated. As we have seen, *chdi* (or rather its ancestor form *thdi*) arose in one particular syntactic environment, namely after a small group of prepositions and conjunctions that triggered aspirate mutation (namely *â* ‘with’, *efo* ‘with’, *gyda* ‘with’, *â* ‘(equative) as’, *na* ‘than’, and *a* ‘and’). From here, it seems to have spread first to focus constructions and to other independent contexts, such as free-standing answers to questions. Welsh distinguishes strong from weak pronouns: strong pronouns occur in contexts without agreement, while weak ones occur in contexts associated with agreement. The contexts in which *chdi* first appeared were typical strong contexts, and all dialects that use *chdi* use it in these contexts, namely after non-inflecting prepositions, (1); as the second conjunct in co-ordinated noun phrases, (2); in sentence-initial (fronted) focus position, (3); and in fragment answers, (4) (which can be thought of as a reduced form of the type in (3)).

- (1) efo chdi  
with you  
‘with you’
- (2) fi a chdi  
me and you  
‘me and you’
- (3) Chdi sy ’n gwybod.  
you be.PRS.REL PROG know.INF  
‘It’s you that knows.’
- (4) Pwy neith gymryd o? Chdi.  
who do.FUT.3SG take.INF it you  
‘Who will take it? You.’

However, *chdi* has spread beyond these environments, and is found also in various contexts traditionally associated with agreement. This may be part of a trend towards loss of agreement more generally in Welsh (Willis 2017: 44–46). Thus, we find *chdi* as the subject of complement clauses headed by *bod*, the nonfinite form (verbnoun) of *bod* ‘be’. In place of the more traditional construction illustrated in (5), with second-person singular agreement marker *dy* and pronoun *di*, we now find (6), with no agreement marker and *chdi* alone.<sup>2</sup>

- (5) Dwi ’n gwbod **dy fod di** ’n siarad Cymraeg.  
be.PRS.1SG PROG know.INF 2SG be.INF you PROG speak.INF Welsh  
‘I know that you speak Welsh.’
- (6) Dwi ’n gwbod **bo’ chdi** ’n siarad Cymraeg.  
be.PRS.1SG PROG know.INF be.INF you PROG speak.INF Welsh  
‘I know that you speak Welsh.’

<sup>2</sup> The citation form *bod* undergoes soft mutation (initial /b/ > /v/) to become *fod* in (5), and the final /d/ deletes before the consonant cluster /χd/ in *chdi* in (6).



There are a number of other contexts to which *chdi* has spread. With the object of an inflecting preposition, such as *am* ‘about’, *amdanat ti* ‘about you’ is replaced by *amdana chdi*. With the subject of various auxiliary and modal elements, *oeddat ti* ‘you were’ is replaced by *oedda chdi*; *(by)sat ti* ‘you would (be)’ is replaced by *(by)sa chdi*; *rhaid (i) ti* ‘you must (lit. ‘it is necessary for you)’ is replaced by *rhai’ chdi*; and *byddi di* ‘you will (be)’ is replaced by *by(dd) chdi*. As the object of a nonfinite verbal form, *chdi* replaces doubling of a preverbal agreement marker and postverbal *di* (the pattern is parallel to that in (5) and (6) above):

(7) Dwi            ’n    dy   nabod    di.  
 be.PRS.1SG PROG 2SG know.INF you  
 ‘I know you.’

(8) Dwi            ’n    nabod    chdi.  
 be.PRS.1SG PROG know.INF you  
 ‘I know you.’

Finally, in possessive noun phrases, *chdi* again replaces doubling of a pronominal agreement element and postnominal pronoun:

(9) dy    gar    di  
 2SG car    you  
 ‘your car’

(10) car    chdi  
 car    you  
 ‘your car’

In other contexts, namely those where agreement is secure, *chdi* has made little to no impact and *ti* (or its syntactically conditioned variant *di*) is compulsory in all dialects. Such contexts include the subject of lexical verbs, (11); the subject of the present tense of the verb *bod* ‘be’, (12); and the subject of imperatives, (13).

(11) Cei            di/\*chdi    un.  
 get.FUT.3SG you        one  
 ‘You’ll get one.’

(12) Ble    wyt            ti/\*chdi?  
 where be.PRS.2SG you  
 ‘Where are you?’

(13) Dal            di/\*chdi    ati.  
 keep.IMP.2SG you        to.3FSG  
 ‘Keep at it.’

The spread of *chdi* is documented historically, and the historical trajectory is reflected in current dialect variation. Historically, we know that *chdi* is first attested in different syntactic contexts at different dates. The contexts in which it is attested earliest are also those where it is found today in the widest range of dialects. Willis (2017) investigates the

historical sequence of the spread of *chdi* in the nineteenth and early-twentieth centuries and finds the following order of innovation:

- (14) object of non-inflecting preposition > other independent (focus, *dyna* ‘there is’ etc.) > subject of nonfinite *bod* ‘be’ > after *i* ‘for, to’ > object of other inflecting preposition/subject of finite auxiliary

*Chdi* is absent in all other contexts in the historical data. Willis (2017) also conducts a study of contemporary patterns of geospatial variation by context. He shows that current variation mirrors the historical sequence of events fairly closely, consistent with wave-like diffusion from a single zone of innovation, in the sense that *chdi* is found to its fullest geographic extent only as the object of non-inflecting prepositions and in other independent contexts. In all other contexts, it is found in a subset of this area, with those that innovated earlier showing a wider geographic distribution than more recent ones. The relative order of contexts in dialect variation is the following (based on the intercept values of a global logistic regression, Willis 2017: 50):

- (15) object of non-inflecting preposition *efo* ‘with’ > other independent (focus, *dyna* ‘there is’ etc.) > *i* ‘to’ > object of other inflecting preposition > subj. of *dylai* ‘should’ > subj. of *oedd* ‘were’ > *rhaid* ‘must’ > subj. of conditional > subj. of *bydd* ‘will be’ > obj. of *gan* ‘with’

Of these, *efo* ‘with’ and the independent contexts had already reached their current state in speakers born in the 1920s (Thomas et al. 2000: 555). Apart from the subject of nonfinite *bod* ‘be’, which is not included in (15), and allowing for the fact that (15) includes a number of environments where *chdi* innovated only after the end of the historical period covered in (14), the order of the items is identical in the two implicational hierarchies. Given that the two hierarchies above are based on entirely independent data and methods, we can be reasonably confident that they reflect the actual course of development and geographical distribution of *chdi*. We can therefore treat (15) as the reference distribution, a standard of comparison for the Twitter data.

The first question, then, is whether Twitter data can accurately identify the geographic area in which *chdi* is used. If this question is answered in the affirmative, then a second, more demanding question is whether Twitter data can provide a level of contextual geographic detail comparable to that obtained by traditional questionnaire and/or corpus methods.

### 3.2 Global analysis of the data

First consider the overall distribution of results for all 6,664 tweets identified as containing a second-person singular pronoun. This is given in Table 1.

Before we turn to examine the overall patterns, some discussion of the status of formulas is necessary: formulaic expressions turn out to have a much higher propensity to contain *ti* than *chdi* and were therefore treated as a separate context. This concerned *da* (*iawn*) *ti* ‘(very) good (on) you’, *penblwydd hapus i ti* ‘happy birthday to you’, *Nadolig Llawn i ti* ‘Merry Christmas to you’, *diolch ti* and *diolch i ti* ‘thank you’, *helo ti* ‘hello, you’ and *hwyl ti* ‘bye, you’. In some of these cases the syntactic structure is not clear: is *diolch ti* ‘thank you’ an elided form of *diolch i ti* ‘thank (to) you’ or is it a calque of English *thank you*? If the latter, it is not clear whether *ti* should be treated as the object of the nonfinite verb *diolch* ‘thank’ or as a syntactically independent, unintegrated unit. Even where the structure is fairly clear, these formulas did not pattern at all with their associated

**Table 1:** Overall distribution of second-person singular pronominal forms by syntactic context.

Context	<i>chdi</i>	<i>di</i>	<i>ti</i>	<i>ti/di</i>	total	% <i>chdi</i>
possessor of noun phrase	161	102	109	211	372	43.3
subject of <i>oedd</i> 'was'	45	0	70	70	115	39.1
subj. of conditional aux.	40	1	65	66	106	37.7
object of nonfinite verb	182	74	311	385	567	32.1
independent use (focus etc.)	85	1	179	180	265	32.1
object of infl. prep.	46	1	100	101	147	31.3
unclassified	2	2	3	5	7	28.6
object of non-infl. prep.	73	0	203	203	276	26.4
subject of nonfinite <i>bod</i> 'be'	94	9	258	267	361	26.0
object of <i>i</i> 'to'	155	0	554	554	709	21.9
<i>rhaid (i) ti</i> 'you must'	8	0	30	30	38	21.1
subject of <i>dylai</i> 'should'	7	0	27	27	34	20.6
other	5	7	24	31	36	13.9
formulas	43	0	370	370	413	10.4
subject of <i>bydd</i> 'will be'	7	61	9	70	77	9.1
object of finite verb	11	88	32	120	131	8.4
object of <i>gan</i> 'with'	10	1	132	133	143	7.0
subject in auxiliary deletion	6	5	1885	1890	1896	0.3
subject of lexical verb	1	361	114	475	476	0.2
subject of imperative	0	24	10	34	34	0.0
subject of pres. tense <i>bod</i> 'be'	0	0	461	461	461	0.0
total	981	737	4946	5683	6664	14.7

context: *penblwydd hapus i ti* 'happy birthday to you' clearly exemplifies the context *i* 'to', but was only ever found with *ti*, never with *chdi*, even though *i chdi* 'to you' was common elsewhere. Formulaic expressions are widely recognized to be linguistically conservative and may resist linguistic innovations that have spread to most productive contexts for some centuries. Furthermore, in Wales, such formulas are widely known, in their standard form with *ti*, by people with no other knowledge or only limited knowledge of the language, and may therefore be used in an otherwise English-language context. Obvious instances of performance of Welsh by non-speakers were manually removed from the dataset (e.g. a single formulaic Welsh expression in a tweet otherwise in English or anti-Welsh racist insults from accounts otherwise tweeting in English). However, accounts could not be systematically examined to try to establish whether the user was a competent speaker of Welsh. Thus, it is likely that some use of formulas represented unidentified cases of this, cf. difficulties identified for Twitter research by Jones (2015: 411) arising from performance of African American Vernacular English. Thus, while searching for formulaic expressions in an untagged Twitter corpus might seem like an attractive way of extracting large amounts of data quickly, in practice it seems unlikely to yield satisfactory results.

Since the Twitter data include all instances of second-person singular pronouns, they inevitably cover more syntactic contexts than other studies. Restricting ourselves solely to the contexts included in (15), the following hierarchy emerges from Table 1.

- (16) subj. of *oedd* ‘were’ > subj. of conditional > other independent (focus, *dyna* ‘there is’ etc.) > object of other inflecting preposition > object of non-inflecting preposition > *i* ‘to’ > > *rhaidd* ‘must’ > *dylai* ‘should’ > *bydd* ‘will be’ > *gan* ‘with’

This hierarchy differs from that in (15) in a number of ways. On the positive side, it correctly places *bydd* ‘will be’ and *gan* ‘with’ at the far right edge of the hierarchy, in the correct order: Willis (2017) shows that these two contexts differ sharply from the others in allowing *chdi* over a much narrower geographic range. However, it fails to order correctly the left-hand portion of the hierarchy. In particular, the historically earliest contexts, namely object of non-inflecting preposition and other independent are not identified as the most favourable environments. A number of comments on this are in order.

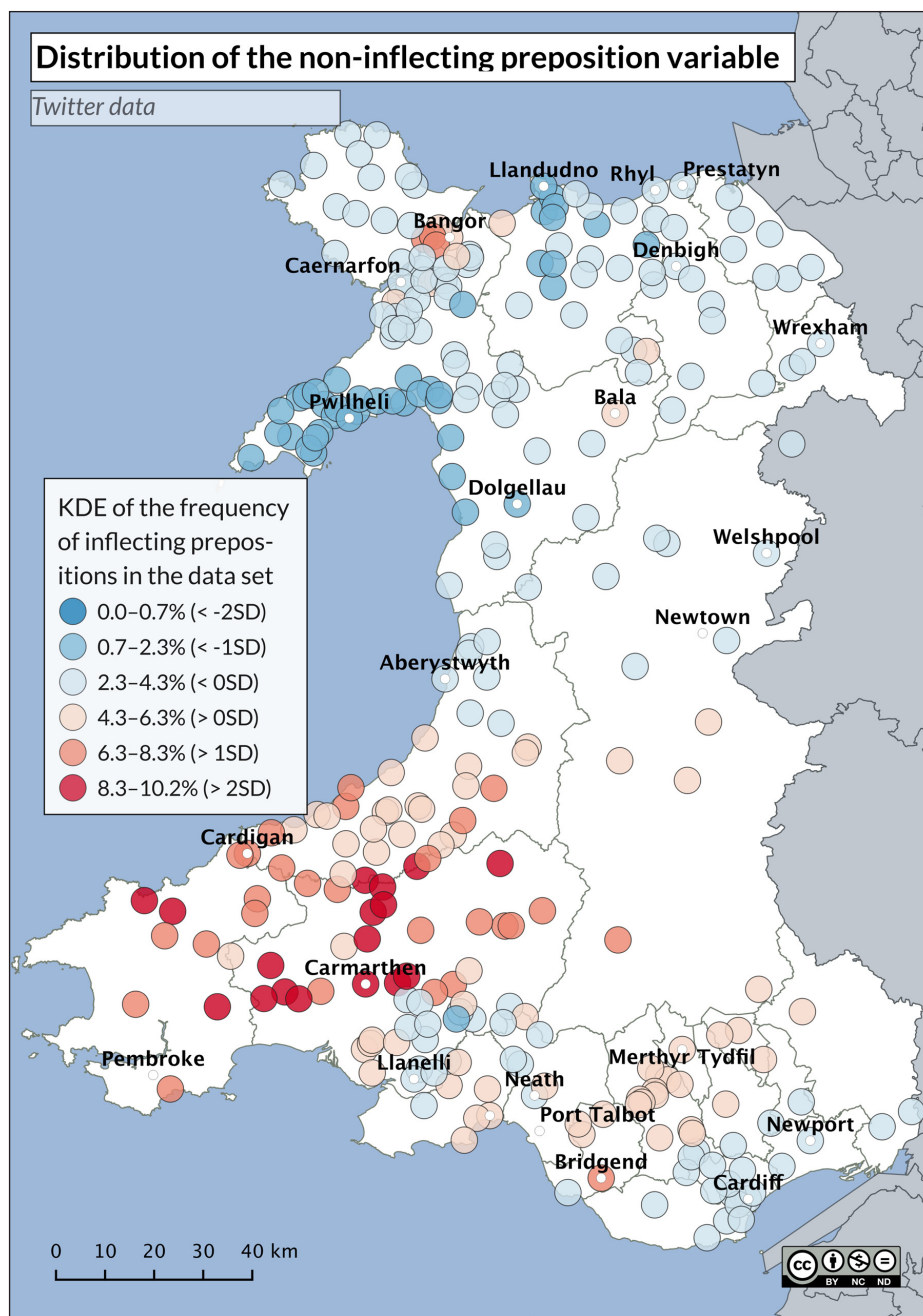
First, the four leftmost environments in (16) are not statistically significantly different from one another, and their relative ordering is therefore rather insecure.<sup>3</sup>

Secondly, the age profile of Twitter users (cf. the statistics given in section 2 above) means that we are, on average, dealing with speakers considerably younger than were used to establish the hierarchies in (14) and (15). Consequently, the data can be interpreted as representing, on average, the language of speakers born in the 1990s. All but the final two contexts in (16) are ones that Willis (2017) found had run to completion in speakers of this age group. Consequently, we would expect the frequency of *chdi* in these contexts to have largely reached its ceiling, this ceiling being set by the proportion of Twitter users coming from the *chdi*-region.

While this is reassuring about the adequacy of the Twitter data, the idea that the change has run to completion in a number of syntactic contexts sits uneasily with some other observed significant differences in the percentages given in Table 1. Some of these other differences seem to be due to another, somewhat surprising fact, namely that the syntactic contexts themselves are not evenly distributed across the map. Above all, the context non-inflecting preposition is itself spatially autocorrelated. Figure 2 shows a *k*-nearest neighbours kernel density estimation (KDE) (bandwidth =  $2\sqrt{\text{dataset size}} = 115.8$  nearest neighbours) of the proportion of data points in the dataset that are instances of the non-inflecting preposition context; that is, it shows the mean frequency of the context across the 115.8 nearest data points to each location.<sup>4</sup> The red areas on Figure 2 are those with above-average frequency of this context, with the darkest red showing areas where the frequency is more than 2 standard deviations away from the mean; the blue areas show the inverse, with areas below the mean. From this, it can be seen that, somewhat counterintuitively, the non-inflecting preposition context is found in the data significantly more frequently in the south than in the north, with a zone of very significantly elevated occurrence across all of the southwest. Since so much of the data for this context comes

<sup>3</sup> Chi-squared results: subj. of *oedd* vs. subj. of conditional  $\chi^2$  (df = 1, n = 487) = 0.620, p = 0.431; subj. of *oedd* vs. independent use  $\chi^2$  (df = 1, n = 380) = 1.774, p = 0.183; subj. of *oedd* vs. obj. of inflecting preposition  $\chi^2$  (df = 1, n = 262) = 1.749, p = 0.186; subj. of conditional vs. independent use  $\chi^2$  (df = 1, n = 371) = 1.086, p = 0.297; subj. of conditional vs. obj. of inflecting preposition  $\chi^2$  (df = 1, n = 253) = 1.400, p = 0.286; independent use vs. obj. of inflecting preposition  $\chi^2$  (df = 1, n = 262) = 0.027, p = 0.870. Despite this, it is not appropriate to merge these positions on the hierarchy. It is often the case that differences between adjacent points on this kind of hierarchy are not significant, while the entire hierarchy is statistically significant. In the current instance, the context of inflecting preposition is not statistically different from non-inflecting preposition or from conditional, but conditional auxiliary and non-inflecting preposition are significantly different. It is not possible to determine statistically whether inflecting preposition should be merged on the hierarchy with non-inflecting preposition or with conditional auxiliary. Consequently it is preferable not to perform either merger.

<sup>4</sup> For further details on KDE as a smoothing procedure, see section 3.3 below.



**Figure 2:** Map of geospatial distribution of the context non-inflecting preposition in the Twitter dataset.

from the south, where *chdi* is not found, it is not surprising that the frequency of *chdi* overall in this context is lower than might be expected.

Once this issue is identified, the reason why it arises can be seen fairly readily. Some non-inflecting prepositions are found only in certain dialects; for instance *fatha* ‘like’ and *efo* ‘with’ are both northern only. However, in these cases, other dialects express the same meaning also with a non-inflecting preposition (*fel* and *gyda* respectively), cancelling out any effect. Conversely, the ‘have’-construction, which varies significantly between dialects, contains a non-inflecting preposition in some dialects but not in others. Thus in southern varieties, the ‘have’ construction uses non-inflecting *gyda* ‘with’ in (17), while the commonest ‘have’ construction in the north uses inflecting *gan* ‘with’ in (18). A third construction using *efo* in (19) is confined to the north, but maps the object possessed to



the object of the preposition, hence ‘you’ is the subject of the verb *bod* ‘be’ in this context and is counted in the statistics accordingly.

- (17) Mae car gyda ni.  
be.PRS.3SG car with us  
‘We have a car.’
- (18) Mae gynnon ni gar.  
be.PRS.3SG with.1PL us car  
‘We have a car.’
- (19) ’Dan ni efo car.  
be.PRS.1PL we with car  
‘We have a car.’

The result is that non-inflecting prepositions are simply commoner in the south than in the north, artificially raising the global frequency of *ti* for this context. This highlights the need to interpret the data geospatially, as will be done in the next section.

### 3.3 Geospatial analysis of the data

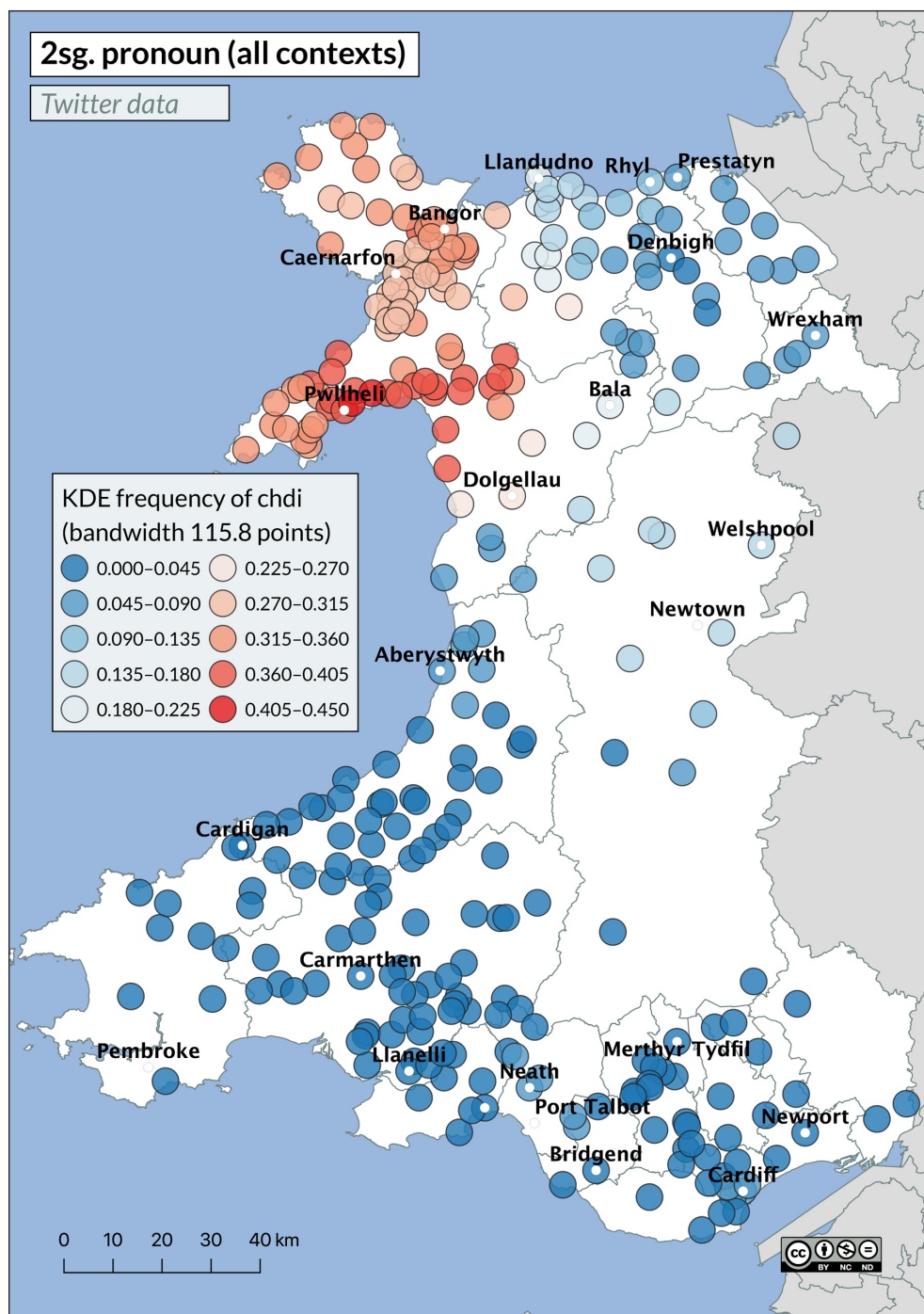
The previous section showed the need to include a geospatial dimension to the linguistic analysis. This is what will now be attempted. Consider first the overall distribution of *chdi*, shown in Figure 3. Here, each circle represents a tweet location; the colour of the circle represents a KDE value for the frequency of *chdi* at that point: KDE calculates the mean frequency of the variant within a kernel centred on the point in question. The size of the kernel is determined by the KDE bandwidth. KDE has previously been used to map dialect variation by Blaxter (2017) and for dialect variation in Twitter data by Jones (2015) and Van Halteren, Van Hout & Roumans (2018). In all cases in the current article, the bandwidth is twice the square root of the subset of data under consideration, in this instance, the nearest 115.8 data points.<sup>5</sup> This method will be repeated for the various syntactic contexts below.

Figure 3 shows the overall pattern for all the data. The frequency of *chdi* never exceeds 45% in any region, because these data include syntactic contexts in which *chdi* is categorically excluded in all varieties. The close similarity to the pattern found in the Welsh Dialect Survey in Figure 1 above is nevertheless clear, with a core region where *chdi* is strongest in the northwest, surrounded by a transitional ring to the east and southeast with moderate frequency of use.

For direct evaluation of the success of the Twitter method, we need to compare each syntactic environment with parallel traditional data. This is done below by comparing the geospatial pattern found in the *chdi* data with data from the Syntactic Atlas of Welsh Dialects. In each case, two maps are provided, constructed according to the same principles: KDE is conducted for the SAWD data in the same way as it was done for the Twitter data.

The data collection for SAWD assumed that *chdi* was not found anywhere in the south; the south is thus blank on the SAWD maps because no questionnaire data was collected there. For the Twitter data, the south is retained, as it is not obvious in advance that the method can correctly rule out the presence of *chdi* in the south; indeed, some instances of

<sup>5</sup> There are no established criteria for bandwidth selection for KDE. Too high a bandwidth risks oversmoothing and obscuring real patterns within the data; too low a bandwidth risks overdifferentiation, so that random fluctuations in the data are wrongly interpreted as representing geographic variation. The current method represents an attempt at a compromise between these two extremes.

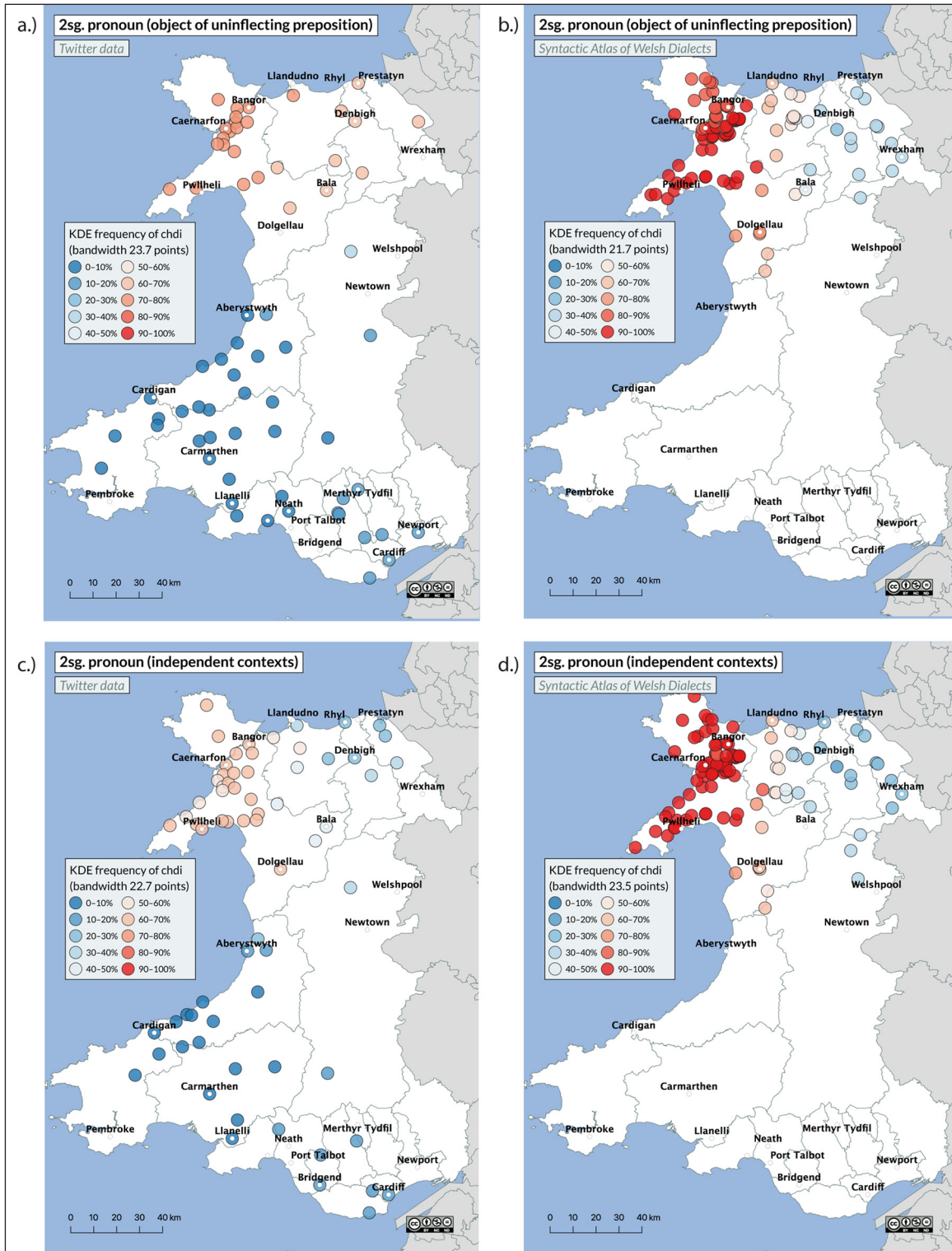


**Figure 3:** Overall distribution of *chdi* in the Twitter data.

*chdi* are localized to points in the south, although KDE smoothing means that these points do not have a significant impact on the final outcome.

An examination of data across the hierarchy of syntactic contexts shows that overall distributions produced by the two methods are rather similar. Figure 4 shows the two historically primary contexts, after a non-inflecting preposition and in independent syntactic contexts (mainly focus and sentence fragments).

Here, as in all cases examined, SAWD always produces higher absolute rates for *chdi* in regions where it is present, presumably because the oral questionnaire that was used focused speakers' attention on spoken usage and further because the Twitter data contain a proportion of material in standard written Welsh (see also discussion of Twitter as a



**Figure 4:** Comparison of Twitter and SAWD results for object of non-inflecting preposition and independent use of second-person singular pronouns.

proxy for spoken data in section 4.2.1 below). The Twitter data for object of non-inflecting preposition are for all such prepositions occurring in the corpus; the SAWD data are for the object of the one such preposition, *efo* ‘with’, that was included in the questionnaire.

For the object of non-inflecting prepositions, the Twitter data overstate the geospatial distribution of *chdi* by attributing majority *chdi* usage to the whole of the northeast. This is due to the relatively small size of the dataset once it is subdivided into different syntactic

contexts: of the 277 data points categorized as non-inflecting preposition, 140 could be localized; and of the 266 categorized as independent use, 129 could be localized. With this size of dataset, mislocalization of one or two data points (for instance, because a user mentions only the place where they now live, but not where they grew up, in their user description) can have a significant impact even once smoothing has applied. This particularly affects the northeast, where there are rather few data points (8–10) for these contexts. This is an issue that is likely to arise from time to time when using Twitter for smaller languages, but its effect would likely be reduced with a larger dataset collected over a longer period of time. For the independent-use context, the data contained no such outlier points and the issue did not arise. The result is an isogloss from the Twitter data that is remarkable close to that found using traditional means.

The two contexts discussed so far are historically primarily; that is, the changes that produced the current dialect distribution occurred in the eighteenth and nineteenth centuries, and there is little evidence of ongoing change in their geographical distribution today. We turn now to consider in turn contexts where there is more ongoing change. Consider first the object of the semi-inflecting preposition *i* ‘to’ and the fully inflecting prepositions (mainly *am* ‘about’, *ar* ‘on’ and *o* ‘from’), shown in Figure 5.

These are contexts where *chdi* emerged in the late-nineteenth century, and where there is some evidence today that *chdi* is still spreading geospatially. For *i* ‘to’ (186 localized points), the area for *chdi* defined by the two methods is again remarkably similar; for inflected prepositions (83 localized points), the Twitter result again slightly overstates the use of *chdi* in the northeast, and for the same reasons as before: the inflected-preposition dataset is too small to be immune to the presence of a small number of points whose localization does not reflect the place where the user in question grew up.

Next consider the subject of the imperfect auxiliary *oedd* ‘were’, the modal *rhaid* ‘must’, and the conditional auxiliary (*bua*)*sa(i)* or *byddai*. These are grouped together because previous research (Willis 2017: 58) has shown them to be undergoing similar rates of change today with respect to the spread of *chdi*, and the Twitter dataset clearly contained too few instances of each context (50 localized tweets for *oedd*, 17 for *rhaid* and 55 for the conditional) for analysis on an individual basis to be viable. Results for these three contexts combined are shown in Figure 6. The Twitter results agree with SAWD in identifying a smaller geographic area for *chdi* in these contexts than for the earlier contexts we have considered. Nevertheless, this is one context where the Twitter results suggest a somewhat wider geographic spread than the SAWD results. This may be due to the fact that these are contexts currently experiencing rapid diffusion of *chdi*, and the Twitter data reflect on average a younger age group, in which *chdi* does indeed have a wider geographic distribution. Closer inspection of the data suggest that the difference is mainly due to *chdi* being the overwhelming form among tweets localized to the Llŷn Peninsula around Pwllheli, while *ti* is the dominant choice (although by no means the only choice) of informants in this region in the SAWD questionnaire. This probably is a genuine instance of diffusion and change in progress.

Finally, consider the two contexts where *chdi* is a very recent innovation found only in a very small geographic area, namely the subject of the future auxiliary *bydd* ‘will be’ and the object of the preposition *gan* ‘with’. These results for these contexts are shown in Figure 7. In both cases, the Twitter dataset is small (47 localized tweets for *bydd* and 79 localized tweets for *gan*), and, consequently, it is difficult to draw firm conclusions. However, the results are consistent with the SAWD questionnaires. The Twitter data agree with SAWD in showing that these are the contexts where *chdi* is most geographically restricted and where it is the minority choice everywhere. Both methods record slightly higher frequencies in the northwest. The greater richness of the SAWD data allows the



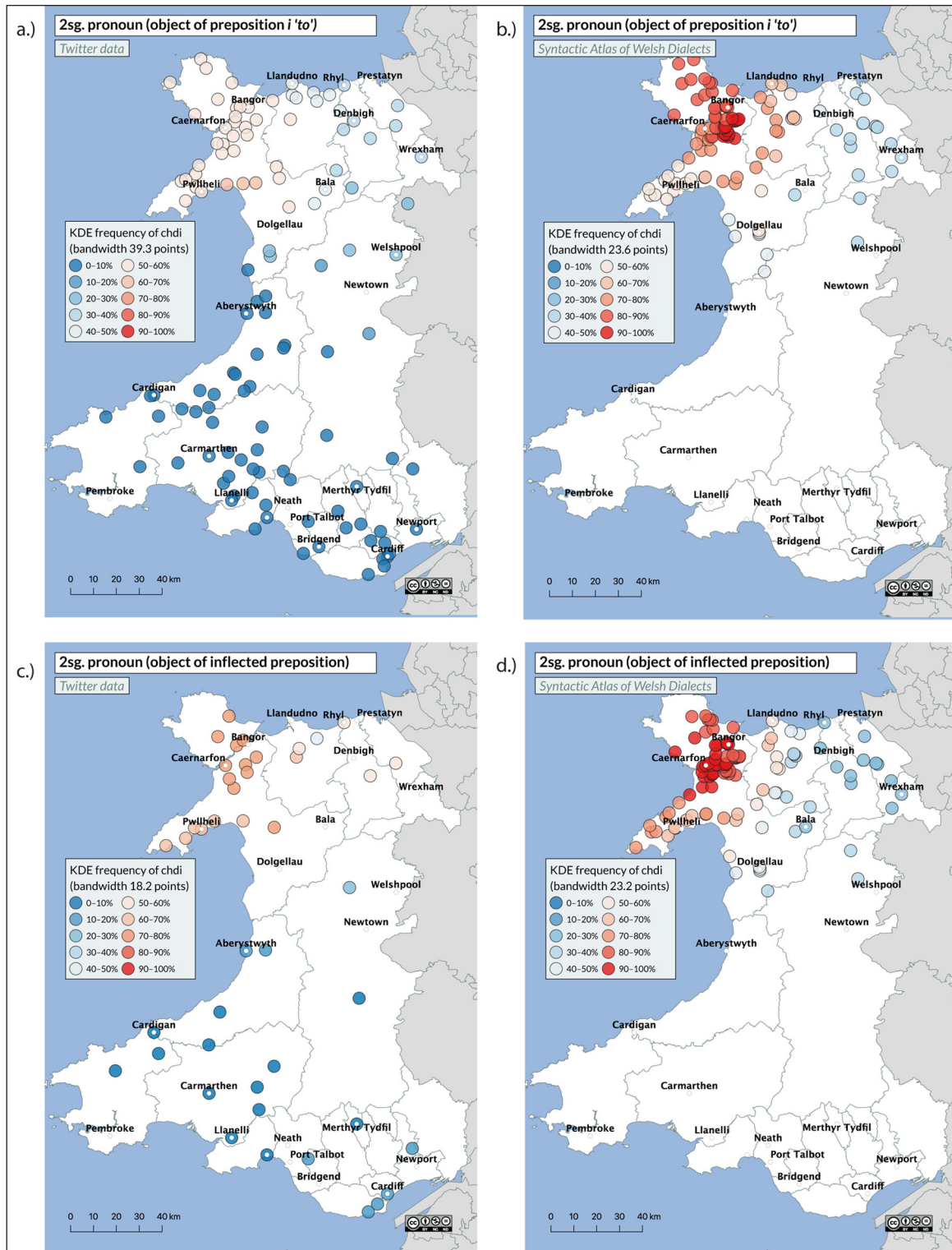
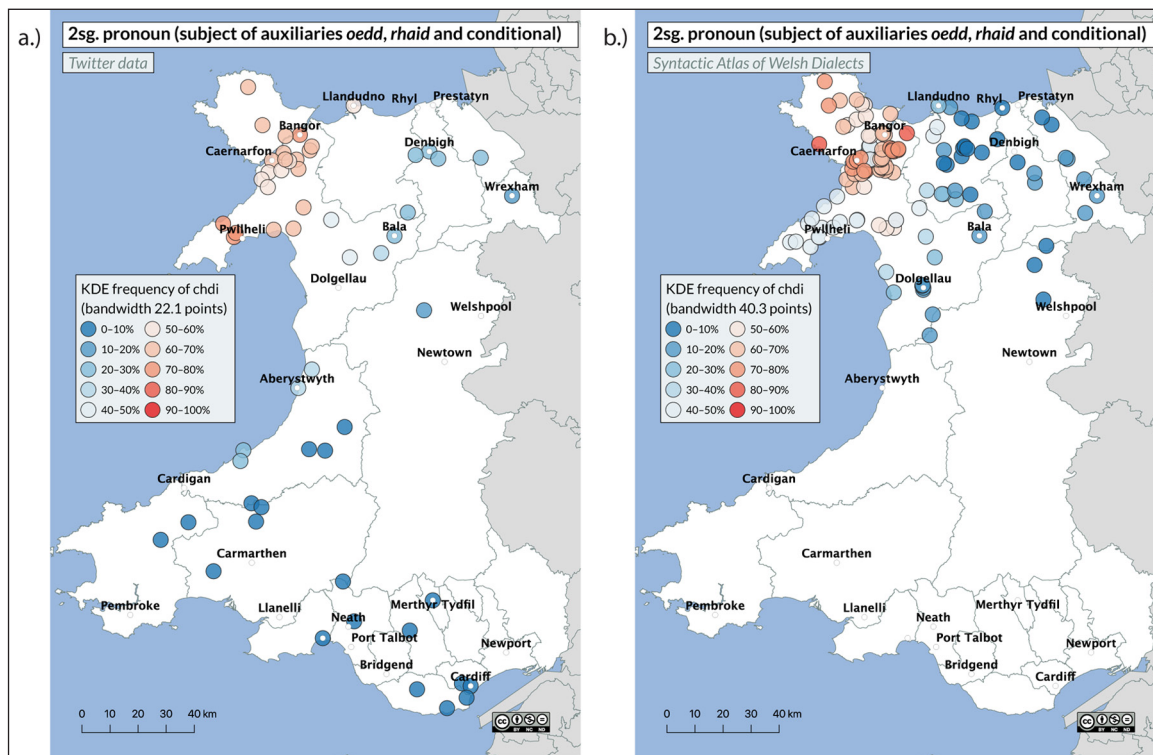


Figure 5: Comparison of Twitter and SAWD results for object of *i* 'to' and object of non-inflecting prepositions.

identification of a centre of innovation in the Caernarfon area. The Twitter data do not contradict this: all the localized tweets for these contexts turn out to be either from the Caernarfon area or from users ultimately from this area who have moved elsewhere (but who ended up being localized to where they currently live).



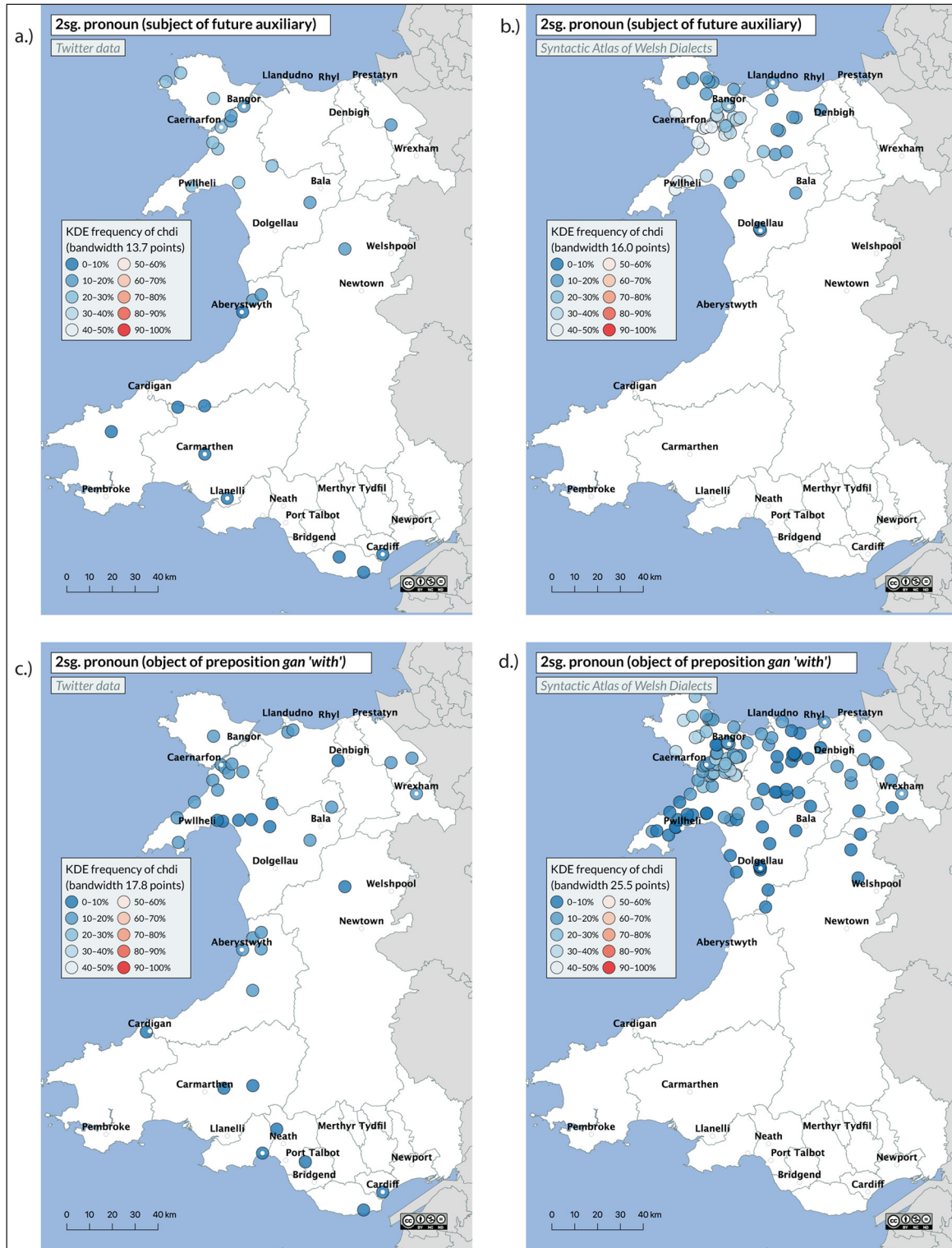


**Figure 6:** Comparison of Twitter and SAWD results for subject of *oedd* ‘were’, *rhaid* ‘must’ and the conditional auxiliary.

### 3.4 Discussion

The overall geospatial distribution of *chdi* established using Twitter data broadly matches that established using traditional means (Figures 1 and 3). While the global analysis of the distribution by syntactic context produces a hierarchy of contexts that differs in points of detail from the traditional one, and does not identify the historically primary contexts, the more fine-grained geospatial analysis conducted in section 3.3 suggests that this is to some extent due to spatial autocorrelation of the syntactic contexts themselves within the dataset. When plotted as maps for individual contexts, the Twitter data largely agree with the result from the earlier SAWD project. Two sources of discrepancy were identified: in a number of cases, insufficient data, either for a particular syntactic context globally, or for a particular geographic region, resulted in oversmoothing or lack of geospatial detail for the Twitter maps; in one case, subject of auxiliaries in Figure 6, the greater geographic extent of *chdi* identified in the Twitter data may plausibly represent real ongoing diffusion that accurately reflects the younger demographic of the Twitter data.

The hierarchy of contexts established by traditional means in (15) above is also largely respected. This can be seen by looking at the estimated smoothed frequencies at various locations, both within and beyond the traditional *chdi*-region, as shown in Table 2. If the hierarchy in (15) is respected, the values should decrease monotonically from left to right within the traditional *chdi*-region. This is broadly the case. There are some exceptions for the four most well-established contexts on the left of Table 2, contexts where the change may well have run to completion within the *chdi*-region. Outside of the *chdi*-region, we find scattered noise in the estimates for the four southern locations (Aberystwyth, Cardiff, Cardigan and Carmarthen), where *chdi* is absent from the local varieties. Aberystwyth and Cardiff have significant student populations; Cardiff, as the capital, has other migration



**Figure 7:** Comparison of Twitter and SAWD results for subject of future auxiliary *bydd* 'will be' and object of preposition *gan* 'with'.

from the north, and, furthermore, as an area of language revitalization, has an emerging new variety with some levelling of features from the south, the north and the literary standard. It is therefore not surprising that Aberystwyth and Cardiff have more noise, in the form of higher estimates for *chdi*, than other southern locations. Finally, Denbigh, in the northeast, shows a similar overall pattern to the northwest, but at rather lower

**Table 2:** KDE values of the frequency of *chdi* at selected locations.

Town/city	obj. of non-inf. prep.	indep.	object of <i>i</i> 'to'	obj. of inf. prep.	subj. of aux.	subj. of fut. aux.	obj. of <i>gan</i> 'with'
Within the traditional <i>chdi</i> -region							
Bala	0.64	0.46	0.43	NA	0.26	0.20	0.11
Bangor	0.75	0.65	0.58	0.79	0.72	0.20	0.15
Blaenau Ffestiniog	0.67	0.48	0.56	NA	0.43	0.25	0.00
Caernarfon	0.75	0.61	0.58	0.79	0.61	0.27	0.17
Llanberis	0.77	0.60	0.57	0.76	0.62	0.25	0.12
Llangefni	0.76	0.67	0.55	0.79	0.64	0.21	0.17
Porthmadog	0.76	0.61	0.65	0.68	0.61	0.27	0.00
Pwllheli	0.79	0.61	0.58	0.68	0.72	0.21	0.10
Outside of the traditional <i>chdi</i> -region							
Aberystwyth	0.07	0.13	0.10	0.11	0.33	0.00	0.10
Cardiff	0.16	0.15	0.03	0.11	0.08	0.00	0.00
Cardigan	0.00	0.00	0.00	0.00	0.03	0.00	0.00
Carmarthen	0.00	0.00	0.00	0.05	0.00	0.00	0.00
Denbigh	0.64	0.25	0.33	0.53	0.26	0.18	0.11

frequencies. As we have seen, this is partly due to lack of data and artefacts of smoothing, although it may also suggest some genuine spread of *chdi* to this area.

#### 4 Auxiliary deletion

We turn now to consider a second variable using the same dataset, namely deletion of auxiliary 'be' in pre-subject position.

##### 4.1 The variable

Deletion of auxiliary *bod* 'be' in the present tense is characteristic of all varieties of spoken informal Welsh today. Thus, in the AuxSVO structure in (20), the auxiliary (*r*)*wyt* may be reduced or omitted entirely.

(20) ((R)w(y)t) ti 'n chwarae pêl-droed.  
 be.PRS.2SG you PROG play.INF football  
 'You're playing football.'

Such deletion is not restricted to the second person singular and is found with most other pronouns, at least in some varieties. It has even been reported for clauses containing lexical NP subjects (Davies 2010: 270).

Deletion is not dependent on the auxiliary being in absolute clause-initial position: it is grammatical in an embedded AuxSVO clause in (21) and in a main-clause *wh*-question in (22).

(21) Mae 'n well na(g rwynt) ti 'n meddwl.  
 be.PRS.3SG PRED better than be.PRS.2SG you PROG think.INF  
 'It's better than you think.'

- (22) Be' (wyt) ti 'n moy'n?  
 what be.PRS.2SG you PROG want  
 'What do you want?'

Auxiliary deletion in the second person singular requires *ti* rather than *chdi* as the subject pronoun (99.6% *ti* in Table 1 above). It thus patterns with its overt counterpart *wyt ti* 'you are' (100.0% *ti*), rather than with the independent-use context (68.0% *ti*). This suggests that auxiliary deletion should be interpreted as involving a real auxiliary that happens to be null, rather than as a construction with an independent pronoun not dependent on any auxiliary. Another argument in favour of this view is that a tag question with a full auxiliary may co-occur with a main clause containing auxiliary deletion (Borsley, Tallerman & Willis 2007: 260–61). If tag questions in some sense involve copying of the auxiliary of the main clause, this would suggest there is a syntactically represented auxiliary in the main clause.

It is agreed in the literature that the availability of auxiliary deletion is conditioned by person and (to a lesser extent) by number (Jones 2004: 101–2; Borsley, Tallerman & Willis 2007: 260–61; Breit 2012). There is a striking degree of variation between dialects in some person–number combinations; for instance, auxiliary deletion in the first person plural is common in the south but very rare in the north.

Davies (2010: 258–335), Davies & Deuchar (2014) and Davies (2016) investigate auxiliary deletion in the second person singular by 28 Welsh speakers from the Siarad Corpus of spoken Welsh (Deuchar et al. 2014; Deuchar, Webb-Davies & Donnelly 2018), 8 of whom grew up in the south and 20 of whom grew up in the north or with a northern background. Davies (2010: 285, 297) finds an overall frequency of auxiliary deletion of 92.8%, with slightly lower levels in older speakers (84.8% in the over-50 group, compared to 93.8% in the under-30 group) and thus age turns out to be a significant predictor of frequency of deletion. However, even some speakers born in the 1920s have deletion with a frequency approaching 100%, indicating that auxiliary deletion is a historically well-established feature of spoken Welsh. Indeed, Willis (2016) argues that the roots of auxiliary deletion go back to the nineteenth century, where it is used in literary representations of second-language Welsh. While Davies (2010: 293) found women to delete auxiliaries at a slightly higher rate than men, this difference was not significant; nor were there significant differences according to the region in which a speaker spent their first year of life (Davies 2010: 295).<sup>6</sup>

The impact of factors other than person and number on the frequency of auxiliary deletion is less well understood. Davies (2010: 303–22) investigated the impact of four factors in the Siarad corpus, failing to find any significant effect for clause type (declarative vs. interrogative), linguality (the presence of absence of code-switching in the clause), or negation. He found an inverse relationship between deletion of the auxiliary and deletion of an aspect marker in the same clause.<sup>7</sup> Breit (2012) administered an acceptability-judgment task to 20 native speakers, almost all from north Wales. He found negligible degradation of acceptability of deletion in negative and interrogative clauses, in line with Davies' findings that these are not factors conditioning variation. With focus fronting, a

<sup>6</sup> Davies asked speakers about where they spent their first year of life because the answer was felt to “reflect the region where that speaker was living during the crucial (i.e. early) stages of language acquisition” (Davies 2010: 295). Of course, where a speaker did not remain in this place during childhood, this question may not provide a reliable guide to the variety acquired, but, in most cases, this is not an issue.

<sup>7</sup> In the Twitter data, the aspect particles were generally retained. Consequently, deletion of aspect markers was not pursued further as a possible conditioning factor.



factor not investigated by Davies, he found a large reduction of acceptability of deletion in clauses with VP-fronting:

- (23) Ffonio 'r gwasanaeth tân ?(wyt) ti.  
 phone.INF the service fire be.PRS.2SG you  
 'Phoning the fire service you are. / You're phoning the fire service.'  
 (Breit 2012: 83)

Such examples leave the sequence deleted auxiliary + subject pronoun in clause-final position and require deletion of the stressed auxiliary, leaving the unstressed pronoun to form the final phonological word of the sentence. This could be a phonological reason to disfavour deletion.

Davies (2010: 323–8) argues that the initial innovation of auxiliary deletion was internally motivated and due to phonological erosion, but that, once it had been innovated internally, it was accelerated by external factors, namely isomorphism between the SV(O) word order that results and the normal word-order pattern in English. This accords well with the nineteenth-century perception of it as a feature of non-native Welsh.

Davies further suggests that dialect differences arose because deletion was favoured in contexts where the auxiliary began with a vowel, thus northern *'dan ni* 'we are' resists auxiliary deletion, while the equivalent southern form *ŷn ni* favours it. Note that both forms are ultimately reductions from Early Modern Welsh *yr ydym ni* (PRT be.PRS.1PL we) > *rydyn ni* > *dyn ni* > *dan ni* in the north and *yr ydym ni* > *rydyn ni* > *rŷn ni* > *ŷn ni* > *ni* in the south. This account thus amounts to saying that reduction has proceeded further in the south than in the north. The reasons for this differential degree of reduction remain unclear.

The same dataset collected to investigate the geographic distribution of *chdi* can be used to investigate auxiliary deletion in the second-person singular. In this environment, auxiliary deletion is common in all dialects of Welsh. Previous studies have not identified geospatial variation to date. The first question addressed by the data here is whether Twitter data provide a useful proxy for spoken data. Clearly, Twitter is a written rather than a spoken medium, but the informal register of Twitter data can make it an attractive proxy for much more difficult to obtain transcriptions of spoken language. In this case, we can compare the frequency of auxiliary deletion in tweets with its frequency in spontaneous speech as reflected in the Siarad Corpus. In addition, we can examine whether existing statements about the effect of linguistic and geographic factors on variation are borne out by the Twitter data.

## 4.2 Data analysis

### 4.2.1 Twitter as a proxy for spoken data

In the Twitter corpus, the global ratio of deletion to non-deletion in the second person singular (including non-localizable tweets) is 1,784: 461 (79.5% deletion). While this is lower than the 92.8% of the Siarad corpus, it suggests a relatively good match between the two methods, especially considering that auxiliary deletion does not occur in standard written Welsh. Results derived from using Twitter data are thus unlikely to be radically different from those using spoken data, although the impact of standard forms within tweets does need to be considered in any analysis, and a gap in absolute values between the two sources will also need to be reckoned with. Further investigation is needed to establish whether the difference between recorded speech and the Twitter corpus is due to the impact of institutional and learner tweeting (both showing influence from the written



standard) or whether it is an inherent property of the (ultimately written) Twitter medium even among users aiming to tweet “as they speak”.

#### 4.2.2 Linguistic factors

Linguistic factors examined in the data were clause type (main or subordinate), force type (declarative, interrogative, focus etc.), and polarity (affirmative or negative). All tweets containing a context for auxiliary deletion were coded for these factors. This allows us to examine the effect of several factors discussed previously in the literature, namely declarative vs. interrogative, not found to be significant by Davies (2010) and not found to lead to degradation in acceptability by Breit (2012: 46); focus vs. non-focus, with focus found to lead to degradation in some contexts by Breit; and negative polarity, not found to lead to substantial degradation by Breit.

For clause type, clauses were counted as subordinate if they were introduced by a complementizer such as *tra* ‘while’, *os* ‘if’ or *pan* ‘when’, or were relative clauses or embedded *wh*-answers. Affirmative complement clauses are formally nonfinite in Welsh. It was assumed that clauses like (24) involve deletion of nonfinite *bod* ‘be’ (rather than of finite *rwyf* ‘are’). Since existing studies have focused on deletion of finite ‘be’, such clauses were excluded from the analysis, rather than being included as subordinate.

- (24) Dwi            ’n    gwybod    \_\_    ti    ’n    ennill.  
 be.PRS.1SG PROG know.INF    you PROG win.INF  
 ‘I know you’re winning.’

For force type, possible values were declarative VSO clause, declarative focus clause, yes–no question, *wh*-question, focus question, and conditional (‘if’-clauses). Focus fronting (with object fronting) is illustrated in (25) and (26).

- (25) Mari (wyt)            ti    ’n    feddwl.  
 Mari be.PRS.2SG you PROG think.INF  
 ‘(It’s) Mari you mean.’

- (26) (Ai/ife) heddiw (wyt)            ti    ’n    feddwl?  
 Q.FOC today be.PRS.2SG you PROG think.INF  
 ‘(Is it) today you mean?’

For polarity, possible values were affirmative and declarative. *Mond* ‘only’ (< *dim ond* ‘nothing but’) and *methu* ‘be unable, not be able’ were treated as affirmative; *heb*, when used as the negative of the perfect marker *wedi*, was treated as negative.

The frequency of auxiliary deletion in each of the syntactic contexts examined is given in Table 3. Rates of deletion varied substantially in the Twitter data from context to context from 92.8% in declarative clauses to 33.3% in focus questions.

The impact of these factors was assessed by implementing a logistic regression model with presence or absence of auxiliary deletion as the binary dependent variable. The results of this model are given in Table 4, with affirmative declarative main clause as the reference level. Positive log odds indicate that a factor level favours auxiliary deletion relative to the reference level, while negative log odds indicate that it disfavours auxiliary deletion relative to the reference level.

The contrast between main and subordinate clauses was not a significant factor in the model (the lower frequency of auxiliary deletion in subordinate clauses reducing largely to the effect of conditionals), while all other factors were significant. All clause types had

**Table 3:** Frequency of auxiliary deletion by syntactic context.

Factor	% deletion	n =
<b>clause type</b>		
main	77.2	1778
subordinate	88.0	467
<b>force type</b>		
declarative	92.8	1105
conditional	81.6	147
yes-no question	66.2	595
wh-question	63.4	352
focus	54.8	31
focus question	33.3	15
<b>polarity</b>		
affirmative	79.6	2132
negative	76.1	113
total	79.5	2245

**Table 4:** Logistic regression model of linguistic factors affecting auxiliary deletion.

Factor	Log odds estimate	Standard error	p-value
Intercept	2.726	0.146	<0.0001
clause type: subordinate	-0.287	0.241	0.234
force type: conditional	-0.881	0.292	0.003
force type: focus	-2.495	0.384	<0.0001
force type: focus question	-3.419	0.567	<0.0001
force type: wh-question	-2.168	0.183	<0.0001
force type: yes-no question	-2.023	0.168	<0.0001
polarity: negative	-0.866	0.252	0.001

a significant inhibiting effect on auxiliary deletion as compared to affirmative declarative main clauses. While the effect of negation and conditional clauses was rather small, the effect of interrogative and focus clause type was substantial. The effect of focus agrees with Breit’s findings. The effect of interrogative clauses is less expected. It should be noted that Davies’s study was based on 648 observations, while the current study is based on 2,245. With a larger and more independent sample size (1,329 Twitter users as against 28 speakers), significant effects are more liable to emerge. The lower overall frequency of auxiliary deletion in the current dataset also means that significant factors are less likely to be hidden by ceiling effects.

To facilitate comparison with traditional work in quantitative sociolinguistics, another presentation of the same model is given in Table 5 using RBrul (Johnson 2009), with the mean of all observations as the reference value and the log odds estimates also transformed into factor weights in the tradition of Sankoff & Labov (1979: 199). Factor weights above 0.5 indicate that a factor level favours auxiliary deletion, and those below 0.5 indicate that it disfavors it relative to the mean probability of deletion over the entire dataset. It is important to bear in mind the fact that the reference level is different in the two

**Table 5:** Rbrul logistic regression model of linguistic factors affecting auxiliary deletion.

Factor	Log odds estimate	n	Proportion	Factor weight
Intercept	0.319	n/a	n/a	0.579 (input probability)
clause type: main	0.143	1778	0.772	(0.536)
clause type: subordinate	-0.143	467	0.880	(0.464)
force type: declarative	1.831	1105	0.928	0.862
force type: conditional	0.950	147	0.816	0.721
force type: focus	-0.664	31	0.548	0.340
force type: focus question	-1.588	15	0.333	0.170
force type: <i>wh</i> -question	-0.337	352	0.634	0.417
force type: yes–no question	-0.192	595	0.662	0.452
polarity: affirmative	0.433	2132	0.796	0.607
polarity: negative	-0.433	113	0.761	0.393

Note: Factor weights that are not statistically significant are given in parentheses.

presentations when interpreting differences between them. See Johnson (2009: 359–362) for a discussion of the differences between these two modes of presentation.

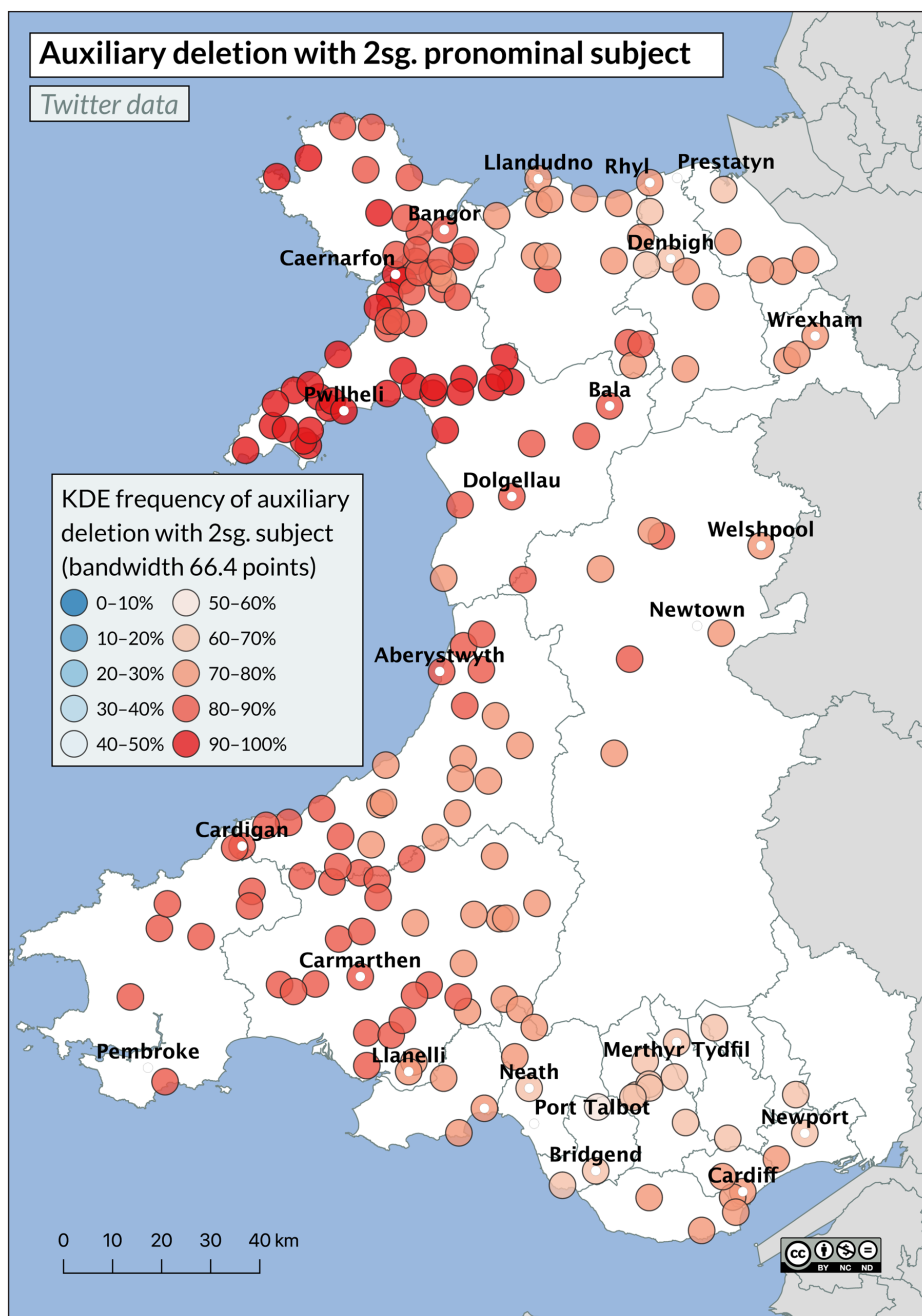
A second model included interactions between subordination, clause type and polarity. Here, two interactions were significant (at  $p < 0.05$ ). Negative subordinate clauses significantly decreased the propensity for auxiliary deletion (coefficient  $-1.622$ , standard error  $0.694$ ,  $p = 0.019$ ). Negative yes–no questions significantly increased the propensity for auxiliary deletion (coefficient  $1.691$ , standard error  $0.683$ ,  $p = 0.013$ ).

#### 4.2.3 Geospatial distribution

Having looked at the impact of linguistic factors in the global distribution, we turn now to the geospatial distribution of auxiliary deletion. In total, of the 2,245 tweets containing either overt or deleted auxiliary ‘be’, 1,100 (49.0%) could be localized at 189 distinct localities. The KDE-smoothed distribution is shown in Figure 8. Auxiliary deletion is the majority option everywhere, although there is some regional variation. The highest values, 80–95% are found across the northwest, with slightly lower values (75–85%) in the southwest, and the lowest values in the east (65–75% in the northeast, 65–72% in the southeast).

The success of this result is more difficult to assess than that the result for *chdi* in section 3.3 above, because the reference distribution against which to compare this result is itself not clearly established in the literature. While Davies (2010) found no statistically significant differences according to the region in which a speaker spent their first year of life, the analysis divided Wales into only two regions (north and south) and the speakers investigated were mostly from the north, so it is possible that such differences were missed.

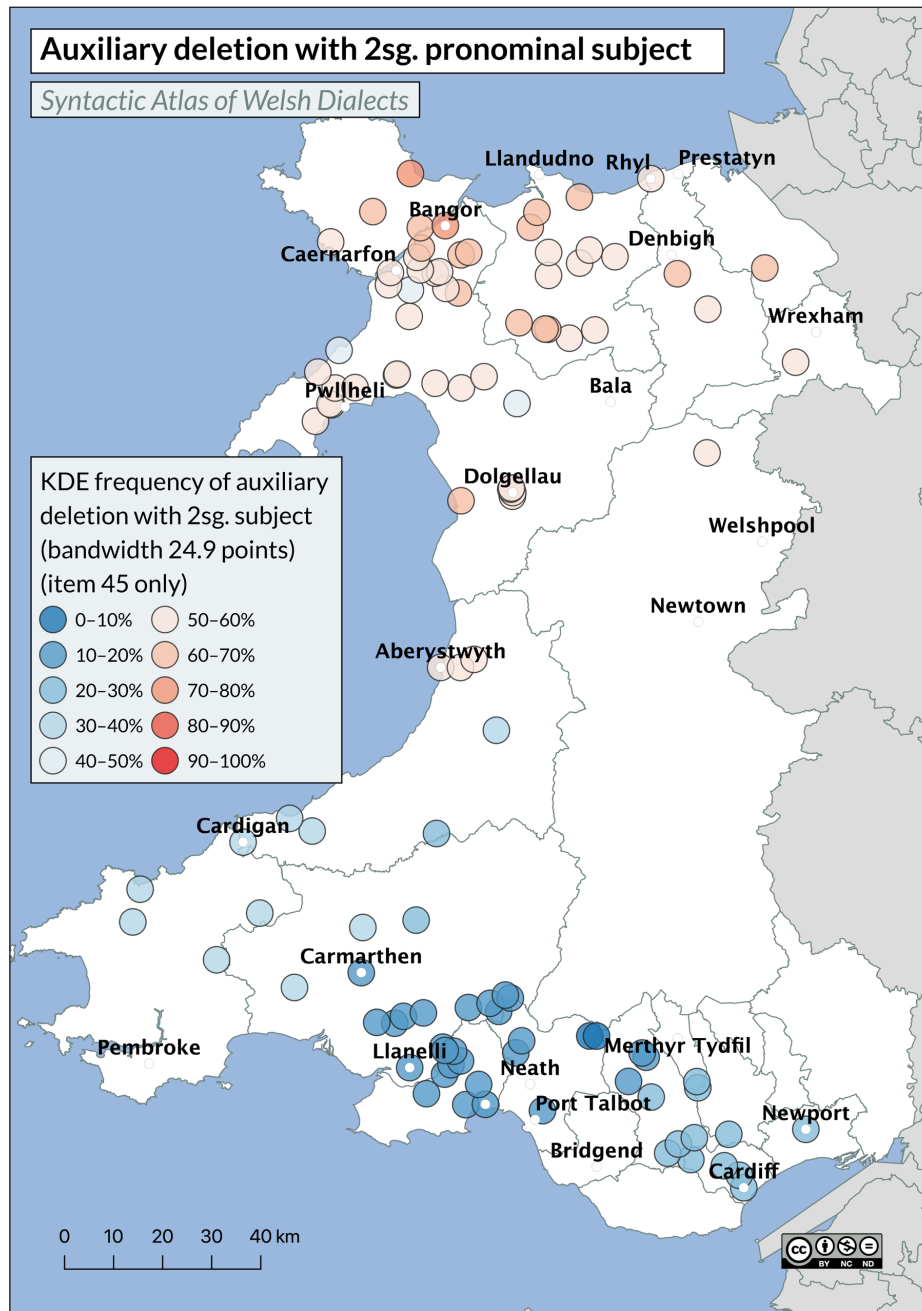
A partial comparison with data from the SAWD questionnaire is possible, although it too is not ideal for current purposes. Nevertheless, such a comparison reveals rather similar overall patterns. No question in SAWD aimed specifically at testing variation in auxiliary deletion in the second person singular. However, the environment arises fortuitously in nine questionnaire items. Unfortunately, these are all in *wh*-questions or in subordinate clauses. Furthermore, most of the questions where the environment does arise were asked only in one region (north or south), which makes the data less than ideal for the current comparison. The only relevant item asked in all areas was question 45 (‘If you’re not happy,



**Figure 8:** Geospatial distribution of second-person singular auxiliary deletion in Twitter data.

don't come.'). No previous study has tested in detail whether a finite subordinate clause of this type is a favouring or disfavouring context for auxiliary deletion. Davies (2010: 316) notes that there are only 4 instances of this context in his materials, with a rate of deletion of 50.0%. This would in principle be a very low rate of deletion and would suggest that this is a strongly disfavouring context for auxiliary deletion. However, as Davies notes, the rarity of the context makes any conclusions difficult to draw. Nevertheless, the possibility that this is a disfavouring context should be borne in mind when interpreting the results.

A KDE plot of the responses for auxiliary deletion in this item (question 45) in SAWD is shown in Figure 9. The overall rate of auxiliary deletion is rather low, at 57.4% of 155 observations. This is either because the questionnaire-based interview favoured its retention or, in line with the discussion above, because the syntactic environment tested is itself one that favours retention. Geographically, we find the highest rates in the north,



**Figure 9:** Geospatial distribution of second-person singular auxiliary deletion in SAWD data.

with 50–75% in the northwest and 50–65% in the northeast; rates are lower in the south, with 20–40% in the southwest, 25% around Cardiff and 10–20% in the central south. The north–south distinction is rather similar to that found in Twitter in Figure 8, albeit at lower absolute levels. An east–west division is clear in the SAWD data in the south, as in the Twitter data. This east–west effect is stronger in the Twitter data than in the SAWD data.

A reasonable hypothesis is that the east–west effect is due to the linguistic impact of language revitalization: in the east, closer to England, Welsh is a minority language and more dependent on revitalization efforts and Welsh-medium education to ensure language maintenance. Such a scenario promotes standard forms, reducing the frequency of colloquial options like auxiliary deletion. If so, it is not surprising to see this effect more strongly articulated in the Twitter data, where L2 speakers cannot be fully removed from



the dataset. Further investigation is needed to establish whether this interpretation can be substantiated by additional research.

## 5 Conclusion

This paper has tested the usefulness of social-media data in examining traditional questions in dialect syntax and sociolinguistics. The three central questions considered have been:

- (i) to what extent do datasets based on Twitter data successfully establish geospatial distributions derived via traditional means?
- (ii) to what extent can Twitter data successfully derive implicational hierarchies of contexts in the same way as studies based on more traditional materials?
- (iii) to what extent can written Twitter data act as a proxy for spoken data?

Two case studies have addressed these questions: the distribution of Welsh second person singular pronoun variants dealt with the first two, while the distribution of auxiliary deletion in Welsh dealt with the first and last of these.

We have seen that overall geospatial patterns in the data are similar to those established by traditional means. Thus, for the second person singular pronoun *chdi*, Figure 3 closely mirrors Figure 1. In the case of auxiliary deletion, where geospatial variation has not been fully established by traditional means, the Twitter data were not out of line with what we know from other sources, and can make a useful contribution to ongoing research when considered alongside those sources.

The overall hierarchy of syntactic contexts that emerges for the second person singular pronoun *chdi* in section 3.3 turned out to be broadly similar to the existing implicational hierarchy in (15), once ceiling effects due to ongoing continuation of change were taken into account. That is, in some cases, differences could be attributed to ongoing change, with Twitter typically reflecting a younger demographic. In some other cases, the quantity of Twitter data in the current study was insufficient once KDE smoothing had been applied.

For auxiliary deletion, the effects of internal linguistic factors uncovered in the data contrasted with the general absence of such effects in existing studies. While focus was found to be significant in inhibiting auxiliary deletion, in line with earlier work, both interrogative clause type and, to a lesser extent, negation, were found to inhibit deletion. These effects were found to be robust and based on a substantially larger dataset than existing work. Given the success of the Twitter data elsewhere, these results should feed in to our broader understanding of the phenomenon at hand.

Finally, in comparison with data from the spoken Siarad Corpus, Twitter data emerged as a good, but not perfect, guide to spoken usage: while auxiliary deletion in the second-person singular occurs with a frequency of 92.8% in spoken corpora, its frequency in Twitter data was 79.5%. Social-media data in this respect occupy a grey area where the distinction between speech and writing is not so clear.

In many cases, it is striking that very different data-collection methodologies produced very similar results. We have considered written corpus data in Twitter, the spoken corpus data of the Siarad Corpus, and the elicited questionnaire data of the Syntactic Atlas of Welsh Dialects and the Welsh Dialect Survey. These striking similarities may suggest that the choice between questionnaire-based and corpus-based methodologies is not as crucial as might first appear. In any case, the findings here demonstrate the general viability of using Twitter data alongside traditional methods to investigate morphosyntactic variation and change.

Finally, the approach adopted here has implications for theoretical developments in language variation and change. Studies using social-media data have the potential to combine the large scale of dialect atlases with the social and linguistic depth of sociolinguistic studies based around a single community. This opens up the possibility that we may be able to answer questions of a theoretical nature that could not be addressed within a single study, for instance, whether linguistic and social conditioning factors are stable across geographic space. This in turn may inform our understanding of both the processes by which innovations spread and the synchronic analysis of the linguistic phenomena under investigation.

## Abbreviations

Glosses follow the Leipzig glossing rules, except for: PRED predicate marker.

## Additional Files

The additional files for this article can be found as follows:

- **Supplementary file 1.** The dataset collected for this study in anonymized form. DOI: <https://doi.org/10.5334/gjgl.1073.s1>
- **Supplementary file 2.** Metadata describing the dataset file. DOI: <https://doi.org/10.5334/gjgl.1073.s2>

## Ethics and Consent

Ethical approval for this research was obtained from the Humanities and Social Sciences Research Ethics Committee of the University of Cambridge.

## Acknowledgements

Parts of this research were presented at the workshop Using Twitter for Linguistic Research, University of Kent, 2016, the 23rd Welsh Linguistics Seminar, Gregynog, 2016, the 50th Annual Meeting of the Societas Linguistica Europaea, Zürich, 2017, the 11th UK Language Variation and Change conference, Cardiff, 2017, and the Cambridge Language Sciences Symposium 2017. I am also grateful to these audiences and to Tam Blaxter, Deepthi Gopal and Adrian Leemann for discussion of and comments on this research.

## Funding Information

This research was carried out as part of the project Investigating the diffusion of morphosyntactic innovations using social media, funded by Economic and Social Research Council research grant ES/P00752X/1. The council's support is gratefully acknowledged.

## Competing Interests

The author has no competing interests to declare.

## References

- Abitbol, Jacob Levy, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot & Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in Twitter: A multivariate analysis. In Pierre-Antoine Champin, Fabien Gandon & Lionel Médini (eds.), *WWW '18 Proceedings of the 2018 World Wide Web Conference*, 1125–1134. New York: Association for Computing Machinery. DOI: <https://doi.org/10.1145/3178876.3186011>
- Bailey, Charles-James N. 1973. The patterning of language variation. In Richard W. Bailey & Jay L. Robinson (eds.), *Varieties of present-day English*, 156–86. New York: Macmillan.
- Blaxter, Tam Tristram. 2017. *Speech in space and time: Contact, change and diffusion in medieval Norway*. Cambridge: University of Cambridge PhD dissertation.

- Borsley, Robert D., Maggie Tallerman & David Willis. 2007. *The syntax of Welsh*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511486227>
- Breit, Florian. 2012. *Constraints on auxiliary deletion in colloquial Welsh*. Bangor: Bangor University BA dissertation.
- Britain, David. 2013 [2002]. Space, diffusion and mobility. In J. K. Chambers & Natalie Schilling (eds.), *Handbook of language variation and change*, 471–500. Chichester: Wiley–Blackwell. DOI: <https://doi.org/10.1002/9781118335598.ch22>
- Brooker, Phillip, Julie Barnett & Timothy Cribbin. 2016. Doing social media analytics. *Big Data & Society* 3. 1–12. DOI: <https://doi.org/10.1177/2053951716658060>
- Chaffey, Dave. 2019. Global social media research summary 2019. Leeds: Smart Insights. <https://www.smartinsights.com/wp-content/uploads/2014/04/Demographic-use-of-social-networks-age-and-gender.jpg>, accessed 24 July 2019.
- Claes, Jeroen. 2017. Cognitive and geographic constraints on morphosyntactic variation: The variable agreement of presentational *haber* in Peninsular Spanish. *Belgian Journal of Linguistics* 31. 30–55. DOI: <https://doi.org/10.1075/bjl.00002.cla>
- Davies, Peredur Glyn Cwyfan. 2010. *Identifying word-order convergence in the speech of Welsh–English bilinguals*. Bangor: Bangor University PhD dissertation.
- Davies, Peredur. 2016. Age variation and language change in Welsh: Auxiliary deletion and possessive constructions. In Mercedes Durham & Jonathan Morris (eds.), *Sociolinguistics in Wales*, 31–60. London: Palgrave MacMillan. DOI: [https://doi.org/10.1057/978-1-137-52897-1\\_2](https://doi.org/10.1057/978-1-137-52897-1_2)
- Davies, Peredur & Margaret Deuchar. 2014. Auxiliary deletion in the informal speech of Welsh–English bilinguals: A change in progress. *Lingua* 143. 224–41. DOI: <https://doi.org/10.1016/j.lingua.2014.02.007>
- Deuchar, Margaret, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto & Diana Carter. 2014. Building bilingual corpora. In Enlli Môn Thomas & Ineke Mennen (eds.), *Advances in the study of bilingualism*, 93–111. Bristol: Multilingual Matters. DOI: <https://doi.org/10.21832/9781783091713-008>
- Deuchar, Margaret, Peredur Webb-Davies & Kevin Donnelly. 2018. *Building and using the Siarad Corpus: Bilingual conversations in Welsh and English*. Amsterdam: John Benjamins. DOI: <https://doi.org/10.1075/scl.81>
- Donoso, Gonzalo & David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. In Preslav Nakov et al. (eds.), *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial'17*, 16–25. Valencia: Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/W17-1202>
- Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In Gosse Bouma & Yannick Parmentier (eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26–30, 2014*, 98–106. Gothenburg: Association for Computational Linguistics. DOI: <https://doi.org/10.3115/v1/E14-1011>
- Eisenstein, Jacob. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19. 161–188. DOI: <https://doi.org/10.1111/josl.12119>
- Eisenstein, Jacob. 2017. Identifying regional dialects in online social media. In Charles Boberg, John Nerbonne & Dominic Watt (eds.), *The handbook of dialectology*, 368–83. Oxford: Wiley–Blackwell. DOI: <https://doi.org/10.1002/9781118827628.ch21>
- Fasold, Ralph W. 1972. *Tense marking in Black English: A linguistic and social analysis*. Arlington, VA: Center for Applied Linguistics.

- Gonçalves, Bruno & David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PLoS ONE* 9, 1–6. e112074. DOI: <https://doi.org/10.1371/journal.pone.0112074>
- Gonçalves, Bruno & David Sánchez. 2016. Learning about Spanish dialects through Twitter. *Revista Internacional de Lingüística Iberoamericana* 14. 65–75.
- Grieve, Jack, Andrea Nini & Diansheng Guo. 2016. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21. 99–127. DOI: <https://doi.org/10.1017/S1360674316000113>
- Grieve, Jack, Andrea Nini & Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46. 293–319. DOI: <https://doi.org/10.1177/0075424218793191>
- Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami & Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2(11). 1–18. DOI: <https://doi.org/10.3389/frai.2019.00011>
- Guy, Gregory & Sally Boyd. 1990. The development of a morphological class. *Language Variation and Change* 2. 1–18. DOI: <https://doi.org/10.1017/S0954394500000235>
- Haddican, Bill & Daniel Ezra Johnson. 2012. Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics* 18. 31–40.
- Hazen, Kirk. 2011. Flying high above the social radar: Coronal stop deletion in modern Appalachia. *Language Variation and Change* 23. 105–137. DOI: <https://doi.org/10.1017/S0954394510000220>
- Hecht, Brent & Monica Stephens. 2014. A tale of cities: Urban biases in volunteered geographic information. In Eytan Adar & Paul Resnick (eds.), *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 197–205. Palo Alto, Calif.: Association for the Advancement of Artificial Intelligence.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59. 244–55. DOI: <https://doi.org/10.1016/j.compenvurbsys.2015.12.003>
- IPSOS Connect. 2017. Tech Tracker. Quarterly release: Q1 2017. <https://www.ipsos.com/en/technology-tracker-q1-2017>, accessed 22 August 2017.
- Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3. 359–383. DOI: <https://doi.org/10.1111/j.1749-818X.2008.00108.x>
- Jones, Bob Morris. 2004. The licensing powers of mood and negation in spoken Welsh: Full and contracted forms of the present tense of *bod* ‘be’. *Journal of Celtic Linguistics* 8. 87–107.
- Jones, Taylor. 2015. Toward a description of African American vernacular English dialect regions using “Black Twitter.” *American Speech* 90. 403–440. DOI: <https://doi.org/10.1215/00031283-3442117>
- Ljubešić, Nikola, Maja Miličević Petrović & Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography* 6. 100–24. DOI: <https://doi.org/10.1017/jlg.2018.9>
- Russ, Brice. 2012. Examining large-scale regional variation through online geotagged corpora. *Paper presented at the American Dialect Society Annual Meeting*, Portland.
- Sankoff, David & William Labov. On the uses of variable rules. *Language in Society* 8. 189–222. DOI: <https://doi.org/10.1017/S0047404500007430>
- Scheffler, Tatjana, Johannes Gontrum, Matthias Wegel & Steve Wendler. 2014. Mapping German tweets to geographic regions. In Josef Ruppenhoffer & Gertrud Faaß (eds.), *Proceedings of the 12th Edition of the Konvens Conference, Hildesheim, Germany, October 8–10, 2014*, 26–33. Hildesheim: Konvens Conference.




- Shoemark, Philippa, James Kirby & Sharon Goldwater. 2017. Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In Julian Brooke, Thamar Solorio & Moshe Koppel (eds.), *Workshop on stylistic variation*, 59–68. Copenhagen: Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/W17-4908>
- Sloan, Luke, Jeffrey Morgan, Pete Burnap & Matthew Williams. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE* 10(3). 1–20. DOI: <https://doi.org/10.1371/journal.pone.0115545>
- Stevenson, Jonathan. 2016. *Dialect in digitally mediated written interaction: A survey of the geohistorical distribution of the ditransitive in British English using Twitter*. York: University of York MA dissertation.
- Strelluf, Christopher. 2019. *Anymore, it's on Twitter: Positive anymore, American regional dialects, and polarity licensing in tweets*. *American Speech* 94. 313–351. DOI: <https://doi.org/10.1215/00031283-7587883>
- Tagliamonte, Sali & Rosalind Temple. 2005. New perspectives on an ol' variable: (t,d) in British English. *Language Variation and Change* 17. 281–302. DOI: <https://doi.org/10.1017/S0954394505050118>
- Thomas, Alan R., Glyn E. Jones, Robert O. Jones, David A. Thorne & Cathair Ó Docharthaigh. 2000. *The Welsh Dialect Survey*. Cardiff: University of Wales Press.
- Upton, Clive & J. D. A. Widdowson. 1996. *An atlas of English dialects*. Oxford: Oxford University Press.
- Van Halteren, Hans, Roeland Van Hout & Romy Roumans. 2018. Tweet geography: Tweet-based mapping of dialect features in Dutch Limburg. *Computational Linguistics in the Netherlands Journal* 8. 138–162.
- Willis, David. 2016. Cyfieithu iaith y caethweision yn *Uncle Tom's Cabin* a darluniadau o siaradwyr ail iaith mewn llenyddiaeth Gymraeg [Translating the language of the slaves in *Uncle Tom's Cabin* and representations of second-language speakers in Welsh literature]. *Llên Cymru* 39. 56–72. DOI: <https://doi.org/10.16922/lc.39.5>
- Willis, David. 2017. Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun *chdi*. *Journal of Linguistic Geography* 5. 41–66. DOI: <https://doi.org/10.1017/jlg.2017.1>

**How to cite this article:** 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): 103.1–33. DOI: <https://doi.org/10.5334/gjgl.1073>

**Submitted:** 19 August 2019    **Accepted:** 14 July 2020    **Published:** 03 November 2020

**Copyright:** © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 