The dataset covers the so-called “dative alternation”. The dative alternation (also referred to as the ditransitive or double-object construction) refers to parallel constructions that have broadly similar meaning but different syntax:

i. “he gave it to the board”
ii. “I gave her my old one”

In i., the verb “give” takes a noun phrase (the pronoun “it”) and a prepositional phrase as arguments (“to the board”), while in ii. the verb takes two noun phrases (“her” and “my old one”) as arguments. In the dataset, we refer to i) as “VNPP” and ii) as “VNN”. We refer to the indirect object role as “recipient” (“her” in i. and “the board” in ii.) and the direct object as “theme” (“it” in i. and “my old one” in ii).

The dataset was collected from the Early-Access Subset (EAS) of the Spoken British National Corpus 2014 [1] for use in a sociolinguistic study of competing linguistic constructions [2]. The corpus is now publicly available via CQPweb at https://cqpweb.lancs.ac.uk/bnc2014spoken.

The dative alternation is a topic of active research in linguistics, but few studies have made datasets available. Meta-studies, creation of specialised NLP tools, and comparisons of results will benefit from better access to this dataset.

Keywords: Linguistics; Linguistic analysis; Sociolinguistics; Speech; dative
In a subsequent step, the concordance results were enriched with metadata annotation, downloaded separately from the corpus interface, and with speaker information, provided as a spreadsheet. This step was automated by means of a Python script which combines the semantic data exported from CQPweb with the data containing the manually annotated syntactic patterns from the corpus. Further data cleaning, primarily for increasing consistency in the annotation, was carried out using R [4]. Both the Python script and the R script are available as supporting files alongside the dataset.

### (3) Dataset description

#### Object name

BNCspoken2014_dative_dataset_v1.csv

#### Format names and versions

Version 1, comma-separated (csv) file

### Creation dates

2016-08-21–2016-10-16

### Dataset Creators

1. Jenset, Gard B.; (data curation, investigation, formal analysis, conceptualisation, software, methodology)
2. McGillivray, Bárbara (data curation, investigation, formal analysis, conceptualisation, software, methodology)
3. Rundell, Michael (data curation, methodology)

### Language

This dataset consists of 1840 observations of transcribed informal spoken British English, along the following 44 variables. Each observation corresponds to an occurrence of the verbs give, lend, show, send, sell, and offer in the BNC Spoken 2014 corpus. Missing values are coded as “NA” for compatibility with R.

### Linguistic variables

| Variable     | Description                                                                 | Example relative to the sentence |
|--------------|------------------------------------------------------------------------------|-----------------------------------|
| **Verb**     | The verb lemma, one of “give”, “lend”, “show”, “send”, “offer”, and “sell”. | send                              |
| **VerbSemTag** | The semantic tag of the verb, obtained from the corpus semantic annotation, based on UCREL [5] semantic analysis system USAS; tags are available at http://ucrel.lancs.ac.uk/usas/semtags.txt | M2 (‘Putting, taking, pulling, pushing, transporting &c.’) |
| **Pattern**  | The observed dative construction, one of “VNPP” or “VNN”                    | VNPP                              |
| **Recipient** | The recipient's noun phrase                                                  | people you don’t see              |
| **RecLen**   | The number of characters in the recipient                                    | 21                                |
| **RecHead**  | The recipient's syntactic head                                               | people                            |
| **RecPrn**   | Boolean defined programmatically based on the semantic tag of the recipient. If the semantic tag is ‘Z8’, the value is TRUE; otherwise, the value if FALSE. | NA                                |
| **RecSemTag** | String with the UCREL [5] semantic tag of the recipient's syntactic head    | S2 (‘people’)                     |
| **AnimateRec** | Boolean indicating whether the recipient’s head is animate (TRUE) or inanimate (FALSE). This was manually annotated | FALSE                            |
| **Theme**    | String with the theme's noun phrase                                          | Christmas cards                   |
| **ThemeLen** | The number of characters in the theme                                         | 15                                |
| **ThemeHead** | String with the theme's syntactic head                                       | cards                             |
| **ThemePrn** | Boolean defined programmatically based on the semantic tag of the theme. If the semantic tag is ‘Z8’, the value is TRUE; otherwise, the value if FALSE. | FALSE                            |
| **ThemeSemTag** | String with the UCREL semantic tag of the theme's syntactic head             | Q1 (‘LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION’) |
| **ThemeField** | First letter of the semantic tag of the theme's syntactic head.              | Q                                 |
| **DefTheme** | Boolean indicating if the theme is expressed as a definite phrase (TRUE) or indefinite (FALSE) | FALSE                            |
| **AnimateTheme** | Boolean indicating whether the theme's head is animate (TRUE) or inanimate (FALSE) | FALSE                            |
| Variable          | Description                                                      | Example                                                                 |
|-------------------|------------------------------------------------------------------|------------------------------------------------------------------------|
| NumSpeakers       | Number of speakers in the conversation                           | Texts with 2 speakers                                                 |
| Location          | Location where the conversation took place                       | Speakers' home                                                         |
| Relation          | Relationship between the speakers in the conversation            | Close family, partners, very close friends                            |
| Subject           | Subject of conversation                                          | Mother and daughter talking about theatre                             |
| Topics            | Topics covered in the conversation                               | Theatre, Disney films, websites, post, Christmas, jobs                |
| ExactAge          | Exact age of the main speaker in the conversation                | 44                                                                     |
| AgeRange          | The age range of the main speaker in the conversation            | 40_49                                                                 |
| AgeRangeMid       | Mid-point of the age range of the main speaker in the conversation | 45                                                                    |
| AgeImputed        | Equals the exact age of the main speaker in the conversation if it is recorded; it is the mid-point of the age range of the main speaker in the conversation, if the age range is recorded but not the exact range; otherwise, NA. This variable is automatically calculated | 44                                                                     |
| Gender            | Gender of the main speaker in the conversation (M or F)          | F                                                                     |
| Nationality       | Nationality of the main speaker in the conversation              | British                                                               |
| BirthCountry      | Country of birth of the main speaker in the conversation         | England                                                               |
| L1                | First language of the main speaker in the conversation           | English                                                               |
| LingOrigin        | Country of linguistic origin of the main speaker in the conversation | England                                                              |
| Accent            | Accent of the main speaker in the conversation                   | South East England                                                    |
| City              | City where the conversation took place                           | High Wycombe                                                          |
| Country           | Country where the conversation took place                         | England                                                               |
| Level1Dialect     | First level of granularity in the categorization of the dialect of the main speaker in the conversation | uk                                                                    |
| Level2Dialect     | Second level of granularity in the categorization of the dialect of the main speaker in the conversation | english                                                              |
| Level3Dialect     | Third level of granularity in the categorization of the dialect of the main speaker in the conversation | south                                                                |
| Level4Dialect     | Fourth level of granularity in the categorization of the dialect of the main speaker in the conversation | southeast                                                            |
| SpeakerHighestQual| Highest qualification of the main speaker in the conversation     | Graduate                                                              |
| Occupation        | Occupation of the main speaker in the conversation               | Team leader                                                           |
| SpeakerSocGrade   | Social grade of the main speaker in the conversation, according to the classification developed by the National Readership Survey ([https://web.archive.org/web/20110303033539/http://www.nrs.co.uk/lifestyle.htm](https://web.archive.org/web/20110303033539/http://www.nrs.co.uk/lifestyle.htm)) | E                                                                     |
| ForeignLangs      | Foreign languages spoken by the main speaker in the conversation  | French–level unspecified; Spanish–level unspecified                    |
| NumUtterances     | Number of utterances of the conversation's main speaker in the whole corpus | 99                                                                    |
| NumWords          | Number of words uttered by the conversation's main speaker in the whole corpus | 1622                                                                  |
(4) Reuse potential

There is a growing trend in linguistics for quantitative research, a trend which is not occurring at the same pace across all branches of linguistics [6]. A natural corollary of this increasing quantitative research is a focus on replicable and reproducible research [7].

True replicability is difficult to achieve in many field-based disciplines and social sciences [7]. A more achievable goal is reproducibility. Reproducibility is clearly important for increasing scientific transparency and accountability. A move towards greater reliance on usage-based theory development can drive convergence in linguistic theory generally [8] as well as in specific sub-fields [6]. Despite some notable exceptions (such as second language acquisition), most linguistic sub-fields do not have a strong tradition for making research data available [7]. Publishing not only corpora and raw data, but also the annotated research datasets means that data can be compared quantitatively across research traditions, or pooled into meta-studies for greater theoretical insights.

For linguistics, and in particular corpus linguistics, the aim of reproducibility requires not only access to raw corpus data, but also to manually retrieved, annotated, and categorised data. Despite advances in computational linguistics, automatic annotation tools still fall short in theoretically important areas such as pragmatics and semantics. In the case of transcribed spoken text, the challenges are compounded by the nature of spoken language. Moreover, parsing tools for automatic syntactic analysis are still not performing as well as on such data as on written text. As a result, manual annotation is in many cases inevitable.

Another effect of the required manual effort is that the annotated research datasets remain comparatively small. From this observation, two further use cases for shared data automatically follow. First, by pooling together different datasets, the resulting increase in statistical power may allow researchers to draw new conclusions based on correlations that remained obscure in smaller datasets. Second, despite great advances in the range of available NLP tools, there are still gaps when it comes to specialised but valuable tasks such as annotating linguistic data for a specific construction. The problem with training data for NLP tools is more commonly associated with historical linguistics [6, 20]. However, much of the freely available NLP data stem from written, not spoken language. Furthermore, any specific task for which training data is required will require specific training data, and such data will often be scarce due to the cost involved in manual annotation.

By publishing the dative alternation data, we contribute to all these reusability scenarios. The dative alternation is a topic of active research in linguistics, not least because it has been studied from different theoretical traditions. The dative alternation is a prominent example of the convergence of different theoretical and empirical research questions in linguistics, providing evidence for the motivations behind the linguistic decisions that speakers make [9]. It is well established that both syntactic and pragmatic factors (especially discourse-new versus discourse-old information) play a role in choosing between the two constructions i. and ii. as shown in [10] and [11]. Later studies have confirmed these findings while adding further nuance. The semantics of the verb arguments also plays a role [12], and there is agreement that, on the whole, the dative alternation is subject to broadly similar constraints across different macro-varieties of English [2, 13–17].

Despite this activity, the dative alternation continues to draw theoretical and empirical attention in linguistics, with a number of relevant and underexplored questions remaining. These include questions of linguistic prototypicality [16], the role of probability in spoken grammar [17], and the role of individual-level sociolinguistic factors [2].

Despite the interest in the dative alternation, few data-sets from the published literature have been made publicly available. One notable exception is the dataset from [11], which was made available in an R package in 2008 [18] and re-used for didactic purposes in [19]. Another recent exception is [17].

By publishing this dataset we contribute to advancing the awareness of the need for reproducibility in linguistics, and specifically the progress of empirical research on the English dative alternation.

Competing Interests

The authors have no competing interests to declare.

References

1. Love, R, Dembry, C, Hardie, A, Brezina, V and McEnery, T. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics. 2017; 22(3): 319–344. DOI: https://doi.org/10.1075/ijcl.22.3.02lov
2. Jenset, G B, McGillivray, B and Rundell, M. The English dative alternation revisited: Fresh insights from contemporary British spoken data. In: Brezina, V, Love, R and Aijmer, K (eds.), Corpus approaches to contemporary British speech: Sociolinguistic studies of the spoken BNC2014. London: Routledge; 2018; 185–207. DOI: https://doi.org/10.4324/9781315268323-10
3. Hardie, A. Using the Spoken BNC2014 in CQPweb. In: Brezina, V, Love, R and Aijmer, K (eds.), Corpus approaches to contemporary British speech: Sociolinguistic studies of the spoken BNC2014.
Jenset and McGillivray: Data from 'The Dative Alternation Revisited: Fresh Insights from Contemporary British Spoken Data'. Journal of Open Humanities Data, 5: 1. DOI: https://doi.org/10.5334/johd.11

Published: 22 August 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 Unported License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Jenset, G B and McGillivray, B 2019 Data from 'The Dative Alternation Revisited: Fresh Insights from Contemporary British Spoken Data'. Journal of Open Humanities Data, 5: 1. DOI: https://doi.org/10.5334/johd.11

How to cite this article: Jenset, G B and McGillivray, B 2019 Data from 'The Dative Alternation Revisited: Fresh Insights from Contemporary British Spoken Data'. Journal of Open Humanities Data, 5: 1. DOI: https://doi.org/10.5334/johd.11

Published: 22 August 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 Unported License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS