Dialect in digitally mediated written interaction: a survey of the geohistorical distribution of the ditransitive in British English using Twitter

Jonathan Stevenson

MA by Research

University of York

Language and Linguistic Science

December 2016

Abstract

Recent research (Gerwin, 2013; Siewierska & Hollmann, 2007; Yáñez Bouza & Denison, 2015) uses historical and contemporary corpora to quantify diachronic and spatial distributions of variants of the ditransitive in British English. Each study pays particular attention to ditransitives with two pronominal objects, where internal factors influencing variation are reduced primarily to the choice of pronoun and verb type. Three variants are attested, a 'prepositional dative' (PDAT - 'send it to me'), a double-object (GTD - 'send me it') and an alternative regionally marked double object construction (TGD - 'send it me'). Corpus evidence reveals the pronominal TGD as the most frequent variant until the beginning of the 19th century, when the PDAT gained preference. The pronominal GTD, now considered canonical, emerges at the beginning of the 20th century. Broad agreement over the geographical distribution of the ditransitive is based primarily on maps drawn from the Survey of English Dialects (SED), but "comprehensive frequency data" (Yáñez Bouza & Denison, 2015, p.248) is lacking. The current project provides detailed frequency data drawn from language use on Twitter which is accurately mapped according to GPS coding. This map shows remarkable crossover with the SED maps, demonstrating both the stability of the geographical distribution over time and the amenability of "interactive written discourse" (Ferrara, Brunner, & Whittemore, 1991) to the expression of dialect. The results detail a large degree of variation in the relative frequency of each variable over physical space. Such variation brings into focus some important questions regarding the nature of a language as conceived in formal linguistic theory and a problematic tendency to 'lump together' large, linguistically diverse regions and treat them as one entity (Siewierska & Hollmann, 2007, p.97). Instead, using statistical tests for difference, the present study groups contiguous regions by the relative probabilistic frequencies of each variant. The results have implications for dialect geography, dialect syntax and recent approaches concerning regionally sensitive probabilistic approaches to grammar (Bresnan & Ford, 2010).

Contents

List of Figures

List of Tables

Acknowledgements

Thanks to Carmen Llamas, my supervisor, for providing the space on her undergraduate module *Methods in LVC*, for this project to get going and giving the courage to experiment. There is nothing like having your own project to motivate you to learn the skills you need to get it done.

Thanks also to Carmen, Paul Kerswill and Ann Taylor who provided clear, insightful and friendly mentoring throughout my undergraduate degree at York, and have continued into postgraduate. Thanks also to Bernadette Plunkett for syntax advice.

Thanks to Tim Shortis who has been a great source of inspiration and insight, for getting me into linguistics in the first place. Thanks to Pia for making me go back to into linguistics. Thanks to Colin for his well timed comment and to Katy for high quality coffee breaks. Thanks to my ridiculously patient and parents and my brother Matt.

Finally thanks to the ESRC and University of York for the generous funding and opportunity.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

"It is the task of sociolinguistics to describe and explain the patterns of variation that occur within a linguistic community given the theoretical limits of this variation uncovered by generative linguistics" (Barbiers, 2005, p.235)

Chapter 1

Introduction

1.1 The pilot study

The current research project is primarily socio-historical/socio-dialectological in nature. It forms the first part of a larger project which extends to PhD level, building on the methodology established during a successful pilot study (Stevenson, 2015).[1] The pilot study was composed of two parts:

1. Examining the geographical distribution of the use of variants of the ditransitive construction in British English using geographically coded Twitter data.

2. Comparing the Twitter data with data drawn from an electronically delivered grammaticality judgement survey.[2]

The focus of the current project is to extend part 1. The extension of part 2, the deployment of grammaticality judgement surveys, will form the first part of the PhD investigation and will complement the current study. Ultimately, it is envisaged that both methodological approaches will be applied concurrently as part of a more targeted study of specific geographic locales revealed to be of interest by the initial more general approach. The survey will additionally provide insight into the underlying syntactic structure of the ditransitive (and other structures) following the methodology employed by Haddican (2010) and subsequently Biggs (2014).

[1] Discussed in greater depth in chapter 4

[2] This follows recent efforts to understand linguistic variation by triangulating several approaches (Siewierska & Hollmann, 2007, p.100).

1.2 Investigation into the ditransitive

Variation in the ditransitive, often referred to as 'the dative alternation' (outlined in section 2.1), has been the focus of considerable academic interest - indeed - as Volk, Bresnan, Rosenbach, and Szmrecsanyi (2013, p.3) report, "it is one of the most studied alternations in the grammar of English". Yet what is known about how the structure is distributed remains limited and fragmentary.

The most comprehensive geographical picture currently available for British English remains that provided by maps drawn from the Survey of English Dialects (SED), which has no data for Scotland or Wales, and charts language use of speakers born at the end of the 19th century. Whilst there has been some recent survey work to provide updated dialect maps, the results are unable to provide the quantitative usage detail and relative frequencies of alternate variants that a corpus approach can offer. A study at Manchester University mimics the SED survey (MacKenzie, Bailey, & Turton, 2014) using surveys distributed by successive undergraduate cohorts as well as the survey deployed in the pilot study (Stevenson, 2015) which involved an online grammaticality judgement task. Additionally, a method to 'crowdsource' dialect data using mobile phone apps has proven successful in German-speaking Switzerland (Leemann, Kolly, Purves, Britain, & Glaser, 2016) with a similar app now available for the UK (Leemann & Blaxter, 2016). The results of these surveys confirm clear regional preferences in the ditransitive which do seem, to an extent, to mirror the SED's findings.

Recent investigations advocating the use of large-scale historical and contemporary corpora incorporating geographical metadata (Gerwin, 2013, 2014; Siewierska & Hollmann, 2007; Yáñez Bouza & Denison, 2015) aim to provide a picture of the distribution of variants over geographical space and historical time. However, whilst traditional corpus data provide some quantitative detail, regions are too broadly defined to allow for a precise presentation of the geographical distribution in present day English (PDE). Additionally, with the exception of Siewierska and Hollmann's (2007) study focusing on the distribution of the ditransitive in the Lancashire dialect, there is still a lack of clear data regarding the relative frequency of variants *in given locations*. Accounting for location-specific variation in syntax is important,

countering a tendency in the syntactic and semantic literature to bundle together large linguistic areas and treat them as one linguistic entity (Siewierska & Hollmann, 2007, p.97).

The bundling together of disparate datasets in the variationist study of syntax is, however, often a necessary step. The relative infrequency of syntactic variants often requires infeasibly large corpora to return meaningful results. This *data problem* (see section 2.3) is compounded when investigating structures like the ditransitive with two pronominal objects (pDit) (e.g. "send him it/send it to him") which tend to occur more frequently in speech than in writing (Biber et al., 1999) as the process of transcribing natural speech data is costly. This leads Siewierska and Hollmann (2007, p.99) to the conclusion that "for the status of the pronouns ... the value of corpus data has been, and is expected to remain, relatively limited". Whilst some data have subsequently emerged on the status of pronouns from historical and contemporary corpus analysis (Gerwin, 2014; Yáñez Bouza & Denison, 2015), this is achieved by pooling together somewhat disparate datasets (see section 2.7.1).

The solution suggested by Hollmann and Siewierska (2006) - "Corpora and (the need for) other methods" - is to marry corpus approaches with traditional grammaticality judgements asking participants to introspect about the validity of a given structure (in their dialect). Doing this allows the investigation of structures that occur infrequently in production, but brings with it all of the issues associated with grammaticality judgements (discussed in section 2.2). A similar triangulation of corpus and introspection data has been suggested as possible way to mitigate an over-reliance on grammaticality judgements as the sole source of linguistic evidence in syntactic theory-building (cf. Adger & Trousdale, 2007; Schütze, 1996; Schütze & Sprouse, 2014).

Whilst combining corpus and judgment data seems to offer a reasonable way to proceed in the investigation of dialect grammar, a new approach to the gathering of dialect data has emerged in the last few years which takes advantage of the massive amount of linguistic output generated on social media. This approach is outlined in the following section.

1.3 Computer-mediated communication (CMC)

It is no exaggeration that CMC has brought about a fundamental shift in communicative practices. Particularly in the past ten years, as devices have grown more portable and capable, they have become progressively more deeply integrated into personal and social life. Crucially for an investigation looking at a feature found mostly in spoken interaction (Biber et al., 1999) (discussed in section 2.4), the interaction that occurs in *conversational* Twitter messages (see section 2.8.1) results in writing that is, much like other forms of messaging,[3] 'conceptually spoken' (Schlobinski, 2005), favouring a language of social proximity. In the case of Twitter messages, the data are publicly accessible and a proportion is geographically locatable.[4] This provides the opportunity to map naturally occurring unmonitored language use on a scale large enough to reveal syntactic patterns at a geographically local level.

Whilst CMC research is an established field, its efficacy for mapping 'real-world' language use is only recently coming to light and is now being used as a serious methodological tool in dialectological research (e.g. Eisenstein, 2017; Jones, 2015). This is a methodologically pioneering approach to dialectology that views CMC as a source of natural, spontaneous dialect data capable of revealing existing, historically robust dialect 'faultlines' (Eisenstein, O'Connor, Smith, & Xing, 2014) with 'fine spatiotemporal resolution and continuity' (Huang, Guo, Kasakoff, & Grieve, 2015, p.1). These are important data not only for traditional dialectology, the practitioners of which are interested in geographical distribution, but also to formal and historical linguists seeking a probabilistic account of the relative frequency with which competing variants and associated objects occur (cf. Bresnan, Cueni, Nikitina, & Baayen, 2007). The answer from these kinds of data is clear - it depends critically on which geographic variety of English is being investigated.

[3] Namely synchronous and semi-synchronous interactive writing

[4] Both using GPS data (infrequent) and user-inputted data (frequent). The nature of the location data available and their being subject to changing rules imposed by Twitter are discussed in section 5.4 and the correlation between user-inputted location and GPS data is tested in section 6.3.

1.4 Aims

By mapping GPS encoded Twitter data, the present study aims to:

1. Provide both geographical resolution and relative frequency data among pronominal ditransitive types (outlined as examples 1a-1c in section 2.1) by region, generalising larger linguistically similar super-regions.

2. In so doing, present evidence that the social and pragmatic context of Twitter creates an environment in which features typically associated with vernacular speech tend to occur.

3. Give additional insight into historical trends brought to light in recent historical corpus investigations (Gerwin, 2013; Yáñez Bouza, 2016; Yáñez Bouza & Denison, 2015).

4. Generate sufficient data to assess the status of *pronoun* and *verb* and their influence on ditransitive choice.

1.5 Roadmap

Section 2.1 outlines variation in the ditransitive construction in English generally and describes the particular variants under consideration in the present study.

Section 2.2 discusses the concept of 'dialect grammar' and situates the current investigation within it.

A discussion of the established problem of gathering sufficient data for the study of syntactic features is presented in section 2.3, as is how leveraging Twitter data may offer a possible solution.

Section 2.4 presents a discussion of the literature concerning differences between speech and writing. This is pertinent to the current investigation. As will be explained, the ditransitive with pronominal objects (pDit) is a feature primarily of spoken, rather than (traditional) written, interaction. The exceptions here are personal correspondence and, centrally to the current investigation, certain kinds of interactive behaviour in CMC.

Sections 2.5 and 2.6 present an overview of recent survey and corpus work on the ditransitive that aims to provide a representation of the distribution of the structure over geographical space and historical time.

Section 2.7.1 looks at constraints on the ditransitive resulting from the choice of object pronoun, whilst section 2.7.2 looks at constraints resulting from choice of verb.

Section 2.8 introduces computer-mediated communication and provides a backdrop against which to present Twitter, the instantiation of CMC from which the data for the current project are drawn.

Section 2.8.1 focuses on recent literature that takes advantage of dialect on Twitter, providing mappable, rich data, with some caveats.

Chapter 4 gives a brief overview of the pilot project which formed the blueprint for the current project and the larger project to follow.

The methodology is detailed in chapter 5 followed by a presentation and analysis of results (chapter 6), which is then followed by an in depth discussion of the findings and how these findings fit in with the current research landscape (chapter 7).

Finally, a detailing of future directions for further study (section 8) is presented.

Chapter 2

Literature review

2.1 Variation in the ditransitive

Variation in the ditransitive has generally been approached as an alternation between two semantically equivalent variants, a *double object construction* where the GOAL precedes the THEME (GTD, 1a) and a *prepositional dative* where the GOAL is contained within a prepositional phrase which follows the THEME (PDAT, 1b). A third alternative double object construction where the THEME precedes the GOAL (TGD, 1c), is also available in a significant area of the Midlands and North-West England (Haddican, 2010; Hughes, Trudgill, & Watt, 2012). Whilst the TGD is acknowledged in the literature, it has often been considered a minority *dialectal* variant (Siewierska & Hollmann, 2007).

(1) Ditransitive variants in British English, with pronominal objects (pDit)[1]

a. send him it (GOAL-THEME Ditransitive: GTD)

b. send it to him (Prepositional Dative: PDAT)

c. send it him (THEME-GOAL Ditransitive: TGD)

A fourth potential variant (pseudo-TGD, 2) discussed in the syntactic literature (Biggs, 2014; Haddican, 2010) is in its surface structure identical to the TGD, but is underlyingly a PDAT with an elided preposition, rather than a GTD with reversed

[1] A note on acronyms used here: unless explicitly stated, in this paper, TGD, GTD and PDAT refer to ditransitives with two pronominal objects.

order.[2]

(2) send it [to NULL] him (Pseudo-TGD)

Although the TGD (1c) is quite widely accepted where both objects are pronominal, it becomes rapidly less acceptable when full NP objects are used (3a-3c).

(3) Ditransitive variants in British English[3]

a. ?she gave the ball her

b. ?she gave it the boy

c. ??she gave the ball the boy

There are established internal causes for the reduction in acceptability of the TGD with full NP objects in Present Day English (PDE), principally relating to weight and information structure (cf. Siewierska & Hollmann, 2007; Yoshikawa, 2006), and this is reflected in the historical record. As Yáñez Bouza and Denison (2015, p.253) demonstrate with full noun phrase (NP) objects, occurrence across a range of historical corpora (see list of corpora used in table A3 in the appendix) is minimal - ranging from 3.3% in Early Modern English (EModE) to (1.3%) in Late Modern English (LModE). The lack of availability of the TGD with full NP objects may explain why the TGD as a whole often fails to feature in discussion of the dative alternation (op. cit.).

It is where both objects are pronominal, however, that certain *alternating* verbs (see section 2.7.2) in British English[4] allow all three orders in (1) and display a high degree of variation that can be observed over historical time and geographical space. In contrast to the low occurrence of the TGD with full NP objects, historically

[2]There is some disagreement on whether there exists such a fourth possible variant or whether what the TGD represents is a relic of what used to be available in full NP form, and which is now available as a 'prefabricated expression' (Yáñez Bouza & Denison, 2015).

[3]The lack of acceptability of the structures in (3) is based on the UK-wide survey put out for the pilot study; see figure A1 in the appendix. Whilst Siewierska and Hollmann (2007) leave these structures as grammatical, but dialectal, this decision is based on the report from Hughes et al. (2012), which appears to lack an empirical base. The pilot survey results also mirror Haddican's (2010) finding for Manchester speakers.

[4]Interestingly, variation does not seem to be in evidence in present day Canadian English, according to Tagliamonte's (2014b) synchronic corpus work

the TGD (1c) appears dominant. Gerwin (2013) shows that throughout the 19th century, the TGD (1c) and PDAT (1b) are virtually the only attested options both in US English and the English spoken in England (as will later be reported in the current study, the situation may be different in Scotland). Yáñez Bouza and Denison (2015) also report the TGD being the dominant form until the 19th century in England, when the PDAT gained preference. Both studies report that the GTD (1a) is virtually unattested in England before the 20th century.[5]

The existence of geographical variation in the distribution of the pDit in present-day English (PDE) is well known, but it is not clear exactly how and where the variation manifests itself. As Hollmann and Siewierska (2006, p.205) report: "there is considerable confusion in the literature as to the presence and distribution of these three constructions in British dialects". Additionally, the extent to which any hypothetical *pseudo-TGD* might contribute to frequency data for the TGD, is unclear. This considered, in their corpus analysis of the Lancashire dialect, Siewierska and Hollmann (2007, p.96) show that rather than being a minor alternative variant, the TGD represents the dominant variant in that region, being twice as common as the GTD.

The discrepancy between the historical record, which shows the TGD all but dying out and contemporary dialect data showing it still alive and well, raises a critical issue. With clear regional variation, shown to be persistent over time, how should such variation be accounted for in historical treatments which necessarily privilege time over space (Britain, 2013)? That is, how can we be sure that a change charted in the historical record is charting a change in the language itself or instead charting the manifestation of regional variation in writing? This issue will be considered in the discussion (chapter 7) and put to future investigation (chapter 8).

[5]This is in contrast to the GTD with one or both objects as full NP, which is highly frequent. There is a notable and potentially revealing exception to this in the 15th century, when as Yáñez Bouza and Denison (2015) show (see figure 2.2), the GTD was used by writers during that period. This is discussed in section 2.5.

2.2 Dialect grammar

The focus of formal linguistics on internal language (I-Language) (Chomsky, 1986) and moving away from the use of external language (E-Language) as a source of data is countered by the fact that the primary methodology advocated to 'access' a speaker's internal language system, namely judgements of acceptability, is itself highly problematic. The context of a given set of judgement tasks is often artificial, as are the test sentences participants are asked to judge. And whilst there are ways to mitigate the artificiality of the experimental context, it is hard to remove the influence of such a context from any answer provided to such tasks. Additionally, judgment tasks are bound, to a greater or lesser extent, to be influenced by the somewhat elusive notion of the 'standard' or 'correct' form (Adger & Trousdale, 2007, p.265). On the other hand, spontaneous natural language data can provide a window onto what people actually do when not under the microscope. And crucially, if a feature can be shown to occur with enough frequency, the position that such a feature is simply noise becomes less tenable (Crisma & Longobardi, 2009). Indeed, as Kortmann (2003, p.64) reports, such data are in fact highly valuable to syntacticians looking at microparametric (language internal) variation. Adger and Trousdale (2007) propose *S-Language* as a third distinction in addition to I/E Language. S-Language aims to capture the externalisation of language specific to individual speech communities (cf. Labov, 1972) and make it available to observational methodology. The concept of a speech community in its original Labovian sense of being defined "by the uniformity of abstract patterns of variation which are invariant in respect to particular levels of usage" (Labov, 1972, p.120) is particularly relevant here. The aggregation of speakers in necessarily size-limited areas (Kerswill, 1994) perpetuates norms and patterns of usage as a result of ongoing face-to-face interaction. The corpus study of dialect grammar, then, serves to validate E/S-Language as a data-source and serves as a connective between dialectology and theoretical linguistics (Hollmann & Siewierska, 2006).

2.3 The data problem

Gaining access to spontaneous natural language data in sufficient quantities to allow statistically significant analysis of syntactic variation is not a straightforward task. This difficulty is due primarily to the fact that a given syntactic feature occurs with much less frequency than, for example, a given phonetic feature in a passage of language data.

Written language, the only option of course for historical texts, has the advantage of being more readily compiled into a searchable database, but is by its nature typically not spontaneous. Spoken language, on the other hand, may be more spontaneous, but such spontaneous speech requires considerable effort to first capture and second transcribe. Despite this, technological and methodological advances in corpus linguistics have permitted access to "sufficiently large amounts of data to enable the study of grammatical phenomena" (Hollmann & Siewierska, 2006, p.204) in spontaneous speech and substantial projects have been undertaken. A principal dataset here is The Freiburg Corpus of English Dialects (FRED), the availability of which (and the potential of such corpora) led Kortmann (2003) to assert

> "Most importantly perhaps, [dialect grammar] can serve as a corrective for typology, which often does not take sufficient care of the striking differences between the grammars of standard (written) and spoken varieties of languages, thus running the risk of comparing apples and oranges, as it were". (Kortmann, 2003, p.63)

Despite these advances, the problem is again compounded when looking at subsets of the data to distinguish patterns constrained by independent variables (e.g. region) or internal constraints (as in the current study, pronouns, verb type). Even a very large dataset, when subdivided in this way, can end up returning a paucity of results. If regional and social variation is to be accounted for in the analysis of internal variation, therefore, there is quite simply a need for more yet more data.

2.4 The pDit as an element of speech

Gerwin (2014, p.123) compares the The Freiburg English Dialect Corpus (FRED)[6] corpus (spoken 1970s) to the LOB corpus[7] (written 1970s) as well as the spoken parts of the British National Corpus (BNC)[8] (BNCreg) (spoken 1990s) and F-LOB[9] (written 1990s). The results show a marked preference for the double object structure (GTD) with GOAL as pronoun and THEME as full NP in speech. Where both objects are full NP, the frequency of occurrence is higher in writing. The prepositional form with two full NP objects also shows up to a far greater extent in writing than in speech.

Of interest to the present investigation, where both objects are pronouns, Gerwin's (2014) results show no instances of TGD or GTD in writing, with only between 1-2% PDAT. This supports the results published by Biber et al. (1999) which show that these structures are almost exclusive to speech and fiction (figure 2.1 below).

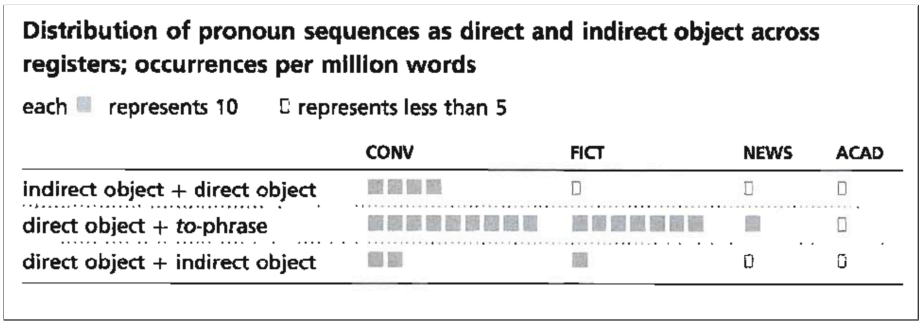

Figure 2.1: Occurrences of pDit in the Longman Grammar of Spoken and Written English

Source: Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). Longman grammar of spoken and written English (Vol. 2) (p.929). Harlow, Essex: Pearson Education.

The finding that the pDit is primarily a feature of interactive speech makes pragmatic sense. Pro-forms, standing in for full noun phrases, add potential ambiguity to a text. This effect is amplified where there are two pro-objects indexing two separate previously identified objects. The general observation that spoken and written grammar differ as a result of "the different properties and functions of the

[6] Available at `http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/`.

[7] Available at `http://www.helsinki.fi/varieng/CoRD/corpora/LOB/`.

[8] Available at `http://corpora.lancs.ac.uk/BNCweb/`.

[9] Available at `http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/`.

two channels" Milroy and Milroy (2012, p.61) is further elaborated by Biber, Gray, and Staples (2016) specifically with regard to pronouns:

> "Conversational participants share time and place, and they normally also share extensive personal background knowledge. As a result, colloquial features like pronouns and vague expressions are common. Referring expressions usually do not need to be elaborated in conversation because the addressee can readily identify the intended reference." (Biber et al., 2016, p.1).

The rarity of pDits in written English fuels Siewierska and Hollmann's (2007) concern that finding sufficient examples of pronominal ditransitives in spoken corpora would require prohibitively large datasets.

2.5 Historical distribution

This considered, one way to increase the amount of data available for study is to pool datasets together. This approach is taken by by Yáñez Bouza and Denison (2015) who combine multiple historical and contemporary corpora (see table A3 in the appendix) to produce an impression of the relative frequencies of each pDit variant over historical time (figure 2.2). As the authors caution, the resulting chart is "convenient for showing general trends but has no statistical validity: the data come from disparate corpora of very varied make-up, including diachronic corpora with non-matching chronological divisions" (Yáñez Bouza & Denison, 2015, p.254).

This acknowledged, the resulting chart (figure 2.2) shows a striking trend: the TGD has been, for most of the historical record, the favoured variant and, counter to the situation described in the previous section, is clearly acceptable in writing. The PDAT is reported as a minority feature ranging from 20% to 30% of total pDit occurrence until the early 18th century.[10] From the early 18th century, there is an apparent change in fortunes: PDAT use rises seemingly at the cost of the TGD, which declines in use at an apparently similar rate. Somewhere around the early 19th century, the PDAT becomes the dominant form, rising to 80% use by the late 20th

[10] With the exception of a spike to $\approx 50\%$ in the early 16th century.

century. Meanwhile, the GTD appears very infrequently in the historical sample used, with an interesting exception in the early 16th century. Its rise to a steady frequency does not occur until the 20th century. The TGD drops in the 20th century and, as the authors point out (p.263), "limitation to dialect use of ...[TGD]... seems to be a twentieth-century phenomenon". The presence of the TGD as a feature of dialect use appears to play against the general trend that sees the feature apparently virtually dying out. As will be discussed in section 2.6, some areas - notably the North West of England - are reported to have majority TGD use (Siewierska & Hollmann, 2007; Yáñez Bouza & Denison, 2015).

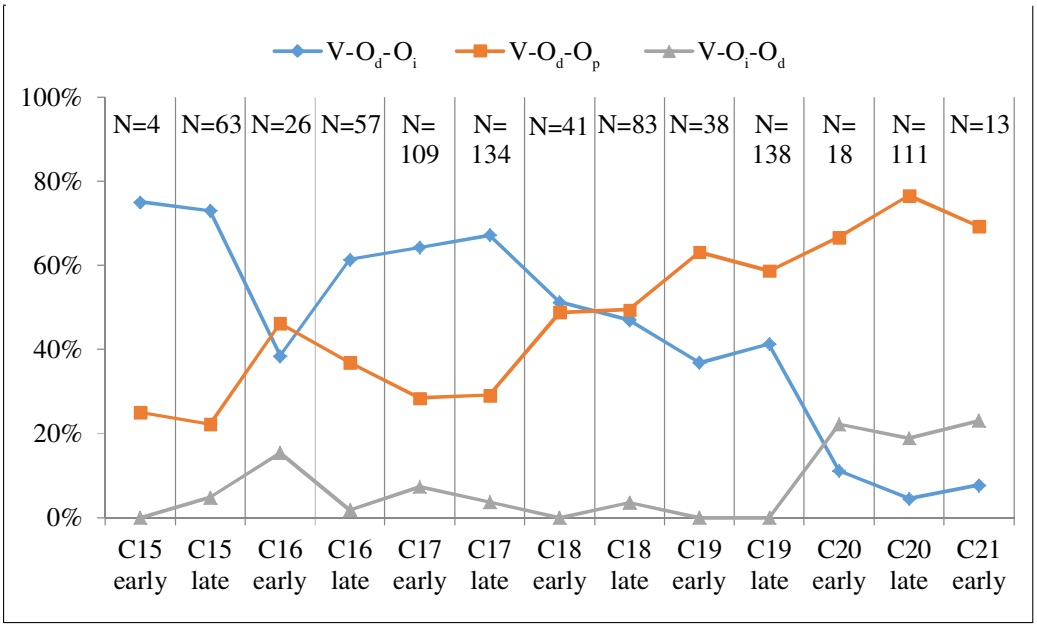

Figure 2.2: Composite graph from Yáñez Bouza and Denison (2015, p.255).
Ditransitives with two pronouns.
VO_dO_i=TGD, VO_dO_p=PDAT, VO_iO_d=GTD
Source: Yáñez-Bouza, N., & Denison, D. (2015). Which comes first in the double object construction? *English Language and Linguistics, 19*(02), p.255.

Gerwin (2014, p.183) also reports much discussion of TGD in use well into the 20th century, raising the point that "the question which needs to be answered is why the (present-day) canonical construction [GTD] manifested at all". Looking at the SED maps, however, it seems likely that the GTD was well established across much of England by the beginning of the 20th century. SED informants were born in the late 19th century, and so would have been in young adulthood by 1910-1920. This

argumentation follows the *apparent time hypothesis*, which assumes that linguistic features are more or less set by young adulthood, and as such the features used by older speakers are indicative of past practice, at least in spoken English.[11]

Additionally, as revealed in the pilot study (Stevenson, 2015) (see chapter 4), the GTD is the *majority* feature in the North East of England and Scotland. It seems unlikely that this should only be the case in the 20th century. Indeed, the existence of the GTD in Scotland, at least, is evidenced by 19th century grammar guides presented by Yáñez Bouza (2016) reporting the GTD as a 'Scotticism' (see section 2.5.1). It would be useful to consult the Corpus of Scottish Correspondence (1500–1715) which charts an analogous period to the Corpus of Early English Correspondence, and check how far back the GTD can be traced in the corpus record for Scotland. Unfortunately the corpus has been under (re)construction for the period of the current investigation.[12]

Focusing on the GTD, the apparent temporary rise in use in the early 16th century may be explained by the disparate nature of corpora used. Searches conducted in the initial stages of the current investigation[13] show a relatively high proportion (see table 2.1 below) of GTD use in the Corpus of Early English Correspondence, which represents the period in question. It may be that what we are seeing in figure 2.2, then, is the use of the GTD in personal letters which are over-represented in this chart for that period.[14]

[11]The application of the apparent time hypothesis is common practice in many sociolinguistic studies. However, it is intended as indicative rather than an absolute measure. Of course, it is *possible* that some SED informants may have switched their use of the pDit later in life.

[12]The project coordinator, Anneli Meurman-Solin, made clear that the corpus would be available soon.

[13]Preliminary searches of the Penn Parsed Corpora were conducted looking at pronominal TGD and GTD. The results are relevant here, but peripheral to the main methodology of the current investigation, the focus of which is the current distribution as revealed on Twitter.

[14]Why we might see this spike in GTD use in these letters is an interesting potential line of enquiry. It may be explained by the geographical locations of the authors.

	count	%
TGD	196	85%
GTD	35	15%
Total	231	100%

Table 2.1: Occurrence of TGD and GTD in the Corpus of Early English Correspondence (EEC)

The fact that the existence of the pronominal GTD alongside PDAT and TGD predates at least the 16th century is interesting to note. It is not a new feature of English; indeed, it has likely been in quite frequent use in speech throughout the history of English across different regions, but has been largely undetectable in the written corpus record.

2.5.1 Prescriptive grammar guides

Evidence that the GTD has been well established in Scotland and nearby regions of England for some time is also provided by a recent analysis of the prescriptive grammar guides of the early 18th century (Yáñez Bouza, 2016; Yáñez Bouza & Denison, 2015). The aim of these investigations is to determine whether such guides may have contributed to the fall of TGD usage in England, at least as represented in the available written texts. The rationale here rests on new evidence that prescriptive guides are able to influence the development of morphological and syntactic features (Yáñez Bouza, 2016). The finding that 18th century guides in fact declared the TGD to be the correct form and the GTD to be a 'vulgar' Scotticism (see extract below) led to the conclusion that these prescriptive guides could not be a cause for the demise of the TGD in England.

> "The Scotch and Engliſh dialects, alſo differ in arrangement.
> *Give me it, ſhow me it.*
> Give it me, ſhow it me." (Sinclair, 1782, p.62) (Italics in the original denoting Scotticism, cited in Yáñez Bouza and Denison (2015))

What Yáñez Bouza (2016) reveals here is interesting: 18th century English prescriptivists wanted to underscore that the TGD should be considered not only En-

glish, but 'correct', *proper English*. The fact, however, that the alternation was salient enough to make the pages of prescriptivists' guides indicates that both forms were in competition at that time. Whilst the guides cannot be said to play a causal role in the fall of TGD use in England, they may provide an early indication that a change towards the GTD was already underway here. Whilst GTD use does not (re)surface in the corpus record until the start of the 20th century (see again figure 2.2), it is not altogether absent either, also being used in the early 16th century.

Of course, grammar usage guides have as their principal target, the (standard) written language and the point of a prescriptive guide is to instruct writers to do one thing and *not* another. But for the alternation to qualify for entry into the guides, there must have been some (perceived) threat from the GTD at the expense of the TGD in written English. The guides in this light may be seen as an effort to resist an impending shift towards the GTD, and — it would seem — to this end, they had some success.

Considering *spoken* English, we know from current dialect data (Siewierska & Hollmann, 2007), of course, that the TGD in fact did not fall out of usage. Similarly, the apparent rise in GTD use in the 20th century, visible in figure 2.2, may be explained by the inclusion of both spoken data for that period, and importantly, a broader spectrum of writers. The democratisation of literacy resulting from the 1871 Education Act, created a new generation of writers. Written data for the 20th century, following this substantial social shift, is arguably going to be more democratically representative of language habits. This again fits with the line of thinking that the GTD was in use in speech in England — speakers of English have the option of flipping the order[15] — but the form failed to surface in writing. Whilst Yáñez Bouza and Denison (2015) are careful to include speech-based data for the pre-20th century record, finding little evidence of GTD use, it is difficult to know how representative these sources are of the everyday speech habits of the time.

[15]That is the linguistic system seems to allow it, dispreference for use is therefore socially motivated.

2.5.2 Geohistorical trends

Any investigation into the geographical distribution of linguistic features is also fundamentally historical in nature. This is implicitly the case as soon as an inquiry asks why a particular distribution occurs. Indeed, providing evidence for the historical investigation of English was reportedly an *explicit* motivation behind conducting the Survey of English Dialects (SED) (outlined in section 2.6.1) (Kretzschmar, 1999, p.274).

Looking at the regional distribution in the historical record, Yáñez Bouza and Denison (2015) examine the available regional metadata of the corpora under investigation. They show that the TGD has been quite established in the North West since the 17th century, comprising 49% of pDit instances, whilst the PDAT stands at 47.9%. The North-East shows 27.1% TGD, 42.4% PDAT and 30.5% GTD for the same time period, though it is unclear what *North East* refers to here (Yorkshire as well as Tyneside/Northumbria). East Anglia is reported as having a particularly high TGD usage, at 77%, although this is mostly from an earlier period (1420-1625). Comparing their historical data to the currently known distribution (explored here in section 2.6), they conclude that "in early English examples such as *give it me* were not confined to the same dialect areas as today" (Yáñez Bouza & Denison, 2015, p.261).

Meanwhile, Gast (2007, p.16), uses an analysis of the SED maps (see figure 2.7, section 2.6.1) to explore possible language contact scenarios as explanations for the reported distribution. Specifically, the possibility that the GTD may have come about due to contact with Old Norse, Gast (2007, p.16) evaluates that "in Old Norse there was a clear tendency towards REC-TH [GTD] order" and that the area where GTD is reported on those maps corresponds roughly to the Danelaw. However, Gast (2007, p.17) also proposes an alternative explanation whereby the GTD may have come about "without any external influence", being explicable solely by its being "favoured by the principle of analogy" with the ordering favoured with full NP objects.

Tagliamonte (2014a, p.302) considers the hypothesis that the development of the PDAT was at least in part due to contact with (Norman) French, speculating

that it may have come about due to a '*change from above*'. Whilst De Cuypere (2014) contends that the PDAT was in common use in Old English prior to the Norman conquest, its use was confined to fewer verbs - interestingly, we see an early distinction here between *give* and *send*. According to De Cuypere (2014, p.2), citing Mitchell (1985), verbs indicating transfer of possession such as "agifan, gifan, sellan 'give' and offrian 'offer'" did not take the PDAT whereas, "cweðan ('to say, speak'), sprecan ('to speak, say, utter'), cleopian ('to call, cry out'), sendan ('to send'), lætan ('to let'), niman ('to take') and bringan ('to bring')" *were* able to take the PDAT. What may have been the case is that Norman influence pressured an extension to the number of verbs able to take PDAT, resulting in an increased frequency of PDAT in areas of high Norman integration.

Such a hypothesis is not out of line with the geographical distribution seen in the pilot data (though interestingly not so much in the SED) - the PDAT is favoured in the South, but drops off progressively as one goes north, falling sharply out of favour north of Yorkshire, beyond which Norman influence was suppressive rather than integrative (cf. the 'Harrying of the North') (see map in figure 2.3).

Figure 2.3: Map of Norman England

Source: `http://www.heritage-history.com/maps/1heurope/eur025.jpg`. [Accessed: 30/11/2016].

More recently, the period in which the TGD is reported to have seen a dramatic fall in usage and the PDAT apparently rose to prominence coincides with large-scale population shifts in England as a result of the Industrial Revolution. Notably, the area of the Midlands and North West saw rapid expansion during this period, as

did London and Glasgow[16]. This can be seen in the map in figure 2.4 and table 2.2 below.

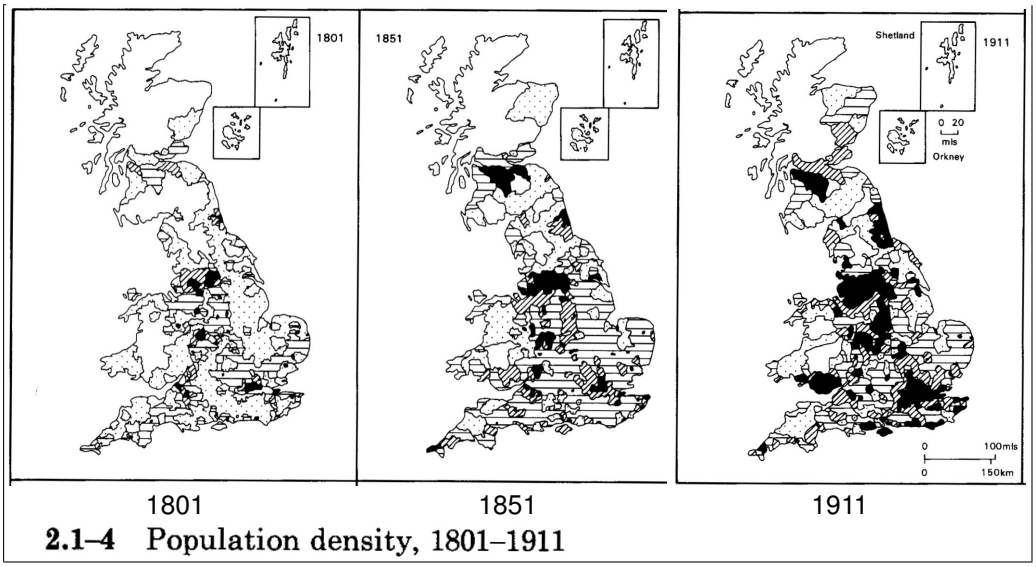

Figure 2.4: UK population density 1801-1911 (Langton & Morris, 1986)

Source: Langton, J.,& Morris, R. (Eds). (1986). *Atlas of industrialising Britain 1780-1914*. London: Methuen.

City	18th Cent.	1801	1831	1851	1901
Glasgow	12K (1725)	77,000	200,000	320,000	762,000
Newcastle		33,322	48,950	80,184	246,905
Leeds	16K (1771)	94,421	183,015	249,992	552,479
Hull		21,280	40,902	57,484	236,772
Manchester	43K (1774)	88,577	205,561	339,483	642,027
Liverpool	30K (1766)	82,430	180,222	320,513	711,030
Sheffield	7K (1736)	60,095	112,408	161,475	451,195

Table 2.2: Northern cities in the 19th century

Source: `http://www.visionofbritain.org.uk`), originally cited in lecture by Paul Kerswill, York 2014

Clearly, such large movements of peoples during the time periods under consideration need to be acknowledged when trying to account for patterns of usage seen

[16]This point about the effect of migration patterns on dialect in industrialising Britain was made in a lecture by Paul Kerswill at York, 2014.

in historical and present day English. Consideration of geohistorical trends will be returned to in the discussion, section 7.3.

2.6 The current geographical distribution

> "No better example exists of a syntactic puzzle than the quite definite regional preferences for the standard *give me it* in northern and eastern England, a non-standard *give it me* in the West Midlands, and an expanded *give it to me* in the south-west, as recorded by SED" (Upton, 2006, p.409)

That there is a clearly defined geographical spread of the pDit is well established in the literature. And variants of the pDit have evoked strong associations with place, identity and correctness, as discussed in the section 2.5.1 regarding prescriptive grammar guides. Certainly, it would appear that the forms still evoke strong regional associations today, specifically relating to the 'north/south divide', as shown in the following Twitter excerpts:

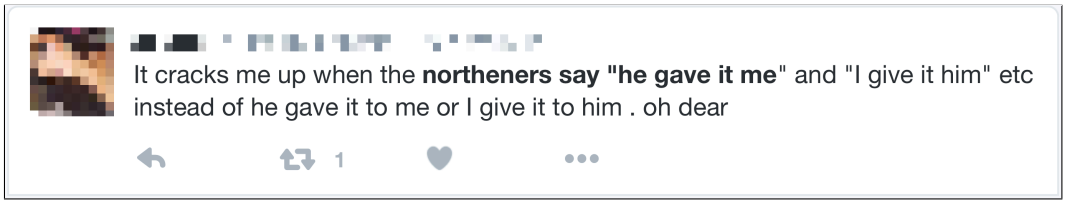

Figure 2.5: Twitter excerpt showing regional identification of TGD (1)

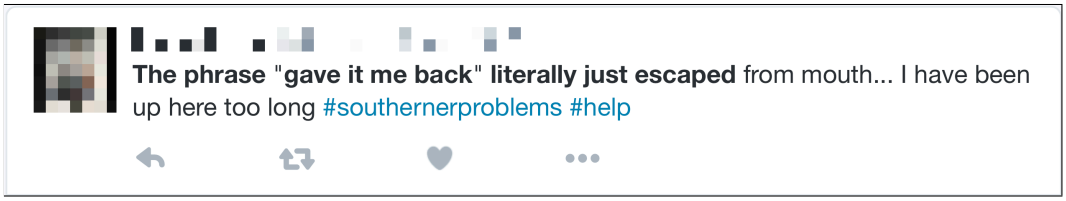

Figure 2.6: Twitter excerpt showing regional identification of TGD (2)

This impression is to a degree mirrored by Hughes et al.'s (2012) qualitative account, reporting on its general conception as a feature of northern speech. They report that the TGD is "very common indeed" (p.19) in the speech of 'educated' speakers of the North of England, and *acceptable* to many people in the South.

This overt reference to the TGD being used by educated speakers of the North of England appears to counter a potential reading of the structure as being class- or education-based. Hughes et al. (2012) also report that the PDAT appears to be the most common form in the South of England, "particularly where the direct object [GOAL] is a pronoun" (p.19). There have been some more systematic and quantitative investigations into the geographical distribution of the ditransitive. These investigations are outlined in the following sections.

2.6.1 Survey of English Dialects (SED)

The most comprehensive data regarding the geographical distribution of the pDit remain that obtained as part of the Survey of English Dialects (SED) (see map in figure 2.7). The data on which this map is based were gathered in the 1950s by interviewing elderly male informants in rural locations ('NORMs').[17] This methodology was designed to provide a window to the past: older speakers who had not come into contact with many 'outsiders' and thus would preserve, in their speech, patterns that were present during their young-adulthood (in this case, then, the beginning of the 20th century). The structure of interest here, the pDit, was elicited from informants using the following question:

> "Jack wants to have Tommy's ball and says to him, not Keep it!, but..." (SED questionnaire, IX. 8.2)

Participants apparently responded with one of the three variants: *give it me*, *give me it* or *give it to me*. The resulting map (figure 2.7) reveals a large area of the West Midlands and the North West of England where the TGD is used.

Immediately obvious when viewing the map is the wide geographical coverage of each isoglossic area. This is in line with Kortmann's (2004, p.2) assertion that "the areas to which morphological and, particularly, syntactic properties may be restricted are typically larger than those for regionally restricted phonological and lexical features". This is interesting to note looking to the current investigation, which will seek to ascertain the extent of such larger regions based on the statistical

[17] NORMs is the acronym given to refer to *Non-mobile Older Rural Males*.

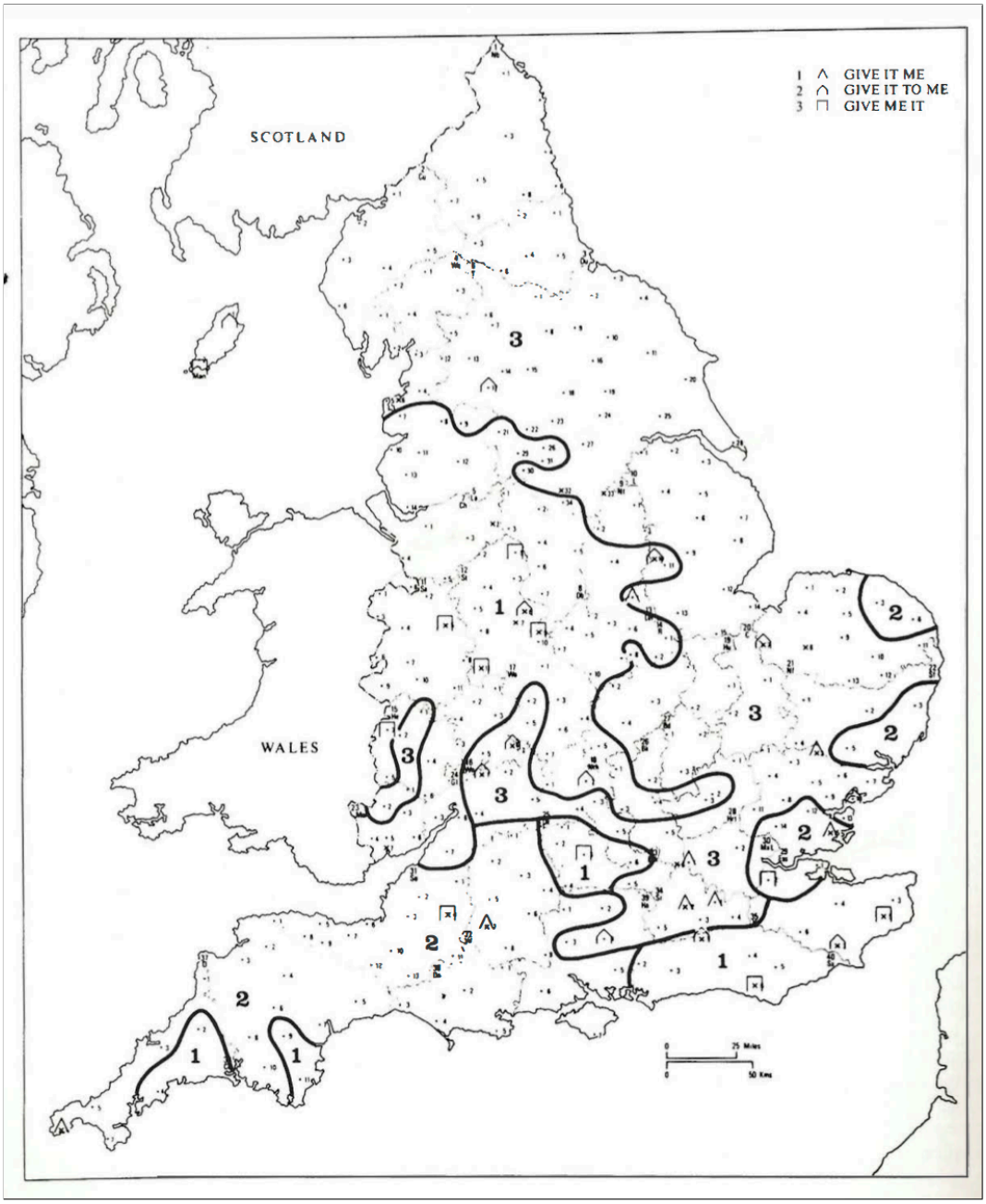

Figure 2.7: Mapped SED data showing areas where different pDit variants were reported in the 1950s survey.

Source: Kirk, J.M (1985). Linguistic atlases and grammar: The investigation and description of regional variation in English syntax. In J.M. Kirk & S. Sanderson (Eds.), *Studies in linguistic geography : The dialects of English in Britain and Ireland (p.132)*. London: Croom Helm

similarity in the three-way relative frequency of each pDit type in conjoining sub-regions (see section 6.3).

Interestingly, following the previous discussion on the origins of the GTD, the map reveals GTD use to be widespread from the South and East of England to the North East and far North West (Cumbria). This again counters an analysis specifying that the GTD is a new feature, instead supporting the proposal that it has long been a feature of speech in British English dialects but tends not to show up in historical corpora. The PDAT is, according to its geographical spread, the minority feature, though it is represented as being dominant in the vicinity of London, the centre of power and linguistic influence, and much of the South West. What this map does not show, (by design) is use in the urban centres. Also not visible are the relative counts of each structure by region: each isoglossic area outlines where a majority of a given feature occurs.

The northern border of the "give it me" area on the SED loosely corresponds to the well-established linguistic heterogloss known as the 'Humber-Ribble' belt (Viereck, 1986), shown in figure 2.8. Given how closely the survey evidence presented above and pilot data (see map in figure 4.1 in section 4.2) correlate with the SED map, it seems likely that whilst many of the distinctions represented in the map may have been levelled out, this border is still active for the pDit. Getting more detailed geographical Twitter data should provide an indication of where the border lies and will open the door to further investigation.

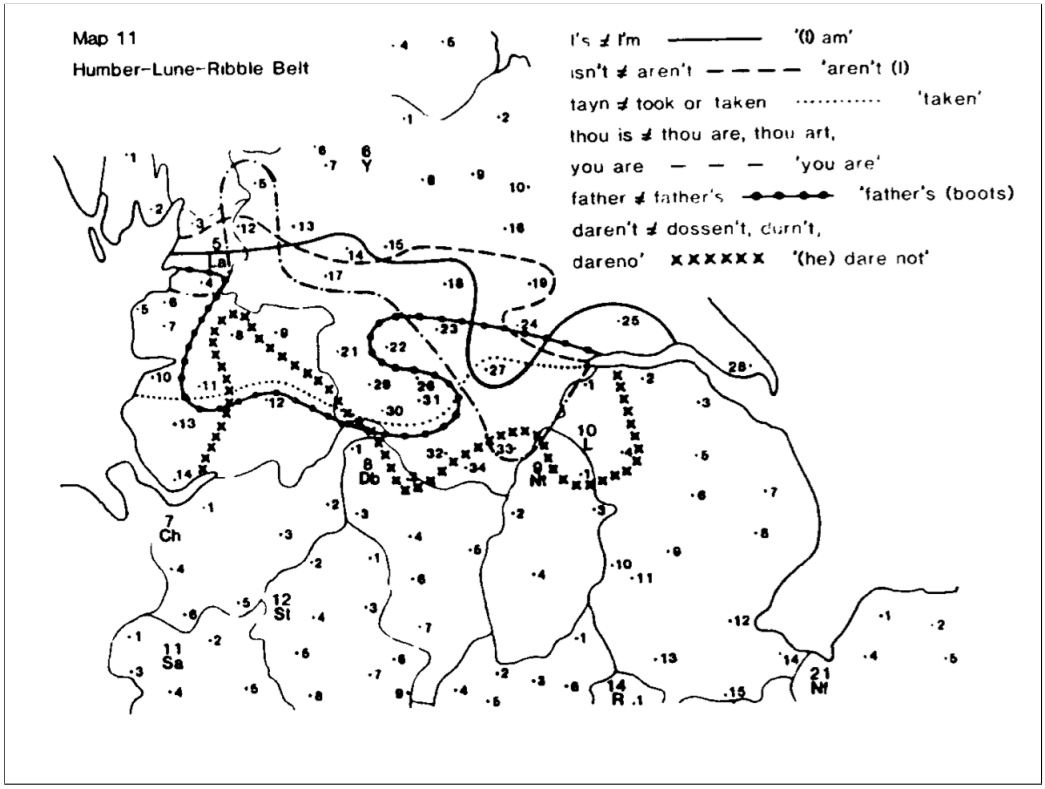

Figure 2.8: Map showing the 'Humber-Lune-Ribble belt'

Source: Viereck, W. (1986). Dialectal speech areas in England: Orton's phonetic and grammatical evidence. *Journal of English Linguistics, 19*(2), 240-257.

2.6.2 FRED and BNC

Gerwin (2013) compares two modern, regionally coded corpora of spoken English - FRED (1970s) and the spoken part of the BNC (1990s) - to generate a picture of the geographical and historically recent distribution of the pDit. This approach is an advance over the previous survey-based approaches. FRED in particular was designed to provide detail of actual dialect usage, specifically targeting dialect grammar (Kortmann, 2004). However, whilst Gerwin's approach may represent an application of the "first empirical foundation for regional analysis" (Gerwin, 2013, p.455), the data are still limited by the amount of data available and the nature of the corpus metadata being worked with. Although FRED offers some geographical precision, in order to gain significant numbers for comparison between the datasets Gerwin uses the broader regional classifications supplied by FRED. For example, the area labelled as 'North' covers the entire northern section of England (see map in figure

2.9).

Figure 2.9: Region boundaries in FRED and the BNCreg

Source: Gerwin, J. (2013). Give it me!: pronominal ditransitives in English dialects. *English Language and Linguistics, 17*(3), p.453.

Whilst the value of the SED data may be partially offset by methodological concerns (see section 2.6.1) and its lack of quantitative detail, its geographical precision and coverage should not be overlooked. And it is clear from the SED maps that the North East patterns very differently from the North West regarding ditransitive use, making the rendering of one broad northern area problematic. With this in mind, Gerwin's (2013) results (reproduced here in figure 2.10) showing a preference for GTD (me it) in 'The North' over TGD (it me) are difficult to interpret. They allow us to quantify that there is still greater TGD use in the North of England as a whole

than the South as a whole, but the result masks the situation found in the SED (and preliminary Twitter data) whereby TGD use in the North West of England is higher than that of GTD, and perhaps more strikingly, that it is non-existent in the North East.

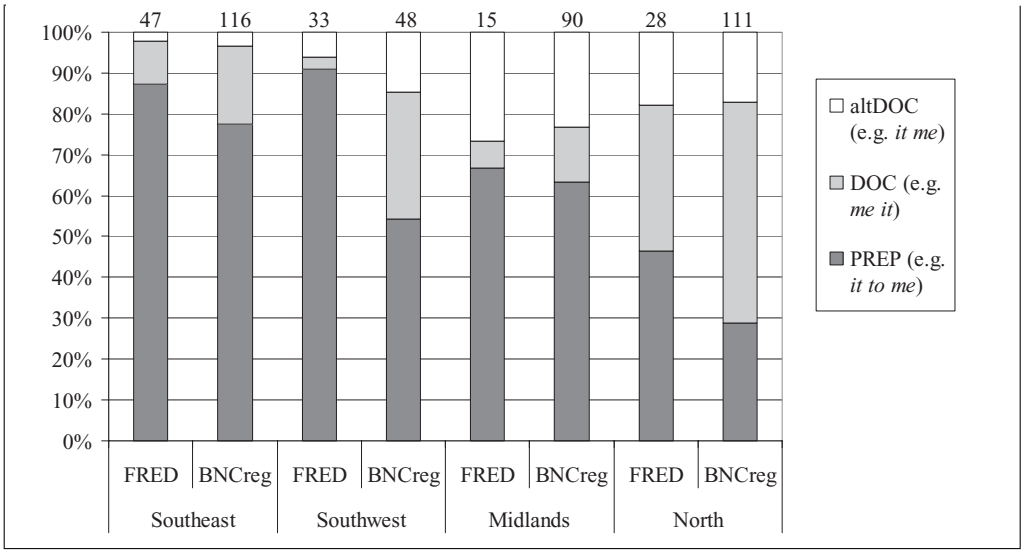

Figure 2.10: "Diachronic comparison of the three pronominal ditransitive constructions" (Gerwin, 2013, p.456)

Source: Gerwin, J. (2013). Give it me!: pronominal ditransitives in English dialects. *English Language and Linguistics, 17*(3), p.456.

Of course, what Gerwin's (2013) results do reinforce is the overall picture that there was a *general increase* in GTD use during the 20th century. They also show a particularly strong rise in GTD use in the south west.

2.6.3 Focus on Lancashire

Siewierska and Hollmann (2007) combine four corpora of spoken English (see appendix) with a focus on the pronominal ditransitive in Lancashire. They find that, counter to what is seen in the standard variety, the TGD is in fact twice as common as the GTD in the Lancashire dialect. This is shown in table 2.3 below.

	count	%
TGD	15	35%
GTD	7	16%
PDAT	20	47%

Table 2.3: Siewierska and Hollmann's (2007) results showing counts and percentages for ditransitives with pronominal objects in their Lancashire dataset.

This finding leads them to the conclusion that:

> "The Lancashire data suggest that even a language-specific double object construction is too simplistic. The form-function mapping in ditransitives in regional dialects should not necessarily be expected to conform to that of the standard variety, and indeed it does not, as is shown most clearly by the theme-recipient variant of the double object construction" (Siewierska & Hollmann, 2007, p.98).

This is in line with the approach to dialect grammar outlined in section 2.2, most notably Adger and Trousdale's (2007) notion of S-language, as revealed by the markedly different relative frequencies in one dialect compared to the standard. The lesson here is essentially that if we are to provide a comprehensive account of the encoding possibilities in a language, we must look further than the standard variety. If this pattern is seen in Lancashire, then what about neighbouring regions?

2.6.4 Manchester dialect project

The Manchester dialect project (MacKenzie et al., 2014) is an ongoing project being conducted at the University of Manchester. It involves successive cohorts of undergraduate students conducting dialect questionnaires based on the SED with friends and family. Responses to asking what people think of the sentence 'give it me' are recorded on a five-point Likert scale ranging from *completely acceptable* to *completely unacceptable*. These are mapped according to the location that the respondent was 'raised in between the ages of 4 and 13'. This map is reproduced in figure 2.11 below.

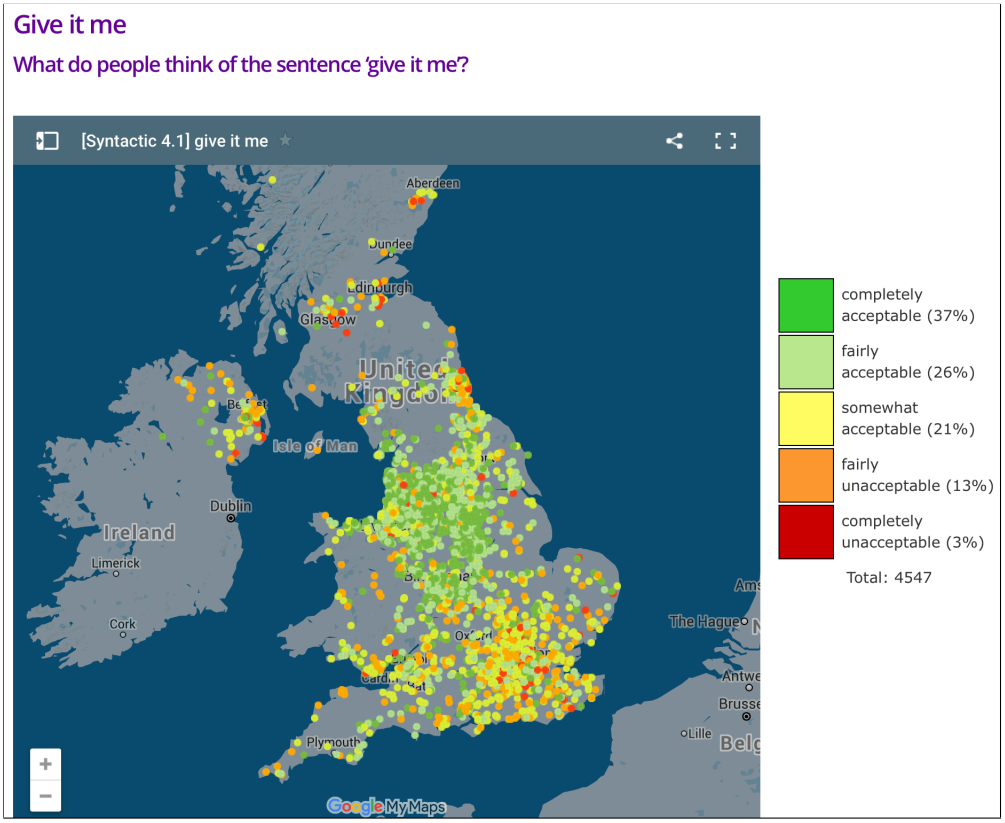

Figure 2.11: Mapped survey conducted by successive undergraduate cohorts at the University of Manchester.

Source: `http://projects.alc.manchester.ac.uk/ukdialectmaps/syntactic-variation/give-it-me/`. [Accessed: 10/12/2016].

As can be seen, the area where the TGD is 'completely acceptable' corresponds to the "give it me" area marked on the SED. This demonstrates the persistence of the feature and suggests that the situation found in Lancashire may extend right across the Midlands, at least as far as its acceptance is concerned. This picture also fits Gerwin's (2013) results shown previously, the 'Midlands' group there showing $\approx 25\%$ TGD in both FRED and BNC.

2.7 Linguistic constraints

2.7.1 Distribution by goal pronoun

As the current investigation is focused on the variation in ditransitives which take two pronominal objects, it is important to consider the extent to which variation is constrained by the choice of pronoun used. As has been established in the literature

(e.g. Yáñez Bouza & Denison, 2015), the vast majority of variation in pronoun use is focused on the GOAL pronoun. The THEME pronoun is, in the vast majority of cases, 'it'.[18]

Consensus in the literature over the extent to which GOAL pronoun choice constrains ditransitive type is not reached. Gerwin (2014) analyses occurrences of RECIPIENT pronouns in FRED and BNCreg. The decision is taken in that study to conflate pronoun categories by number (see figure 2.12 below), primarily on the basis of referential semantics. The reasons given for conflating pronouns are (Gerwin, 2014, p.193):

1. "in most cases it is impossible to determine singular or plural reference for occurrences of the pronoun *you*" .
2. "'me' and 'us' are often used interchangeably with first-person singular reference."
3. "For consistency the third-person category must also be merged."

Gerwin (2014) finds a statistically significant difference between the conflated categories and on that basis reaches the conclusion that the GOAL pronoun does influence ditransitive type, finding that first-person pronouns favour TGD/GTD and third-person ones favour PDAT (see figure 2.12). The conclusion that the pronoun *them* is "especially prone" to favour PDAT "to avoid case ambiguities" (p.196) is based on the fact that it can be "ambiguous as to whether it constitutes a recipient [GOAL] or THEME pronoun" and that "the prepositional construction serves as a disambiguation device here in overtly marking the recipient". This is an interesting conclusion, to which the present study will return in the discussion (section 7.5).

[18]Other options are of course possible and *do* occur with low frequency, but for the purposes of the current investigation, the focus is on the GOAL pronoun.

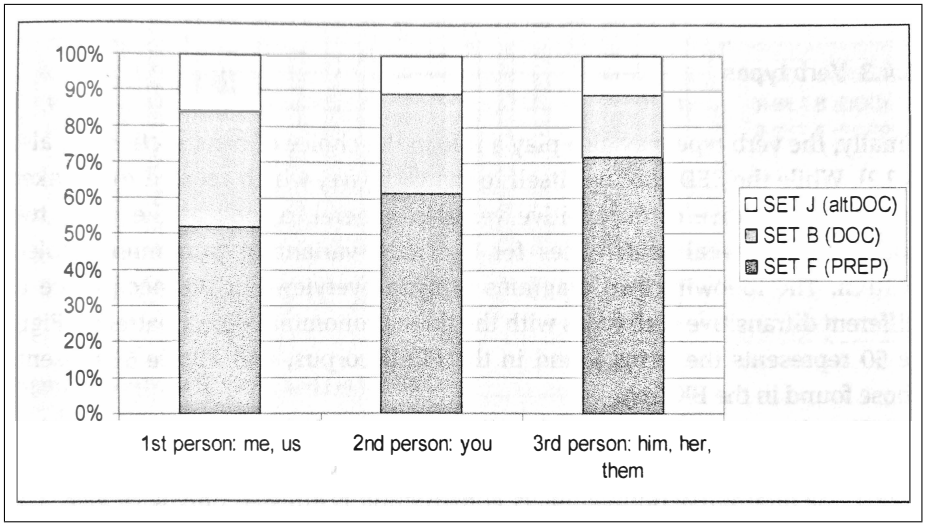

Figure 2.12: GOAL pronouns grouped by person found in combined FRED and BNCreg data.
SET J = TGD, SET B = GTD, SET F = PDAT
Source: Gerwin, J. (2014). Ditransitives in British English dialects (p.195). Berlin: Walter de Gruyter.

Yáñez Bouza and Denison (2015) observe the influence of GOAL object choice on ditransitive type in their 'it-dataset'. This dataset is a subset of all the corpora used in their study (see table A3), only containing instances of the ditransitive verbs with *it* as THEME object and pronoun as GOAL object. As already discussed, whilst combining all the corpora was necessary to get sufficient data to view trends, using a dataset that spans such a long time period (1410-2011) is clearly problematic. Any conclusion based on these data must therefore be tentative. Nonetheless, Yáñez Bouza and Denison (2015) and Gerwin's investigations represent the only substantial data on the issue.

2.7.2 Distribution by verb

The effect of verb on ditransitive choice has been much discussed. It is established that some verbs exclusively take either PDAT or GTD (Levin, 1993) and that those verbs also take TGD (Haddican, 2010).[19] For the purposes of the current investigation, which focuses its search on *give* and *send*, it is enough to note that regarding pDit constructions, *give* and *send* are considered as belonging to the set of alternat-

[19]A number of verbs have been identified as 'alternating', and are reported by Levin (1993). These include among others GIVE VERBS: *feed, give, lease, lend, loan, pay, etc.* and SEND VERBS: *send, forward, hand, mail, ship, etc.*.

ing verbs, and together constitute by far the most commonly occurring of that set, making up about half of the verbs used across Yáñez Bouza and Denison's (2015) cross-corpus study. Additionally, both *send* and *give* are generally considered to pattern similarly with the pDit types they encode. For example, Siewierska and Hollmann (2007) find that between *give*, *send* and *show*, there is no effect of the verb on choice of ditransitive, in their Lancashire data.

Tagliamonte (2014a, p.309) does, however, find a difference in her Canadian dataset, finding that "Verbs other than *give* appear more frequently with the PD across the board". Meanwhile, Gerwin (2014, p.110) asks "why *give*, of all ditransitive verbs, displays a decrease in the double object construction even though the double object construction is on the increase for most other ditransitive verbs". This considered, it will be worth investigating whether there is any difference between *send* and *give* and whether there is regional variation in that difference.

As discussed in section 2.5.2, there was apparently a distinction between *give* and *send* in Old English, the former not being able to take the PDAT. The spreading to the situation found in PDE, where both verbs are considered alternating may have happened at different rates in different places, *if* such spreading happened as a result of pressure from contact with Norman French.

2.8 Computer-mediated communication (CMC)

> "Over the last century, developments in telecommunications have made possible new communicative modalities that blend the presuppositions of spoken and written language." (Baron, 1998a, 134)

Since its inception, scholars have noted that some forms of written CMC, particularly where interaction is synchronous, or near-synchronous, display certain properties and characteristics ordinarily associated with speech (cf. Werry & Herring, 1996). As Yates (1996, p.46) reports, "the *mode* of CMC as a communications medium is neither simply speech-like nor simply written-like. Though CMC bears similarities in its textual aspects to written discourse, it differs greatly in others, *namely pronoun and modal auxiliary use*"(italics not in original).

The shared space in which "interactive written discourse" (Werry & Herring,

1996, p.47) occurs generates a kind of virtual reality in which an imagined conversation is enacted. In this space, virtual objects referenced in previous 'messages' or 'posts' enter into the shared consciousness of the interactants and can thus be referred to using pronominals. This can be seen in the extract from a conversation that took place on Internet Relay Chat (IRC), an early form of CMC (figure 2.13 below). The pronoun '1' in the turn by 'torex', "Hodgy i got 1 u will like" refers to the referring expression "this pic" in the previous turn by 'Hodgy'.

Funchat.log:
* torex is in england
<^^Sun^^> <---Tennessee, USA
<^^Sun^^> and you tinas?
<tinas> Germany
* ^^Sun^^ pokes Hodgy!.....Still with us??
<Hodgy> yea!
<Hodgy> lol
<^^Sun^^> your so quiet
<Hodgy> was checkin out this pic a girlsent me
<Hodgy> lol
<^^Sun^^> uh huh! i see
* Hodgy is a sucker for girls
<^^Sun^^> lol
<Hodgy> lol
<torex> Hodgy i got 1 u will like
<Hodgy> <---- been single too long
<Hodgy> lolololol

Figure 2.13: Extract from IRC chat, circa 1999.

Also clearly speech-like in this example from IRC (figure 2.13) are the use of short turns and supportive non-clausal 'utterances' - e.g. "and you tinas", "Still with us??". Of course simulated laughter with the use of "lol", now ubiquitous, plays a important communicative function, used much as it would in face-to-face speech, in this way re-supplying social cues which are absent owing to the lack of physical co-presence.

The CMC literature has tended to focus on the ways that the new technology itself is influencing language use (Squires, 2016). More recently, however, attention has turned towards how language use online reflects language use 'in real life' (e.g. Eisenstein, 2013). In early CMC, users typically did not know each other 'in real

life' and much was made of users' freedom to generate any identity they chose, free from the constraints of their physical presence (Donath, 1999). The situation today has evolved: users often know each other personally and digitally-based interaction in this case serves to augment pre-existing social relationships (Shortis, 2016).[20] As was observed early on in the CMC literature, "many ordinary individuals possess a compensatory 'literary' capability to project their personality into writing destined for the computer screen." (Feenberg, 1989, p.1), cited by Baron (1998b). This idea is important: 'ordinary individuals' are enfranchised to produce micro-literary outputs, through which they will attempt to project their identity. If a speaker, then, says "give it me" in their speech, they are likely to do the same in conversational messaging.

2.8.1 Twitter for dialect study

Twitter was originally setup as a 'microblogging' platform, with the purpose of broadcasting messages to an internet-wide audience, but its re-appropriation for 'conversation and collaboration' was swift (Honeycutt & Herring, 2009, p.1). A proportional increase of 'tweet as conversation' over 'tweet as broadcast' is shown by a doubling to around 30% conversational tweets from the 12.5% found previously by Java, Song, Finin, and Tseng (2009).

Page (2012a) similarly identifies different forms of tweet, distinguishing communicative intention in types of Twitter message. Public facing, one-to-many messages add to a public dialogue but are not characterised by interaction between interlocutors in the way that conversational one-one/one-few messages are. In these kinds of interaction, message authors have usually only one or two recipients in mind, resulting in conversational practice more akin to face-to-face interaction. Conversational Twitter messages such as these (see figures 6.1 and 6.2 in section 6.2) are much closer to other forms of digitally mediated communication.

Whilst communication in Twitter is not strictly synchronous, it *is* fast-paced (Page, 2012b) enough to create the sense of a conversation in process (p.183). There is an expectation of fast response. As Honeycutt and Herring (2009, p.7) report, "most conversations that occur in Twitter appear to be dyadic exchanges of three

[20] Also including SMS text messaging.

to five messages sent over a period of 15 to 30 minutes".

Although the degree to which users on Twitter know each other in real life is unclear when Twitter is used for conversation, with the inclusion of profile pictures and locations there is the fostering of an environment where real identity is reinforced rather than reinvented. Instead of using language to construct new identities, it appears that users choose to draw on the regionally-specific linguistic repertoires they already use in their everyday speech practice in the imagined dialogue. This is evidenced by recent studies that have shown how writing in social media "displays influence from structural properties of the phonological system" (Eisenstein, 2013, p.1), a finding which has been leveraged to map social media phonetic respellings to patterns of migration in the United States (Jones, 2015). Gabriel Doyle has shown that syntactic features acquired using the Twitter API correspond to data gathered by traditional methodology for features like "*needs done*" (Doyle, 2014). David Willis in his work constructing an atlas of Welsh syntax (Willis, 2013) explains how he uses Twitter as a diagnostic tool to gauge where structures are being used - "as people tweet much as they speak" (District, 2013). Similarly, in the pilot study for the current project, TGD use on Twitter lines up with maps drawn from the SED (see figure 4.1).

It is worth considering here that, particularly when sending and receiving messages on a handheld device, there may be little perceived difference between different social media platforms. To the user, an incoming Twitter message, it is conjectured here, is likely not interpreted very differently from an incoming SMS or other message (e.g. Facebook). In this way, the author of a message has in mind only the recipient as audience rather than the internet as a whole, and designs their linguistic output accordingly. It is, of course, well established that the audience that an author believes they are engaging with influences language choice (cf. Bell, 1984). Authors of social media texts in this context are engaged in the fabrication of a shared interactive space that mimics face-to-face communication, and whilst participants may be physically distant they are "in imagined close social proximity" (Shortis, 2015, p.489). This goes beyond the idea of social media texts being loosely defined as 'non-standard'.

The cumulative result of this kind of interaction is a massive and expanding body

of mappable, natural, unmonitored, speech-like data, onto which we have a window via Twitter's public APIs, as discussed in section 5.3. And as Doyle (2014) concludes, such results, when compared to data gathered using traditional methodologies, are "tightly correlated with existing gold-standard studies at a fraction of the time, cost, and effort" (Doyle, 2014, p.98).

2.9 Summary

The literature review presented here reveals the high level of interest among scholars in the ditransitive construction. Research on dialect grammar is an emergent field, and a sensitivity to regional variation in syntax offers an important counterpoint to some of the more general approaches to grammar, at the level of *language* often taken in theoretical linguistics.

Running through the literature is a consciousness of the distinction between speech and writing. This is of particular relevance to the pDit, which is shown to be a feature primarily of spoken English that is little used in writing. This confinement of the pDit to spoken rather than written English, combined with the relative scarcity of syntactic variants in general, makes the size of corpora needed to capture distribution patterns prohibitively large. This 'data problem' may find resolution through the use of Twitter data due to their proximity to speech data, combined with the sheer volume of the data available.

The historical distribution of the pDit has been reported by several recent studies which use corpus approaches to chart the changing use of pDit variants over time. These studies challenge the canonical view that the pronominal TGD is a minority dialectal feature of peripheral importance, and highlight the fact that until the 19th century the TGD was the favoured variant. Additionally, the TGD is shown to still be favoured in present day English dialects.

Linguistic constraints and the effect of *verb* and GOAL pronoun choice on pDit type have been discussed in the literature, but conclusions can only be tentative. The data problem surfaces again here, when subsetting the dataset by internal constraints. Any possibility of regional variation in the distinction between *sent* and *gave* is not captured, as datasets necessarily span across regions. It may be that the

nature of data gathering from Twitter would provide enough data to get a clearer idea of how each ditransitive variant patterns across *verb* and *pronoun*.

Chapter 3

Research Questions

In the light of the aims set out in section 1.4, and the material covered in the literature review, the following research questions are put forward. Aim 2, to evidence the speech-like nature of Twitter, should be satisfied by virtue of its applicability to the problem at hand.

1. **What is the geographical spread of the pDit, and how do its variants (TGD, GTD and PDAT) pattern relative to each other by region?**
2. **How do the Twitter data relate to historical and contemporary corpora?**
3. **Is there a difference between *give* and *send* in choice of pDit type and is any difference regionally distinct?**
4. **What is the effect of pronoun choice on ditransitive use?**

Chapter 4

Pilot Study

4.1 Introduction

As explained in the introduction, the pilot study (Stevenson, 2015) was composed of two parts - (1) a corpus study of Twitter messages and (2) a grammaticality judgement survey. The results of this pilot study are sketched briefly here.

4.2 Twitter messages

Figure 4.1 on page 56 shows Tweets retrieved for the pilot containing the TGD structure (sent|gave **it** me|him|her|them|us) overlaid onto Kirk's (1985) rendering of SED data. As can be seen here, there is a remarkable correlation with the isoglossic area 'give it me' on the SED map, showing the resilience of the structure over time and providing some validation of the use of Twitter as a data source.

The breakdown of each pDit type by area based on the data from the pilot is present in figure 4.2. This shows clear regional preferences for each pDit type. The methodology used to gather the data for the pilot is replicated in the present study, with an important difference: the pilot data was located using the far more plentiful, but potentially error-prone, *user-inputted* data, whilst the present study uses only GPS data. The GPS data is more accurate, but much less frequently available. This is discussed in more detail in the methodology (chapter 5). What is useful about having these two sets of data is that they can be statistically correlated to test the extent to which user-inputted data matches GPS data. The results of this

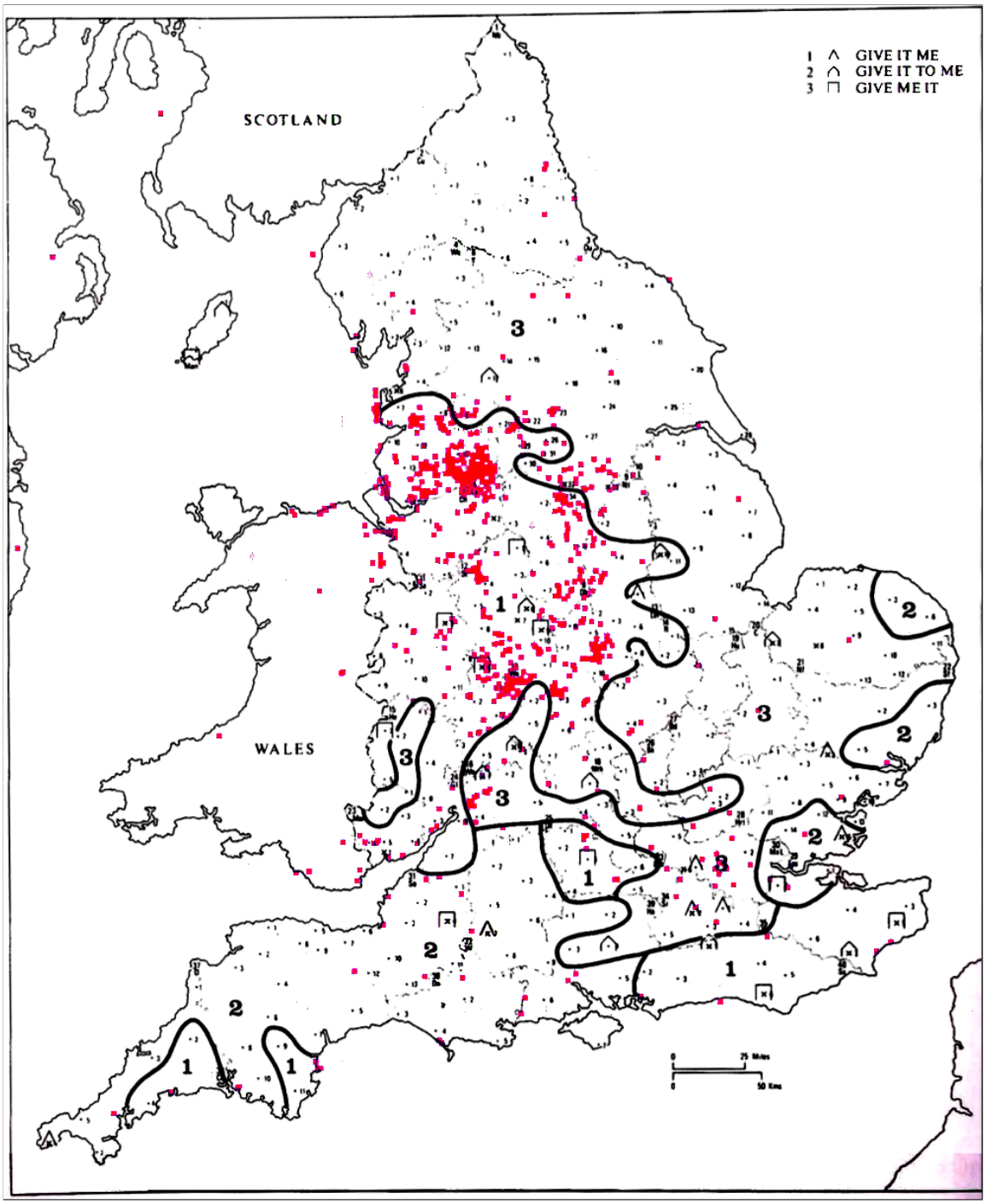

Figure 4.1: Tweets containing TGD structure overlaid on Kirk's 1985 rendering of SED data.

correlation are presented in section 6.3.

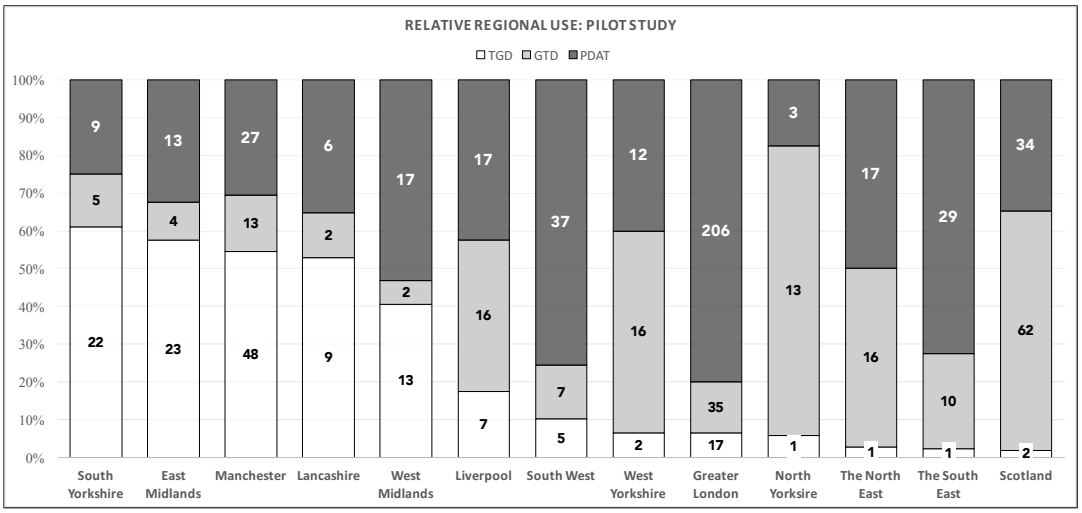

Figure 4.2: Relative use of ditransitive by region as found in pilot study.

4.3 Survey

The survey involved, amongst other things, a set of online acceptability judgement tasks with a range of test sentences based on the survey used by Haddican (2010).[1] The survey was sent to various university departments around the country and was forwarded on to undergraduate students. In total, there were around 140 completed surveys. Of relevance to the current investigation is the regional distribution of responses to the sentence containing the TGD: "Its a scanner/Printer thing. Someone gave it me but..". The pattern revealed here corroborates what is found in the Twitter results, showing East Midlands (EM), Greater Manchester (GM), South Yorkshire (SY) and West Midlands (WM) as areas of high mean TGD acceptance. Meanwhile East (E), North East (NE), North Yorkshire (NY) and The South East (SE) show a low mean acceptance. West Yorkshire also displays a high acceptance, but a notably lower one than neighbouring regions. This is interesting, as West Yorkshire lies geographically in between where the TGD is largely accepted and where it is not, and as such is a likely *transition zone*.

[1]The pilot survey is still live and available at `https://eSurv.org?s=LIHHJF_29bbbe0c`.

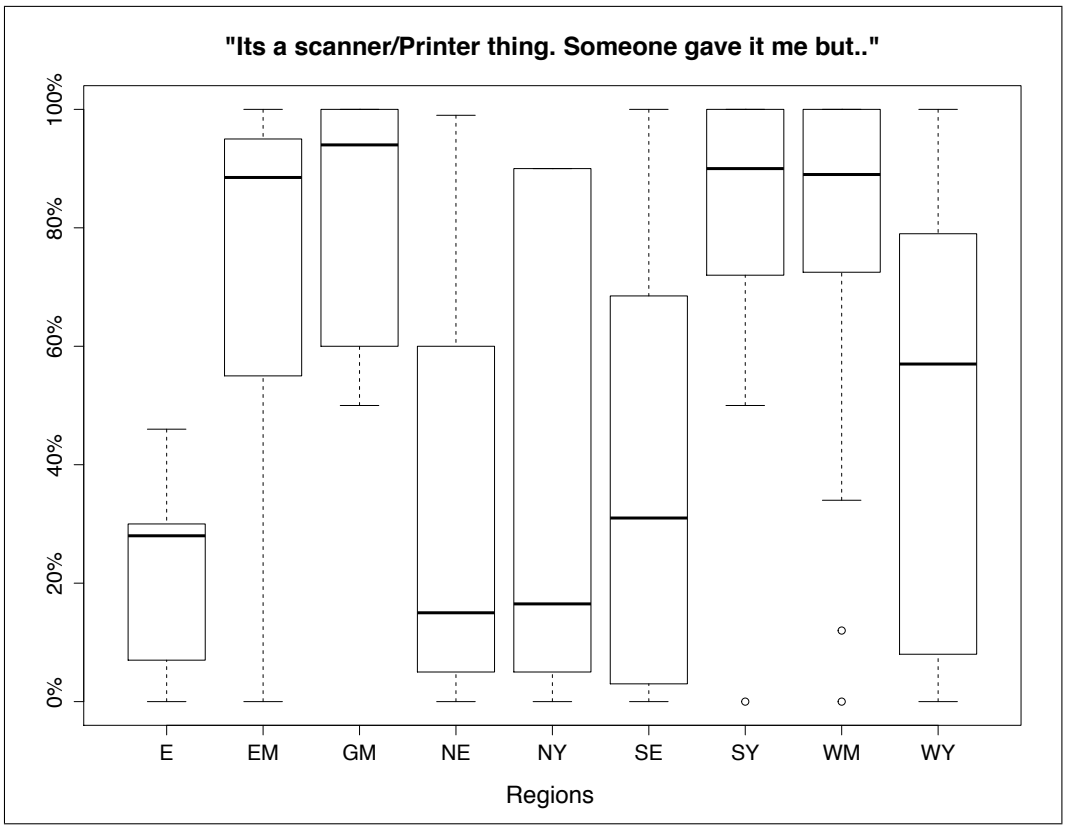

Figure 4.3: Box-plot displaying mean acceptance of the sentence "Its a scanner/Printer thing. Someone gave it me but.." by region, from (Stevenson, 2015).

E='East', EM='East Midlands', GM='Greater Manchester', NE='North East', SE='South East', SY='South Yorkshire', WM='West Midlands', WY='West Yorkshire'.

Chapter 5

Methodology

5.1 Introduction

The principal aim of the present investigation is, using Twitter, to gain enough data to be able to show the relative frequencies of each pronominal ditransitive (pDit) type at a geographically local level. This, it is anticipated, will provide additional insight into historical trends and data on the effect of internal constraints, and will validate Twitter as a viable data source. The methodology replicates that used in the pilot study.

5.2 Twitter as corpus

Whilst the 'data problem' discussed in section 2.3 has meant that researchers investigating dialect grammar have often needed to pool together disparate datasets, using Twitter as a corpus offers to supply the quantities of data needed whilst maintaining geographical distinction. There are, however, some clear limitations: the sample population is not representative of the general population, favouring young, urban speakers of "the interior classes" (Jones, 2015, p.408) and there is a paucity of metadata available on individual users. Gender can be inferred only on a case-by-case basis from usernames and profile pictures, but social class, age, and occupation data are not available. As Eisenstein (2017, p.2) explains, a "quantitative analysis of Twitter text ... can describe only a particular demographic segment within any geographical area". With this acknowledged, it is worth noting that selective

sampling in areal linguistics has traditionally been a particularly challenging task. Practical considerations mean that dialectologists have been able at best to seek to gain "a general view of a complicated situation in a reasonable time" (Kurath, 1973, p.1). Additionally, as Doyle (2014, p.105) points out, having a sample that is biased towards a more youthful demographic (see figure 5.1 below) may actually be beneficial when looking for signs of linguistic innovation or persistence.

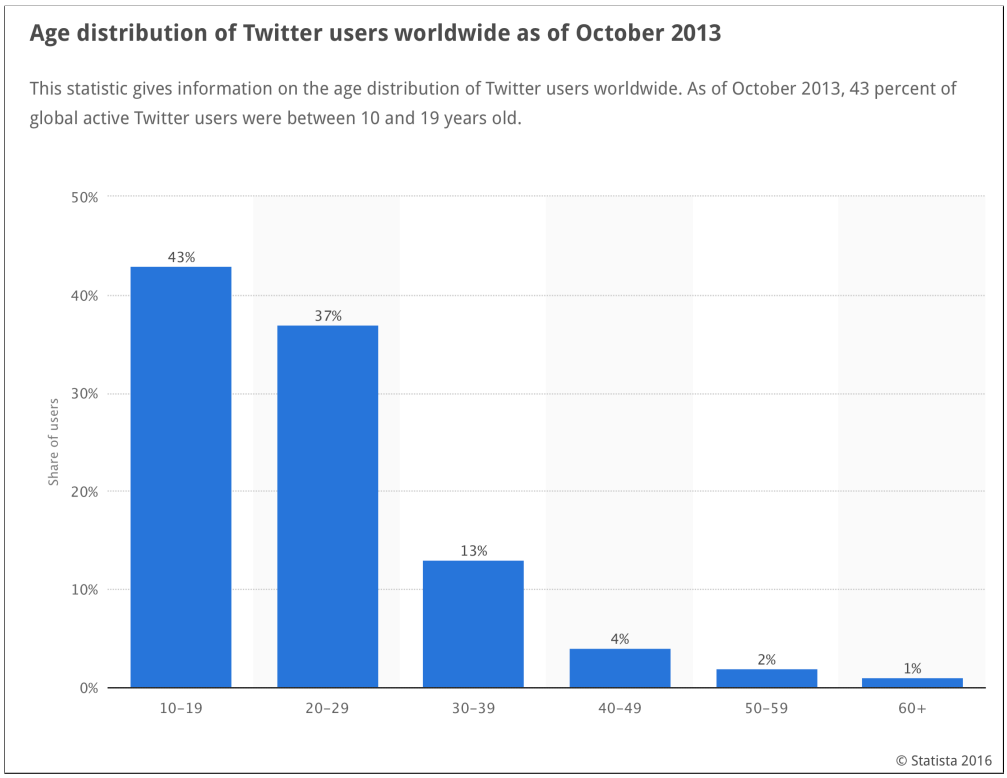

Figure 5.1: Age distribution of Twitter users worldwide (Statista, 2013).

Source: `https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/`. [Accessed: 10/12/16].

5.3 Twitter APIs and subsetting of data

Access to Twitter data is provided through three APIs (Application Program Interfaces). These APIs are essentially a set of tools that can be accessed and activated by user-created programs. Each API grants different levels of access to the Twitter data. Two APIs (Search and Stream) are publicly accessible with only a developer Twitter account needed to gain access. Anyone is free to create a developer

account, but the data accessible are restricted to about one percent of the total moving through Twitter's servers. A third API (Firehose) allowing access to the entire data stream is available at a cost, and is managed via Twitter's monetised commercial front end 'GNIP'.

The Stream API provides a continuous stream of data, whereas the Search API provides access to the past seven days. Additionally, the Search API biases results towards 'relevance' over 'completeness'. This has been shown to manifest as a bias towards more 'central' Twitter users who have a greater number of followers and 'mentions', that is, users who are more linked with other users (González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2012).[1] Whilst the Stream API offers a more representative slice of the total Twitter sample, storing the returned data requires a machine to be left continually 'on' and connected.

For the current study, the Google Spreadsheet interface 'TAGS' (Hawksey, 2014) (discussed in section 5.5) was used, and this interface uses the Search API. The main advantage of using TAGS/Search API combination is that doing so allows the results of a given search to be returned to a Google Spreadsheet which can be left running for months at a time, without any user maintenance or the need for a local machine to be left powered on.

There is a risk that using one of the free-to-use APIs will not return enough results in a given time period to allow for sufficient subsetting of the data to, for example, see the distribution of ditransitive variants by region *and* verb/pronoun. What the results will show, however, is whether it is worth paying for the full access provided by the Firehose API. Additionally, Firehose access would enable faster results, comparison across time, and access to the full range of user types.

5.4 Geolocation and Twitter's changing rules

Both Search and Stream APIs allow the user to request that only Twitter messages occurring within a given range of a geographical location be returned. It is possi-

[1] The degree to which the biased sample returned by the Search API might affect linguistic behaviour is unclear. For example, it may be the case that users with more followers are more likely to be more style conscious in their linguistic output, and, perhaps as a result more normative, though any effect overall is likely to be marginal.

ble to specify, for example, to return all messages that occur within 300 miles of Nottingham, and thereby cover most of the mainland UK. Two sources of location metadata are potentially attributed to a given tweet and available in the results returned by the API. One source is user-inputted data. This is the location that a Twitter user enters into their profile when they set up an account and it can be changed at will (it is not clear how regularly a user might update their location). The other source is that provided by a phone's GPS chip. Access to the GPS chip has to be granted by the user.[2] Only a small minority of Twitter messages contain GPS metadata. Comparing the number of messages retrieved for the pilot, in which user-located tweets were used, to the number for the current study, we see a stark contrast. Getting a comparably sized dataset took 15 days using user-inputted location data rather than 14 months using geolocation data.[3]

At the time the pilot study was conducted, the Search API could use user-inputted location alongside GPS data and include relevant results within the geographically-bounded search query (by a process known as *reverse geocoding*). This resulted in a relatively large dataset combining mostly reverse-geocoded tweets with a small number of GPS-encoded tweets. However, the Search API suddenly stopped reverse geocoding results in November 2014, *only* returning results with GPS data within the defined area. For the current study, then, only GPS-encoded Tweets are used. This meant it took several times longer to gather a comparable amount of data. Fortunately, using TAGS and Google's always-on web apps, the search could be left to repeat at given intervals indefinitely. In this way, 14 months' worth of data were gathered, providing enough for the initial purposes of the current study.[4]

[2]The way this is managed has gone through some changes. Originally, it was a global setting - once a user had enabled access to the GPS on their phone, it would remain activated and would globally encode all subsequent tweets. The current policy requires the user to individually tag each tweet with GPS information, similar to Facebook's 'checking in' feature, which lets users share their location at a particular venue. The new policy is, however, implemented in the app installed on the user's phone and not on Twitter's server. This means that anyone using the older version of the app (that is, who has not yet updated) still globally broadcasts the GPS data.

[3]The proportion of GPS to user-located tweets is now even smaller due to the changes in the way the user supplies GPS data.

[4]As will be explained in section 6.1, actually, the limitation on data gathered as a result of the changed rules here does prevent subsetting the data to the extent that would be desirable.

More recently (as of 12.04.16) functionality has been restored to the Search API, enabling it to use user-inputted location. This does not affect the data gathered for the current study, which was gathered in the intervening period when only GPS tweets got through. However, the search query is still running in Google Sheets, and at the time of writing, returning around five tweets a day containing TGD, compared to one or two a week when the functionality was removed. Of course, this is returning a much larger dataset, and it may be worth comparing the results of this new data in a later study. Additionally, the user-inputted data and the GPS-only data show a strong correlation (see section 6.3). This is a good result, which validates the future use of only user-inputted data or a combination of user-inputted and GPS data.

5.5 TAGS

The TAGS web-app (Hawksey, 2014), based in 'Google Sheets' (Google's online spreadsheet software) is freely available online.[5] Once the user has set up appropriate accounts,[6] they can enter the search terms they are looking for in the search field (see figure 5.2 below). Individual words and strings placed in double inverted commas can be searched for, and multiple terms can be included by separating each with the Boolean operator 'OR'. Additionally, the location can be specified as an area defined by the radial distance to a given point.

[5] `https://tags.hawksey.info/get-tags/`.

[6] Twitter developer account, linked to Google account; instructions are provided on the website.

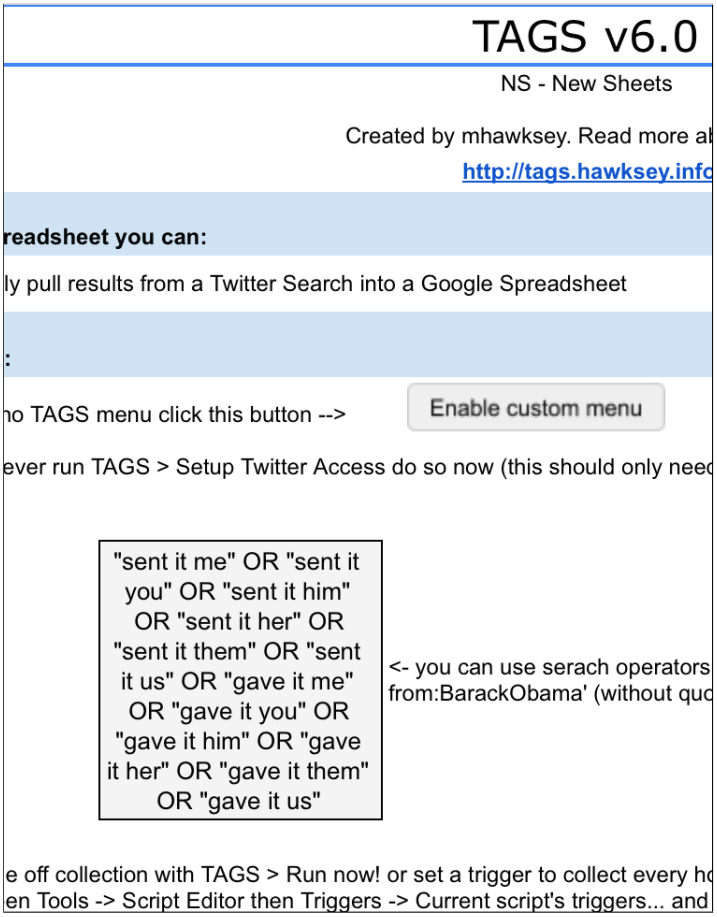

Figure 5.2: Example search string used on TAGS web-application.

The search function interfaces with Twitter's Search API (see section 5.3), which returns any Tweet from the preceding seven days that contains the text strings specified in the initial search field, along with accompanying metadata. This output is used to populate a separate Google Spreadsheet containing a row for each Tweet and columns for accompanying metadata. An example TAGS sheet can be seen in figure 5.3. Note the high occurrence of duplicates. This is a quirk of the way Twitter operates, but duplicates can be easily removed.

5.6 Search terms and structures to exclude

The focus for the current study was on two verbs, GIVE and SEND. These are the most common two verbs occurring with "it+pronoun" found in Yáñez Bouza and Denison's (2015) study. There, SEND and GIVE together comprise 49% of all

	A	B	C	D	E	F	G	H
1	id_str	from_user	text	created_at	time	user_location	geo_coordinates	user_lang
2	814875	[redacted]	Quality game of darts. Well done @GaryAnderson180 pleasure to watch. You gave it you're all @ChizzyChisnall and was involved in a CRACKER 🎯	Fri Dec 30 16:48:57 +	30/12/2016 16:48:57	Wigan, England		en
3	814667	[redacted]	@MuffaFekkinD or you can just gave it me	Fri Dec 30 02:59:51 +	30/12/2016 02:59:51	England, United Kingdom		en
4	814610	[redacted]	@Burforders female cat wasn't bothered the other one started doing mad arched sidewalks....never gave it him again	Thu Dec 29 23:14:00	29/12/2016 23:14:00	Sheffield		en-gb
5	814608	[redacted]	@abbeyrhurst you've gave it me you slag	Thu Dec 29 23:05:22	29/12/2016 23:05:22	Coalville		en
6	814608	[redacted]	@abbeyrhurst you've gave it me you slag	Thu Dec 29 23:05:22	29/12/2016 23:05:22	Coalville		en
7	814587	[redacted]	@alycxne *** replied that when I sent it them and it's ********	Thu Dec 29 21:40:50	29/12/2016 21:40:50	Manchester, England		en
8	814574	[redacted]	Can't believe it's my sisters birthday on Saturday and I haven't got her anything yet 😭 (I bought her stuff and gave it her for Christmas y)	Thu Dec 29 20:53:05	29/12/2016 20:53:05	United Kingdom		en
9	814509	[redacted]	@eandrewsx you obvs gave it him, you think he's peng, admit it I won't laugh x	Thu Dec 29 16:34:51	29/12/2016 16:34:51	East Midlands, England		en-GB
10	814491	[redacted]	@wildconrad It's utterly disgusting - my sister gave it me 😳	Thu Dec 29 15:21:05	29/12/2016 15:21:05	Bury St Edmunds, England		en
11	814466	[redacted]	My grandad brought this then gave it me to get me into football at 3 years old https://t.co/dyEO53WcyS	Thu Dec 29 13:41:06	29/12/2016 13:41:06	Birmingham		en
12	814443	[redacted]	@MancJosh mate sent it me 😂	Thu Dec 29 12:12:17	29/12/2016 12:12:17	Manchester, England		en
13	814291	[redacted]	From "Singing in the Rain" to Singing in the Grave, Debbie Reynolds always gave it her all. Rest in Peace https://t.co/LvwLkNbHRm	Thu Dec 29 02:06:43	29/12/2016 02:06:43	Cuckfield, England		en
14	814278	[redacted]	RT @annalouiseadams: This guy gave me his old iPod touch for £30 but when I went to give him the £30 he gave it me for free 😔 https://t.co/…	Thu Dec 29 01:13:22	29/12/2016 01:13:22			en
15	814276	[redacted]	This guy gave me his old iPod touch for £30 but when I went to give him the £30 he gave it me for free 😔 https://t.co/QQNHqe5fLB	Thu Dec 29 01:05:51	29/12/2016 01:05:51	London, England		en
16	814276	[redacted]	This guy gave me his old iPod touch for £30 but when I went to give him the £30 he gave it me for free 😔 https://t.co/QQNHqe5fLB	Thu Dec 29 01:05:51	29/12/2016 01:05:51	London, England		en
17	814238	[redacted]	RT @Becky_1595: Love how @sadiesummersx bought a bra that didn't fit her and gave it me speed bump tits 😂	Wed Dec 28 22:37:46	28/12/2016 22:37:46	manchester		en
18	814236	[redacted]	Love how @sadiesummersx bought a bra that didn't fit her and gave it me speed bump tits 😂	Wed Dec 28 22:27:11	28/12/2016 22:27:11	Manchester, England		en
19	814236	[redacted]	Love how @sadiesummersx bought a bra that didn't fit her and gave it me speed bump tits 😂	Wed Dec 28 22:27:11	28/12/2016 22:27:11	Manchester, England		en
20	814188	[redacted]	@SharleneBoothe @94JBieber21 @RealBieber22 yeah I gave it him he's my little teddy bear why ?	Wed Dec 28 19:15:29	28/12/2016 19:15:29	England		en
21	814154	[redacted]	Eh, named three characters from Nevada and they gave it me	Wed Dec 28 17:00:41	28/12/2016 17:00:41	Nottingham, England		en

Figure 5.3: Example of TAGS output sheet

occurrences from their "it-dataset".[7] Further, the decision was taken for the current study to search for instances of the verbs as they occurred in their past-tense forms *gave* and *sent*. Using the past tense has a few advantages: it avoids common phrases that might skew the results (e.g. 4a);[8] it maintains a more consistent aspect (e.g. 4b); and the verb morphology is unchanged by subject agreement (e.g. 4c), resulting in consistent phonology at the word boundary.

(4) Structures avoided by using past-tense forms of the verb

a. give it to me

b. I will give it you (future)/he gives it you (present)

c. I give you it/he gives you it

Additionally, there is an issue with sentences involving second-person pronoun 'you' occupying the same place as the object-GOAL position, but actually representing the start of a new clause. Examples include expressions such as e.g. "I sent it you fool" and "if you've sent it you can't change it". Such cases were manually removed. Of the 49 instances of 'sent it you', only two false positives of this kind were found.

The expression "give it one's all" was problematic when expressed with the female third-person singular 'her', which is orthographically identical to its genitive form. So instances of "gave it her all" were searched and removed from the dataset.

Finally, in the search terms, focus was on strings using only the full, standard orthographic convention, for example 'you' and not 'u' and 'them' and not 'em'. The main reason for doing this was simply that the search field on TAGS is limited to a maximum number of characters. A way around this might be to set up additional TAGS searches, run them concurrently and consolidate the resulting outputs. It is hard to say how many more tweets this would capture, but given the high incidence of shortenings on social media (particularly regarding 'u'), it would likely be significant.

The final search strings are displayed in table A4 in the Appendix.

[7]The "it-dataset" is the term Yáñez Bouza and Denison (2015) use to refer to a restricted subset of the twelve corpora they used (see their table A3) allowing only "it+pronoun".

[8]Example 4a returned a high frequency of direct quotes of song lyrics, for example "give it to me baby".

5.7 Defining geographical regions

Defining regions for the purpose of dialectological research is problematic. The concepts of space and place are critical to an understanding of language and behaviour (Britain, 2013). However, whilst it is acknowledged here that there is a strong case to be made for problematising traditional notions of space, the approach taken in this study is to use politically defined regions as a starting point.

Place names, addresses and postcodes were automatically generated from the GPS coordinates encoded into the tweets *reverse geocoding*. This can be done using a number of online services; the service chosen for the current project was *Maplarge.com*.[9] Coding the tweets by region was done semi-automatically, with a certain degree of manual intervention required. Manual intervention involved systematically looking up place names and finding the political region to which they belong.[10] A map of the regions as defined here can be seen in figure 5.4.

[9] `http://maplarge.com/reversegeocoding`.

[10] This proved the quickest way with the current dataset, but there are likely to be ways to fully automate this process with a larger dataset.

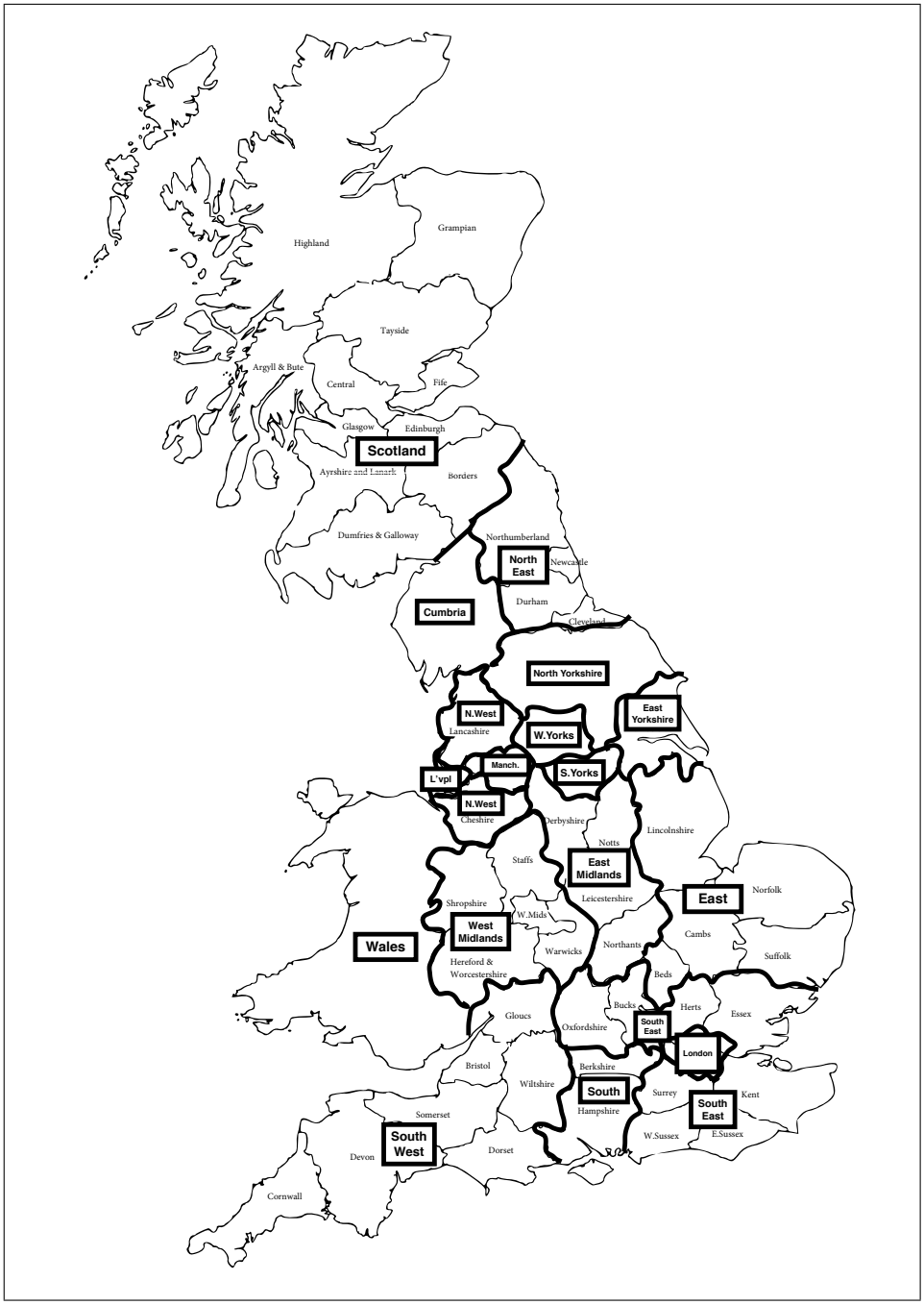

Figure 5.4: Map of regions used.

Source: Adapted from `http://listofmaps.com/map-regions-of-england/`. [Accessed: 28/12/2016].

Once data for political regions have been gathered, each region can be compared statistically to all other regions. The results of the pilot study appeared to show that certain conjoined regions displayed not only similar levels of one variant, but that *all three* variants were proportionally similar. Further to this, there also appears to be

a core set of regions that display similarly high levels of TGD usage, with peripheral regions displaying progressively less usage as one moves further away from the core. Having grouped statistically similar conjoined regions into larger 'super-regions', each super-region can then be compared to every other.

5.8 Batch Geo

The free version of BatchGeo (BatchGeo, 2012) was used to place Twitter messages onto a Google map, organised by *ditransitive type*. The resulting output (see figure 6.6 in the results, section 6.3) represents each Twitter message with a single point on the map. The points are dynamically grouped together into 'pie-chart' representations of the wider area and separated back out to single points as the user interacts with the map to zoom out or in. When processing the data on the *BatchGeo* website, there are a number of configurations the map creator can apply. Each point can be configured to display a range of data in a *context window* when clicked on by the end-user. This data could include the link provided by the raw data pulled from Twitter back to the original tweet in context on the user's Twitter page. In the interests of safeguarding user anonymity, this link was not included. The level of 'zoom' was also limited so that the end-user could not see the exact location of a given message. Data finally included in the context window are displayed in table 5.1 below.

Category	Description
text	The anonymised Twitter message
gender	Gender of user when available
verb	The main verb (*SENT* or *GIVE*)
pronoun_2	GOAL pronoun
type	Ditransitive variant (*TGD*/*GTD*/*PDAT*)

Table 5.1: Data selected to be provided in the *context window* when a point is clicked on map.

The map provided by Batch Geo gives an immediate sense of the distribution of the three variants across the UK. As such, it provides a starting point for breaking down the data and functions as a useful tool for the researcher to see potential areas of interest. Its transparency also provides an accessible 'way-in' to the data for

people new to the study.

5.9 Summary

The methodology outlined here describes the process used to gather data from Twitter for the purposes of analysing the geographical distribution of the pDit.

Twitter as a corpus has considerable benefits — namely large amounts of geographically locatable, spontaneous and unmonitored language data — but it also has some limitations. It is, after all, only one text-type, and the sample population is heavily skewed towards a young and urban demographic. Additionally, using Twitter for gathering data of this kind can also be somewhat precarious — research conducted using this method is always at the mercy of Twitter's changing rules.

There are different ways of accessing Twitter data, via the various APIs (Search, Stream and Firehose). Each has its own merits: search and stream are free, but data are heavily bottlenecked and Firehose is potentially costly. Online tools such as the TAGS web-app used in this study offer to mitigate some of the issues in accessing Twitter data by providing a more user-friendly interface and the ability to save data directly to a remote server, negating the need to leave a local machine powered on.

Once data have been gathered, there is a time-consuming process of filtering the data, eliminating false positives and other unwanted artefacts. Regions need to be defined, and the dataset coded with those regions, as well as verb and pronoun type, to allow for frequency comparisons between sets. Interactive maps can be automatically generated by inputting the geographical coordinates into an online service (BatchGeo).

Chapter 6

Results and analysis

6.1 The nature of the dataset

Despite what has been said about the *data problem* and the potential to overcome it by using large datasets drawn from Twitter, the dataset captured here is considerably smaller than it might have been. After cleaning of duplicates, there was a total of 1416 tweets containing the strings searched for (see table A4 in the Appendix). This was drawn over a period of 14 months from November 2014 to March 2016.

The reasons for the lower data count are looked at in table 6.1 below and section 5.4 about Twitter's ever-changing rules. However, enough data are still provided to generate meaningful statistics, and the resolution of the issues presented in table 6.1 is straightforward. The first problem is essentially a matter of cost, and the second has (at least for the time being) resolved itself.[1]

One way to mitigate the effect of having lower counts is to pool data into larger sets. As will be shown, the data are shown to pattern into three super-regions that pattern similarly.

[1]The caveat here is that whilst being able to map using user-location data dramatically increases the amount of data, that data tends to be less 'clean' - that is, they contain more duplicates, commercialised messages, etc. Additionally, of course, not all user-inputted location data necessarily point to a physical location; a user can put any text they wish in the user-location box, and do. Some users, for example, use it to supply their sexual orientation, others fictional locations. However, this 'noise' can be cleaned and after some processing, as shown by the similarity of the data drawn for the pilot survey (which used user-inputted data) to the current data, the majority of user-inputted data appears to correspond to where the user actually is.

Problem	Solution
Data were gathered using the public search-API which only provides access to ≈ 1% of the total datastream running through Twitter.	Pay for access to *firehose*-API.
Data were limited to GPS-encoded Tweets, which represent (during the period in which the data were gathered) ≈ 4% of the ≈ 1% available on the search-API.	Twitter has reinstated ability of API to match user-inputted location data (see 5.4)

Table 6.1: Causes of smaller dataset, and resolutions.

6.2 Conversation threads

The majority (76% - see table 6.2) of Tweets in the dataset are responses to another tweet, and so fit into the *tweet as conversation* category. This compares to Page's (2012a) finding that the general trend for 'ordinary users' (that is, non-celebrity users) was 48% 'broadcast' messages (public-facing) and 42% conversational addressed messages. Similarly, Eisenstein's (2017) dataset of 114 million geotagged messages involves "more than 40% of messages... addressed to another user".

A general trend showing an increase in the conversational use of Twitter does not account for the 76% found in the present data, which is better explained by the skewing of the current dataset to consist only of messages containing pDit. Pronominal elements, as discussed above (see section 2.4), function as part of a shared dialogue, referring to previously identified entities in the (virtual) world.

	In response		Not in response		Total
No of Tweets	1078	(76%)	338	(24%)	1416

Table 6.2: Proportion of tweets sent in response to another tweet in the Twitter dataset.

As discussed in section 2.8.1, messaging to an individual user is a fundamentally different activity from messaging to a general audience.

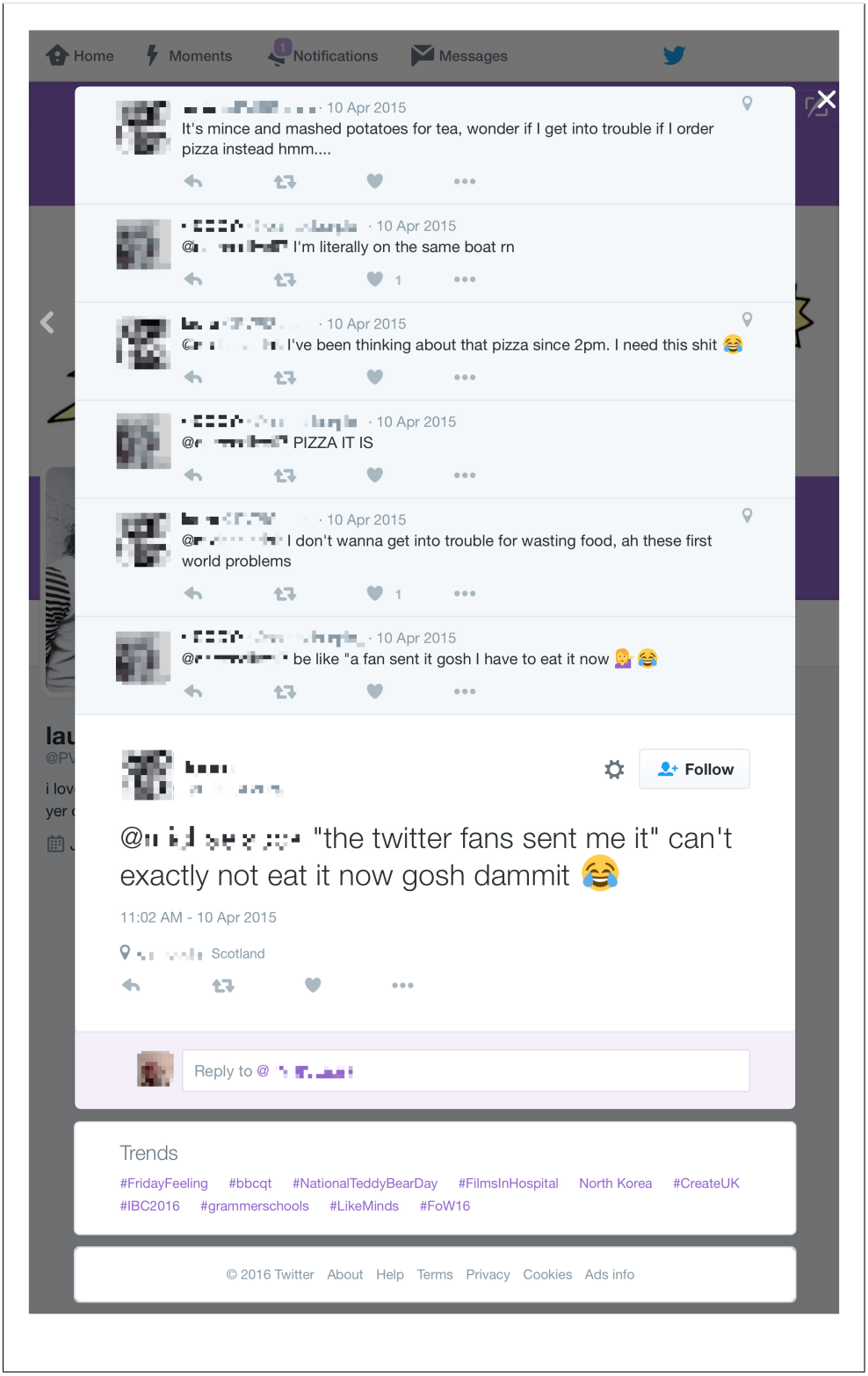

Figure 6.1: Example conversation thread in Twitter.

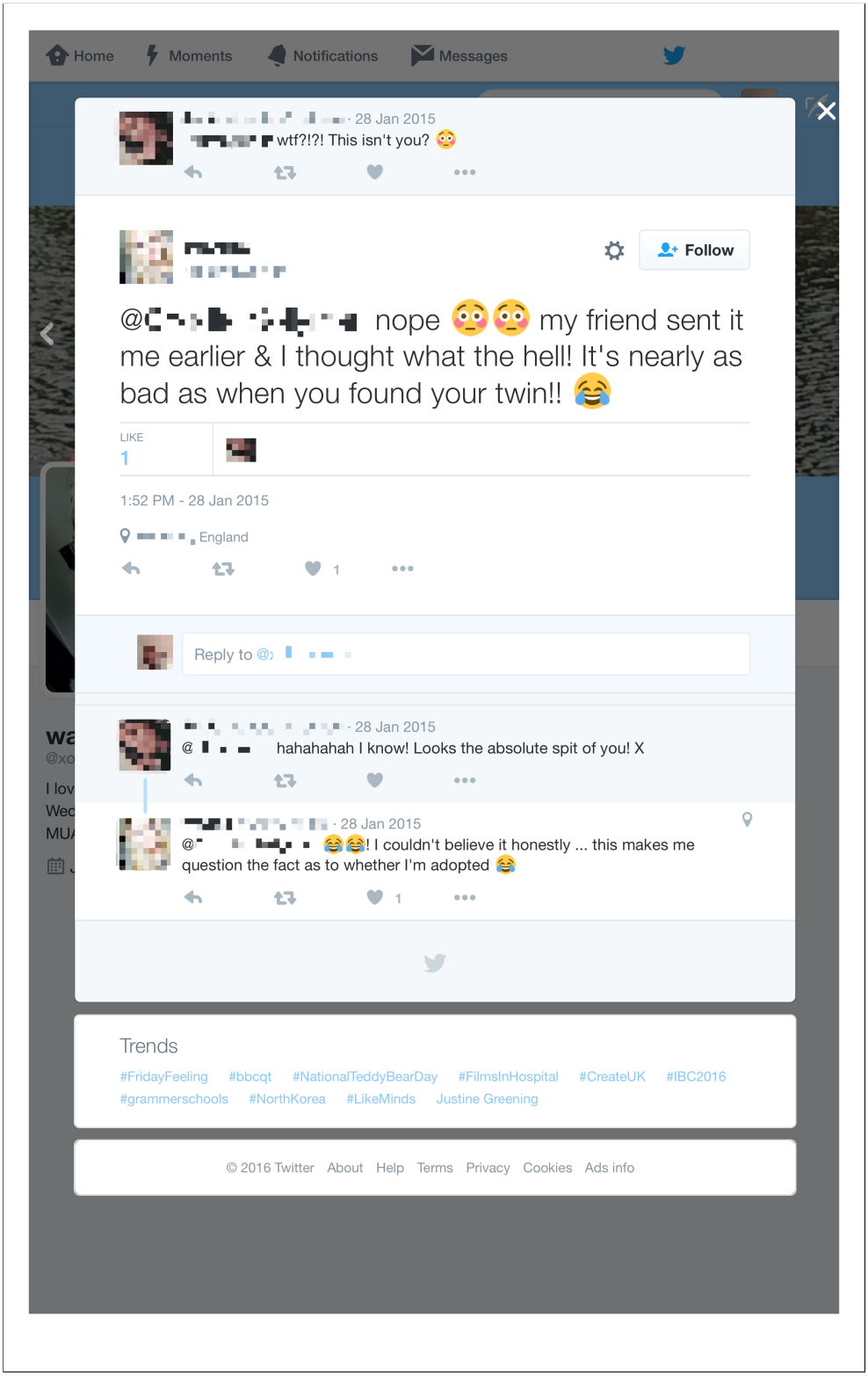

Figure 6.2: Example conversation thread in Twitter.

6.3 Geographical distribution and correlation with pilot data

The results for the current dataset are displayed in figure 6.3 below. The data are ordered from highest TGD use to lowest.

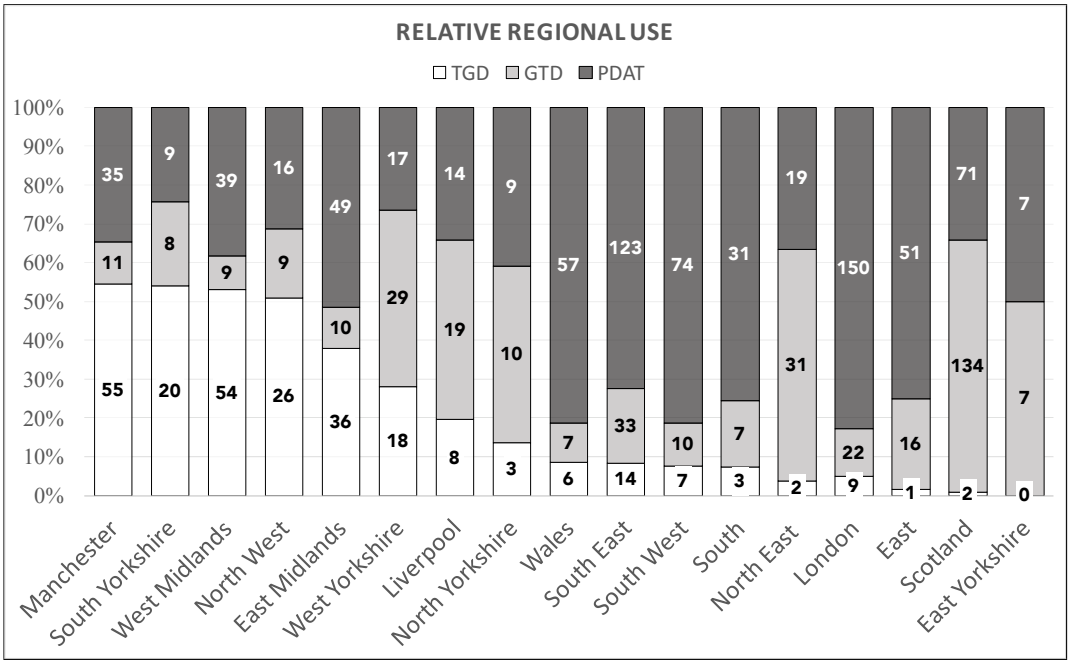

Figure 6.3: Regional variation of pronominal ditransitive use.

The scatterplot in figure 6.4 shows that the distribution found in the current dataset patterns with remarkable similarity to the data gathered in the pilot. A bivariate correlation test returns the correlation between the two datasets as R^2 Linear = 0.593 and $p<0.01$.

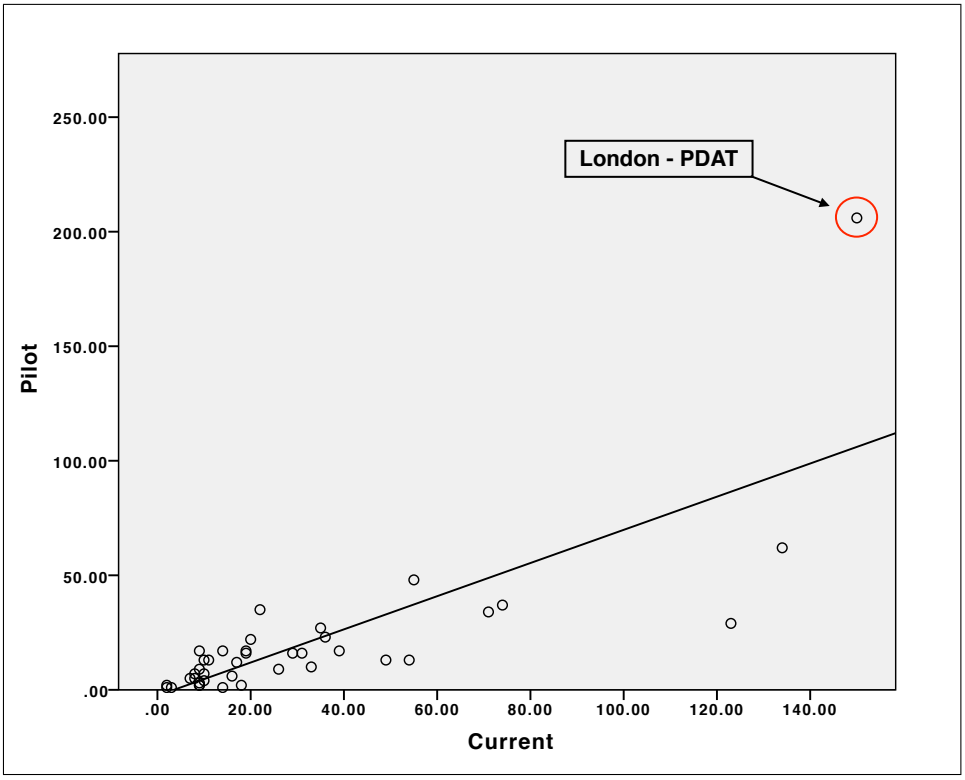

Figure 6.4: Scatterplot displaying correlation between the pilot data and the current data.

The bivariate correlations were achieved by comparing each pDit type in each region to the equivalent across both datasets using SPSS. This can be seen in the table below (table 6.3), which shows the first three regions.

region	type	pilot	current
East Midlands	emid_gtd	4	10
	emid_pdat	13	49
	emid_tgd	23	36
Liverpool	liv_gtd	16	19
	liv_pdat	17	14
	liv_tgd	7	8
London	lon_gtd	35	22
	lon_pdat	206	150
	lon_tgd	17	9

Table 6.3: First three regions shown in correlation table used for bivariate analysis. emid='East Midlands', liv='Liverpool', lon='London'.

This correlation shows both that the patterns we are seeing here are robust and that there is a good correlation between user-inputted location (as used in the pilot) and geo-tagged location (as used in the present study). There is only one marked outlier, the rate of PDAT in London, circled in red in figure 6.4. It is unclear why there is this discrepancy. It may be that London represents a special case as the capital, whereby users who are not actually based there still mark it as their location on Twitter, thus skewing the results. If London is removed from the correlation, the fit between the two datasets is even more apparent, as can be seen in figure 6.5, here R^2 Linear = 0.668 and p=0.01.

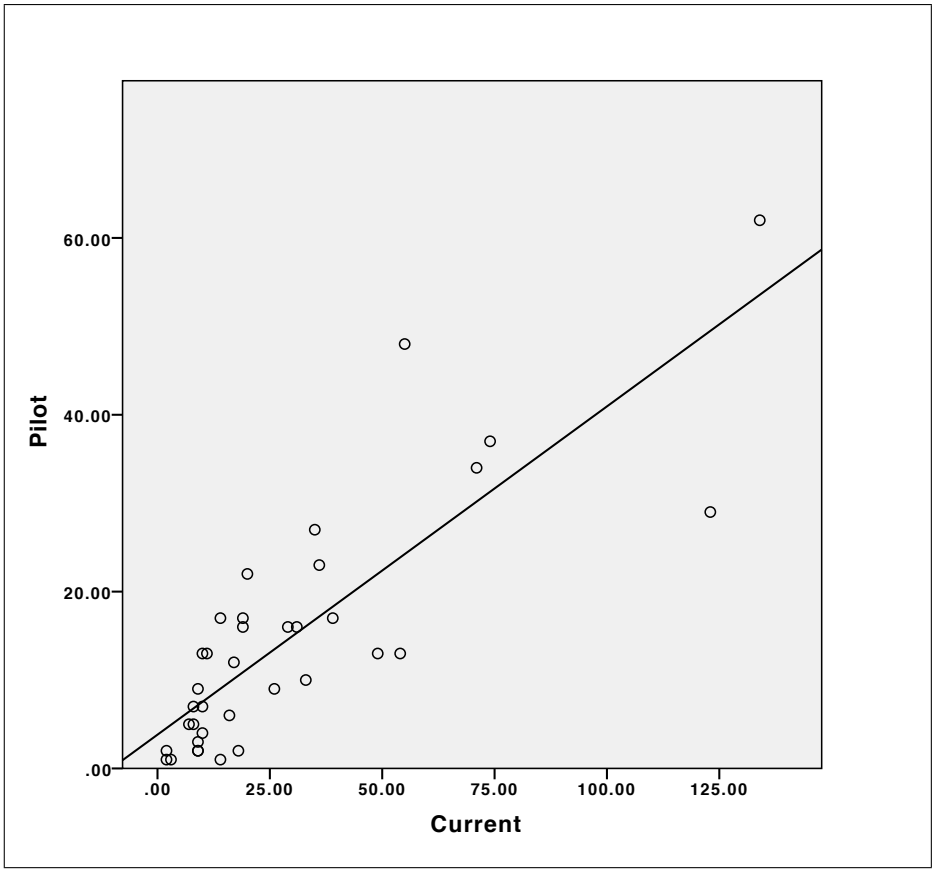

Figure 6.5: Scatterplot displaying correlation between the pilot data and the current data, with London removed.

This acknowledged, we see that otherwise the apparent advantage of using geo-tagged location — that each variant and its relative frequency can be accurately mapped — is to a large extent nullified; user-inputted data, once processed to remove

un-mappable results,[2] are on a par with geo-located Tweets. This is fortunate as GPS data are, as a result of recent changes[3] to the Twitter app (see section 5.4), increasingly scarce.

The automatically mapped results (using *Batchgeo*) of the compiled geocoded Twitter messages can be seen in figure 6.6 on page 79. The relative frequencies by region are also displayed on the map in figure 6.7 on page 80.

[2]Some user-inputted location data may, for example, state *at the bar* or *the Moon*.

[3]Occurring after the data for the current study were gathered.

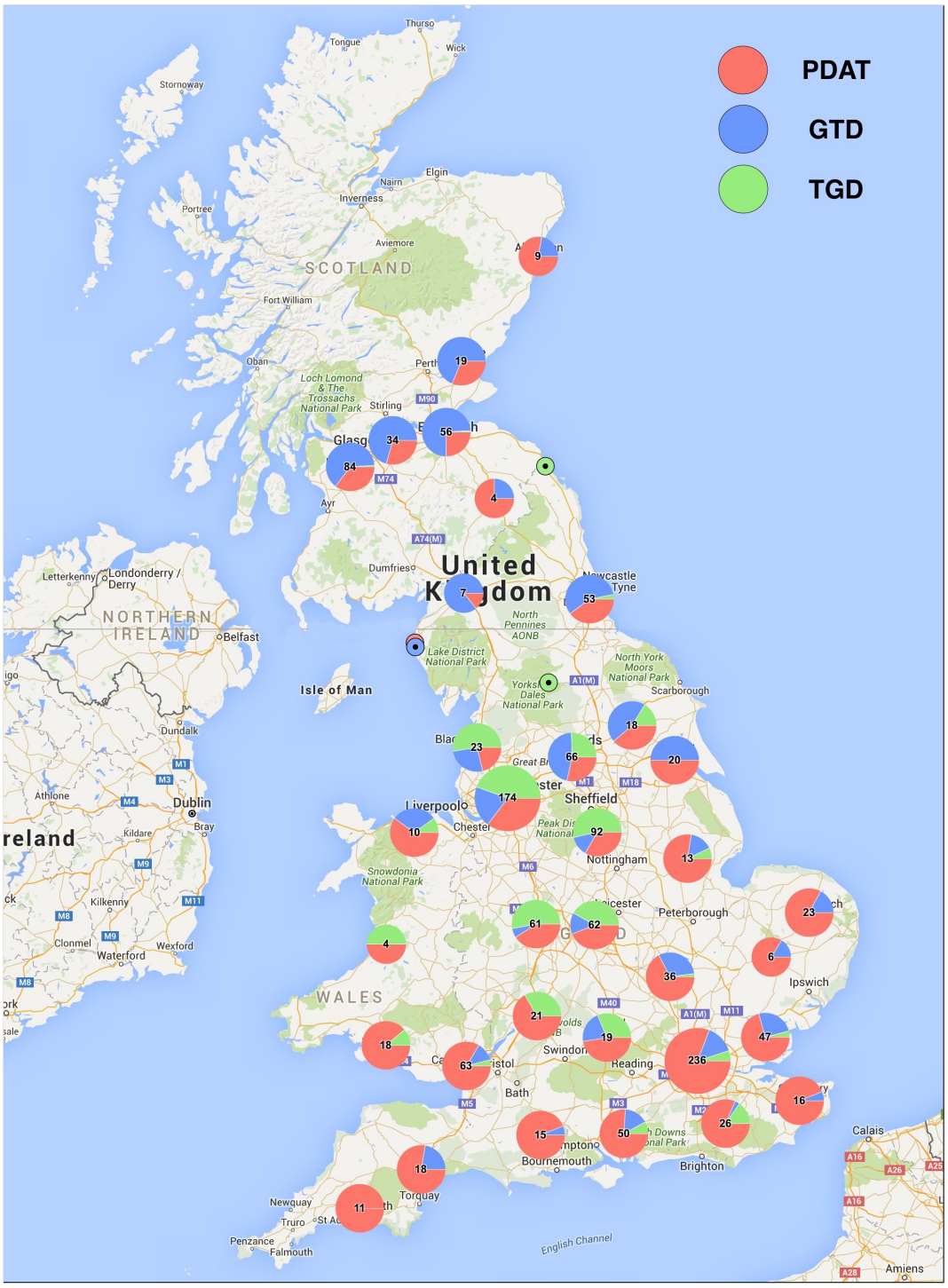

Figure 6.6: Distribution of ditransitive types used in UK Twitter messages.

Interactive map available at `https://batchgeo.com/map/6dbc125a32bdcb9c037727f03eed1114`.

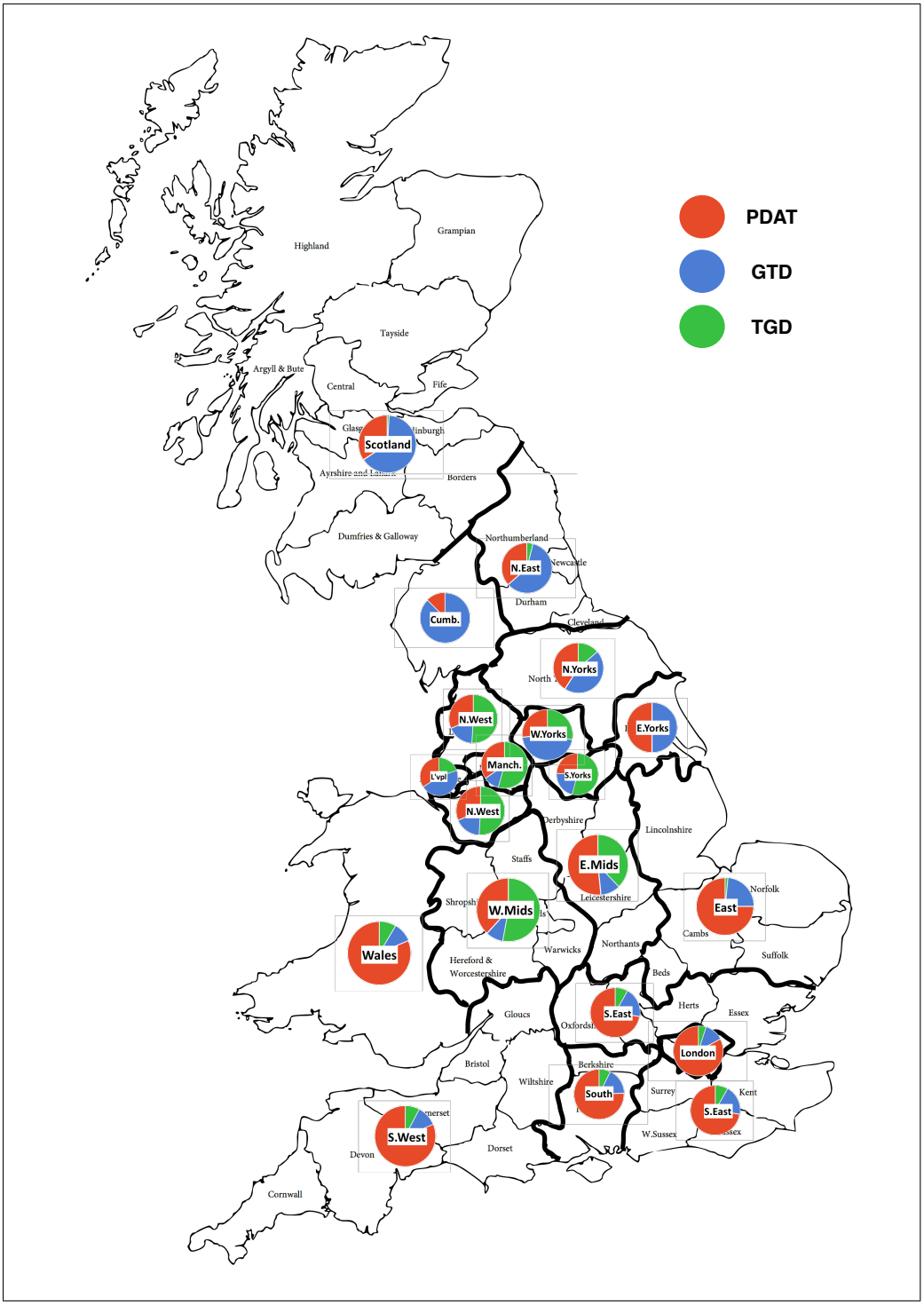

Figure 6.7: Distribution of ditransitive types, counts per region.

6.3.1 Super-regions

Whilst the results show a pronounced difference in ditransitive distribution across the UK, it is possible to group regions into similarly patterning sets (see figure 6.8 and table 6.4). The contingency table (table 6.4) compares each region to each other region for similarity by chi-square analysis. Unsurprisingly, most regions show little similarity to each other. What is of more interest in the current analysis is where two regions are not significantly different from each other. Cells showing this are highlighted on the table.

Importantly, most of the regions that show this lack of difference from each other form part of one of three distinct super-regions formed on the basis of (1) lack of significant difference and (2) adjacency. These regions are described below and shown in figure 6.8.

- GROUP A (high GTD): Scotland and the North East (and Cumbria[4]) form a northern region characterised by high GTD use ($\approx 75\%$), low PDAT ($\approx 25\%$) and the TGD being all but absent.
- GROUP B (high TGD): Manchester, South Yorkshire, West Midlands, North West and East Midlands form a central region characterised by a three-way mix, with TGD $\approx 50\%$, GTD $\approx 10 - 20\%$ and PDAT $30 - 50\%$.
- GROUP C (high PDAT): Wales, South East, South West, South, London and East form a southern region characterised by high PDAT ($\approx 75\%$), low GTD ($\approx 10\%$) and lower still TGD ($\approx 8\%$).

West Yorkshire was a marginal case, although it could have been grouped with GROUP B as it was not significantly different from the adjacent North West region, and so the decision was taken to exclude it. This was done on the basis that it had a much higher rate of GTD ($\approx 50\%$) than the other regions in the group. Liverpool patterned similarly to West Yorkshire but the lack of adjacency prevents them from

[4]Cumbria was originally included as part of the North West region and was as a result not included in the table as a separate entity. But based on the ratio of GTD to PDAT, it likely also patterns with GROUP A.

being considered as a region. North and East Yorkshire may form another group but were excluded on the basis of low raw counts.

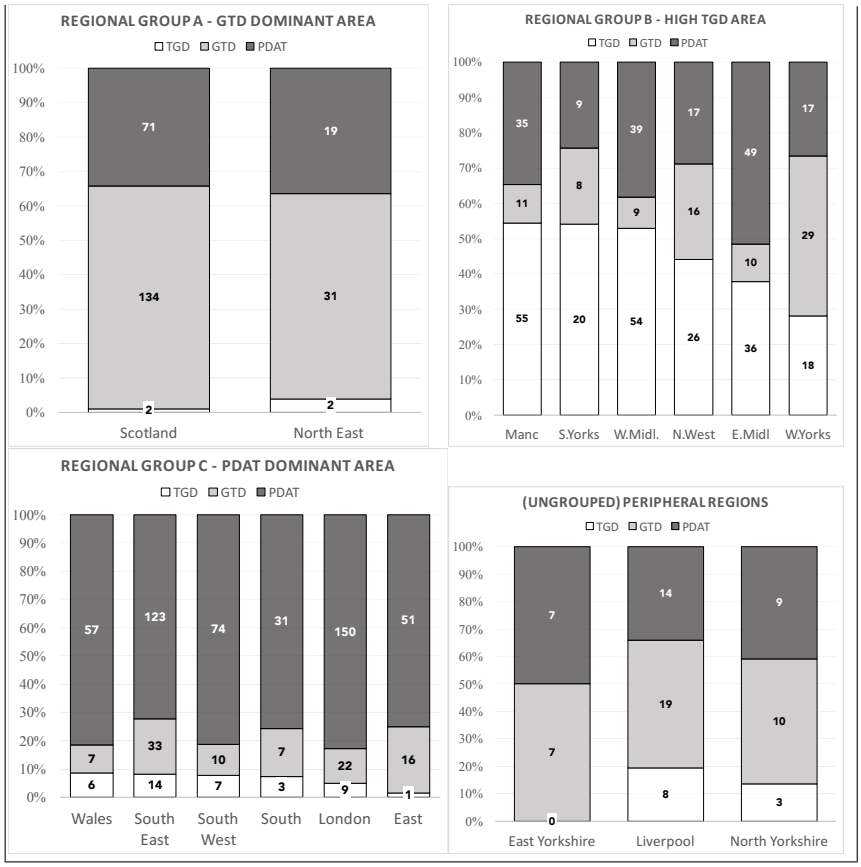

Figure 6.8: Regional variation grouped into similarly patterning super-regions

6.4 *Gave* vs *sent*

Next, each super-region was tested for difference by chi-Square analysis between *gave* and *sent* on the choice of pDit type. The counts for each group are presented in the chart in figure 6.9 below. TGD for both verbs was excluded from Group A for the chi-Square analysis, as each constituted <1% of the data for the region and so was deemed to be anomalous.

	Manchester	South Yorkshire	West Midlands	North West	East Midlands	West Yorkshire	Liverpool	North Yorkshire	Wales	South East	South West	South	North East	London	East	Scotland	East Yorkshire
Manchester		3.127	.420	6.997	6.170*	26.1**	25.375**	19.221**	41.246**	71.438**	50.780**	27.268**	53.321**	93.006**	51.494**	146.746**	20.432**
South Yorkshire			5.164	.917	8.545*	7.827*	10.534**	9.594**	35.736**	49.905**	42.096**	24.591**	30.192**	72.071**	43.934**	109.094**	12.483**
West Midlands .				9.603**	4.546	30.045**	28.693**	21.963**	37.365**	68.053**	46.631**	25.396**	57.064**	87.555**	49.826**	148.305**	22.874**
North West					10.681**	5.015	7.066*	6.554*	36.974**	46.992**	43.165**	23.363**	25.128**	71.616**	39.710**	92.969**	9.595**
East Midlands						25.705**	22.099**	16.313**	19.215**	35.431**	24.565**	13.132**	45.749**	50.021**	30.908**	115.594**	17.502**
West Yorkshire							1.241	2.522	40.879**	41.518**	46.508**	24.373**	11.863**	69.965**	35.878**	52.826**	6.043*
Liverpool								.465	26.069**	21.176**	28.989**	14.233**	6.021*	41.126**	21.387**	31.068**	3.448
North Yorkshire									15.657**	9.342**	16.537**	7.59*	2.831	20.700**	11.199**	16.046**	2.106
Wales										3.187	.075	1.186	34.248**	1.310	7.399*	66.407**	13.913**
South East											3.205	.178	31.375**	5.618	3.973	81.173**	7.635*
South West												.933	38.271**	.849	6.935*	75.975**	14.025**
South													17.177**	1.175	2.889	34.719**	6.46*
North East														52.066**	17.935**	2.463	1.223
London															6.080*	111.865**	14.955**
East																35.164**	4.132
Scotland																	1.497
East Yorkshire																	

Table 5.4: Contingency table showing chi-square value between each region *p<.05 **p<.01 df=2

Shaded areas not significantly different

Green=high TGD Blue=high GTD Red=high PDAT Yellow=mix

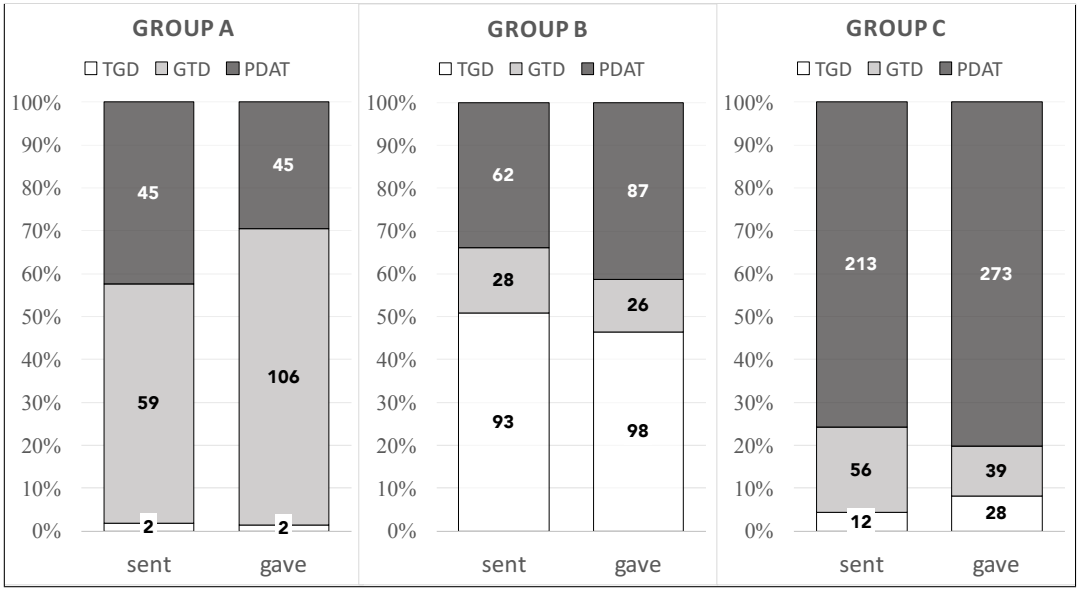

Figure 6.9: Counts for each pDit type and verb by super-region.

The results of the Chi-Square analysis are presented in table 6.5 below.[5]

Gave vs *Sent*	p-value	chi-square	df
Group A	.027*	4.891*	1
Group B	.298	2.422	2
Group C	.003**	11.346**	2

Table 6.5: Chi-square analysis results comparing difference between verb type and pDit type by super-region.

This is an interesting result. There is no significant difference in respect of pDit choice between verb type in GROUP B in this sample. But the difference between verb type in both GROUP A and GROUP C is significant ($p<.05$ and $p<.01$). Moreover, the variable driving the difference is different in each group. For GROUP A, PDAT is proportionally higher for *sent*, whereas in GROUP C, PDAT is lower for *sent*, with GTD showing an increase. There is also a notable increase in TGD use in GROUP C with *gave*, but the numbers are still relatively low. Again, if there is an effect of verb type on pDit choice, such a constraint appears to be geographically variable.

[5]Note that for ease, p-values are also reported here as the degrees of freedom differ between Group A and the other two groups, making Chi-Square figures potentially confusing.

6.5 GOAL pronoun

The order of objects has been reported to be affected by the property of the GOAL pronouns used. The results from the Twitter dataset are presented in figure 6.10 and table 6.6 below.

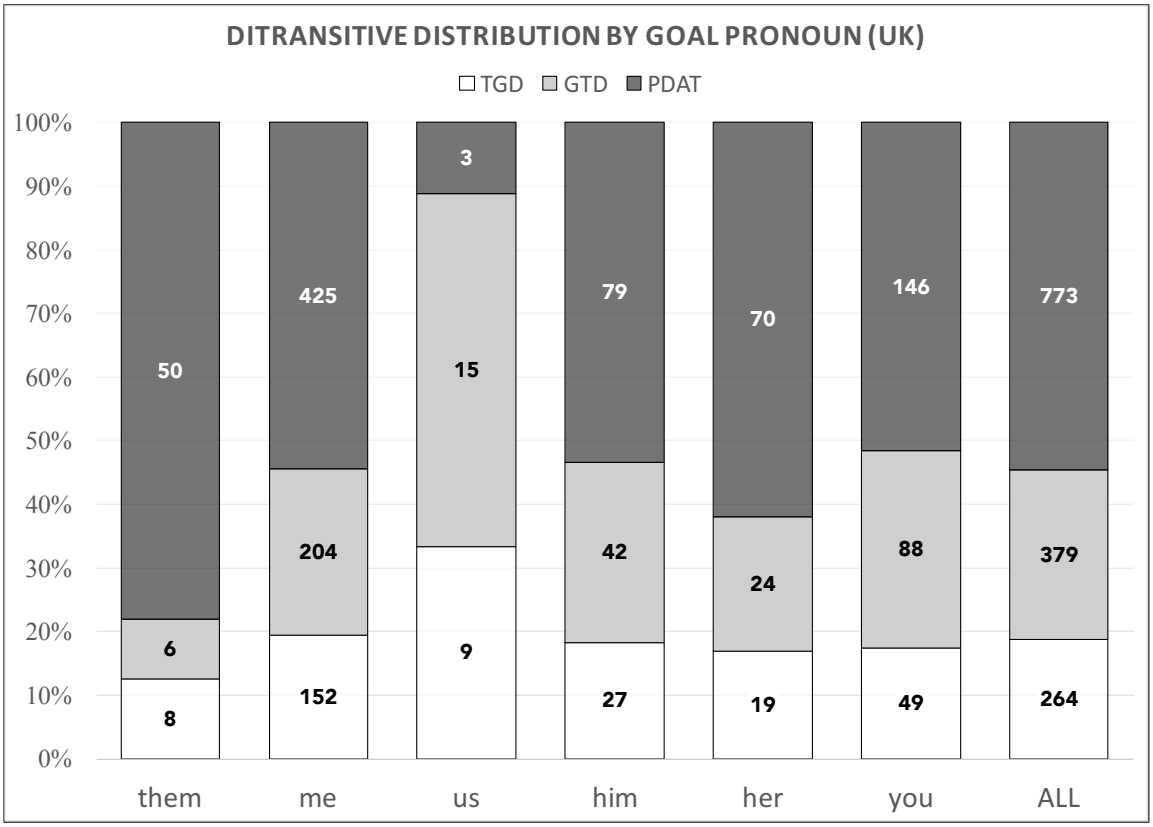

Figure 6.10: Variation by pronoun, from Twitter corpus.

In the Twitter dataset generalised across the UK, object pronoun 'THEM' patterns significantly differently from the average for all other object pronouns (apart from 'HER') at ($p<.01$). Object pronoun 'US' patterns differently from the average for all object pronouns, and the difference is significant ($p<.01$).

	THEM	ME	US	HIM	HER	YOU
THEM		14.103**	36.602**	12.156**	5.469	16.333**
ME			20.103**	.360	2.295	2.690
US				16.412**	22.987**	16.222**
HIM					2.190	.345
HER						4.374
YOU						

Table 6.6: Contingency table showing statistical difference between categories **p<.01.

The *data problem* is still present here. The number of results returned for some of the pronouns is lower than ideal.[6]

Additionally, numbers are not high enough to allow for subsetting by region or super-region. However, it is likely that high enough frequencies would be achieved following the resolution of the issues discussed in table 6.1, in the introduction to the results.

6.6 Summary

The dataset retrieved from Twitter for the current investigation took longer than might be expected, and the amount of data was also smalles than it could have been. There are, however some easy fixes to this problem, and part of the problem was a result of Twitter changing the way it operated, something that has since been rectified. This acknowledged, it should be noted that the data here represent only the 'hits' — that is, the parts of the corpus that contain the relevant strings. The overall size of the dataset from which these 'hits' were gathered is, of course, many times bigger.

Often overlooked when using Twitter for linguistic analysis is the fact that not all Twitter messages are equal — some messages are written for a general audience, whilst others are addressed to an individual as part of a more free-flowing conver-

[6]The general rule in chi-square tables is that all cells should contain a minimum of 5 observations. However, in larger tables, it has been shown that it is valid to have some cells with value lower than 5, provided that such cells make up less than 20% of all cells (Field, 2009, p.695).

sation. The current dataset displays a weighting towards tweets as part of these *conversation threads*.

The geographical distribution of the pilot Twitter data which used 'user-inputted' information show a strong correlation with the distribution shown in the current Twitter dataset, which uses GPS data automatically added to each message. This indicates that the more plentiful user-inputted data may be relied on in future studies.

The geographical spread of pDit variations may be grouped into similarly patterning *super-regions* by comparing each region to each other using chi-square analysis. Following this, the data for distribution by *verb* according to *super-region* and *pronoun* may be analysed for statistical difference.

This process found a significant difference between *sent* and *gave* in two of the super-regions. GOAL pronouns *them* and *us* are also argued to pattern significantly differently from the other GOAL pronouns for the pDit type with which they co-occurred.

Chapter 7

Discussion

7.1 Introduction

The body of knowledge regarding the distribution of the pDit, to which the present investigation has aimed to add, is — as was said in the introduction — *limited and fragmentary*. It has been explained that the main reason for this, despite recent efforts to ameliorate the situation, is a lack of data, particularly when constraining the investigation to only include ditransitives with two pronominal objects, a structure that is primarily found in speech. The principal aim has been to provide this data by taking advantage of Twitter's public API. Having done this, the discussion inevitably turns to ask first whether such data can reasonably be considered to represent the situation 'on the ground' - that is, between speakers in everyday conversation - and if so, what these newly uncovered patterns might tell us. As discussed, the geohistorical implications are at the surface of an analysis of such data.

The current section recalls the research questions outlined in chapter 1 (chapter 3) and addresses each one in the light of the data presented in the results (chapter 6).

7.2 Geographical distribution

The primary aim of the present investigation, represented by **research question 1** (RQ1 is repeated below), is to provide detailed resolution regarding the relative

frequencies of each pDit type by region.

RQ1: What is the geographical spread of the pDit, and how do its variants (TGD, GTD and PDAT) pattern relative to each other by region?

The argument here is that the results presented in figure 6.3 (section 6.3), detailing the relative counts of Twitter messages containing the relevant construction, are representative of the relative frequencies of each pDit type in actual speech in these areas — at least, that is, the speech produced by the demographic who use Twitter, and following the nature of the users singled out by the Search API (see section 5.3), the demographic who are *heavy* Twitter users. This should be considered carefully when interpreting the results. However, as has been shown by recent studies (e.g. Doyle, 2014; Eisenstein, 2017; Jones, 2015), the data drawn from Twitter do repeatedly fall where they would be expected to fall with respect to established dialect surveys. The data for the current study were also shown to be consistent with previous data, correlating ($p<.01$) with the data from the pilot study. Additionally, the distribution of the TGD found in Twitter fits with the SED map, and also and fits with the survey conducted for the Manchester dialect project (MacKenzie et al., 2014).

Comparing the pilot survey results (figure 4.3, section 4.3), to the Twitter results also reveals a striking correspondence. The areas in the survey displaying >80% mean acceptance of TGD (East Midlands, Manchester, South Yorkshire and West Midlands) match the areas in the Twitter results that show $\approx$ 50% usage, and correspond to the *super-region* identified as 'GROUP B'. Notably, West Yorkshire shows a diminished mean acceptance rating ($\approx$ 60%), which corresponds to the lower rates of use reported on Twitter, supplying further evidence that West Yorkshire lies in the transition zone. Likewise, survey respondents from the South-East — which exhibits lower TGD occurrence in the Twitter corpus — also show a weaker acceptance of the TGD.[1]

[1]The present study's focus is on comparing the corpus data to the Twitter data, and as such a fuller comparison between responses to an updated survey and an expanded Twitter dataset are left here to follow in the PhD. The close correspondence between the pilot survey and present Twitter

This border region is roughly in line with the isoglossic boundary known as the Humber-Ribble belt, discussed in section 2.6.1. The black dotted line in the detail from the BatchGeo map, shown in figure 7.1, estimates a possible location of this border. To the south of the line, the TGD (green) predominates, whilst north of the line, GTD (blue) and PDAT (red) seem to be favoured.

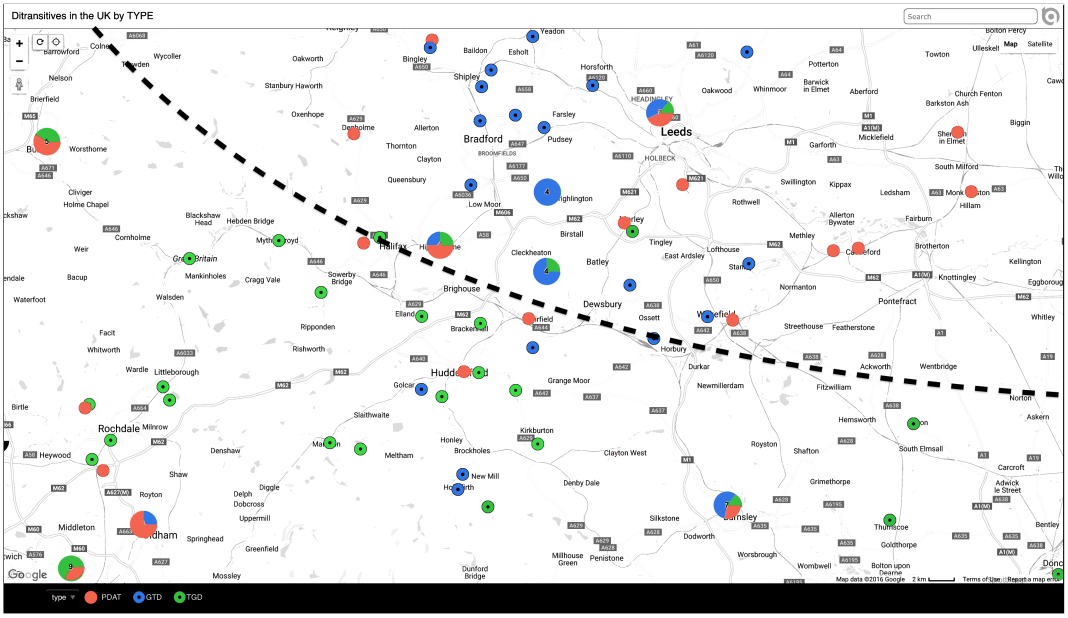

Figure 7.1: Detail of West Yorkshire from BatchGeo interactive map indicating possible border / 'transition zone'.

The most striking aspect of the geographical distribution presented here is the uniformity of the distribution over wide geographical areas. This aspect is discussed in the following sections.

7.3 Relating the current picture to the historical distribution

This section relates to **research question 2** (RQ2 is repeated below). First is a brief discussion on the nature of CMC as written speech, followed by an in-depth exploration of how the current picture, as represented on Twitter, relates to the historical and contemporary corpora introduced in the literature review.

data gives some indication as to what the deeper investigation will reveal.

RQ2: How do the Twitter data relate to historical and contemporary corpora?

7.3.1 Computer-mediated communication and written speech

Whilst historical linguistics has necessarily relied on written data, such written data tend to be out of step with spoken data. Written language is also traditionally more susceptible to forces of prescription. And whilst it might be true that Twitter is over-representative of the *interior classes*[2] (Jones, 2015), it is certainly more socially representative than the correspondence available from the literate classes in the historical record.

It seems, following the literature on CMC, and from the data presented here, that certain subtypes of CMC do indeed engender a form of written data that reflect spoken habits. The speech-like nature of conversational Twitter messages is evidenced by the occurrence of the pDit. In addition, comparing the example Twitter conversations (figures 6.1 and 6.2) to the example from IRC (figure 2.13), there are striking similarities. The same factors indicated as being speech-like components of IRC are also present in Twitter. Short turns, non-clausal elements and supportive paralinguistic cues of simulated laughter ('lol', 'hehe' and emojis) are all present here. It seems that this behaviour is quite unselfconscious, occurring as a spontaneous response to the affordances of the medium.

Meanwhile, the widespread dissemination of digital communications devices has resulted in the democratisation of access to the written medium as a means of interpersonal communication. As a result, written personal communication has moved from what was once a more distinct formal practice, with its associated registers and grammars, to a radically different situation. Today, people *en masse*, *speak* to each other with written words. Whilst CMC has developed its own standards and affordances, it has separated itself from the confines of the print-based era subject to hierarchically imposed standardisation - what Shortis (2016, p.488) refers to as a "post-print, post-standardisation written system". In this space, the impedance to patterns of dialectal expression is dramatically lowered. There is, as a result, a renewed freedom - indeed *pressure* - for people to write how they would speak,

[2]Meaning 'middle classes'.

and this is what they appear to do. The resulting Twitter data can thus provide a powerful insight into the linguistic behaviour of ordinary people, insofar, of course, as Twitter users are 'ordinary'.

7.3.2 Comparison to the corpus record

What the corpus record shows us can be deceptive. An apparent trend showing the reduction in use of a certain feature may be better explained as a shift towards the standard written form, or — given regional variation — a shift in the regional origin of the corpus data. The historical trend reported by Yáñez Bouza and Denison (2015) shows the TGD falling almost completely out of usage, at least in the standard dialect. As the authors report, "variation has been gradually restricted in present-day standard British English to patterns (1)[GTD] and (2)[PDAT], and that pattern (3)[TGD] is confined to certain dialects only" (Yáñez Bouza & Denison, 2015, p.262). This is supported by Siewierska and Hollmann's (2007) finding of majority TGD use in Lancashire and Gerwin's (2014) finding of a higher general rate of use in the north of England.

The analysis of the Twitter corpus shows that the situation found in Lancashire is indeed the situation found across a large area of the Midlands and North West of England, including large urban settlements: Manchester, Birmingham, Sheffield, Nottingham, and to an extent, Leeds. Whilst this was to an extent already known, despite recent efforts (Gerwin, 2013, 2014) traditional corpus methods fell short of quantifying the extent and precise whereabouts of TGD use.

What the Twitter data do here, then, is to give considerable new detail to the trends indicated by the historical and contemporary data. It is likely that overall TGD use did fall, as reported by the overall historical record, but it is crucial — following the call from dialect grammar studies — to recognise that such a picture cannot account for the language as a whole. The actual decline of the TGD in English generally is certainly, given the continued high usage by a youthful Twitter population, considerably less steep.

Similarly, the reported ascendency of the PDAT in the historical record is matched by the current picture for the South and South East but not in the Midlands, and not in Scotland or the North East.

7.4 Super-regions and syntactic persistence

This section addresses both **research questions 1 and 2** by looking at how geographical frequency distributions relate to what we know of distributions from the corpus record. The geohistorical implications, first outlined in the literature review (section 2.5.2), are revisited and explored.

Section 6.3.1 showed how statistical analysis of frequency data by region could be used to infer larger *super-regions* of similarly patterning subregions. The discovery of these super-regions is intriguing, and how long they have existed is unclear. An indication of their longevity is, however, provided by the apparent robustness of the patterns of syntactic occurrence. This robustness is underwritten by the wide geographical spread of each super-region, with each super-region containing a wide spread of distinct dialects.

The finding that adult speakers in interaction tend to repeat the syntactic structures that have occurred in the immediately preceding utterances (Bock, 1986) may also help explain the maintenance of pDit type across generations. This is supported by the finding that children are sensitive to the probabilistic patterns of syntactic variation in their linguistic milieu such that their language production is predictable (De Marneffe, Grimm, Arnon, Kirby, & Bresnan, 2012). Distinct dialects then, may perpetuate similar frequencies of syntactic variation across generations, whilst phonological and lexical features may be more prone to change.

Following this line of thinking, it would seem that the main disruption to the syntactic patterns found would be the result of large-scale historical events of migrations and contact. Such migrations may be relatively recent, as with the movement of people into the rapidly expanding cities of industrialising Britain in the 19th century, which may explain the pattern linking the areas in Group B, or more distant in time as with the patterns of Viking and Norman conquest and settlement, which may explain Group A. Pertinent here is Fabian's (1983) concept of "typological time" oriented around "socioculturally meaningful events" (Fabian, 1983, p.23). These ideas are explored in the following subsections.

7.4.1 Scotland and the North East (and Cumbria) (GROUP A: high GTD)

The high use of GTD in Scotland might, if Gast (2007) is correct in his analysis, be an indication of increased Old Norse contact. Scots (which of course intersects with Standard Scottish English), is known to have its roots in the Northumbrian dialect of Old English, which saw considerable influence from Old Norse before subsequently being cut off from Norman-controlled England and developing independently during the Middle Scots period. Whilst Kopaczyk (2013) warns that the immediacy of historical events may not coincide with linguistic periods (due to the slower nature of language change), it seems plausible that the high rates of GTD seen in Scotland could find root here. That is, the patterns seen today are the result of Scotland, and the North East retaining a feature — or more precisely, the increased tendency to use a feature — that has been eroded in the southern parts of England, particularly those southern parts that saw a high level of integration with Norman rule.

Such a working hypothesis, that the high GTD use in Group A is the result of the lack of Norman influence, is of course dependent on the assumption that it was Norman influence that drove the expansion of verbs that could take PDAT. Some evidence of this may be found in the Twitter verbs data.

Research question 3 asked:

> **RQ3: Is there a difference between *give* and *send* in choice of pDit type and is any difference regionally distinct?**

The finding of a significant difference between the verbs *sent* and *gave* in Group A — with *sent* showing a marked increase in occurrence with PDAT when compared to *gave*, which shows greater relative preference for GTD — is potentially revealing here. As reported in section 2.5.2, the PDAT construction was not found for Old English verbs corresponding to *give* ('agifan, gifan') but was found with *send* ('sendan'). This may also explain Tagliamonte's (2014a) finding, discussed earlier, that verbs other than *give*, in her Canadian and UK English datasets, prefer PDAT "across the board" (p.309).

Further evidence to support this idea might come from the historical analysis of a greater number of ditransitive verbs across Middle English and Middle Scots,

and it would need to inspect more closely the encoding patterns of French verbs at that time. But given the evidence presented by the two ditransitives here (*send* and *give*), this seems a worthwhile pursuit.

7.4.2 The Midlands and the North West of England (GROUP B: high TGD)

Group B, the area characterised by high TGD usage, involves several of the major industrial cities which each saw rapid expansion and population increases during the 18th and 19th centuries (see table 2.2 in section 2.5.2). If the TGD was in widespread use during this time, as is suggested by the corpus record, mass migration into these cities during this period might provide an explanation for its maintenance there, whilst in the south it continued to lose prominence.

Yáñez Bouza and Denison's (2015) finding that the TGD was once found in high numbers outside of this area is intriguing. Whilst it is unclear exactly what region the area labelled 'The North East' in their data refers to, the use of the TGD in Tyneside and further north — in speech at least — seems unlikely considering the current complete lack of usage reported on Twitter. It may be that its use here was confined to more formal written texts, or that the North East area corresponds to North Yorkshire, which is more in keeping with the current picture. The historical data for the frequency of TGD in East Anglia are, at 77%, particularly different from that found today. East Anglia is, in the contemporary Twitter data, part of 'GROUP C', the *high PDAT* group (discussed below), which is characterised by low TGD use. By this account, it does appear that the overall geographical area in which the TGD is used has reduced considerably.

7.4.3 The South and East England (GROUP C: high PDAT)

Correspondingly, the geographical spread of GROUP C, showing high PDAT occurrence, is quite distinct from the spread shown by the SED. The SED map shows the PDAT to be preferred in the South West, but not in much of the South East and East, which are mostly reported here as GTD areas.

The discrepancy between the Twitter data and the SED map for the South and East of England needs to be looked at more closely. It may be that the type of

informant for the SED (NORMs) — who were chosen specifically because they were thought to represent the most conservative linguistic features — results in a picture that is biased towards the GTD, which — following the working hypothesis supplied by Gast (2007) — may be the older form. Again here, we return to the idea that the GTD is an old feature, existing alongside TGD in the South of England long before its rise to 'canonical' status in the 20th century. Whist there seems to be little evidence of GTD use here in the corpus record prior to the 20th century, the GTD did surface here in the late 16th century corpora. The view taken in this paper is that both TGD and GTD orders have been available to many speakers in the history of English, just as both are still available to many present-day English speakers. The reason for the lack of corpus evidence for its use in the period between the 17th and 19th century may have more to do with its social status as being 'un-English', following 18th century prescriptive guides. As discussed in the literature review, it seems that GTD use was deemed high enough to warrant the TGD being prescribed as *proper* English, and the GTD as a 'vulgar Scotticism'. [3]

Returning to the PDAT, Yáñez Bouza and Denison's (2015) data are more comparable to the current picture. Their 'South East' as well as 'South West' both report high PDAT occurrence (96.3% and 87% respectively). However, as discussed in the previous section, their historical data for East Anglia are predominantly TGD. Again, it seems that the picture presented by the overall historical trend — the TGD being replaced by the PDAT — fits the data for the South East.

Returning again to **RQ3**, the finding that for the effect of *verb* on pDit type there is an inverse situation to that found in Group A is curious. It is mirrored by Gerwin's (2014) finding that *give* shows a decrease in the GTD (discussed in section 2.7.2), whilst all other ditransitive verbs show an increase. There can be no clear answer without fuller investigation, but for now it is perhaps enough to say that, in the South at least, *give* is undergoing a pragmatic shift independent of other regions, which is resulting in an increase in use.

[3] As a side note, it may be that a rise in social prestige of the GTD can be attributed to a rise in first migration and later social status through the course of the 19th century of the Scottish in England, particularly in London. This is discussed in Smout (2005), and may warrant further investigation.

7.5 Distribution by pronoun

This section addresses **research question 4**:

> **RQ4: What is the effect of pronoun choice on ditransitive use?**

> "The morpho-phonological status of the pronominal THEME and RECIPIENT in the two double object patterns is of considerable interest as this may have a direct bearing on how the two patterns are to be dealt with in a model of grammar." (Siewierska & Hollmann, 2007, p.97)

Whatever mechanism is proposed to underlie the generation of each ditransitive type, selection of type appears to be sensitive to phonological factors. These factors are explored here.

The discussion of the distribution by pronoun needs to be prefaced by recognition again of the fact that occurrences in the current Twitter corpus for some of the pronouns are still quite low. This means that the current analysis had to follow the measures taken by previous quantitative investigations by pooling all regions across the dataset together into one UK-wide superset. Again, the ideal here would be to differentiate the dataset by region or 'super-region', so as to survey variation in the patterning of ditransitive types between individual locales. As explained in section 2.3, this is something that would be relatively straightforward, and is certainly worthy of future investigation (chapter 8).

As reported in the literature review, (section 2.7.1), there have been until this point several quantitative assessments of pronoun choice on pDit type. Gerwin (2014) combines BNCreg and FRED corpora and then conflates pronouns by number (see again figure 2.12), whilst Yáñez Bouza and Denison (2015) make an assessment based on their 'it-dataset' (described in section 2.7.1). Additionally, Siewierska and Hollmann (2007) look at the structure in their Lancashire corpus. Gerwin's comparison of the conflated groups reaches the conclusion that pronoun choice does influence pDit type, with first-person pronouns favouring TGD/GTD and third-person pronouns favouring PDAT, and that this distinction is statistically significant ($p<.05$).

The decision to conflate the categories is motivated primarily by the reasoning that there is no way of distinguishing between 'you' singular and 'you' plural, and therefore to mirror this already conflated category, it is necessary to conflate 'him'/'her' with 'them' and 'me' with 'us'.

However, for the current dataset, the use of 'you' in the expression 'sent it you' is usually determined to be singular by the tendency of the interaction in Twitter conversational messaging, from which the vast majority of the Twitter dataset is drawn, to be targeted at *one* other user.[4] More importantly, once we are freed from an obligation to merge the categories by number, it is possible to take into account any potential individual phonological distinction.

Whilst the Twitter dataset is written text, as has been discussed throughout, it shares aspects of speech practice. As shown in the literature review, this has been taken advantage of in recent studies that use Twitter to map phonological patterns. As Eisenstein (2013, p.1) reports, "social media displays influence from structural properties of the phonological system". It seems reasonable in the light of this to consider phonological factors underlying the patterning of pronoun and pDit type in the current Twitter dataset. Specifically, the methodological decision to conflate the pronoun 'us' (the only true vowel-initial dative pronoun included here[5]) with 'me' (the only pronoun to not be in a position to lose its initial consonant) is problematic. Meanwhile, 'him' and 'her' will frequently lose the initial glottal fricative /h/ and thus potentially become vowel-initial in production. The initial palatal approximant /j/ in 'you', as may also behave more like a vowel in production. 'Them' may also have its initial consonant deleted to 'em' (although less frequently) and is the only pronoun with both word-initial and word-final consonants (notwithstanding the

[4]Clearly, it is still true that 'you' (and 'us', particularly in the North East) can reference singular and plural entities in the world, and to that end, treating 'you' as one pronoun is potentially problematic, but doing so has the advantage of enabling the analysis of the behaviour of the other pronouns as individual entities and mitigates the perhaps larger issue of creating artificial pronoun categories. Also worth noting is the fact that 'you guys' and 'yous/youse' (in Scotland and the North East) are frequently used to distinguish between number in the second-person, and 'you' is likely to be singular in most cases.

[5]Of course 'it' would also be, and perhaps should have been included for completeness, although "send it it" is likely a rare occurrence.

initial glottal fricative in 'him' which as just mentioned is often lost, and 'her' with rhotic 'r'). Taking these factors into account it is possible to arrive at a grouping by phonology which is quite distinct from the grouping by semantics (person/number) applied by Gerwin (2014).

The results shown in figure 6.10 and table 6.6 show that in the Twitter sample gathered here, four of the pronouns individually pattern with remarkable similarity: 'me', 'him', 'her' and 'you' each exhibit no significant difference in their distribution. This similarity, importantly for the current discussion, stays within pronoun number - all are *singular* (with possible plural 'you' as potential caveat). The plural pronouns 'us' and 'them' are shown to pattern significantly differently from all the singular pronouns, with the exception of 'her'. Meanwhile, each plural pronoun is revealed to pattern differently from the others in its set, 'us' showing a preference for GTD/TGD and 'them' a preference for PDAT. It is an intriguing result in itself that 'her' should pattern differently from 'him' to the extent that 'her' is not significantly different from 'them', but 'him' *is* significantly different from it.

The preference of 'them' for the PDAT matches Gerwin's (2014) finding (mentioned in the literature review) that "third-person recipients [GOALS], especially plural *them*, are prone to a PREP-encoding [PDAT]" (p.196). The fact that 'her' is similar in distribution to 'them' perhaps provides support for number groupings. However, the same can not be said for 'him', which is significantly different from 'them' ($p<.01$). Additionally, the evidence presented here does not find the singular third-person pronouns to pattern differently from first-person or second-person pronouns 'me' or 'you'.

Gerwin's (2014) conclusion that 'them' prefers PDAT to "avoid case ambiguities" (p.196) is one possibility, though it seems unlikely. As Siewierska and Hollmann (2007, pp.95-96) point out, "case recoverability problems are most likely to occur when both of the pronouns are animate", which is seemingly a rare occurrence, not being attested at all in their corpus. However, it seems likely that factors relating to weight (longer pronouns being *heavier*) are likely also at play. 'Them', after all, is the 'heaviest' of the pronouns, and following the pattern of 'quantitative

harmonic alignment'[6] is more likely to occur after the shorter and lighter pronoun 'it', favouring TGD and PDAT constructions.

Further, Siewierska and Hollmann (2007) draw on Larson's (1988, p.364) claim that "in the canonical double object construction a pronominal recipient preceding a pronominal THEME must be necessarily unstressed" (Siewierska & Hollmann, 2007, p.96) to discuss the potential ambiguity of case in production between 'them' when shortened to 'em' and 'him' in this unstressed position. This combination of factors could explain the preference to move 'them' to a position where it can more easily be stressed and thereby, to an extent, avoid case ambiguity.

7.6 Conclusions

The current project was essentially exploratory in nature. It set out to define in detail the present geographical distribution of variants of the pDit using Twitter and in so doing demonstrate the applicability of Twitter to the task at hand, its speech-like nature, and its promise for future dialectal research.

It was hoped that providing such a detailed map might shed light on geohistorical trends and consolidate recent traditional corpus studies, while also providing data on the status of pronouns and a possible distinction between the two verbs under investigation.

To a large extent, these aims have been addressed successfully. The data presented here challenge a problematic tendency to 'lump together' large, linguistically diverse regions and treat them as one entity (Siewierska & Hollmann, 2007, p.97). However, the data throw up many more questions than they answer and there are many possibilities for further investigation. Some of these are explored in the following chapter.

[6] Quantitative harmonic alignment is defined by Bresnan and Ford (2010, p.181) as "the existence of a statistical pattern in which, all else being equal, animate, definite, pronominal, discourse-accessible, and shorter arguments tend to precede inanimate, indefinite, nonpronominal, less discourse-accessible, or longer arguments in both of the dative constructions".

Chapter 8

Future directions

8.1 More data

As was discussed (section 5.4), not as much data were gathered as had been anticipated. Whilst there were enough data to provide substantial new insight into the geohistorical distribution of the pDit, more data would be beneficial. The ideal would be, getting enough data to view regionally-specific variation in linguistic/pragmatic constraints, and being able to compare the patterns found across regions. Again, as discussed, one of the reasons for the reduction in data gathered was the restriction to only GPS-encoded tweets for this dataset. Now that this restriction has been lifted, using the public Search API alone will gather approximately ten times the amount of data in the same time period.

Alternatively, paying for Firehose access using Twitter's commercial data service 'GNIP' would allow instant viewing of much larger datasets, and the capability of viewing historical Twitter data. This route will be investigated in the upcoming PhD.

8.2 Using Python and a 'Part of speech tagger' (POS)

The method used to gather data in the current study, TAGS, had the advantage of running on Google's servers and not requiring a machine to be left running. However, the main advantage of building a Python script to access the API and leaving it running for several weeks/months is that doing so allows for the gathering

of all Twitter data coming through the API rather than just the data containing a given search string. These data can then be tagged using a part of speech tagger (POS). There are several POSes that have already been 'trained' on Twitter data. Having a fully POS-tagged dataset of *all* tweets sent over a given period (one to two months) would allow for full comparison with historical data. Crucially, the full set of possible ditransitives could be searched and compared.

In addition, the dataset could be searched for other structures. Given what has been argued in this paper regarding the amenability of "digitally mediated vernaculars" (Shortis, 2016, p.487) to host aspects of spoken vernaculars, such a dataset would likely be a powerful resource for dialectologists. Datasets have already been gathered and tagged, of course, and these datasets are sitting on local machines in universities in the UK, and are prevented from being shared by Twitter's sharing policies. A way round this might be to produce a 'front-end' that permits searches of a dataset without access to the dataset itself. Another possible solution would be to co-author a paper with someone who already has a tagged dataset.

8.3 Expansion of the pilot survey

As explained in the introduction, the second part of the pilot study (discussed in chapter 4, section 4.3) involved the distribution by email of an online grammaticality judgement survey. The results of this survey provided a dataset which could be compared against the Twitter data. A more extensive survey, distributed more widely and involving the gathering of more metadata, would be beneficial. As discussed in the literature (Cornips & Corrigan, 2005; Siewierska & Hollmann, 2007), the comparison of grammaticality judgement data and corpus data offers us the opportunity to arrive at a more comprehensive understanding of regional variation in syntax/semantics/pragmatics.

8.4 Probabilistic syntax and structural persistence

The probability with which a given structure will occur in speech has been shown to inform the predictive capacity of speakers in their choice of a given variant (cf. Bresnan et al., 2007). Whilst such approaches have tended to neglect regional vari-

ation, recent studies by Bresnan and Ford (cf. 2010) have explored probabilistic differences within varieties of a given language. It has also been shown that children use distributional data to make linguistic choices and in so doing, perpetuate those distributions (De Marneffe et al., 2012).

An obvious application of the current data is to such a probabilistic account, allowing for a predictive syntax that is regionally sensitive and is likely to be detectable in the individual in psycholinguistic studies along the lines of that of Bresnan and Ford (2010). For example, a speaker from the area found to exhibit high TGD use is likely to anticipate — in a measurable way — the occurrence of the TGD in example sentences presented in an experimental setting.

The border region between Leeds and Huddersfield[1] and the region between Manchester and Liverpool warrant closer investigation. Using the this kind of psycholinguistic study alongside more traditional sociolinguistic interviews may be revealing.

8.5 Semantics, pragmatics and regional variation

Where several semantically equivalent features appear in a pattern of stable variation, we expect to see pragmatic distinctions in their use. We would expect, following the results of the current study, that such pragmatic distinctions would be particular to each super-region. Anecdotally, for example, the PDAT to a Scottish speaker has been reported as belonging to a more formal register, whereas in the South there is no overt distinction relating to formality.

Deeper corpus analysis of texts from each region looking at the pragmatic contexts of each pDit type would be an obvious way to examine regionally-specific pragmatic distinctions like this. Additionally, controlled experimental study might allow us to detect pragmatic distinctions.

[1]Shown in the extract from the *Batchgeo* map in figure 7.1 in the discussion.

8.6 Corpus of Early English Correspondence (PCEEC) and the Corpus of Scottish Correspondence (CSC)

A closer inspection of historical corpora, particularly comparing the Corpus of Early English Correspondence (PCEEC) to the Corpus of Scottish Correspondence (CSC), might provide some validation of the working hypotheses developed in the discussion. If the effect of Norman French on Old and Middle English were to have catalysed the spread of the PDAT to a greater range of verbs, this might be supported by such analyses. In addition, following Kopaczyk (2013), it would be beneficial to consider more carefully the situation of societies — that is, their status in relation to one another, economic power, political relations etc. — through the historical periods in question.

8.7 Final thoughts

The project completed here for the degree of Masters by Research shows how the present day habits of speakers engaged in online communications on Twitter, a 21st century technology, can reveal robust linguistic patterns that stretch back across time, providing a window to the past. The project represents the first part of a bigger, more ambitious PhD project that pairs Twitter data with judgement data and on-the-ground fieldwork.

Appendices

Corpora used for study of pronominal ditransitives (Gerwin 2013)			
Corpus	Period	Size (m)	Content
Freiburg English Dialect Corpus (FRED)	1968-2000	2.5	Interviews (mean age = 75)
British National Corpus (BNCweb)	1980-1993	100	spoken, fiction magazines, newspapers academic
Corpus of Historical American English (COHA)	1810-2010	400	fiction, magazines newspapers, non-fiction

Table A1: List of corpora used for study of pronominal ditransitives in Gerwin (2013)

Corpora used for study of pronominal ditransitives (Siewierska & Hollmann 2007)			
Corpus	Period	Size (t)	Content
Freiburg English Dialect Corpus (FRED)	1968-2000	250	23 spoken interviews
British National Corpus (Lancashire)	1980-1993	150	ten spoken texts
Survey of English Dialects Incidental recordings (Lancashire)	1950s-1960s	22	spoken texts
Helsinki Corpus of British English Dialects (Lancashire)	1970s-Present	50	spoken texts

Table A2: List of corpora used for study of pronominal ditransitives in Siewierska & Hollmann (2007).

Corpora used for study of pronominal ditransitives (Yañez-Bouza and Denison 2015)			
Corpus	Period	Size (m)	Content
Corpus of Early English Correspondence (PCEEC)	1410-1695	2.16	letters
Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)	1500-1710	1.74	multi-register
Salamanca Corpus	1500-1951	1.25	dialect literature
Corpus of English Dialogues (CED)	1560-1760	1.18	speech-related registers
A Representative Corpus of Historical English Registers (ARCHER 3.2)	1600-1999	1.96	multi-register
Penn Parsed Corpus of Modern British (PPCMBE)	1700-1914	0.95	multi-register
Corpus of Late 18th-Century Prose	1761-1790	0.30	letters
Corpus of Nineteenth Century English(CONCE)	1800-1900	0.99	multi-register
Corpus of Late Modern Prose	1861-1919	0.10	letters
Helsinki Archive of Regional English Speech - Cambridge Sampler (HARES-CAM)	1970s-1980s	0.18	interviews
Freiburg English Dialect Corpus 1970-99 Sampler (FREDS)	1970-1999	1.01	interviews
Diachronic Electronic Corpus of Tyneside English (DECTE)	1960s-70s 1990s, 2001-11	0.81	interviews

Table A3: Corpora used for study of pronominal ditransitives in Yañez-Bouza and Denison (2015).

Type	String (entered in main search box)	Location (entered in 'script editor')
PDAT	"sent it to me" OR "sent it to you" OR "sent it to him" OR "sent it to her" OR "sent it to them" OR "sent it to us" OR "gave it to me" OR "gave it to you" OR "gave it to him" OR "gave it to her" OR "gave it to them" OR "gave it to us"	"geocode": "52.95478319999999, -1.1581085999999914,300mi"
GTD	"sent me it" OR "sent you it" OR "sent him it" OR "sent her it" OR "sent them it" OR "sent us it" OR "gave me it" OR "gave you it" OR "gave him it" OR "gave her it" OR "gave them it" OR "gave us it"	"geocode": "52.95478319999999, -1.1581085999999914,300mi"
TGD	"sent it me" OR "sent it you" OR "sent it him" OR "sent it her" OR "sent it them" OR "sent it us" OR "gave it me" OR "gave it you" OR "gave it him" OR "gave it her" OR "gave it them" OR "gave it us"	"geocode": "52.95478319999999, -1.1581085999999914,300mi"

Table A4: Search strings used in TAGS.

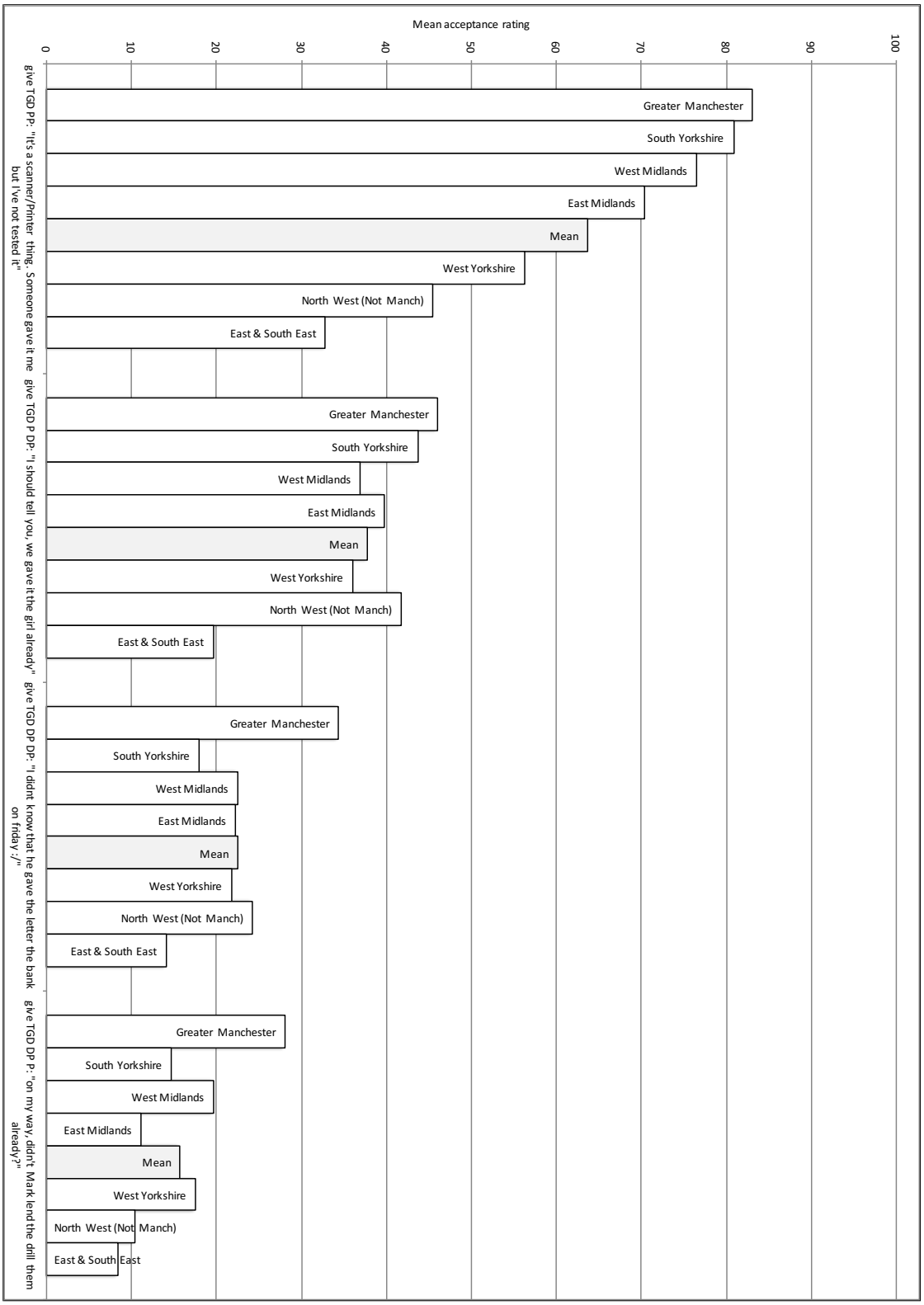

Figure A1: TGD acceptability scores in England by sentence type and region.

References

Adger, D., & Trousdale, G. (2007). Variation in English syntax: Theoretical implications. *English Language*, *11*(02), 261-278. doi: 10.1017/s1360674307002250

Barbiers, S. (2005). Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics*. In K. P. Corrigan & L. E. A. Cornips (Eds.), *Syntax and variation: Reconciling the biological and the social* (p. 233-264). Amsterdam; Philadelphia: John Benjamins.

Baron, N. S. (1998a). Letters by phone or speech by other means: the linguistics of email. *Language & Communication*, *18*(2), 133-170.

Baron, N. S. (1998b). Writing in the age of email. *Visible Language*, *32*(2).

BatchGeo, L. (2012). *Batchgeo.* Retrieved 01-06-17, from `https://batchgeo.com`

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145-204.

Biber, D., Gray, B., & Staples, S. (2016). Contrasting the grammatical complexities of conversation and academic writing: Implications for EAP writing development and teaching. *Language in Focus*, *2*(1), 1-18.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Harlow, Essex: Pearson Education.

Biggs, A. (2014). Passive variation in the dialects of Northwest British English. University of Zurich. Paper presented at the 3rd Conference of the International Society for the Linguistics of English (ISLE). Retrieved from `http://www.isle-linguistics.org/resources/Biggs--passive_variation--2014.pdf`

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355-387. doi: 10.1016/0010-0285(86)90004-6

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In *Cognitive foundations of interpretation* (p. 69-94). Amsterdam: Royal Netherlands Academy of Science Amsterdam.

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

Britain, D. (2013). Space, diffusion and mobility. In D. Britain, J. K. Chambers, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (p. 471-500). Chichester: Wiley-Blackwell.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Cornips, L., & Corrigan, K. (2005). Convergence and divergence in grammar. In P. Auer, F. Hinskens, & P. Kerswill (Eds.), *Dialect change: convergence and divergence in European languages* (p. 96-134). Cambridge: Cambridge University Press.

Crisma, P., & Longobardi, G. (2009). Change, relatedness, and inertia in historical syntax. In P. Crisma & G. Longobardi (Eds.), *Historical syntax and linguistic theory* (p. 1-13). Oxford: Oxford University Press.

De Cuypere, L. (2014, 11). The old English to-dative construction. *English Language and Linguistics*, *19*(01), 1-26. doi: 10.1017/s1360674314000276

De Marneffe, M.-C., Grimm, S., Arnon, I., Kirby, S., & Bresnan, J. (2012). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes*, *27*(1), 25-61. doi: 10.1080/01690965.2010.542651

District, T. (2013, 5). University of cambridge research. Retrieved from `http://www.cam.ac.uk/research/features/welsh-twitter-capturing-language-change-in-real-time`

Donath, J. S. (1999). Identity and deception in the virtual community. In J. S. Donath, M. A. Smith, & P. Kollock (Eds.), *Communities in cyberspace* (p. 27-58). London: Routledge.

Doyle, G. (2014). Mapping dialectal variation by querying social media. In *Presented at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (p. 98-106). Gothenburg, Sweden: Association for

Computational Linguistics.

Eisenstein, J. (2013). Phonological factors in social media writing. In *Proceedings of the NAACL/HLT 2013 workshop on language analysis in social media (LASM 2013)* (p. 11-19). Atlanta, GA: LASM.

Eisenstein, J. (2017). Identifying regional dialects in online social media. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *The handbook of dialectology* (p. 368-383). Wiley-Blackwell.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLOS ONE*, *9*(11), e113114.

Fabian, J. (1983). *Time and the other: How anthropology makes its object.* New York: Columbia University Press.

Feenberg, A. (1989). The written world: On the theory and practice of computer conferencing. In R. Mason & A. Kaye (Eds.), *Mindweave: Communication, computers, and distance education* (p. 22-39). Oxford: Pergamon Press.

Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication*, *8*(1), 8-34.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE Publications.

Gast, V. (2007). *I gave it him* - on the motivation of the alternative double object construction in varieties of British English. *Functions of Language (special issue: Ditransitivity)*, *14*(1), 31-56.

Gerwin, J. (2013). Give it me!: pronominal ditransitives in English dialects. *English Language and Linguistics*, *17*(3), 445-463.

Gerwin, J. (2014). *Ditransitives in British English dialects* (Vol. 50). Berlin: Walter de Gruyter.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012). Assessing the bias in communication networks sampled from Twitter. arxiv preprint. *arXiv preprint arXiv:1212.1684*.

Haddican, W. (2010). Theme–goal ditransitives and theme passivisation in British English dialects. *Lingua*, *120*(10), 2424-2443.

Hawksey, M. (2014). *Tags.* Retrieved 2016-05-23, from `https://tags.hawksey.info/get-tags/`

Hollmann, W., & Siewierska, A. (2006). Corpora and (the need for) other methods in a study of Lancashire dialect. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), 203-216.

Honeycutt, C., & Herring, S. C. (2009). Beyond microblogging : Conversation and collaboration via twitter. In *Proceedings of the 42nd Hawaii international conference on system sciences* (p. 1-10).

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2015). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*. doi: 10.1016/j.compenvurbsys.2015.12.003

Hughes, A., Trudgill, P., & Watt, D. (2012). *English accents & dialects: An introduction to social and regional varieties of English in the British Isles* (5th ed.). London: Hodder Education.

Java, A., Song, X., Finin, T., & Tseng, B. (2009). Why we twitter: An analysis of a microblogging community. In H. Zhang et al. (Eds.), *Advances in web mining and web usage analysis* (pp. 118–138). Berlin, Heidelberg: Springer.

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using black Twitter. *American Speech*, *90*(4), 403-440.

Kerswill, P. (1994). *Dialects converging: rural speech in urban Norway*. Oxford: Oxford University Press.

Kirk, J. M. (1985). Linguistic atlases and grammar: The investigation and description of regional variation in English syntax. In J. M. Kirk & S. Sanderson (Eds.), *Studies in linguistic geography : The dialects of English in Britain and Ireland* (p. 130-135). London ; Dover, N.H.: Croom Helm.

Kopaczyk, J. (2013). Rethinking the traditional periodisation of Scots. In R. McColl Millar & J. Cruickshank (Eds.), *Selected papers from the forum for research on the languages of Scotland and Ulster, Aberdeen, July 2012.*

Kortmann, B. (2003). Comparative English dialect grammar: a typological approach. *Fifty Years of English Studies in Spain (1952: 2002). A Commemorative Volume, Santiago de Compostela: University of Santiago*, 63-81.

Kortmann, B. (2004). Why dialect grammar matters. *The European English Messenger*, *XIII*, 24-29.

Kretzschmar, W. A. (1999). The future of dialectology. *Leeds Studies in English*,

30, 271.

Kurath, H. (1973). *Studies in area linguistics.* Indiana University Press.

Labov, W. (1972). *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.

Langton, J., & Morris, R. (Eds.). (1986). *Atlas of industrializing Britain 1780–1914.* London: Methuen.

Larson, R. K. (1988). On the double object construction. *Linguistic Inquiry*, *19*(3), 335-391.

Leemann, A., & Blaxter, T. (2016). *Cambridge app maps decline in regional diversity of English dialects.* Retrieved 2016-12-06, from `http://www.cam.ac.uk/research/news/cambridge-app-maps-decline-in-regional-diversity-of-english-dialects`

Leemann, A., Kolly, M.-J. J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing language change with smartphone applications. *PLOS ONE*, *11*(1), e0143060. doi: 10.1371/journal.pone.0143060

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago: University of Chicago Press.

MacKenzie, L., Bailey, G., & Turton, D. (2014). *Our dialects: Mapping variation in English in the UK.* Retrieved from `http://projects.alc.manchester.ac.uk/ukdialectmaps/`

Milroy, J., & Milroy, L. (2012). *Authority in language: Investigating standard English* (2nd ed.). Abingdon: Routledge. (Original work published 1985)

Mitchell, B. (1985). *Old English syntax* (Vol. 1). Oxford: Clarendon Press.

Page, R. (2012a). The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication*, *6*(2), 181-201. doi: 10.1177/1750481312437441

Page, R. (2012b). *Stories and social media: Identities and interaction.* Abingdon: Routledge.

Schlobinski, P. (2005). Mündlichkeit/Schriftlichkeit in den neuen Medien. In L. Eichinger & W. Kallmeyer (Eds.), *Standardvariation: Wie viel variation verträgt die deutsche Sprache?* (p. 126-142). Berlin: Walter de Gruyter.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments*

and linguistic methodology. Chicago: University of Chicago Press.

Schütze, C. T., & Sprouse, J. (2014). Judgement data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (p. 27-50). Cambridge: Cambridge University Press.

Shortis, T. (2015). *Orthographic practices in SMS text messaging as a case signifying diachronic change in linguistic and semiotic resources* (Doctoral dissertation, The UCL Institute of Education, University College London). Retrieved from `http://discovery.ucl.ac.uk/1485733/1/Shortis_FINAL_POST-VIVA_THESIS_240416_compressed.pdf`

Shortis, T. (2016). Texting and other messaging: Written system in digitally mediated vernaculars. In V. Cook & D. Ryan (Eds.), *The Routledge handbook of the English writing system* (p. 487-511). Abingdon: Routledge.

Siewierska, A., & Hollmann, W. (2007). Ditransitive clauses in English with special reference to Lancashire dialect. In M. Hannay & G. J. Steen (Eds.), *Structural-functional studies in English grammar: in honour of Lachlan Mackenzie* (Vol. 83, p. 83-102). Amsterdam; Philadelphia: John Benjamins Publishing.

Sinclair, J. (1782). *Observations on the Scottish dialect.* London: Printed for W. Strahan, and T. Cadell.

Smout, T. C. (2005). *Anglo-Scottish relations from 1603 to 1900* (Vol. 127). Oxford: Oxford University Press.

Squires, L. (2016). Computer-mediated communication and the English writing system. In V. Cook & D. Ryan (Eds.), *The Routledge handbook of the English writing system* (p. 471-486). Abingdon: Routledge.

Stevenson, J. (2015). *Send it me later: investigating geographical variation in the acceptability of the theme-goal ditransitive* (BA dissertation, University of York).

Tagliamonte, S. (2014a). A comparative sociolinguistic analysis of the dative alternation. In R. T. Cacoullos, N. Dion, & A. Lapierre (Eds.), *Linguistic variation: confronting fact and theory* (pp. 297–318). New York: Routledge.

Tagliamonte, S. (2014b, June). Sociolinguistics for computational social science. Baltimore, Maryland, USA. Paper presented at the 52nd Annual Meeting of the Association for Computational Linguistics.

Upton, C. (2006). Modern regional English in the British Isles. In L. Mugglestone (Ed.), *The Oxford History of English* (p. 379-414). Oxford: Oxford University Press.

Viereck, W. (1986). Dialectal speech areas in England: Orton's phonetic and grammatical evidence. *Journal of English Linguistics*, *19*(2), 240-257. doi: 10.1177/007542428601900206

Volk, C., Bresnan, J., Rosenbach, A., & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, *3*(30), 382-419.

Werry, C. C., & Herring, S. (1996). Linguistic and interactional features of internet relay chat. In *Computer-mediated communication: linguistic, social and cross-cultural perspectives* (p. 47-63). Amsterdam: John Benjamins Publishing.

Willis, D. (2013). *Syntactic atlas of Welsh dialects.* Retrieved 2016-11-06, from `http://lion.ling.cam.ac.uk/david/sawd/index.html`

Yáñez Bouza, N. (2016). "May depend on me sending it you": Double objects in early grammars. *Journal of English Linguistics*, *44*(2), 138-161. doi: 10.1177/0075424216630793

Yáñez Bouza, N., & Denison, D. (2015). Which comes first in the double object construction? *English Language and Linguistics*, *19*(02), 247-268. doi: 10.1017/s136067431500012x

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing: A corpus based study. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (Vol. 39, p. 29-46). Amsterdam: John Benjamins Publishing.

Yoshikawa, F. (2006). The periphrastic dative and the Wycliffite bible. *Studies in the Humanities and Sciences*, *36*(2), 33-54.