Journal of Open Humanities Data

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

DATA PAPER

Data from 'The Dative Alternation Revisited: Fresh Insights from Contemporary British Spoken Data'

Gard B. Jenset[1] and Barbara McGillivray[2]

[1] Independent researcher, GB

[2] The Alan Turing Institute/University of Cambridge, GB

Corresponding author: Gard B. Jenset (gjenset@gmail.com)

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

(1) Overview

Repository location

https://figshare.com/articles/BNCspoken2014_dative_dataset_v1_csv/7353164.

Context

This data was produced as part of the research for a publication on the English dative alternation for spoken data [2].

(2) Methods

The EAS is composed of transcripts of spontaneous conversations, recorded in the period 2012–2014 and contains over 4 million tokens. The corpus also contains rich metadata about the speakers and the context of the conversation [1].

The EAS provides a rich opportunity for studying linguistic phenomena in a deeper sociolinguistic context. The dataset presented here deals with the so-called English dative alternation. To identify such constructions, we manually queried the EAS via the CQPweb interface, an online corpus query and analysis system [3]. The queries were carried out for six frequent verbs that occur with both dative alternation patterns [2]: *give*, *lend*, *show*, *send*, *offer*, and *sell*.

Although these queries used the EAS, they can be reproduced in the full BNC2014 corpus via the metadata tag *Sample release inclusion* (available in CQPweb and the underlying XML). The queries produced six intermediate sets of results with concordance lines containing a limited surrounding context for each occurrence of the target verb in the corpus. These intermediate result sets were saved manually as separate spreadsheet files.

The raw files were manually examined, and the rows that did not correspond to either of the constructions in i. or ii. were filtered out. Examples of omitted results are phrasal verbs like "give up" and idioms such as "give a shit". The remaining results were manually annotated for two syntactic patterns exemplified in i. and ii. Additionally, the syntactic head of the noun phrase arguments (e.g. "board" in the phrase "the board") were manually identified. The corpus markup does not include annotation for the relevant features we required. Experiments with automated syntactic annotation using tools trained on written English data did not yield good results, and for lack of appropriate training data a manual annotation process was followed. Moreover, we manually annotated the lemmas of recipients and themes with information about animacy. The resulting file is available as a supporting file alongside the dataset.

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, Barbara (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 *just send Christmas cards ... to people you don't see from year to year*
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

Metadata

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. This variable is automatically calculated	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)	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

License
CC BY 4.0

Repository name
Figshare

Publication date
2018-11-16

(4) Reuse potential

There is a growing trend in linguistics for quantitative research, a trend which is not proceeding at the same pace in 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 statistical 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 datasets 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*.

2018; 27–30. London: Routledge. DOI: https://doi.org/10.4324/9781315268323-4
4. **R Core Team.** R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/.
5. **Rayson, P, Archer, D, Piao, S** and **McEnery, A M.** The UCREL semantic analysis system; 2004.
6. **Jenset, G B** and **McGillivray, B.** Quantitative historical linguistics: A corpus framework. Oxford: Oxford University Press; 2017. DOI: https://doi.org/10.1093/oso/9780198718178.001.0001
7. **Berez-Kroeker, A L, Gawne, L, Kung, S S, Kelly, B F, Heston, T, Holton, G,** et al. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*. 2018; 56(1): 1–18. DOI: https://doi.org/10.1515/ling-2017-0032
8. **Geeraerts, D.** Methodology in cognitive linguistics. In: Kristiansen, G, Achard, M, Dirven, R and Ruiz de Mendoza, F (eds.), *Cognitive linguistics: Current applications and future perspectives*. 2006; 21–49. Berlin: Mouton De Gruyter.
9. **Arppe, A, Gilquin, G, Glynn, D, Hilpert, M** and **Zeschel, A.** Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora*. 2010; 5(1): 1–27. DOI: https://doi.org/10.3366/cor.2010.0001
10. **Arnold, J E, Losongco, A, Wasow, T** and **Ginstrom, R.** Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*. 2000; 76(1): 28–55. DOI: https://doi.org/10.1353/lan.2000.0045
11. **Bresnan, J, Cueni, A, Nikitina, T** and **Baayen, R H.** Predicting the dative alternation. In: Bouma, G, Kraemer, I and Zwarts, J (eds.), *Cognitive foundations of interpretation*. 2007; 69–94. Amsterdam: Royal Netherlands Academy of Arts & Sciences.
12. **Jenset, G B** and **Johansson, C.** Lexical fillers influence the dative alternation: Estimating constructional saliency using web document frequencies. *Journal of Quantitative Linguistics*. 2013; 20(1): 13–44. DOI: https://doi.org/10.1080/09296174.2012.754597
13. **Bresnan, J** and **Hay, J.** Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua*. 2008; 118(2): 245–59. DOI: https://doi.org/10.1016/j.lingua.2007.02.007
14. **Bresnan, J** and **Ford, M.** Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*. 2010; 86(1): 168–213. DOI: https://doi.org/10.1353/lan.0.0189
15. **Kendall, T, Bresnan, J** and **Van Herk, G.** The dative alternation in African American English: Researching syntactic variation and change in a conglomerated corpus. *Corpus Linguistics and Linguistic Theory*. 2011; 7(2): 229–44. DOI: https://doi.org/10.1515/cllt.2011.011
16. **Bernaisch, T, Gries, S T** and **Mukherjee, J.** The dative alternation in South Asian English (es): Modelling predictors and predicting prototypes. *English World-Wide*. 2014; 35(1): 7–31. DOI: https://doi.org/10.1075/eww.35.1.02ber
17. **Szmrecsanyi, B, Grafmiller, J, Bresnan, J, Rosenbach, A, Tagliamonte, S** and **Todd, S.** Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics*. 2017; 2(1). DOI: https://doi.org/10.5334/gjgl.310
18. **Baayen, R H.** languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". [Internet]: 2013. Available from: https://CRAN.R-project.org/package=languageR.
19. **Baayen, R H.** Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press; 2008. DOI: https://doi.org/10.1017/CBO9780511801686
20. **McGillivray, B.** Methods in Latin computational linguistics. Leiden: Brill; 2014.

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/.

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

OPEN ACCESS