SCP Volume 9 (2024)
Special Issue: Noun phrase and pronominal syntax in medieval and early modern Celtic languages
doi: https://doi.org/10.14746/scp.2024.9.3

THE DIACHRONY OF WELSH SUBJECT PRONOUNS

MARIEKE MEELEN
https://orcid.org/0000-0003-0395-8372

University of Cambridge

DAVID WILLIS
https://orcid.org/0000-0003-0755-9248

University of Oxford

ABSTRACT

In many languages, independent pronouns become reduced to inflectional affixes which are ultimately lost, resulting in the creation of new independent pronouns. The loss of null subjects therefore often goes hand in hand with a loss of agreement morphology on the verb. Inflectional morphology has remained virtually unchanged from the Middle Welsh period up to the present day, but whereas null subjects were frequently found in the earliest period, in Present-day spoken Welsh overt pronouns are generally preferred. In this article we present a pilot study of the history of subject pronouns in Welsh based on six annotated texts from the Parsed Historical Corpus of the Welsh Language (PARSHCWL) from three different time periods (fourteenth, sixteenth and eighteenth centuries), as well as in translated and non-translated texts. We show that null subjects are favoured in all periods and use a mixed-effects logistic regression model to test which factors have an effect on whether the subject pronoun is overt or null and if this distribution changes over time.

Keywords: null subjects, subject pronouns, Middle Welsh, information structure, verbal agreement.

1. Introduction

It is widely agreed that the development of subject pronouns proceeds along a cycle.[1] The cycle begins with reduction of an independent pronoun to an inflectional affix, thereby giving rise to subject–verb agreement morphology; subsequently this inflection is lost, leading to replacement by new independent pronouns in what is known as the agreement cycle (Givón 1976, Siewierska 1999, Van Gelderen 2011: 37–85):[2]

(1) independent pronoun > unstressed pronoun > clitic > affix > zero

The idea in (1) goes back to the long-established intuition that languages participate in a morphological cycle (Humboldt 1825), gaining agglutinative morphology via grammaticalisation of formerly independent words (in this case, pronouns), before becoming synthetic as individual morphs merge, with these morphs eventually eroding to return us once again to an analytic system.[3]

The central part of this cycle creates new inflectional elements (inflectionalisation) out of pronouns and is widely discussed in grammaticalisation literature. Once its subject pronouns become verbal affixes, a language gains the possibility of leaving the subject unexpressed, that is, as a null subject. This leads to the idea that there is a correlation between the availability of null subjects in a language and the 'richness' of its verbal subject agreement (Taraldsen 1980), which became known as Taraldsen's Generalisation (Jaeggli 1981: 134, Teixeira 1986: 82, Roberts 2014: 115–116). Thus, from a diachronic point of view, subject pronouns develop into agreement inflections, leading to loss of the need to express subject roles via separate pronouns, but the distinctions between inflections are lost

[1] This research was funded by AHRC–DFG UK–German collaborative research grant AH/V00347X/1 'The history of pronominal subjects in the languages of northern Europe' and British Academy–Leverhulme Trust Small Research Grant SRG18R1\181450 'Developing a Welsh Historical Treebank'. This support is hereby gratefully acknowledged.

[2] Developments running counter to this trend are not unknown, and have been widely discussed in the literature on degrammaticalization (Norde 2009). Within Celtic, particularly notable is the debonding of Irish verbal suffixes to become subject pronouns such as *muid* 'we', instantiating affix > unstressed pronoun (Doyle 2002, Diertani 2011: 174–188, Willis 2017: 33–34). Nevertheless these are far less frequent than the pattern in (1).

[3] For a recent overview of the history of these ideas, see Haspelmath (2018). Note, however, that Haspelmath rejects central parts of the traditional narrative, reinterpreting the changes as a spiral rather than a cycle, and rejecting both the coherence of the traditional analytic–synthetic distinction and the requirement that agglutinating forms go through a synthetic phase before analyticity is restored.

over time, whether through regular sound change or through analogy, giving rise once more to the need for subject pronouns to disambiguate the now poorly distinguished inflections.[4]

Many European languages have witnessed the second of these shifts in the course of their recorded histories. Loss of null subjects has often coincided with the loss of subject-agreement inflectional endings on the verb. Thus, across Germanic and in French (Adams 1987), verbal agreement has become less distinctive and subject pronouns have become largely obligatory. Consider the present-tense paradigm in the history of German of the verb 'seek, look for', given in Table 1 (based on Braune and Reiffenstein 2004: 260–261). While Old High German manifests six distinct verbal forms, only four remain in Modern German. Correspondingly, early Old High German permitted null subjects, while, in later varieties, including Modern German, referential null subjects have disappeared, so that, today, only remnants of the former system remain in the form of expletive null subjects (Axel 2005).

Table 1. Present-tense paradigms of a regular (weak) verb *suochen/suchen* 'seek' in the history of German.

	Old High German	**Modern German**
1 sg.	suochu	suche
2 sg.	suochis	suchst
3 sg.	suochit	sucht
1 pl.	suochemēs	suchen
2 pl.	suochet	sucht
3 pl.	suochent	suchen

Null subjects have similarly become more restricted in the history of Russian, perhaps in line with loss of verbal inflection in the expression of the past (Meyer 2009), although in the absence of extensive loss of verbal inflection in other tenses.

[4] This is often, as here, presented as a pull chain: the phonological deficiency of the verbal endings forces speakers to use overt pronouns to avoid misunderstanding; however, it can also be conceived of as an inflationary push chain: more and more frequent use of pronouns makes the endings redundant. In the Welsh case, we see increasing use of subject pronouns without loss of verbal inflection, which is more consistent with the second scenario (cf. Haspelmath 2018: 112–113, who argues that the second scenario is the correct one more generally). Consider also Jespersen's (1924: 213) scenario: 'In course of time, however, it became more and more usual to add the pronoun even when no special emphasis was intended, and this paved the way for the gradual obscuration of the sound of the personal endings in the verbs, as these became more and more superfluous for the right understanding of the sentences.'

Some languages have introduced subject pronouns in contexts where verbal morphology has become impoverished, but retained null subjects with rich morphology. This is true of Irish (Roma 2000). Some Romance varieties (Northern Italian, Franco-Provençal, and some Occitan dialects) also favour full pronouns with tense–aspect forms that are non-distinctive with respect to the person of the subject (Manzini and Savoia 2005).[5]

This cycle is traditionally framed in terms of phonological status and morphological form (phonetic erosion and renewal) or in terms of analogy (loss of verbal distinctions due to analogical levelling of person–number affixes), but there are also potential interactions with changes in the discourse function (information structure) of the items involved: overt pronouns, only used for marked or contrastive discourse functions, such as a change in topic, come to be used in a wider range of discourse functions as languages proceed along the cycle. There is thus a complex interaction between phonology, morphology, syntax and information structure which is not yet fully understood.

In this paper we present a pilot study on subject pronouns in the history of Welsh to highlight a peculiar problem in the agreement cycle. In Present-day Spoken Welsh, subject pronouns are normally required to be overt in contexts where, in earlier stages of the language, null subjects were frequently found. Formal literary Welsh is like these earlier stages in also making frequent use of null subjects. Unlike in the cases just outlined, however, loss of null subjects has not been accompanied by significant loss in distinctiveness of verbal person–number inflection. Indeed, the relevant inflectional endings have remained largely unchanged from the medieval period until the present day, as shown in Table 2. The only loss of richness is the merger of the first- and third-person plural, both *gwel*(*s*)*on* in Table 2, and even these forms remain distinct (as *gwelsom* and *gwelsant* respectively) in the most formal written register. The varieties thus do not differ very significantly in the richness of their verbal morphology, but do differ in the availability of null subjects.

---

5 Roberts (2019: 29–31) notes that these varieties typically tend to have subject clitics as well, but argues that these clitics are not subject pronouns but heads (cf. Rizzi 1986; Poletto 2000). Roberts (2005: 52) observes the resemblance between agreement markers (which are like Romance subject clitics under his analysis) and the equivalent pronouns in Welsh (except for the third-person singular) and suggests that the original inflectional endings have been replaced by what used to be pronouns in the history of Welsh (cf. Roberts 2005: 169). As Roberts (2019: 389) points out, however, various questions remain with this analysis, for example regarding the series of strong pronominal first- or second-person pronouns that cannot occur in postverbal subject position with an agreeing verb (**Canais fi.* 'I sang'). For the present pilot study, we therefore take a step back and return to a more detailed description of the variation in the historical data, which will form an important step to solving remaining puzzles in Welsh agreement marking in general.

Table 2. Representative paradigms of a regular verb in the history of Welsh.

	Middle Welsh		**Modern Welsh**	
	pres.	past	pres.-fut.	past
1 sg.	gwelaf	gweleis	gwelaf	gwelais
2 sg.	gwely	gweleist	gweli	gwelaist
3 sg.	gwel	gweles	gweliff/gwelith	gwelodd
1 pl.	gwelwn	gwelsom	gwelwn	gwel(s)on
2 pl	gwelwch	gwelsawch	gwelwch	gwel(s)och
3 pl.	gwelant	gwelsant	gwelan	gwel(s)on

There are other cases where languages have lost or restricted their use of null subjects without extensive loss of verbal inflection. Icelandic has retained the rich verbal morphology of Old Norse, but lacks referential null subjects (Kinn, Rusten and Walkden 2016). In Old and Middle French, for example, there is a large temporal lag between the loss of null subjects and loss of agreement (Ranson 2009; Roberts 2014; Schøsler 2002). Simonenko, Crabbé and Prévost (2019) conducted a statistical analysis of agreement inflection and null subjects in a corpus of historical French and concluded that there was an increase in overt personal pronominal subjects over time, but that this increase was uniform across old and new inflectional endings. In addition, they found that the spread of new syncretic inflectional endings in French was uniform across clauses with null and overt subjects. Carvalho and Child (2011) reached the same conclusion for Spanish, as did Nagy and Heap (1998) for Francoprovençal.

The current study offers a first step to investigating how this state of affairs has come about by investigating the availability of null subjects in an ongoing diachronic, annotated corpus of Welsh texts from the medieval and early modern periods. We begin (section 2) by outlining the texts that were selected for the corpus in this present pilot study and the annotation methodology used to develop it. In Section 3, we present examples and a regression model of the data, followed by detailed discussion in Section 4. In Section 5 we conclude and present suggestions for further research.

2. Methodology and corpus description

The texts included in the corpus for the current study are listed in Table 3. Texts were chosen so as to cover a period from classical Middle Welsh to the end of the early modern period, and to control in large part for text type. All the texts are essentially narrative in nature, whether the nature of the narrative is religious or secular. In order to assess whether translation from other languages might have

had an impact on the rise of overt subjects in writing, a mixture of translated and non-translated ('native') texts was included. Note, however, that this is not an entirely fixed distinction, as *Cronicl Hywel ap Syr Mathew*, while not being a direct translation of any particular English source, deals with historical events for which source material in English will have been used in its compilation.

Table 3. Texts included in the study.

text	**type**	**source**	**date**	**words**[6]
Pwyll Pendefig Dyfed (selection) (*Pwyll*)	non-trans. narrative	ms. Peniarth 4	*c*. 1350	3919
Breuddwyd Pawl (*BP*)	translated religious	ms. Jesus 119	1346	1688
Marwolaeth Mair (*MM*)	translated religious	ms. Jesus 119	1346	3411
Cronicl Hywel ap Syr Mathew (selection) (*CHSM*)	non-trans. historical	ms. Peniarth 168	1589–90 (comp. 1568)	8233
1588 Bible (Gen. 37: 39-41, I Samuel 16–20, Matt. 26–28, II Cor. 1–9) (*B1588*)	translated religious	printed	1588	17681
Gweledigaetheu y Bardd Cwsc (selection) (*GBC*)	non-trans. narrative	printed	1703	4356
TOTAL				**39288**

All corpus texts were preprocessed and annotated with part-of-speech (POS) tagging and syntactic parsing in conformity with the principles of the ongoing Parsed Historical Corpus of the Welsh Language (PARSHCWL) (Meelen and Willis 2021, 2022), of which they will form a part. These conventions are adapted from those of the various Penn-style historical corpora, such as the Penn–Helsinki Parsed Corpus of Middle English Prose (Kroch and Taylor 2000). *BP*, *MM* and *Pwyll* follow the edition in *Rhyddiaith Gymraeg: Welsh Prose 1300–1425* (Thomas, Smith and Luft 2007–17); *B1588*, *CHSM* and *GBC* follow the edition in *Corpws Hanesyddol yr Iaith Gymraeg: A Historical Corpus of the Welsh Language 1500–1850* (Willis and Mittendorf 2004). While the final corpus aims for a broader

[6] Note that the number of words listed here reflects the number of words that have been fully annotated and therefore used in the present pilot study. For most texts this is only a part of the text, but the two shorter texts (*Breuddwyd Pawl* and *Marwolaeth Mair)* were annotated entirely.

coverage of text-type and time period, it was felt that the range of both should be limited for this pilot study, both for practical reasons and in order to limit the possibility of variation due to register. Thus, the corpus, in its current form, ends in the early eighteenth century, and we cannot say anything about the development of null subjects after this period. It was expected that the early modern period would be a crucial one for change, although this expectation is not borne out by the results, which suggest that examination of further texts, particularly colloquial ones, from the seventeenth century onwards will be needed to provide a full understanding of how null subjects came to be restricted in Welsh.

According to the PARSHCWL system, the pronouns that appear as subjects, namely simple independent and dependent affixed pronouns in the traditional system (Evans 1964: 49–58), are tagged as PRO. They receive a further extension if they are conjunctive (PROC) or reduplicated (PROR) in form. Pronouns that act as subjects are parsed into subject noun phrases annotated as NP-SBJ. Where an overt subject is absent, a null subject pronoun is marked in the parse as *pro* within an NP-SBJ. Examples are given in (1) and (2).[7]

(1) *Düw llün gwedi hynny i marchokaodd ef …*
Monday after DEM PRT ride.PAST.3SG he
o Sowthampton I Winsiestr
from Southampton to Winchester
'The Monday after this he rode … from Southampton to Winchester'
(*CHSM* 235r.3–235v.2)

(2) *a thrannoeth yn ieuengtit y dyd kyuodi a oruc* *pro*
and next.day in youth the day get.up.INF PRT do.PAST.3SG
'and the next day at dawn he got up'
(*Pwyll*, Jesus 111, 175rb.23–24)

The goal of this pilot is to measure how frequently null and overt subject pronouns are used and to establish the factors that favour the choice of one over the other. All finite main and subordinate clauses containing pronominal subjects, whether null or overt, were therefore extracted from the parsed corpus texts using CorpusSearch2 (Randall, Taylor and Kroch, 2005). A sample query is given in Appendix A.

A number of clause types were excluded from analysis because they do not allow a full choice between null and overt pronominal subjects.

---

7 Glossing follows the conventions for this volume, except: FOC = focus marker, PROG = progressive marker, QU = question particle, SUP = superlative.

First of all, Subject–Verb (SV) orders were excluded. Example (3) shows an SV order with an overt subject pronoun *ni* 'we'. Where the subject precedes the verb, the preverbal particle *a* appears in immediately preverbal position. With this order, the subject cannot be left null: there is no equivalent grammatical sentence beginning simply with *hefyd* 'also' followed by the preverbal particle *a*.

(3) *Hefyd ni a vynnwn Ddoctor Moor a Doctor*
also we PRT want.PRES.1PL Doctor Moor and Doctor
Crispin … yn rhydd
Crispin PRED free
'Also we want Doctor Moor and Doctor Crispin … free'
(*CHSM* 232r.27–29)

This lack of variability means that we exclude such orders from the analysis. It is true that a null subject in this context is possible in certain coordination structures. Thus, in (4), it can be argued that there is an empty element between *ac* 'and' and the preverbal particle *a*.[8] The presence of such an element is suggested by the need to explain why the preverbal particle is *a*, which is normally triggered by a preceding subject or object.

(4) *ar Trwmpeter a ddaüth wrth Vwlen iat*
and.the trumpeter PRT come.PAST.3SG to Boulogne gate
ac a ganodd I Drwmpet
and PRT play.PAST.3SG his trumpet
'and the trumpeter came to Boulogne gate and played his trumpet'
(*CHSM* 228r.4–5)

However, in the corpus, these cases are treated as null topics marked as *top* (Meelen and Willis 2022: 24–28). While they would need to be dealt with in a full analysis of null subjects as well, they were excluded from this pilot study in which coordinate sentences like the above were not taken into account. This means that the following analysis is restricted to cases where the subject is clearly postverbal, either because it is overt in postverbal position, or because the behaviour of preverbal particles or other aspects of word-order allow us to establish unambiguously that we are dealing with verb–subject word order and that the null subject is in postverbal position.

---

[8] In this example, the 'gap' in the second conjunct is coreferential with a preverbal subject in the first conjunct, but this is not a requirement, and second-conjunct gaps may be coreferential with elements fulfilling other grammatical functions or with multiple elements ('split antecedents'), see Willis (1997).

Various nonfinite clauses were also excluded from analysis, because these clauses, often lacking any kind of subject-agreement morphology, require an overt subject. Thus, for instance, subjects of absolute clauses, such as (5), where there is no verb at all, are always overt, and do not need to be predicted. These were simply not collected by the search query.

(5) *ac ynte wrtho ehün ynn y kanol*
and he.CONJ at.3SM REFL.3SG in the middle
'and he on his own in the middle/with him on his own in the middle'
(*CHSM* 235v.12–13)

A number of main clause types were collected, but are excluded from the analysis below for parallel reasons:

(i) parenthetical quotative clauses, such as those with the quotative particle/verb *heb*, which are very common in narrative, particularly in *Pwyll*, were excluded because they require an overt subject;
(ii) clauses with expletive subjects, because these are generally overt;
(iii) imperative clauses, because they mostly lack an overt subject and it is not clear whether a null subject should be posited here;
(iv) responsive clauses (answers to yes–no questions), because they require a null subject;
(v) sentence tags containing finite verbs, because they too generally require a null subject;
(vi) clauses with impersonal verbs in *-ir*, *-wyd* etc., because these cannot have an overt subject.

The goal then is to predict the choice between a null or overt subject pronoun on the basis of other linguistic characteristics of the clause or text, restricting ourselves to factors that could be implemented within the constraints of a pilot project.

The data set was thus annotated for a number of features that could plausibly be relevant for this choice, both language-internal, such as type of subject, person, number, etc., and language-external, such as year of production.[9] We will now consider each of these in turn.

[9] Because of the detailed set of POS tags, internal features like person, number and gender (for objects of prepositions) could be extracted semi-automatically. Similarly, pronoun type (regular overt, conjunctive, or reduplicated) can be derived from the POS label, while null subjects can be found by searching for terminal nodes with the empty-category *pro* within a syntactic phrase labelled NP-SBJ.

The language-internal factors for consideration are:
(a) clause type: main vs. subordinate
(b) person of the subject: first vs. second vs. third
(c) number of the subject: singular vs. plural

We first of all focus on clause type, because the existing literature suggests that, in certain verb-second (V2) languages, the frequency of null subjects is higher in main clauses than in subordinate clauses. This was first observed for Old French (Adams 1987: 2). Since then, it has also been observed for various historical Germanic varieties, namely Old Saxon (Walkden 2014: 190), Old Swedish (Håkansson 2013: 170–173), Old High German (Axel 2007: 310) and Old English (Walkden 2013: 163–164, Rusten 2015: 67–69). For Middle English, there are very few instances of null subjects altogether, but Walkden and Rusten (2016: 450–451) note that here too, subordinate clauses have the lowest proportion of null subjects. On the other hand, this asymmetry is either absent or poorly attested for Old Italian (Poletto 2020: 331, 346–347).

Intuitively, this is perhaps surprising, as subjects of subordinate clauses are more likely to have a recent antecedent, an antecedent in the main clause being quite probable, and it is well-known that discourse-old elements tend to be expressed using null subjects in null-subject languages. Indeed, in Modern Italian (but not in many historical varieties), the null subject of a subordinate clause is generally interpreted as coreferential with the subject of the main clause, while an overt pronominal subject is generally interpreted as referentially distinct (Poletto 2020: 326). This might be expected to give rise to a higher frequency of null subjects in subordinate contexts. We would therefore like to know where the history of Welsh sits within this space of variation, particularly as Middle Welsh is a V2 language, while Modern Welsh is not.

The next language-internal factor for investigation is person of the subject. Here too existing literature on other languages suggests that it might play a role. For English, Berndt (1956: 65–68) showed that first- or second-person subjects are less likely to be null than third-person subjects (see also Walkden 2013: 164–166; Rusten 2015: 69–71). Various work (Kinn 2016, Stausland Johnsen 2016) agrees that null subjects are significantly more frequent in the third person than in the first and second persons in Old and Middle Norwegian.

The final internal factor is number. Walkden and Rusten (2016: 440) take number into account in their investigation of Middle English null subjects, but note that whether the subject is singular or plural does not appear to have a consistent effect outside the combination with person. In Middle English poetry, they report that plural subjects are more likely to be null, but the reverse is the case for prose texts (Walkden and Rusten 2016: 457). The number of examples of plural subjects in the first- or second-person singular is exceedingly small in

Middle English, however. Again, it would be useful to establish whether such effects play a role in the development of null subjects in the history of Welsh.

External factors may also influence the choice between null and overt subject pronouns. The following factors were considered:

(a) period/year of production (manuscript or early printed book)
(b) interlinguality: non-translated vs. translated

The choice of year of production as a factor is primarily because we are interested in detecting a possible change in the use of null subjects. The quantity and variety of texts in Old Welsh, from the eleventh century or earlier, is sufficiently limited that it was felt that these could not be included in a pilot study. The focus on the period from Middle Welsh to the eighteenth century gives us a manageable time range over which a change in the frequency of null subjects might be expected to occur. For descriptive statistics, the texts are divided into three periods: Middle Welsh (up to 1450), Early Modern Welsh (1450–1650) and Modern Welsh (after 1650); for the regressions, year of production is treated as a continuous variable.

The second external factor, interlinguality, is based on the long-established observation that translated medieval Welsh texts differ linguistically from texts composed originally in Welsh (see Luft 2015 for a thorough overview of the history of the field). Differences in agreement patterns, for instance, the tendency in translated texts for postverbal plural subjects to trigger plural agreement on their verbs, provide one example where Latin originals seem to have exerted an influence on Welsh translations (Evans 1971: 56).[10] It might be anticipated, therefore, that a similar kind of phenomenon might be at work with the expression of subject pronouns, with Welsh texts translated from Latin, a null-subject language, underusing overt pronouns, while texts translated from French and, later, English, both non-null-subject languages might be expected to underuse them. Furthermore, the linguistic properties of medieval Welsh translations, along with translation techniques themselves, have been the subject of renewed interest in the past few years with a number of publications devoted to the topic (Parina 2018a, b). Parina (2018b) indeed examines various linguistic features of one of the texts in our corpus, namely *Breuddwyd Pawl*, suggesting that it contains certain syntactic features that betray its Latin origin. Nevertheless, the role of translation can be complex: Parina (2022), for instance, argues that

[10] rNote, however, recent research that has questioned the traditional view, either by failing to establish clear differences between translated and non-translated texts (Meelen and Nurmio 2019 on adjective agreement) or by attempting to reinterpret differences found in translated texts as reflecting an elevated register of written Middle Welsh, rather than as translation errors (Plein 2018: 269–271 on verbal agreement).

demonstrative relatives, although frequently cited as a feature of 'translationese' in the medieval and early modern periods, have a distribution regulated by complex factors, and bring new distinctions into the language. For all of these reasons, it was thought useful to consider translation as a factor in this study.

Having introduced the factor that might condition variation, in the next section we turn to consider the results, including a regression analysis, and analyse the effect of the external and internal factors above on whether there is an overt or null subject in the clause.

2.3. Results

In this section, we first consider the results of our corpus queries illustrated with relevant examples for each factor. We then present a fixed-effects regression analysis of these results to determine which factors have an effect on the possibility of using a null pronoun as a subject.

2.3.1. Corpus results and examples

First, it is clear that both null and overt pronouns are used at all time periods, as shown in example (6) for null and (7) for overt pronouns.

(6) a. *ac* *ny* *s* *keffynt* *pro*
and NEG ACC.3 get.IMPF.3PL
'and they didn't get it'
(Middle Welsh, *BP*, Jesus 119, 129r.23)

b. *honn* *a* *gawsom* *pro*
PROX.SF PRT find.PAST.1PL.
'we found this'
(Early Modern Welsh, *B1588*, Genesis 37: 32)

c. *a* *phwy* *ydynt* *pro* *?*
and who be.PRES.3PL
'and who are they?'
(Modern, *GBC* 10.24–25)

(7) a. *ac* *nẏ* *chẏuarchaf* *i* *well* *it*.
and NEG greet.PRES.1SG I better to.2SG
'and I don't greet you'
(Middle Welsh, *Pwyll*, Peniarth 4, 1rb.19–20)

b. *Ac yno y bu efe yn y carchar-dy.*
and there PRT be.PAST.3SG he in the prison
'And he was there in prison.'
(Early Modern Welsh, *B1588*, Genesis 39: 20)

c. *siccra' oll yw ef o hono.*
certain.SUP all be.PRES.3SG he of.3SM
'He is the most certain of all of it.'
(Modern Welsh, *GBC* 13.12–13)

Table 4 shows an overview of the frequency of null and overt pronouns by text and time period; and Figure 1 visualises the same data with texts in roughly chronological order. The highest proportion of null subjects is found in *Cronicl Hywel ap Syr Mathew* and the lowest in *Breuddwyd Pawl*. The latter, however, is a very short text with a very low number of pronominal subjects overall, so direct comparison with any of the other texts is difficult. From a diachronic point of view, the proportion of null subjects increases from 62% in Middle Welsh to 73% in Early Modern and, finally 78% in Modern Welsh. As this pattern deviates from that in many other European languages, we will consider these results in particular in further detail in Section 4.

Table 4. Frequency of null vs. overt pronouns by text and time period.

Text	**null**	**overt**	**null**	**overt**
Pwyll	115	75	61%	39%
Breuddwyd Pawl	17	14	55%	45%
Marwolaeth Mair	68	35	65%	35%
total Middle Welsh	**200**	**124**	**62%**	**38%**
Cronicl Hywel ap Syr Matthew	73	14	84%	16%
1588 Bible	66	38	63%	37%
total Early Modern	**139**	**52**	**73%**	**27%**
Gweledigaetheu y Bardd Cwsc **(Modern Welsh)**	**75**	**21**	**78%**	**22%**
TOTAL	**414**	**197**	**68%**	**32%**

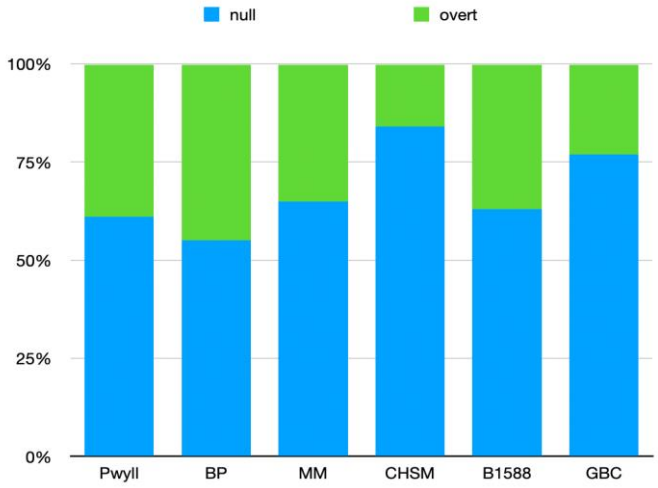

Fig. 1. Frequency of null vs. overt pronouns by text in chronological order.

When it comes to translated vs. non-translated texts, we see a similar pattern. Both null and overt subjects are attested as shown in (8) for null and (9) for overt pronouns:

(8) a. *Ac yno I bü* *pro* *noswaith*
and there PRT be.PAST.3SG evening
'And he was there for a night.'
(non-translated, *CHSM* 220v.10)

b. *A dywedut a orugant* *pro*
and say.INF PRT do.PAST.3PL
'And they said'
(translated, *MM* 140)

(9) a. *A pha ỏlat ŷd hanỏŷt titheu*
and which country PRT originate.PRES.2SG you
oheni.
from.3SF
'And which country are you from?'
(non-translated, *Pwyll*, Peniarth 4, 1v.10–11

b. *Paham yd ochy di pawl*
why PRT sigh.PRES.2SG you Paul
'Why do you sigh, Paul?'
(translated, *BP*, Jesus 119, 130v.9)

Table 5 shows the breakdown between non-translated and translated texts. The proportion of null subjects is slightly higher in non-translated (71%) than in translated texts (63%).

Table 5. Frequency of null pronouns in translated. vs. non-translated texts.

Text type	**null**	**overt**	**null**	**overt**
non-translated	263	110	71%	29%
total translated	151	87	63%	37%

The next factor under investigation is clause type, main vs. subordinate clause, shown in (10) for null and (11) for overt pronouns. As noted above (section 2), only those clauses that could potentially have an overt subject pronoun were included.

(10) a. *ac* *nit* *oedẏnt* *pro* *unllef.*
and NEG be.IMPF.3PL one.cry
'And they were not in unison.'
(main, *Pwyll*, Peniarth 4, 1ra.22)

b. *a* *chẏmer* *ẏ* *march* *kẏntaf* *[a*
and take.IMPV.2SG the horse fastest PRT
ỏẏpẏch *pro*]
know.PRES.SBJ.2SG
'And take the fastest horse you know.'
(subordinate, *Pwyll*, Peniarth 4, 4rb.8–9)

(11) a. ny phedrussaf ynhev
NEG hesitate.PRES.1SG I
'I will not hesitate.'
(main, *MM*, Jesus 119, 72r.6)

b. *ac* *nẏ* *ỏnn* *i* [*pỏẏ* *ỏẏt* *ti*.]
and NEG know.PRES.1SG I who be.PRES.2SG you
'And I don't know who you are.'
(subordinate, *Pwyll*, Peniarth 4, 1va.5–6)

As can be seen from Table 6, both types of clause allow frequent use of both null and overt subject at all periods. Overall, null subjects are slightly more frequent in main clauses (69%) than in subordinate clauses (66%), with null subjects in the majority in both. However, the direction of the difference is entirely due to the data from the Early Modern Welsh period, where main clauses (80% null) are much more favourable to null subjects than subordinate clauses (52% null). Both the 1588 Bible translation and the *CHSM* also have a larger proportion of main

clauses in general, compared to other periods; the small number of remaining subordinate clauses is therefore not a particularly useful guide to broader usage. Overall, no clear pattern emerges.

Table 6. Null vs overt pronouns by clause type and time period.

Text	**null**	**overt**	**null**	**overt**
Middle Welsh main	110	80	58%	42%
Middle Welsh subordinate	90	44	67%	33%
Early Modern main	113	28	80%	20%
Early Modern subordinate	26	24	52%	48%
Modern Welsh main	49	15	77%	23%
Modern Welsh subordinate	26	6	81%	19%
TOTAL main	**272**	**123**	**69%**	**31%**
TOTAL subordinate	**142**	**74**	**66%**	**34%**

The final factors under investigation were person and number of the subject pronoun. Again, examples of both null and overt subjects cover the entire range of person-number combinations, with some examples shown in (12) for null and (13) for overt subjects:

(12) a. *ac yn y Twr gwynn I tiriodd* *pro*
and in the tower white PRT arrive.PAST.3SG
'and he came ashore at the White Tower.'
(third-person singular, *CHSM* 229r.13)

b. *ac yno i büon* *pro* *ynghylch chwe mis*
and there PRT be.PAST.3PL around six month
'And they were there for about six months.'
(third-person plural, *CHSM* 219v.31–220r.1)

(13) a. *Yna y dy6at hi vrth yr angel*
then PRT say.PRES.3SG she to the angel
'Then she said to the angel'
(third-person singular, *MM*, Jesus 119, 70v.4)

b. *Ac vrth hynny y bydant wy*
and at that PRT be.FUT.3PL they
yn y poen hwnn hyt dyd brawt.
in the pain this until day judgement
'And because of that they will be in this pain until the Day of Judgement.'
(third-person plural, *BP*, Jesus 119, 130r.6–7)

Table 7 shows the distribution by person and number. There is a preference for null subjects everywhere except in the second-person plural, and, overall, this preference is stronger in the plural than in the singular. The number of data points for some person–number combinations is extremely low, however, so that more data is needed to confirm whether this is a significant pattern. The 92% rate of null subjects in the third-person plural is striking, however, and we will come back to this in our discussion in Section 4 below.

Table 7. Null vs. overt pronouns by person and number.

	subject	null	overt	null	overt
	first	83	44	65%	35%
SG	**second**	44	30	59%	41%
	third	163	97	63%	37%
Total SG		**290**	**171**	**63%**	**37%**
	first	18	10	64%	36%
PL	**second**	5	7	42%	58%
	third	101	9	92%	8%
Total PL		**124**	**26**	**83%**	**17%**

2.3.2. Regression analysis

We conducted a mixed-effects regression analysis to establish which of the above-mentioned factors have a significant effect on the choice of subject pronoun. Results for all coefficients are presented in Figure 2. The model attempts to predict the presence of an overt pronoun; hence, a negative effect indicates a preference for a null subject. This mixed-effects model includes a term for the random effect of source, that is, the individual text; this is necessary because our data come from a relatively small number of texts, each of which may have its own particular preference for or against null subjects, not in line with its broader characteristics as being a text from a given period or reflecting a given text type. Year (date of manuscript or publication) is treated as a continuous variable, with 1300 treated as year 0 to make interpretation of the result more intuitive.

The intercept (-0.27) indicates the log odds of an overt subject for the reference conditions, which were defined as a first-person singular main clause in a non-translated text from the year 1300; the log odds value of -0.27 is equivalent to a predicted probability of 0.43.[11] A number of individual factors

[11] The log odds for a given context are calculated by taking the intercept value and adding the values of any factor levels which diverge from the reference levels. From there, given log odds of x, probability = $e^x/(1+e^x)$. Thus, for instance, the predicted log odds of an overt subject in a third-person plural main clause in a non-translated from 1400 is the intercept (-0.2747) + third person (-0.2385) + plural (-1.0250) + 100 years (-0.1793) = -1.7175. This corresponds to a predicted probability of 0.152; that is, an overt pronoun is quite unlikely in this context.

under investigation have a significant additional impact. The strongest effect is for number, with plural subjects having a highly significant ($p < 0.001$) and strongly negative effect (of size -1.02) on the prediction.

There are two further significant effects. The first of these is year, with each additional century decreasing the log odds of an overt subject by -0.18 (-0.0018 per year), significant at the 0.05 level ($p = 0.04$). Finally, there is the marginally significant impact of a text being translated. This factor has a positive impact on the likelihood of an overt subject, with an effect size of +0.43. Note that, in this case, our confidence in the existence of the effect is relatively weak: this factors reaches significance only at the 0.10 level ($p = 0.09$).

The effects of the other factors, person and clause type, are minimal and far above any significance threshold. This confirms the unclear picture for these factors in the descriptive statistics set out above.

The AIC of this model is 744.0; this is a comparative measure of model goodness which we can use to compare other models.

```
Random effects:
 Groups Name        Variance Std.Dev.
 source (Intercept) 0.03257  0.1805
Number of obs: 611, groups:  source, 6

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -0.2747208  0.3047624  -0.901   0.3674
translatedyes                   0.4277444  0.2492011   1.716   0.0861 .
year                           -0.0017930  0.0008718  -2.057   0.0397 *
clause_typesub                 -0.0492838  0.1908515  -0.258   0.7962
personsecond                    0.1199385  0.2915908   0.411   0.6808
personthird                    -0.2385327  0.2153435  -1.108   0.2680
relevel(number, ref = "sg")pl -1.0249987  0.2447456  -4.188 2.81e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 2. First mixed-effects regression model to predict presence of overt pronouns.

The best regression model is one with all non-significant factors removed. This then is a model which includes effects for time (year), number (singular vs. plural), and the translation status of the text. Other factors, being non-significant, are therefore not present in this model, which is given in Figure 3.

In this model, number remains the strongest predictor (effect size -1.07 for plural subjects, $p < 0.001$), while year remains a significant, albeit weaker factor, later texts disfavouring overt subjects, log odds falling at the same rate as before (-0.0018 per year, -0.18 per century, $p = 0.06$). The effect of a text being translated remains similar (+0.46, $p = 0.09$). The AIC of this model is 740.4, an improvement on the previous model.

```
Random effects:
 Groups Name        Variance Std.Dev.
 source (Intercept) 0.04843  0.2201
Number of obs: 611, groups:  source, 6

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.4242321  0.2715588  -1.562   0.1182
translatedyes                  0.4623146  0.2736623   1.689   0.0912 .
year                          -0.0017855  0.0009465  -1.886   0.0592 .
relevel(number, ref = "sg")pl -1.0745110  0.2425308  -4.430 9.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 3. Second mixed-effects regression model to predict presence of overt pronouns (non significant factors removed).

Finally, we noted above the fact that the effect of number seems largely due to the preference for null subjects in the third-person plural. For this reason, a model with an interaction between number and person was considered. If a term for this interaction is added, the translation status of the text ceases to be statistically significant at all, but the overall model does improve. This model is presented in Figure 4. Here, the impact of year remains much the same, but the effect of number is almost entirely reduced to the impact of the third-person plural: the effect of (plural) number itself is not significant (effect size = +0.01, p = 0.98), nor is the effect of third person (effect size = +0.12, p = 0.63), but the additional effect of the combination of third person and plural is strong and highly significant (effect size = -1.98, $p < 0.001$). The AIC of this model is 725.6, better than either of the two previous models.

```
Random effects:
 Groups Name        Variance Std.Dev.
 source (Intercept) 0.1011   0.3179
Number of obs: 611, groups:  source, 6

Fixed effects:
                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -0.226177   0.332218  -0.681 0.495993
year                                       -0.002188   0.001146  -1.909 0.056240 .
personsecond                                0.036290   0.312015   0.116 0.907407
personthird                                 0.116476   0.239266   0.487 0.626395
relevel(number, ref = "sg")pl               0.013185   0.446894   0.030 0.976463
personsecond:relevel(number, ref = "sg")pl  0.899949   0.791123   1.138 0.255305
personthird:relevel(number, ref = "sg")pl  -1.962806   0.583295  -3.365 0.000765 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 4. Best mixed-effects regression model to predict presence of overt pronouns (with person–number interaction).

We will now turn to discuss the implications of these results in detail.

3.4. Discussion

The results overall show that there is a preference for null subjects in Welsh texts at all time periods under investigation. Regardless of which factor we focus on, whether year/time period, clause type or person–number features, null subjects almost always outnumber overt subject pronouns. Only second-person plural contexts have a higher number of overt pronouns, but on a very small sample in the current data set (7 overt vs. 5 null subjects), a minimal difference that needs to be revisited with more data available.

3.4.1. Person and number effects

The most striking result is the effect of number: plural subjects are significantly more likely to be null than singular subjects. This effect is essentially due to the behaviour of third-person plural subjects. In three out of the six texts (one from each time period, namely *Pwyll*, *MM* and *GBC*), all of the instances of third-person plural pronouns are null and there are only 9 instances of overt third-person plural subjects in total in the other three texts in our data set. The picture is very different in the third-person singular, with two Modern texts (*B1588* and *GBC*) showing a larger number of overt pronouns. Although, in some languages, third-person plural null subjects are used to express indefinite readings of the subject ('people in general'), this is not responsible for the pattern in the Welsh data, where none of the third-person plural instances involved an indefinite reading (see Meyer 2011: 175–248 for discussion of this phenomenon in the history of Russian, Polish and Czech).

These results raise the question of whether there might be a person–number hierarchy in operation in our data, with less clearly individuated person–number combinations showing a higher propensity for null subjects, thus plur. > sing. for number. Third-person plural is the least individuated person–number combination, since it refers to multiple individuals not present in the speech event. It has been established for various varieties of Spanish that null subjects are more likely in the plural (Cameron 1992: 175 for Puerto Rican Spanish); however, in these studies, first-person plural rather than third-person plural shows the greatest tendency to omit the subject pronoun. Cameron (1992: 179–180) suggests that plural subjects are more likely to be mentioned in or inferable from the previous context, and that it is this that is responsible for these patterns. Ingham (2018: 247) makes the same suggestion for Old French. These remain tentative suggestions at this stage; it will be possible to test them more fully for the Welsh texts once a larger body of data has been collected and information-structure annotation for the texts has been completed.

A similar hierarchy for person (perhaps third > first > second) is plausible in principle, and this is the order than we find in our data, but the effect is not significant. Furthermore, the number of first and second-person plural subjects in our data set is low, limiting any conclusions. More data will have to be annotated to find out whether individuation of the subject referent and/or participant have a meaningful effect on the distribution of null subjects or whether this is perhaps a by-product of other factors, such as information structure. Being functionally motivated, we might expect all of the hierarchies just mentioned to operate cross-linguistically, unless other factors intervene in particular languages.

We have seen that these results bear some similarity with those for Modern Spanish. Old French, as examined by Vance (1997) and Prévost (2011), shows greater use of null subject pronouns in the third person than in the first person.

Work on Middle English has used a very similar methodology to that adopted here. Hence a statistical comparison is possible and potentially insightful, although it must be borne in mind that null subjects are far rarer in Middle English (with a frequency under 2% up to 1250 and under 1% thereafter) than in Middle and Early Modern Welsh (> 60% null subjects across our texts), and that this may be the reason behind any differences between the languages. Walkden and Rusten (2016: 461) report a stronger effect for person than for number. As in our Welsh corpus, null pronouns are favoured for third-person subjects, but, unlike in Welsh, they are disfavoured in the first person, suggesting a hierarchy third > second > first.[12] Furthermore, for number, a weakly significant factor in their analysis, singular rather than plural subjects favour null pronouns. Walkden (2014: ch. 5) reaches similar conclusions for a variety of other early Germanic languages, including Old English.

3.4.2. Diachrony of null subjects

As we saw in the introduction, the rise of overt subject pronouns is often observed to proceed in parallel with increasing 'impoverishment' of agreement morphology. When it comes to the history of Welsh, recall from Table 2 above that the inflectional paradigm largely remains unchanged from the medieval period up to the present day. Our mixed-effects model shows a weakly significant result for year of production, with earlier texts showing a slight preference for overt subjects compared to later ones. Formal literary Welsh still retains null subjects (Borsley, Tallerman and Willis 2007: 60), but in Present-day Spoken Welsh there is a strong preference for overt subject pronouns (Borsley et al. 2007:

[12] Note that they report the reverse effect in prose and poetry, but, since the Welsh corpus contains only prose texts, the same comparison cannot be made.

34). Our results, however, do not show a gradual loss of null subjects over time; in fact, we see a slight increase in null subjects from a rate of 62% null pronouns in Middle Welsh, to 73% and 78% in Early Modern and Modern Welsh (i.e. early-18th century) respectively and a negative value for year in the regression.

It is difficult to draw any firm conclusions based on the six texts in this pilot study, but there are a number of factors that remain to be investigated. The most pertinent involve the effects of text type, genre and, in particular, register. Today, there are large differences between spoken and literary forms of Welsh with respect to null subject. The development of register and its interconnection with the emergence of a standard written Welsh variety from the sixteenth century onwards is likely to play a significant role. It may be impossible to establish change in the spoken language through the largely literary texts used for the pilot. Comparison with text types closer to the spoken language may be necessary to have access to the emergence of a gap in usage between written and spoken varieties. We can tentatively hypothesise that the increase in frequency of null subjects over time in the pilot texts reflects the emergence and impact of the standard, with writers increasingly diverging from spoken usage, steadily incrementing the difference between the written and spoken language as null subjects become a mark of literary style. As a broader range of texts are included in the corpus, a more complex picture may emerge. Further study of syncretism in the Present-day Spoken Welsh paradigm in relation to the choice of subject pronoun may also shed more light on the reasons for loss of null subjects in speech.

3.5. Conclusion

This article has presented a pilot study of the history of subject pronouns in Welsh based on six annotated texts from the Parsed Historical Corpus of the Welsh Language (PARSHCWL). The texts were selected to represent a balance in translated vs. non-translated prose from three different time periods ranging from Middle Welsh to the early-eighteenth century. A mixed-effects logistic regression model was developed to test which factors have an effect on whether the subject pronoun is overt or null. Although a wide range of factors were tested (translated vs. non-translated text; main vs. subordinate clause; year of production; person; number), the best regression model included only year and person–number (a significant interaction term between the two), with third-person plural subjects strongly favouring null pronouns, and earlier texts exhibiting a weaker preference for overt pronouns. The frequency of null pronouns increases up until the most recent text in our corpus dating from the early-eighteenth century. This increasing preference for null subjects is in line with Present-day literary Welsh, but stands in stark contrast to the favouring of overt pronouns in Present-day Spoken Welsh.

For future research, in addition to extending the corpus to include more annotated texts from a wider range of time periods, genres and registers, it would be worthwhile to investigate other internal factors, such as animacy and discourse factors, for instance, information status and topic chains.

ABBREVIATIONS

B1588	1588 Bible translation
BP	*Breuddwyd Pawl*
MM	*Marwolaeth Mair*
Pwyll	*Pwyll Pendefig Dyfed*
CHSM	*Cronicl Hywel ap Syr Mathew*
GBC	*Gweledigaetheu y Bardd Cwsc*

REFERENCES

Adams, Marianne. 1987. From Old French to the theory of pro-drop. *Natural Language and Linguistic Theory* 5: 1–32.

Axel, Katrin. 2005. Null subjects and verb placement in Old High German. In Stefan Kepser and Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*. Berlin: Mouton de Gruyter, 27–48.

Axel, Katrin. 2007. *Studies on Old High German: Left sentence periphery, verb placement and verb-second*. Amsterdam: John Benjamins.

Berndt, Rolf. 1956. *Form und Funktion des Verbums im nördlichen Spätaltenglischen*. Halle: Max Niemeyer.

Borsley, Robert D., Maggie Tallerman and David Willis. 2007. *The syntax of Welsh*. Cambridge: Cambridge University Press.

Braune, Wilhelm and Ingo Reiffenstein. 2004. *Althochdeutsche Grammatik I: Laut- und Formenlehre*. Tübingen: Max Niemeyer Verlag.

Cameron, Richard. 1992. *Pronominal and null subject variation in Spanish: Constraints, dialects, and functional compensation*. Philadelphia: University of Pennsylvania doctoral dissertation.

Carvalho, Ana M. and Michael Child. 2011. Subject pronoun expression in a variety of Spanish in contact with Portuguese. In Jim Michnowicz and Robin Dodsworth (eds.) *Selected Proceedings of the 5th Workshop on Spanish Sociolinguistics*, 14–25.

Doyle, Aidan. 2002. Yesterday's affixes as today's clitics. In Ilse Wischer and Gabriele Diewald (eds.), *New reflections on grammaticalization*. Amsterdam: John Benjamins, 67–81.

Diertani, Chaya Eliana Ariel. 2011. *Morpheme boundaries and structural change: Affixes running amok*. Philadelphia: University of Pennsylvania PhD dissertation.

Evans, D. Simon. 1964. *A grammar of Middle Welsh*. Dublin: Dublin Institute for Advanced Studies.

Evans, D. Simon. 1971. Concord in Middle Welsh. *Studia Celtica* 6: 42–56.

Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In Charles N. Li (ed.), *Subject and topic*. New York: Academic Press, 149–188.
Håkansson, David. 2013. Null referential subjects in the history of Swedish. *Journal of Historical Linguistics* 3, 2: 155–191.
Haspelmath, Martin. 2018. Revisiting the anasynthetic spiral. In Heiko Narrog and Bernd Heine (eds.), *Grammaticalization from a typological perspective*. Oxford: Oxford University Press, 97–115.
Humboldt, Wilhelm von. 1825. Über das Entstehen der grammatischen Formen, und ihren Einfluss auf die Ideenentwicklung. *Abhandlungen der historisch-philologischen Klasse der königlichen Akademie der Wissenschaften zu Berlin aus den Jahren 1822 und 1823* 1822/23: 401–430.
Ingham, Richard. 2018. Topic, focus and null subjects in Old French. *Canadian Journal of Linguistics* 63. 242–263. https://doi.org/doi: 10.1017/cnj.2017.48.
Jaeggli, Osvaldo. 1981. *Topics in Romance syntax*. Dordrecht: Foris.
Jespersen, Otto. 1924. *The philosophy of grammar*. London: George Allen & Unwin.
Kinn, Kari. 2016. Referential vs. non-referential null subjects in Middle Norwegian. *Nordic Journal of Linguistics* 39: 277–310.
Kinn, Kari, Kristian A. Rusten and George Walkden. 2016. Null subjects in Early Icelandic. *Journal of Germanic Linguistics* 28: 31–78.
Kroch, Anthony and Ann Taylor. 2000. Penn–Helsinki Parsed Corpus of Middle English Prose. www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html.
Luft, Diana. 2015. Tracking *ôl cyfieithu*: Medieval Welsh translations in criticism and scholarship. *Translation Studies* 9: 168–182. https://doi.org/doi.org/10.1080/14781700.2015.1118404.
Manzini, Maria Rita and Leonardo Maria Savoia. 2005. *I dialetti italiani e romanci: Morfosintassi generativa*, vol. 1. Alessandria: Edizioni dell'Orso.
Meelen, Marieke and Silva Nurmio. 2019. Adjectival agreement in Middle and Early Modern Welsh native and translated prose. *Journal of Celtic Linguistics* 21: 1–28
Meelen, Marieke and David Willis. 2021. Towards a historical treebank of Middle and Early Modern Welsh, part I: Workflow and POS tagging. *Journal of Celtic Linguistics* 22: 125–154.
Meelen, Marieke and David Willis. 2022. Towards a historical treebank of Middle and Modern Welsh: Syntactic parsing. *Journal of Historical Syntax* 6, 5: 1–32.
Meyer, Roland. 2009. Zur Geschichte des referentiellen Nullsubjekts im Russischen. *Zeitschrift für Slawistik* 54: 375–397.
Meyer, Roland. 2011. *The history of null subjects in North Slavonic: A corpus-based diachronic investigation*. Regensburg: University of Regensburg habilitation thesis.
Nagy, Naomi and David Heap. 1998. Franco-Provençal null subjects and constraint interaction. *Chicago Linguistics Society* 34: 151–166.
Norde, Muriel. 2009. *Degrammaticalization*. Oxford: Oxford University Press.
Parina, Elena. 2018a. The good, the bad and the translator: The concept of predestination in a Middle Welsh translation of the *Elucidarium*. *Indo-European Linguistics and Classical Philology* 22: 1003–1012. https://doi.org/doi:10.30842/ielcp230690152272.
Parina, Elena. 2018b. Multiple versions of *Breuddwyd Pawl* as a source to study the work of Welsh translators. *Studia Celto-Slavica* 9: 79–100.
Parina, Elena. 2022. Relative clauses with overt marking in Early Modern Welsh. *Journal of Historical Syntax* 6, 10: 1–23.

Plein, Kerstin. 2018. *Verbalkongruenz im Mittelkymrischen*. Berlin: Curach Bhán.
Poletto, Cecilia. 2000. *The higher functional field: Evidence from northern Italian dialects*. New York: Oxford University Press.
Poletto, Cecilia. 2020. Null subjects in Old Italian. In Rebecca Woods and Sam Wolfe (eds.), *Rethinking verb second*. Oxford: Oxford University Press, 325–347.
Prévost, Sophie. 2011. Expression et position du sujet pronominal en français: Évolution en français. In Jacques François and Sophie Prévost (eds.), *L'évolution grammaticale à travers les langues romanes*. Leuven: Peeters, 13–33.
Randall, Beth. 1999. *CorpusSearch: A linguistic search program*. Philadelphia: Drexel University MSc dissertation.
Randall, Beth, Ann Taylor and Anthony Kroch. 2005. *Corpussearch 2*. http://corpussearch.sourceforge.net/credits.html.
Ranson, Diana L. 2009. Variable subject expression in Old and Middle French prose texts: The role of verbal ambiguity. *Romance Quarterly* 56: 33–45.
Rizzi, Luigi. 1986. On the status of subject clitics in Romance. In Osvaldo Jaeggli and Carmen Silva-Corvalan (eds.), *Studies in Romance linguistics*. Dordrecht: Foris, 391–419.
Roberts, Ian. 2005. *Principles and parameters in a VSO language: A case study in Welsh*. New York: Oxford University Press.
Roberts, Ian. 2014. Taraldsen's Generalization and language change: Two ways to lose null subjects. In Peter Svenonius (ed.), *Functional structure from top to toe: The cartography of syntactic structures*, 9. Oxford: Oxford University Press, 115–148.
Roberts, Ian. 2019. *Parameter hierarchies and universal grammar*. Oxford: Oxford University Press.
Roma, Elisa. 2000. How subject pronouns spread in Irish: A diachronic study and synchronic account of the third person + pronoun pattern. *Ériu* 51: 107–57.
Rusten, Kristian A. 2013. Empty referential subjects in Old English prose: A quantitative analysis. *English Studies* 94, 8: 970–992.
Rusten, Kristian A. 2014. Null referential subjects from Old to early Modern English. In Kari E. Haugland, Kevin McCafferty and Kristian A. Rusten (eds.), *'Ye whom the charms of grammar please': Studies in English language history in honour of Leiv Egil Breivik*, vol. 4. Oxford: Peter Lang, 249–270.
Rusten, Kristian A. 2015. A quantitative study of empty referential subjects in Old English prose and poetry. *Transactions of the Philological Society* 113, 1: 53–75.
Schøsler, Lene. 2002. La variation linguistique: le cas de l'expression du sujet. In Rodney Sampson and Wendy Ayres-Bennett (eds.), *Interpreting the History of French, A Festschrift for Peter Rickard on the occasion of his eightieth birthday*. New York: Rodopi, 187–208.
Siewierska, Anna. 1999. From anaphoric pronoun to grammatical agreement marker: Why objects don't make it. *Folia Linguistica* 33, 2: 225–251.
Simonenko, Alexandra, Benoit Crabbé and Sophie Prévost. 2019. Agreement syncretization and the loss of null subjects: Quantificational models for Medieval French. *Language Variation and Change* 31, 3: 275–301.
Stausland Johnsen, Sverre. 2016. Null subjects, preproprial articles, and the syntactic structure of Old Norwegian pronouns. *Norsk Lingvistisk Tidsskrift* 34: 183–217.
Taraldsen, Knut Tarald. 1980. *On the Nominative Island Condition, vacuous application and the* that-*trace filter*. Bloomington, Ind.: Indiana University Linguistics Club.

Teixeira, Raquel Figueiredo Alessandri. 1986. *Zero anaphora in Brazilian Portuguese subjects and objects: Morphological and typological considerations*. Berkeley, Calif.: University of California, Berkeley PhD dissertation.

Thomas, Peter Wynn, D. Mark Smith and Diana Luft (eds.). 2007–17. *Rhyddiaith Gymraeg: Welsh Prose 1300–1425*. Cardiff: Cardiff University. http://www.rhyddiaithganoloesol.caerdydd.ac.uk.

Vance, Barbara. 1997. *Syntactic change in medieval French: Verb-second and null subjects*. Dordrecht: Kluwer.

Van Gelderen, Elly. 2011. *The linguistic cycle: Language change and the language faculty*. Oxford: Oxford University Press.

Walkden, George. 2013. Null subjects in Old English. *Language Variation and Change* 25: 155–178.

Walkden, George. 2014. *Syntactic reconstruction and Proto-Germanic*. Oxford: Oxford University Press.

Walkden, George. 2016. Null subjects in the Lindisfarne Gospels as evidence for syntactic variation in Old English. In Julia Fernández Cuesta and Sara M. Pons Sanz (eds.), *The Old English glosses to the Lindisfarne Gospels: Language, author and context*. Berlin: Mouton de Gruyter, 237–254.

Walkden, George and Kristian A. Rusten. 2017. Null subjects in Middle English. *English Language and Linguistics* 21, 3: 439–473.

Willis, David. 1997. Clausal coordination and the loss of verb-second in Welsh. *Oxford Working Papers in Linguistics, Philology and Phonetics* 2. 151–72.

Willis, David. 2017. Degrammaticalization. In Adam Ledgeway and Ian Roberts (eds.), *The Cambridge handbook of historical syntax*. Cambridge: Cambridge University Press, 28–48.

Willis, David and Ingo Mittendorf (eds.). 2004. *Corpws Hanesyddol yr Iaith Gymraeg: A Historical Corpus of the Welsh Language 1500–1850*. Cambridge: University of Cambridge. http://www.celticstudies.net.

Appendix A
An example of a CorpusSearch query

Figure 5 shows a sample query for extracting all Verb-Subject (VS) main clauses with overt pronominal subjects from the parsed corpus.

```
node: CP*
define: …/definitions.def
query: (NP-SBJ* iDominates PRO|PROC|PROR) AND
(!embedded_clause Dominates NP-SBJ*) AND (NP-SBJ* HasSister
finite_verb) AND (NP-SBJ* Precedes finite_verb)
```

Fig. 5. CorpusSearch query file to extract main clauses with overt pronominal subjects following verbs (Verb-Subject order).

CorpusSearch queries are text files (saved as .q) that specify which sentences to extract from a corpus that is annotated in the style of the Penn Parsed corpora (Randall 1999). Queries first specify the search domain, in this case the root node of the constituency tree, labelled CP. The asterisk wildcard indicates that all types of clause (CP) should be searched, that is, main/matrix clauses (CP-MAT), subordinate complement clauses (CP-SUB), etc. This is followed by a path that leads to a definitions file, in which shortcuts for phrases are specified.

```
embedded_clause:
CP-SUB*|CP-REL*|CP-THT*|CP-FOC|CP-QUE-SUB*|CP-
SMC*|CP-ADT*|CP-ADV*|CP-FRL*|CP-CMP*
finite_verb: VP*|VB*|BE*-*|GT*-*|DO*-*
```

Fig. 6. A CorpusSearch definitions file.

This file, shown in Figure 6, contains shortcuts for ‘embedded_clause’ and ‘finite_verb’. The former shortcut defines an embedded clause as including any type of subordinate complement clause (CP-SUB*), any type of relative clause (CP-REL*) and so on. The shortcut ‘finite_verb’ ensures that verbal nouns and other non-finite verb forms are excluded from the query, as it is limited to POS tags like VB* (any lexical verb, but not VN ‘verbnoun’), BE*-* (any tensed form of ‘be’), DO*-* (any tensed form of ‘do’) etc. With these definitions in place, the query specifies which sentences to extract. In this query, which is for clauses

containing overt subjects, the search procedure looks for instances of any subject (NP-SBJ*) immediately dominating ('iDominates') an independent pronoun (PRO), a conjunctive pronoun (PROC), or a reduplicated pronoun (PROR). In addition, this query is limited to main clauses, which is specified by the further requirement that the subject must not ('not' indicated as '!') be dominated by a node representing an embedded clause. Non-finite clauses are excluded by specifying that the subject must have a sister node that is a finite verb. Finally, we restrict ourselves to VS orders with the requirement that a finite verb precedes the subject ('finite_verb Precedes NP-SBJ*'. This query will then yield an output file with all sentences in the corpus that meet these criteria.