

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

### Supplementary Materials: Description of the dataset

The dataset includes data for two variables described below. Each data point (row) represents a single tweet containing a second-person singular pronoun. All data points are relevant for Variable 1. Only a subset is relevant for Variable 2. Some data fields (**clause\_type**, **force\_type** and **polarity**) are relevant only for Variable 2, and only those data points relevant for Variable 2 are specified for those fields.

## 1 Variable 1: Form of 2sg. pronoun (chdi)

### 1.1 Short description of the variable

In certain contexts, the second-person singular (informal) pronoun may be either *ti* (or its grammatically conditioned alternant *di*) or *chdi*:

- (1) Chdi / Ti sy 'n ennill.  
2SG be.PRS.3SG.REL PROG win.INF  
'It's you that's winning.'

This pronoun is not marked for case and so this variation may occur in a wide variety of syntactic environments. For further details, see Willis (2017).

### 1.2 Metadata fields

Each data point is marked up in the following way:

**data\_id**: a unique identifier for each data point.

**user\_id**: a unique identifier for each user account.

**variant\_narrow**: the form of the pronoun used, categorized into **ti**, **di** or **chdi**.

**variant\_wide**: the form of the pronoun used, categorized into **ti** or **chdi** (*di* treated as a form of *ti*).

**context\_narrow**: the grammatical context in which the pronoun is found. The possible values are:

**a (and)**: the pronoun is in a co-ordinated noun phrase following *a* 'and'.

**a (with)**: the pronoun is the object of the non-inflecting preposition *â* 'with'.

**am**: the pronoun is the object of the inflecting preposition *am* 'about'.

**ar gyfer**: the pronoun is the object of the complex preposition *ar gyfer* 'for'.

**ar ol**: the pronoun is the object of the complex preposition *ar ôl* 'after'.

**ar**: the pronoun is the object of the inflecting preposition *ar* 'on'.

**at**: the pronoun is the object of the inflecting preposition *at* 'to'.

**auxdrop**: the pronoun is the subject of a present-tense clause in which a finite form of the auxiliary *bod* 'be' (typically *wyt* '(you) are') has been deleted.

**auxdrop/bod:** the pronoun is in a context where it could either be the subject of a present-tense clause in which a finite form of the auxiliary *bod* ‘be’ has been deleted; or it could be the subject of a nonfinite *bod*-clause; for instance:

- (2) Dw i 'n gwybod \_\_\_\_ ti/chdi 'n iawn.  
 be.PRS.1SG 1SG PROG know.INF 2SG PRED right  
 ‘I know that you’re right.’

For further details, see Variable 2 below.

**bod:** the pronoun is the subject of a nonfinite *bod*-clause:

- (3) Dw i 'n gwybod bod ti/chdi 'n iawn.  
 be.PRS.1SG 1SG PROG know.INF be.INF 2SG PRED right  
 ‘I know that you’re right.’

**bu:** the pronoun is the subject of the perfect of the verb ‘be’ *bu-*.

**buasai:** the pronoun is the subject of a conditional form of the verb ‘be’ of the (*bua*)*sai* type.

**bydd:** the pronoun is the subject of the future of the verb ‘be’ *bydd-*.

**byddai:** the pronoun is the subject of a conditional form of the verb ‘be’ of the *byddai* type.

**co:** the pronoun is the complement of the presentative particle *co* (< *acw* ‘yonder’), as in *co ti/chdi...* ‘there you are...’.

**cyn:** the pronoun is the object of the non-inflecting preposition *cyn* ‘before’.

**da:** the pronoun appears in the phrase *da (iawn) ti* ‘good for you (lit. good (very) you)’.

**ddaru:** the pronoun is the subject of the past-tense auxiliary *ddaru*.

**diolch i:** the pronoun is in the phrase *diolch i ti/chdi* ‘thank (to) you’.

**diolch:** the pronoun is in the phrase *diolch ti/chdi* ‘thank you’.

**dros:** the pronoun is the object of the inflecting preposition *dros* ‘over’.

**dylai:** the pronoun is the subject of the modal verb ‘should’ *dylai*.

**dyma:** the pronoun is the complement of the presentative particle *dyna* (cf. French *voici*), as in *dyma ti/chdi...* ‘here you are...’.

**dyna:** the pronoun is the complement of the presentative particle *dyna* (cf. French *voilà*), as in *dyna ti/chdi...* ‘there you are...’.

**efo:** the pronoun is the object of the non-inflecting preposition *efo* ‘with’.

**eisiau:** the pronoun is in a construction with (*gweld*) *eisiau* in the sense ‘miss’, for instance:

- (4) O’n i 'n gweld eisiau ti.  
 be.IMPF.1SG 1SG PROG see.INF absence/need/want 2SG  
 ‘I missed you.’

**elided i-clause:** the pronoun is the subject of a nonfinite clause of a type normally marked in writing by *i* ‘to’, but *i* is not present:

- (5) Dw i isio (i) ti/chdi ddod heno.  
 be.PRS.1SG I want (to) 2SG come.INF tonight  
 ‘I want you to come tonight’.

**fatha:** the pronoun is the complement of *fatha* ‘like’.

**fel:** the pronoun is the complement of *fel* ‘like’.

**focus:** the pronoun appears in clause-initial focus-fronting position, as in (1) above.

**gan:** the pronoun is the object of the inflecting preposition *gan* ‘with, by’.

**greetings:** the pronoun is in a formulaic greeting not listed elsewhere in this list, mostly *hwyl ti* ‘bye (to?) (you)’.

**gyda:** the pronoun is the object of the non-inflecting preposition *gyda* ‘with’.

**heb:** the pronoun is the object of the inflecting preposition *heb* ‘without’.

**heibio:** the pronoun is the complement of *heibio* ‘(go) past’.

**i:** the pronoun is the object of the semi-inflecting preposition *i* ‘to’.

**imperative:** the pronoun is the overt subject of an imperative verb.

**independent:** the pronoun stands in an independent position: as an utterance or sentence fragment on its own (*Ti/Chdi?* ‘You?’), sometimes as the first element in a conjunct (*Ti/Chdi a fi?* ‘You and me?’); in metalinguistic use (e.g. *galw rhywun yn ‘ti’* ‘call someone “ti”’, use informal pronouns with someone’); or as a predicate after the predicate marker *yn*:

- (6)      sw'n                    i            'n            ti  
             be.COND.1SG 1SG    PRED    2SG  
             ‘if I were you’

**lexical verb:** the pronoun is the subject of a lexical verb (i.e. a verb other than the ones listed separately in this list).

**meddai:** the pronoun appears after the quotative marker/verb *meddai* ‘say’.

**mo:** the pronoun is a negative object marked by the inflecting preposition *mo* ‘not of’.

**na:** the pronoun is the complement of *na* ‘than’.

**Nadolig Llawen i:** the pronoun is in the phrase *Nadolig Llawen i ti/chdi* ‘Merry Christmas to you’.

**o flaen:** the pronoun is the object of the complex preposition *o flaen* ‘in front of’.

**o:** the pronoun is the object of the inflecting preposition *o* ‘of, from’.

**object:** the pronoun is the direct object of a finite verb.

**oedd:** the pronoun is the subject of the imperfect of the verb ‘be’ *oedd*-.

**other:** use of the pronoun is comprehensible but does not fit into any of these categories.

**penblwydd hapus i:** the pronoun is in the phrase *Penblwydd Hapus i ti/chdi* ‘Happy Birthday to you’.

**possessive:** the pronoun is the possessor of a noun phrase (e.g. *dy gar di/chdi* ‘your car’).

**rhaid:** the pronoun is in a construction with *rhaid* ‘necessity, must’, *rhaid (i) ti/chdi* ‘you must’.

**s-negative:** the pronoun is the subject of a negative auxiliary in *s*-, for instance:

- (7)      So                    ti            'n            mynd.  
             NEG.AUX 2SG    PROG go.INF  
             ‘You aren’t going.’

**tag:** the pronoun is the subject in a tag question.

**trwy:** the pronoun is the object of the inflecting preposition *trwy* ‘through’.

**tu ol i**: the pronoun is the complement of the preposition *tu ol i* ‘behind’.

**unclassified**: it was impossible to determine the function of the pronoun; whether a pronoun was intended; or whether the tweet was intended to be in Welsh.

**verbnoun**: the pronoun is the object of a verbnoun (infinitive/nonfinite verb):

- (8) Dw i wedi gweld ti.  
 be.PRS.1SG 1SG PERF see.INF 2SG  
 ‘I have seen you.’

**wrth**: the pronoun is the object of the inflecting preposition *wrth* ‘to, by’.

**wyt**: the pronoun is the subject of a present-tense clause in which the finite form of the auxiliary *bod* ‘be’ (namely *wyt* ‘are (2sg.)’) is retained in some form.

**yli**: the pronoun is the complement of the particle *yli* ‘look’ (< *gweli* ‘you see’).

**yn erbyn**: the pronoun is the object of the complex preposition *yn erbyn* ‘against’.

**yn**: the pronoun is the object of the inflecting preposition *yn* ‘in’.

**context\_wide**: an amalgamation of the categories listed for context\_narrow (Table 1).

context_wide	contains context_narrow
auxdrop	auxdrop, auxdrop/bod
bod	bod
bydd	bydd
conditional	buasai, byddai
dylai	dylai
formulas	da (iawn), diolch, diolch i, greetings, Nadolig Llawen i, penblwydd hapus i
gan	gan
i	i, elided i-clause
imperative	imperative
independent	a (and), co, dyma, dyna, focus, independent, meddai, yli
inflected preposition	am, ar, at, dros, heb, mo, o, trwy, wrth, yn
lexical verb	lexical verb
object	object
oedd	oedd
other	ar gyfer, ar ol, bu, ddaru, eisiau, o flaen, other, s-negative, tu ol i, yn erbyn
possessive	possessive
rheid	rheid
unclassified	unclassified
uninflected preposition	a (with), cyn, efo, fatha, fel, gyda, heibio, na
verbnoun	verbnoun
wyt	wyt

Table 1. Mapping from context\_narrow to context\_wide.

The following fields are valued only for data points which were successfully localized to a specific geographic location:

**x**: eastings projected (in metres) according to the Ordnance Survey coordinate system.

**y**: northings projected (in metres) according to the Ordnance Survey coordinate system.

**xjitter, yjitter**: random values from -1000 to 1000 to create jittered coordinates up to 1km from actual value.

**x\_jittered, y\_jittered**: coordinates jittered to prevent data points being plotted on top of each other in maps.

**map\_to**: name of the place (in English or Welsh) to which the data point has been localized (corresponds to the x and y coordinate fields).

**localization\_source**: the basis for this localization, either UserLocation (the localization derives from the user location field of the Twitter metadata) or UserDescription (the localization derives from the user description field of the Twitter metadata).

### 1.3 Search procedure

This data set contains all instances of either form of the pronoun (including informal written variants such as *t* for *ti* or *tn* for *ti'n* (pronoun + aspect marker) etc.).

## 2 Variable: 2sg. auxiliary deletion (auxdel)

### 2.1 Short description of the variable

Auxiliary ‘be’ may be omitted in a present-tense AuxSV(O) clause:

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

This may occur in various clause types (both main and subordinate) and is not restricted to clause- or sentence-initial position. The phenomenon is referred to in the literature as auxiliary deletion (Davies 2010: 263). This data set contains all instances of present-tense AuxSV(O) in the second person singular in the corpus, whether the auxiliary is present or omitted, coded for clause type, force type and polarity. For further details, see Borsley & Jones (2001: 17–18), Jones (2004: 101–2), Borsley, Tallerman & Willis (2007: 260–61), Davies (2010: 258–335), Breit (2012), Davies & Deuchar (2014), and Davies (2016).

### 2.2 Variants (context\_narrow and context\_wide)

**Possible values** of variant\_narrow relevant to this variable are **auxdrop**, **auxdrop/bod** and **wyt**. Data points with other values are not relevant to this variable. Possible values of variant\_wide are **auxdrop** and **wyt**. Note that these values are contexts with respect to Variable 1 and variants with respect to Variable 2, hence the discrepancy between the field name and its description here.

**Classification method.** A tweet is classified as **auxdrop** if it contains a present-tense AuxSV(O) clause in which no auxiliary is present. A tweet is classified as **wyt** if an auxiliary is present, irrespective of the spelling of the auxiliary (variant forms included *wyt*, *rwyt*, *wt*, *w* and *yw*).

In some cases, it was unclear whether the full form of the clause would have contained a finite or a non-finite form of the auxiliary. This is the case with the following example:

- (10) Dwi 'n gwybod \_\_\_\_ ti 'n barod.  
 be.PRS.1SG PROG know.INF you PRED ready  
 'I know you're ready.'

Prescriptively, this context would require the full form to be restored as nonfinite *bod*, and hence excluded from the data set. However, overt finite forms do occur in this context in the data set, hence it was decided to retain them but code them separately to indicate the uncertainty as to the correct analysis. In these cases, the tweet is coded as auxdrop/bod in variant\_narrow, and as auxdrop in variant\_wide. The second coding scheme is appropriate if they are considered to alternate with a finite form.

## 2.3 Conditioning factors

The three factors considered were clause type, force type and polarity. Values for these factors are given only for data points classified as auxdel, auxdel/bod or wyt for context\_narrow.

### 2.3.1 Clause type (clause\_type)

**Possible values: main, subordinate.**

Relative clauses are coded like other subordinate clauses and not coded separately. Clauses introduced by *gobeithio* 'hope, hopefully' and *efallai* 'perhaps, it may be' are treated as subordinate as they can be followed by a nonfinite verb and thus they fall into the category **auxdrop/bod** discussed above:

- (11) Gobeithio/Efallai ti 'n iawn.  
 hopefully/perhaps you PRED right  
 'Hopefully/perhaps you're right.'

Tag questions are included in the overall dataset, but given the value **tag** for context\_narrow and therefore excluded from this variable. Note that the clause to which a tag question is attached is coded as (declarative) main.

### 2.3.2 Force type (force\_type)

**Possible values: conditional, declarative, focus, focus question, yes-no question and wh-question.** Examples are given below:

- (12) Os (wyt) ti 'n barod...  
 if be.PRS.2SG you PRED ready  
 'If you're ready...' (**conditional**)
- (13) (Wyt) ti 'n iawn.  
 be.PRS.2SG you PRED right  
 'You're right.' (**declarative**)
- (14) Fi (wyt) ti 'n meddwl!  
 1SG be.PRS.2SG you PROG think.INF  
 'Me you mean!' (**focus**)

- (15) Un doniol (wyt) ti.  
 one comical be.PRS.2SG you  
 ‘You’re a funny one.’ (**focus**)
- (16) (Ai) fi (wyt) ti ’n meddwl?  
 Q.FOC 1SG be.PRS.2SG you PROG think  
 ‘You mean me?/Is it me you mean?’ (**focus question**)
- (17) (Wyt) ti ’n dod?  
 be.PRS.2SG you PROG come.INF  
 ‘Are you coming?’ (**yes-no question**)
- (18) Pryd (wyt) ti ’n dod?  
 when be.PRS.2SG you PROG come.INF  
 ‘When are you coming?’ (**wh-question**)

The codes **focus** and **focus question** indicate that some element has been fronted, as in (14). This includes copular constructions of the type in (15). The fronted element need not be the subject of the auxiliary. Embedded *wh*-answers (‘I know what you’re doing’) are coded as declarative not as *wh*-questions.

### 2.3.3 Polarity (polarity)

**Possible values: affirmative, negative.**

Clauses containing markers of sentential negation such as a negative auxiliary (*dwyt* ‘aren’t’), *ddim* ‘not’ and/or a negative indefinite (*neb* ‘no one’, *dim byd* ‘nothing’) are coded as negative. In some cases, somewhat arbitrary decisions about what counted as a negative element had to be made. The elements *mond* ‘only’ (< *dim ond* ‘nothing but’) and *methu* ‘be unable, fail’ were not deemed sufficient on their own for a clause to be coded as negative; the aspectual marker *heb* ‘without’ + nonfinite verb (negative perfect) was treated as the negative of the perfect marker *wedi* and coded as negative. Hence, (19) and (20) would be coded as affirmative, and (21) as negative.

- (19) Ti mond yn ifanc.  
 you only PRED young  
 ‘You’re only young.’
- (20) Ti ’n methu gweld fi yn y llun.  
 you PROG be.unable.INF see.INF 1SG in the picture  
 ‘You can’t/you’re unable to see me in the picture.’
- (21) Ti heb neud e ’to.  
 you PERF do.INF 3MSG yet  
 ‘You haven’t done it yet.’

## 2.4 Search procedure

This data set contains all instances of the second-person singular pronoun (including informal written variants such as *t* for *ti* or *tn* for *ti’n* (pronoun + aspect marker) etc.).

## References

Borsley, Robert D. & Bob Morris Jones. 2001. The development of finiteness in early Welsh. *Journal of Celtic Language Learning* 6. 9–20.

- Borsley, Robert D., Maggie Tallerman & David Willis. 2007. *The syntax of Welsh*. Cambridge: Cambridge University Press.
- Breit, Florian. 2012. Constraints on auxiliary deletion in colloquial Welsh. Bangor University BA dissertation.
- Davies, Peredur. 2010. Identifying word-order convergence in the speech of Welsh–English bilinguals. Bangor University PhD dissertation.
- Davies, Peredur. 2016. Age variation and language change in Welsh: Auxiliary deletion and possessive constructions. In Mercedes Durham & Jonathan Morris (eds.), *Sociolinguistics in Wales*, 31–60. London: Palgrave MacMillan.
- Davies, Peredur & Margaret Deuchar. 2014. Auxiliary deletion in the informal speech of Welsh–English bilinguals: A change in progress. *Lingua* 143. 224–41.
- Jones, Bob Morris. 2004. The licensing powers of mood and negation in spoken Welsh: Full and contracted forms of the present tense of *bod* ‘be’. *Journal of Celtic Linguistics* 8. 87–107.
- Willis, David. 2017. Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun *chdi*. *Journal of Linguistic Geography* 5. 41–66.