How lexical merger can drive grammaticalization
Lieven Danckaert, Liliane Haegeman, Sophie Prévost

To cite this version:
Lieven Danckaert, Liliane Haegeman, Sophie Prévost. How lexical merger can drive grammaticalization: Third person pronouns from Latin to Old French. Journal of Historical Syntax, The University of Konstanz, 2021, 10.18148/hs/2021.v5i16-25.66. hal-03506435

HAL Id: hal-03506435
https://hal.archives-ouvertes.fr/hal-03506435
Submitted on 2 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The aim of this paper is to elucidate the genesis of Romance third person personal pronouns deriving from Latin ille and ipse, with special reference to the history of French. Drawing a parallel with the taxonomy of phonological mergers from Labov (1994), we argue that the Late Latin competition between ille and ipse was resolved through a series of lexical mergers (i.e. the opposite of better known lexical splits). Concretely, we propose that strong personal pronouns (such as French lui) arose through merger of ille and ipse, to the effect that the union of the feature sets of the latter two elements was transferred to the newly formed category. In contrast, weak pronouns (like French il) only retain the intersection of the feature sets of ille and ipse. By creating new functional categories, lexical merger thus acts as a driving force behind grammaticalization.

1 Introduction

1.1 Pronouns and determiners from Latin to Romance

As is well known, Latin did not have definite articles; in contrast, all of the Romance daughter languages do have this category, which in most cases is derived either from the Latin demonstrative pronoun ille, or, less often, from the intensifying reflexive ipse (key references from the very bulky literature

* Many thanks to the audience at DiGS 21, two anonymous reviewers, and Johanna Wood for much-appreciated feedback.
on this subject include Meader 1901, Trager 1932, Abel 1971, Selig 1992, Giusti 2001, Carlier & De Mulder 2010, Ledgeway 2015, Lüdtke 2015, Ledgeway & Smith 2016; in particular on the remnants of ipse, see Aebischer 1948 and Sornicola 2008, 2012). The grammaticalization of definite articles is arguably related to the afterlife of Latin ille and ipse in contexts other than that of adnominal determiners, namely third person personal pronouns. Aspects of this development were studied in, among others, Wanner (1987), Sornicola (1998, 2001), Schøsler & Strudsholm (2013), Prévost (2018) and Danckaert & Prévost (submitted). Relatively few studies have been devoted to these two developments in conjunction (but see e.g. Renzi 1997; Vincent 1997; Boral Hertzenberg 2015).

As noted in Cappellaro (2016: 722–723), one question which has received relatively little attention is how exactly true third person personal pronouns (i.e. a category distinct from, and perhaps in opposition to, demonstrative pronouns) came into being in the history of Latin/Romance. At first sight, the relevant development does not seem to be a ‘classical’ case of grammaticalization: for one thing, just as is the case with the development of definite articles, we are not dealing with a lexical category that develops into a functional one. In addition, some of the newly created items, in particular strong personal pronouns (on which, see Section 1.3) are also not phonologically, and perhaps not even semantically, reduced with respect to the pre-existing categories. With respect to this last point, Cappellaro (2016: 723) correctly points out that the change is not simply one from ‘deictic’ to ‘anaphoric’ (pace Renzi 1997): Latin demonstratives could express anaphoric reference, and strong third person pronouns such as present-day Italian lui and French lui can still be used deictically.

The present paper aims to shed some light on the origins of third person personal pronouns in the Romance language family (with special reference to French): by applying insights from the literature on phonological change to the realm of morphosyntax, we propose that the merger of two competing (in the sense of Kroch 1989, 1994) functional categories, in this case ille and ipse, can give rise to the creation of new functional material. We suggest that recognizing diachronic processes of this type may deepen our understanding of certain cases of grammaticalization.

1.2 Third person reference in Latin

To overtly express third person reference, speakers of Latin could make use of a series of pronouns which are traditionally labelled ‘demonstratives’ (is, hic, iste, and ille); other options include the reflexive intensifier ipse (‘self, the very’) and the identity term idem ‘the same’, to which we could add what
is known as the ‘connecting relative’ *qui* (on which, see Danckaert 2012: 181–185, and references cited there). Importantly, the relative frequencies of these elements do not remain stable over time. Väänänen (1981: 121) summarizes the development from Classical to Late Latin as follows (forms marked with an asterisk do not survive in Romance; the arrows show which item took the place of which):

\[(1) \quad \text{IS}^* \quad \text{HIC}^* \quad \text{ISTE} \quad \text{ILLE} \quad \text{IPSE} \quad \text{IDEM}^*\]

In Section 3 of this paper, we will offer corpus evidence confirming that this basic picture is correct, supplementing it with some new observations about the syntactic distribution of the bare (i.e. non-adnominal) nominative forms of these pronouns. As pointed out in Danckaert (2012: 217–223), there is every reason to believe that the traditional view that the elements listed in (1) are not personal pronouns is indeed correct. Cardinaletti & Starke (1999a: 284) list a number of diagnostics to distinguish (overt) personal pronouns from demonstratives. One of them is that the former are subject to Principle B of the Binding Theory, and the latter to Principle C. In other words, only personal pronouns can appear in the c-command domain of a co-referential category. For example, in Latin we never find demonstrative pronouns acting as the subject of a complement clause, in a configuration whereby the relevant demonstrative is coreferent with the matrix subject. In Old French on the other hand, this construal can readily be found:

\[(2) \quad (a) \quad \text{Oliver} \quad \text{sent} \quad [\text{qu’} \quad \text{il} \quad \text{est} \quad \text{a mort} \quad \text{nasfret}] \quad \text{Oliver} \quad \text{feel.PRS.3SG} \quad \text{that} \quad \text{be.PRS.3SG} \quad \text{to death} \quad \text{wounded} \quad \text{‘Olivier feels that he is wounded to death.’} \]

\[\text{(Roland, ca. 1100, v. 1965)}\]

1 For general discussion of pronominal reference in (the history of) Latin, the interested reader is referred to Pieroni (2010) and Bordal Hertzenberg (2015).
2 Given its overall low token frequency, in what follows we will not further take into account IDEM.
3 This claim is in need of some qualification, as in certain circumstances it does in fact seem possible for demonstratives to be c-commanded by a co-indexed constituent (see e.g. Hinterwimmer & Brocher 2018 (on German)). However, this observation does not affect our argument, which is that the absence of third person pronouns in ‘Principle B’ configurations can be taken to mean that the relevant items were not personal pronouns.
To the best of our knowledge, in the entire Latinity the counterpart of (2) involving ille, is, hic or iste is never attested: though strictly speaking an argumentum ex silentio, this observation does suggest that these elements are demonstratives, not personal pronouns.

Note that the non-attestation of embedded demonstratives c-commanded by a co-indexed category is not due to the fact that Latin was a pro-drop language. In languages that allow for zero anaphora and which have third-person personal pronouns, the latter are often dispreferred in the c-command domain of coreferential expressions, pro always being the default option in those environments. Only when pro-drop is for some (pragmatic) reason dispreferred is an overt (strong) personal pronoun felicitous (for relevant discussion, see e.g. Lobo & Silva 2015: 197–199). We can illustrate this point by means of the Spanish examples in (3)-(4) (from Larson & Luján 1989: 6, their (27)–(28)). The question-answer pair in (3) starts with a long-distance object question; therefore, the embedded object in (3b) can be said to be a narrow focus. In the same clause, the subject is part of the presupposition: for this reason pro is strongly preferred over an overt pronoun. In (3b), the hash-sign indicates pragmatic infelicitousness, rather than ungrammaticality.

This exchange is minimally different from that in (4): here the a-example is a long-distance-subject question, and as a result, the embedded subject in the corresponding b-example acts as a new information focus. When the intended coreference relations obtain, this interpretation requires the overt lexicalization of the embedded pronoun:
How lexical merger can drive grammaticalization

(4) (a) Quién cree Juan [que t ganará el premio]?
who think.prs.3sg Juan that win.fut.3sg the award
‘Who does John think will win the award?’

(b) Juan cree [que #∅/el ganará el premio]
Juan think.prs.3sg that he win.fut.3sg the award
‘John thinks that #he/HE will win the award.’

We can conclude that there is no principled ban in null subject languages on strong personal pronouns appearing in the c-command domain of a coreferential element. As a result, we can maintain that the absence of this pattern in third person environments in the Latin corpus suggests that the language only had demonstrative third person pronouns. This conclusion is confirmed by the fact that the first and second person singular pronouns ego ‘I’ and tu ‘you’ can in fact be c-commanded by a coreferential category:

in.truth why prt say.prs.3sg voice.prs.1sg you.dat.pl that I.nom hear.prf.1sg voices.acc angels.gen in heavens.abs.
“Sanctus” in praise.abs. Domini proclaiming nom.m.pl
‘He said: “I tell you in truth that I personally heard angels’ voices in heaven proclaiming “Sanctus” to praise the Lord”.’

(Greg. Tur., vit. patr. 16.2, late sixth c. ce)

(6) quid enim ∅ non dicis, [quod tu benigne]
why prt not say.prs.2sg that you.nom.sg kind.adv
iam in ea operatus es]?
already in dem.abs.f.sg worked.nom.m.sg be.prs.2sg
‘Why don’t you say that you have already worked kindly in her?’

(Greg. M., in cant. 44, late sixth c. ce)

Finally, and very importantly, the coreference restriction just illustrated does not hold foripse, no doubt because the original function of this element is not

4 The examples in (5) to (7) were all drawn from later authors, who frequently use finite complement clauses, in environments where earlier writers would have used an AcI (i.e. an infinitival clause with an accusative subject). It is not easy to find similar examples from Early or Classical Latin, but we can safely assume that this is due to the scarcity of finite, and non-obviative, complement clauses in earlier texts, not to the properties of Early and Classical Latin personal pronouns. On the Early, Classical and Late Latin corpus samples that we will investigate in this study, see Sections 2.1–2.2.
that of a (deictic or anaphoric) pronoun; rather, in an example like (7) *ipse* presumably acts as a secondary predicate co-indexed with a phonologically null subject (on *ipse*, see e.g. Bertocchi 2000 and Pieroni 2007; see Gast 2006 for cross-linguistic parallels). In any event, there is no reason to assume that *ipse* was at any point subject to Principle C. As we will argue below, this point is quite important for the later development of Romance third person pronouns.

(7) *unde probamus quod *∅* *iam nouerat* whence *prove.PRS.1SG* that *already know.PLFV.3SG*

[quia *∅* *ipse* *baptizaret* in *spiritu* that *REFL.NOM.M.SG* baptise.IPfv.SBJV.3SG in *spirit.ABL sancto]*?

hol.y.ABL

‘How can we prove that he already knew that he would himself baptise in the Holy Spirit?’

(Aug., in euang. Ioh. 124.8, ca. 407 ce)

1.3 Aims and structure of the paper

The main goal of the paper is to elucidate how two types of Romance third person personal pronouns came into being. We will focus in particular on (historical) French, but our analysis should apply to other Romance varieties as well. For reasons of space, we will only consider subject pronouns. More precisely, adopting the three-way typology of pronouns developed in Cardinaletti & Starke (1999b), we will be concerned with strong and weak subject pronouns, and thus not with subject clitics such as those found in many Northern Italian dialects (see e.g. Poletto 2000 and Poletto & Tortora 2016). Examples of the two types of pronouns that we will be concerned with include French *lui* and Italian *lui* (which are both strong), and French *il* and Italian *egli* (which are weak).

The core idea to be worked out is that the competition (Kroch 1989, 1994) between Late Latin *ille* and *ipse* was resolved through various types of lexi-

---

5 We will also not be concerned with the related development of the rise of definite articles in Romance, which as mentioned also derive from *ille* and *ipse*.

6 Cardinaletti & Starke (1999b) list the following differences between strong and weak pronouns: (i) only the former can be coordinated (with another strong pronoun, or with a full lexical DP), (ii) weak but not strong pronouns may be phonologically reduced, (iii) only strong pronouns can appear as complements of prepositions, and (iv) for prosodic reasons, weak pronouns have to surface in dedicated functional projections in the clause, whereas strong pronouns are allowed to occur VP-internally as well as in right- and left-peripheral positions. The authors also propose that there are semantic differences between the two categories: we refer to the original paper, as well as to Cardinaletti & Starke (1999a) for full discussion.
How lexical merger can drive grammaticalization

cal merger of these two items, rather than through extinction or specialization
of one element, and the concomitant generalization of the other. In particu-
lar, drawing upon the typology of phonological mergers proposed in Labov
(1994), we propose that ille and ipse underwent ‘merger by expansion’ to
yield strong personal pronouns, and that a second operation of ‘merger by
approximation’ resulted in the creation of weak pronouns; thirdly, the phono-
logical shape of a newly formed pronoun is always the result of ‘merger by
transfer’ (which can either go towards ille or to ipse). Our analysis thus sug-
gests that at least in some cases, the need to resolve competition between two
(and perhaps even more) lexical items may drive morphosyntactic change,
and thus grammaticalization. If on the right track, our analysis also entails
that the parallels between phonological and lexical change (at least as far as
changes affecting the feature composition of functional vocabulary items is
concerned) may be deeper than previously thought.

Before we proceed, we would like to emphasize that the genesis of third
personal personal pronouns is – though not unrelated to – independent of
another major development, namely the loss of pro-drop in certain parts of
the România. The independence of the two developments clearly comes to
the fore when we consider that all Romance languages have third person per-
sonal pronouns, but only some (most notably Gallo-Romance) varieties lost
the possibility of having referential null subjects. However, in Section 5 we
briefly speculate on the syntax of pro-drop in Old French.

The paper is structured as follows. In Section 2, we present the results
of a series of corpus studies documenting the frequency and distribution of
pronominal subjects in Latin and Old French. The descriptive generalizations
emerging from these corpus studies constitute the empirical foundation of the
following discussion. We pay special attention to the situation in Old French,
where the differential behaviour of main and embedded clauses makes it
somewhat difficult to properly interpret the data. In Section 3, we turn to
the pragmatic status of ille and ipse in Late Latin, pointing out certain dis-
tributional and interpretive differences between these two third person refer-
ence devices. In Section 4, we first elaborate on various types of phonological
mergers, as discussed in Labov (1994) and related work. We go on to apply
this apparatus to the development of third person reference markers in the
history of Latin/Romance. In Section 5, we briefly touch upon an issue that
is arguably related to – but as mentioned above, clearly distinct from – the
rise of weak personal pronouns in the history of French, namely the demise
of pro-drop. Section 6 concludes the paper.
2 Subject pronouns from Latin to Old French: The lay of the land

The aim of this section is to formulate a number of empirical generalizations concerning the frequency and syntactic distribution of overt personal and demonstrative pronouns acting as the syntactic subject of finite clauses. When quantifying the rate of these items, we will use the incidence of clauses with a referential null subject as a baseline, but we repeat that our focus is never on (the syntax of) pro-drop per se: rather, what we are interested in is a diachronic comparison between pronominal third person subjects on the one hand, and first and second person subjects on the other.

Although the broad outlines of the data that we are concerned with are generally known, there is at present no reliable corpus study to document the finer details of the relevant developments. As we will show, the corpus data do in fact reveal some rather surprising contrasts between (Late) Latin and Old French. In particular, our results show that the two languages behave very differently with respect to the parameters ‘Clause type’ ([± root clause]), and to a lesser extent ‘Person’ (operationalized as [± third person subject]). The basic generalizations are the following. First, in Late Latin overt subject pronouns (personal pronouns in the case of first and second person, demonstratives and *ipse* for the third person) are generally more common in main than in embedded clauses. In Old French, we find exactly the opposite. Second, in Latin, Person effects only play a relatively minor role when it comes to predicting the incidence of overt pronominal subjects. This is most notably the case in Late Latin, where overt demonstrative pronouns and overt personal pronouns occur at similar rates (at least when acting as a grammatical subject). In Old French on the other hand, overt first and second person pronouns are generally more common than their third person counterparts. Crucially, closer inspection reveals complex interactions between the factors ‘Clause type’ and ‘Person’ in Old French, to the effect that this last generalization does not hold in embedded clauses.

2.1 Early and Classical Latin

To gauge the incidence of overt subject pronouns in Early and Classical Latin we analysed a corpus consisting of the seven text samples listed in Table 1.7 8

One important factor we took into account when putting together this corpus

---

7 With the term ‘Early Latin’, we refer to the language of all Latin texts dating from before 100 BCE. The label ‘Classical Latin’ is reserved to the period of ca. 100 BCE until ca. 120 CE (roughly speaking from Cicero until Suetonius). In other words, the term ‘Classical’ is not used with any stylistic connotation.

8 All texts for the corpus study reported on in Section 2.1 were taken from the Brepols database, available online at www.brepolis.net.
How lexical merger can drive grammaticalization

<table>
<thead>
<tr>
<th>Author</th>
<th>Text</th>
<th>Date (approx.)</th>
<th>N of pronominal subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plautus</td>
<td>Amphitruo</td>
<td>185 BCE</td>
<td>1068</td>
</tr>
<tr>
<td>Terence</td>
<td>Andria</td>
<td>165 BCE</td>
<td>1027</td>
</tr>
<tr>
<td>Cicero</td>
<td>In Verrem II.1</td>
<td>70 BCE</td>
<td>1032</td>
</tr>
<tr>
<td>Livy</td>
<td>Ab Urbe condita</td>
<td>5 CE</td>
<td>658</td>
</tr>
<tr>
<td>Seneca</td>
<td>Ad Lucilium 1–25</td>
<td>50 CE</td>
<td>605</td>
</tr>
<tr>
<td>Petronius</td>
<td>Satyricon</td>
<td>60 CE</td>
<td>1001</td>
</tr>
</tbody>
</table>

Table 1 Corpus of Early and Classical Latin texts.

is whether or not a given text contains enough tokens of all Person-Number combinations. For this reason we left aside among other things technical treatises, which typically contain little or no second persons (singular and plural), nor many first person verbs (in particular plurals, and to a lesser extent first person singulars).

Only finite clauses with a pronominal (null or overt; personal or demonstrative) subject were taken into account. For third person contexts, we included the pronouns *is, hic, iste, ille* and *ipse*, but we took care to only count the latter as a genuine subject pronoun if the ‘reinforced reflexive’ reading illustrated in (7) is not available.\(^9\) We excluded all clauses which appear as a second, third etc. conjunct and which contain what appears to be a null subject coreferential with that of the first conjunct. The reasons for this is that the relevant tokens are structurally ambiguous, in that they are amenable to a phrase structure analysis with ‘low’ coordination, i.e. of two (extended) VPs out of which a subject has been extracted in an across-the-board manner. Clauses with an impersonal finite verb such as *oportet* ‘it befits’ and *licet* ‘it is allowed’ were also excluded from our dataset, as in this environment there is no alternation between null and overt subjects. The same methodological choices were made for the two other data sets that we will report on shortly (with the proviso that the set of overt subject pronouns is of course different in Old French: see Section 2.3 for further clarifications).

Our findings are summarized in Figure 1, where we have plotted the coefficients of a (fixed effects) logistic regression, modeling the likelihood for

\(^9\) For the time being, we will not further differentiate between these five items, but we refer to Section 3 for additional discussion of their relative rate and distribution.
Figure 1  Rate of overt pronominal subjects in Early and Classical Latin (*ca.* BCE - 60 CE): Effects of ‘Clause type’ and ‘Person’.

The basic generalizations that emerge from the statistical analysis can be summarized as follows. First, overt subjects are overall more likely to be found in third person contexts (cf. the solid lines in Figure 1; estimated log odds ($\beta = 0.429$, $p < 0.001$), and less likely in embedded clauses (cf. the thicker lines in Figure 1; $\beta = -0.628$, $p < 0.001$). In addition, there is a significant interaction between the factors ‘Date’ and ‘Clause type’, which says that the observed discrepancy between main and embedded clauses becomes stronger over time ($\beta = 0.003$, $p < 0.001$). A main effect of ‘Date’ turned out not to be significant, nor did any of the other interaction terms between the
How lexical merger can drive grammaticalization

afore-mentioned predictors. In any event, as can be deduced from Figure 1, though significant, the size of the three effects just mentioned is overall relatively small.

Note, in passing, that it is not immediately obvious how the relative preponderance of overt third person subjects is to be understood.\(^{10}\) It is tempting to interpret this effect in terms of the Accessibility Marking Scale from (Ariel 1990: 73), according to which demonstratives tend to refer to less accessible referents than (personal) pronouns. Assuming that less accessible referents need more explicit coding, the relatively high frequencies of bare demonstrative subjects would indeed be expected. On the other hand, as we will see later, pronominal subjects in Old French main clauses are sensitive to Person effects which are quite different from the ones discussed here. This suggests that Person-asymmetries related to the frequency of occurrence of various types of pronouns are not always to be explained in pragmatic terms only (or at least not in terms of Accessibility). Clearly, what is needed at this point is a detailed study of the pragmatic properties of various (overt and covert) pronouns and their antecedents. From a methodological point of view, the multifactorial, corpus-based approach pursued in Bordal Hertzenberg (2015) seems like a good starting point to design such a large-scale study, preferably with some diachronic depth. Needless to say, such an enterprise goes well beyond the scope of this study (but see Section 3 of this paper, as well as Danckaert & Prévost submitted for some initial remarks).

Turning to the observed Clause type effect, nothing in what we have said thus far makes us predict that there should be any difference between main and embedded clauses, but as we will see, this factor is particularly important when it comes to modeling the transition from Latin to French.

2.2 Merovingian Latin

Let us now turn to pronominal subjects in Late Latin. We analysed a sample of seven Merovingian and Early Carolingian Latin texts dating from between 520 and 835 CE, which were all composed in the geographical area roughly corresponding to present-day France. We used the text editions from the Monumenta Germaniae Historica (viz. the Scriptores rerum Merovingicarum, cf. Krusch & Levison 1884–1920). A full description of this sub-corpus is given in Table 2.

\(^{10}\) We are indebted to a reviewer for very insightful discussion of the issues discussed in this paragraph.
As above, we ran a logistic regression on this dataset with ‘Date’, ‘Clause type’ and ‘Person’ as fixed effects. The results we obtained are summarized in Figure 2.

In the model from which we extracted the coefficients plotted in Figure 2, only an interaction term between Person and Date (β = 0.002, p = 0.04) and one between Clause type and Date (β = -0.002, p = 0.03) came out as (borderline) significant at the 0.05 level. As to the factor Person, the obtained result would mean that originally first and second person subjects were more likely to be overt, but that towards the early ninth century this is no longer true (compare the trajectories of the two dotted and the two solid lines in the graph: they both stay parallel to each other, but the two pairs cross each other around 650 CE). Similarly, the two thicker lines and the two thinner ones also cross each other, which is indicative of an interaction effect between Clause type and Date to the effect that, in our later texts, it is no longer generally the case that overt pronominal subjects are preferred in main clauses. However, in addition to the p-values reported above not being far below the 0.05 threshold, the estimated log odds (i.e. the β-coefficients) are particularly low, which means that the relevant significant effects are only very small. In sum, there are some reasons to doubt that the effects plotted in Figure 2 are reliable. To have a better idea of what is going on in this sub-corpus, we ran a more powerful mixed-effects analysis, without a time variable, but with the individual source texts as random effects. The strongest such model we could build is one with random intercepts for ‘Text’ and random slopes for the two levels of the factor ‘Clause type’, as well as a fixed effect for Clause type. A fixed

---

Table 2: Corpus of Late Latin texts.

<table>
<thead>
<tr>
<th>Text</th>
<th>Date (approx.)</th>
<th>N of pronominal subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Vita Genovefae</em></td>
<td>6th c.</td>
<td>259</td>
</tr>
<tr>
<td>Gregory of Tours, <em>Historiae</em> 6</td>
<td>580–585</td>
<td>856</td>
</tr>
<tr>
<td>Fredegar, book 4</td>
<td>658–660</td>
<td>568</td>
</tr>
<tr>
<td><em>Vita Eligii</em>, book 1</td>
<td>ca. 670–680</td>
<td>453</td>
</tr>
<tr>
<td><em>Liber Historiae Francorum</em></td>
<td>727</td>
<td>674</td>
</tr>
<tr>
<td><em>Vita Trudonis</em></td>
<td>784–791</td>
<td>408</td>
</tr>
<tr>
<td><em>Vita Galli Walahfrido</em></td>
<td>833</td>
<td>631</td>
</tr>
</tbody>
</table>

---
main effect for Person did not turn out to be significant, neither when considered in isolation, nor when added to the model with significant Clause type-effects. The estimated log odds for the factor ‘Clause type’ indicates that overt pronominal subjects are dispreferred in embedded clauses ($\beta = 1.247$, $p < 0.001$).

Let us then compare the results of the two Latin sub-corpora. The role of the factor ‘Clause type’ seems to remain pretty constant, in that overt pronominal subjects are always more frequent in main clauses. In contrast, the role of ‘Person’ seems to change in the transition from Classical to Late Latin: the once robust contrast between first and second person on the one hand, and third person on the other, seems to have evaporated, or at least to have become very weak, in the Merovingian period. We can interpret this observation to mean that at least from a quantitative perspective (i.e. as far as rates of usage are concerned), the various third person pronouns start to behave more and more like true personal pronouns, a result which is, of course, entirely

Figure 2  Rate of overt pronominal subjects in Late Latin (ca. 550–833 CE): Effects of ‘Clause type’ and ‘Person’. 
expected (given what we know about later developments). In Section 3, we will further elaborate on the discourse interpretation of in particular \textit{ille} and \textit{ipse} in Late Latin (but as mentioned we will not attempt a systematic comparison between the conditions of usage associated with third and first/second person pronominal subjects).

2.3 (Early) Old French

Finally, to study subject pronouns in Old French, we analysed 11 texts dating from \textit{ca.} 1000 to 1250, which are listed in Table 3.\footnote{For this part of the corpus study we used texts from the \textit{Base de Français Médiéval} (http://bfm.ens-lyon.fr/).} As to the overt subject pronouns, we took into account only personal pronouns (for a complete list of possible forms that Old French subject pronouns could take, see Buridan \textit{2000: 408}), leaving aside the various newly formed demonstratives (on which, see Guillot \textit{2015}, Guillot-Barbance \textit{2017}). To maximally ensure that the Old French data can be compared with the results obtained from our Latin corpus samples, here too we did not take into account clauses with an impersonal verb. As shown in Figure 3, in the first centuries of the second millennium the rate of overt subjects changes very quickly.

\begin{table}[h]
\centering
\begin{tabular}{lll}
\textbf{Text} & \textbf{Date (approx.)} & \textbf{N of pronominal subjects} \\
\hline
Passion & \textit{ca.} 1000 & 304 \\
\textit{Vie de Saint Alexis} & \textit{ca.} 1050 & 471 \\
\textit{Chanson de Roland} & \textit{ca.} 1100 & 600 \\
\textit{Eneas} & \textit{ca.} 1155 & 607 \\
Beroul, \textit{Tristan} & late 12\textsuperscript{th} c. & 540 \\
Chretien de Troies, \textit{Yvoine} & 1180 & 580 \\
R. de Clari, \textit{La Conqueste de Constantinople} & after 1205 & 434 \\
\textit{Aucassin et Nicolete} & late 12\textsuperscript{th}/early 13\textsuperscript{th} c. & 601 \\
\textit{Queste del Saint Graal} & 1225 & 580 \\
J. Renart, \textit{Roman de la rose ou de G. de Dole} & \textit{ca.} 1228 & 591 \\
J. Sarrasin, \textit{Lettre à Nicolas Arrobe} & 1249 & 92 \\
\hline
\end{tabular}
\caption{Corpus of (Early) Old French texts.}
\end{table}
How lexical merger can drive grammaticalization

![Graph showing probability of overt S-Pron in 4 conditions over time.]

**Figure 3** Rate of overt pronominal subjects in (Early) Old French (ca. 1000–1250): Effects of ‘Clause type’ and ‘Person’.

Most importantly, and in strong contrast with Latin, our data show that in Old French overt pronominal subjects are generally preferred in embedded clauses (cf. the thicker lines in Figure 3), and dispreferred in main clauses. This confirms a claim that goes back to at least Foulet (1930) and Franzén (1939). In addition, embedded clauses initially favour overt third person pronouns. However, after ca. 1200, Person distinctions no longer play a role in embedded clauses. In main clauses, overt third person pronouns are always less frequent than their first and second person counterparts, and this tendency becomes stronger over time. To give a more precise idea of the data summarized in Figure 3, in Table 4 we give the full details of the logistic model on the basis of which this plot was produced.

We would like to interpret the discrepancy between Old French main and embedded clauses to mean that the language developed a new mechanism to license null subjects in main clauses, which explains why the demise of pro-drop is delayed in this syntactic environment. We will offer some speculation...
about the nature of this new licensing mechanism in Section 5. In contrast, we take it that embedded null subjects are the (disappearing) remnants of the type of null subject that was available in Latin, but crucially, we still have to explain why in Old French embedded clauses, overt third person personal pronouns are initially more frequent than first and second person pronouns, despite the fact that the former were not inherited as such from the Latin mother language.

### 2.4 Taking stock: the explananda

In order to answer the main research question of this paper (how did Romance third person personal pronouns come into being?), we need to take into account a number of considerations. First, as documented in great detail in Cappellaro (2016), in many modern Romance varieties the paradigm of strong (or tonic, to use the author’s terminology) third person personal pronouns contains descendants of both ille and ipse. We would like to follow Cappellaro (2016: 725) (see also Cappellaro 2018: 125–134) in interpreting this state of affairs to mean that in Late Latin, ille and ipse competed to express third person reference. In addition, as mentioned in the introduction the emergence of third person personal pronouns cannot easily be treated as a run-of-the-mill case of grammaticalization, in particular not when strong pronouns are considered: not only is the change one from one functional (rather than lexical) category to another, it is also not clear in which sense the newly formed strong personal pronouns are semantically, structurally and/or phonologically reduced with respect to their demonstrative/reflexive ances-
How lexical merger can drive grammaticalization

tors. In contrast, the genesis of weak (atonic) pronouns in Old French constitutes a more likely case of ‘classical’ grammaticalization, as the newly created category clearly is reduced (with respect to demonstrative as well as strong personal pronouns), both structurally and phonologically (see Cardinaletti & Starke 1999a,b). A final question that remains to be answered is whether Old French weak pronouns were derived directly from Latin *ille*, or rather from (Old French) strong personal pronouns: we will come back to this issue in Section 4.

3 Discourse properties of Late Latin *ille* and *ipse*

Let us now have a closer look at the interpretive properties of *ille* and *ipse* in Late Latin. As has often been pointed out, despite there being a clear sense in which the two items ‘compete’ for the same function in the relevant period, *ille* and *ipse* are certainly not fully equivalent (see among others Selig 1992; Vincent 1997). Some new corpus findings suggest that this is indeed correct: concretely, *ille* gradually develops a particularly strong preference for appearing in main clauses, whereas in Late Latin the most frequent third person pronoun in embedded clauses is *ipse*. The relevant quantitative data we observed in the two subcorpora introduced in Sections 2.1 and 2.2 are summarized in Table 5. Among other things, the data in this table confirm that, in Late Latin, *ille* is by far the most frequently used third person pronominal reference device. We also indicate whether or not the distribution of a given pronoun is significantly skewed towards main or embedded clauses; the expected frequencies used for the $\chi^2$-tests were calculated on the basis of the total number of (finite) main and embedded clauses with an (overt or otherwise) pronominal subject in each period. The significant results appear in the shaded cells.

<table>
<thead>
<tr>
<th></th>
<th>Early and Classical Latin</th>
<th>Late Latin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N main</td>
<td>N embedded</td>
</tr>
<tr>
<td><em>Hic</em></td>
<td>107</td>
<td>49</td>
</tr>
<tr>
<td><em>ille</em></td>
<td>70</td>
<td>39</td>
</tr>
<tr>
<td><em>ipse</em></td>
<td>27</td>
<td>39</td>
</tr>
<tr>
<td><em>is</em></td>
<td>50</td>
<td>37</td>
</tr>
<tr>
<td><em>iste</em></td>
<td>44</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 5  Distribution of Latin third person subject pronouns.
Late Latin *ille*’s preference for main clauses suggests that in the relevant period, the pronoun typically functions as a syntactic topic, a type of constituent whose distribution is known to be restricted in embedded clauses (Emonds 1970; Hooper & Thompson 1973; Ross 1973; Haegeman 2012). This observation can be taken to mean that *ille* took over some of the functions of its erstwhile competitors *hic* and *is*, which in Danckaert (2012: 207–257) were shown to readily function as left-peripheral topics in Classical Latin main clauses, but not in embedded ones. Another factor conspiring against the usage of *ille* in embedded clauses may have been the fact that the pronoun was subject to Principle C (cf. Section 1.2).

In our Late Latin corpus, bare *ille* in subject function typically indicates a topic shift, very often indicating a change of speaker (compare Frascarrelli & Hinterhölzl (2007)’s notion of ‘Shifting Topic’). Consider, for example, the little stretch of narrative discourse in (8), which features three discourse referents, namely Wiomadus, Egidius, and a group of Franks whose identity is not further specified. In this example, none of the three occurrences of subject *ille* shifts the topic away from one that was newly introduced by means of a full NP, but rather, an *ille*-topic either succeeds another *ille*-topic, or an instance of zero anaphora:

(8)  ∅_i_ hortabatur Egidio_j_ [aliqus_ Francos]_k_ encourage.ipfv.3sg Egidius.acc some.acc Franks.acc
dolose oppremere. ille_j_ audiens
cunning.adv oppress.inf dem.nom.m.sg hearing.nom.m.sg
consilium eius_i_, acruius coepit
advice.acc dem.gen.m.sg sharp.comp.adv begin.prf.3sg
oppremere eos_k_. illi_i_ uero in timore
oppress.inf dem.acc.m.pl dem.nom.m.pl prt in feat.abl
ac seditione uersi, iterum consilium a
and uproar.abl turned.nom.m.pl again advice.acc from
Wiomado_j_ expetierunt, qualiter ∅_k_ facere
Wiomadus.abl ask.prf.3pl how do.inf
deberten. at ille_j_ dixit
have.to.ipfv.sbjv.3pl but dem.nom.m.sg say.prf.3sg
eis_k_ [...]
dem.dat.m.pl

12 The data in Table 5 also suggest that that *hic* was originally especially fit to mark discourse topics (i.e. a well-known main clause phenomenon), given the early preference for bare nominative forms of *hic* to occur significantly more frequently in main clauses than in embedded domains. Early and Classical Latin *is*, on the other hand, seems to have been less narrowly specialized.
How lexical merger can drive grammaticalization

‘He (viz. Wiomadus) encouraged Egidius to cunningly oppress some of the Franks. He followed his advice, and started to oppress them relentlessly. Out of fear, the Franks then revolted and asked Wiomadus for advice as to what they should do. He answered to them: [...]’

(Liber Historiae Francorum 7, 727 CE)

Alternatively, *ille* is a more neutral anaphoric marker indicating topic continuity (as in (9)), a usage which may correspond to the notion of ‘Familiarity Topic’ from Frascarelli & Hinterhölzl (2007).13

(9)  
\[
\begin{align*}
\text{[DP Stilla, [CP quem comitem superius]i, nominavit]i, apparuit, ita ut in circuitu call.PRF.1SG appear.PRF.3SG so that in surrounding.ABL eiusi magna nigrido esset; et DEM.GEN.F.SG great.NOM blackness.NOM be.IPfv.SBJV.3SG and illa, tamquam se in foramen aliquod DEM.NOM.F.SG as.if REFL.ACC in opening.ACC some.ACC posita, inter tenebras relucebat. placed.NOM between darkness.ACC shine.IPfv.3SG
\end{align*}
\]

‘A star, which I earlier called a comet, appeared, in such a way that everything around it was cloaked in thick blackness. And this star shone a bright light through the darkness, as if it were placed in some opening in the sky.’

(Greg. Tur., Hist. 6.14, ca. 595 CE)

On the other hand, subject *ipse* is either used with its original Classical Latin force (roughly meaning ‘he […] himself, she […] herself etc.’), or it indicates that some familiar discourse participant is explicitly being contrasted with the current discourse topic (as in (10), where the topic changes from the churches destroyed by Clovis to Clovis himself), yielding a configuration which is more strongly contrastive than that of the topic shifting usage of *ille* illustrated in (8).

13 Another typical function of Late Latin *ille* often discussed in the literature is the one coined definite Erstennung (‘definite first mention’) in Selig (1992), which often involves (an NP modified by) *ille* acting as the head of a relative clause. However, at least as far as bare (i.e. non-adnominal) *ille* in subject function is concerned, this usage is not particularly frequent in the Late Latin corpus we have investigated for this study.
At that time many churches were looted by Clovis’ army. He was a fanatic, and a pagan.’

(Liber Historiae Francorum 7, 727 CE)

We conclude that although there may have been some functional overlap between \textit{ille} and \textit{ipse}, the two pronouns were clearly not freely interchangeable. With this established, we return to the later developments of Romance pronominal paradigms.

4 \textbf{Lexical mergers in pronominal paradigms}

As documented in great detail in Labov’s work on sound changes in American English (see \cite{Labov1994} for a comprehensive synthesis), many instances of phonological change involve the merger of two previously distinct phonemes into one ‘atomic’ category. Apparently less frequently, phonological change goes in the other direction, yielding a phonological split. In the realm of lexical change, lexical splits are of course very well documented. However, much less has been said about the merger of two distinct lexical items, be it content (open class) or function (closed class) words. The guiding idea behind our analysis is that lexical merger is exactly what happened to \textit{ille} and \textit{ipse}.\textsuperscript{14}

4.1 \textbf{A typology of phonological mergers}

\cite{Labov1994:321–323} discusses three “radically different” mechanisms of merger in phonology (for seminal discussion, see also \cite{Trudgill1978}). Focusing, in particular, on vocalic mergers, Labov first distinguishes ‘merger by expansion’, which he describes as a process through which “[t]he

\textsuperscript{14} In addition to that, we certainly also have to assume that one or more lexical splits took place: whereas both \textit{ille} and \textit{ipse} could act either as adnominal determiners or as free-standing pronouns, Romance definite articles and third person personal pronouns are clearly distinct. It is in all likelihood necessary to assume another split inside the category of personal pronouns, which would account for the fact that Romance object clitics invariably derive from \textit{ille} not \textit{ipse} (\cite{Vincent1997}).
phonetic range of the new phoneme is roughly equivalent to the union of the range of the two phonemes that merged”. As an example, Labov cites the development of tense /oh/ and lax /o/ phonemes in Eastern Pennsylvania, as studied in Herold (1990): whereas older speakers use the two variants in clearly distinct phonetic contexts, younger speakers randomly interchange them. A second type is ‘merger by approximation’, defined as “the gradual approximation of the phonetic targets of two phonemes until they are nondistinct”, whereby the resulting phoneme “may show a mean value intermediate between those of the original two.” In the latter case, the output of this process is distinct from both input phonemes.¹⁵ According to the description of Lennig (1978), this is what was happening to the vowels /a/ and /ɑ/ in the variety of French spoken in Paris in the mid-1970s. A third and final type of change is called ‘merger by transfer’, which is “a unidirectional process in which words are transferred gradually from one phonemic category to another. As a rule, it is not consistent with a result that shows an intermediate phonetic form.” Labov refers to Milroy (1980) for discussion of this type of merger in the history of Belfast English (on the differences between merger by approximation and merger by transfer, see also Trudgill & Foxcroft 1978).

For reasons of space, we cannot here further elaborate on the various case studies which have led to the establishment of this typology: for full discussion, we refer to Labov’s original work, and also to Babel, McAuliffe & Haber (2013), who provide illustrations which nicely visualize the relevant processes (see in particular Figure 1 on p. 2 of their article).

4.2 Lexical mergers of pronouns

In a nutshell, we propose that once language learners started to analyse ille andipse as competing for the same function, namely that of a relatively neutral third person reference device, the two items were merged together. In order to resolve the interpretive clash between the discourse properties not shared by both elements, merger happened in such a way that the resulting lexeme (let us call it lui) retained features of both ille andipse. More precisely, we can say that lui ends up being endowed with – roughly speaking – the union

¹⁵ According to Labov (1994: 321), merger by approximation may also result in a phoneme whose realizations do, in fact, have the same mean value as one of the input phonemes: an example of this would come from Greek, where no less than eight originally distinct phonemes were eventually merged into one front vowel /i/ (cf. Labov 1994: 229). As pointed out in Johnson (2007: 4), this subtype of merger by approximation may formally be very similar to merger by transfer: the differences between the two processes reside in, among other things, the role of sociolinguistic factors (which are stronger in the case of merger by transfer, and above the level of conscious awareness) and the rate of change (merger by transfer being slower, cf. Labov 1994: 323).
of the feature sets of *ille* and *ipse*: the relevant process is therefore an instance of merger by expansion. This development affected the whole of Romance, witness the fact that all Romance varieties have third person personal pronouns distinct from demonstratives (Cappellaro 2016). In addition, in Old French (and other varieties which developed weak pronouns), a second lexical item (*il*) was created which ended up retaining only features that were originally shared by *ille* and *ipse* (say roughly the intersection of the feature sets of those two items). This last process can be characterized as an instance of merger by approximation. Finally, in all cases, the phonological shape of a newly created pronoun was determined through merger by transfer; witness the fact that despite the considerable amount of variation in Romance pronominal paradigms, no mixed forms seem to exist containing one or more phonological segments inherited from *ille*, and one or more that derive from *ipse*. Let us then discuss the genesis of strong *lui* and weak *il* in more detail, starting with the former.

For the sake of concreteness, and abstracting away from features which are not essential to the analysis (such as ‘Number’), let us assume the following sets of properties to characterize Late Latin *ille* and *ipse*:

\[(11) \begin{array}{ll}
(a) & [+ 3^{rd} \text{ Person}, + \text{ Topic continuity}, + \text{ Deixis}, \text{ DP}] \quad \text{**ille**} \\
(b) & [+ 3^{rd} \text{ Person}, + \text{ Contrast}, – \text{ Deixis}, \varphi P] \quad \text{**ipse**}
\end{array} \]

Let us briefly comment upon these feature sets. First, we take it that the ‘+ 3^{rd} \text{ Person}’ property is obvious; suffice it to add that we could also assume that ‘3^{rd} \text{ Person}’ is in fact the absence of Person (following a tradition going back to Benveniste 1966: 251–257). In a similar vein, we also don’t want to exclude that the notion of ‘Person’ can (or has to) be decomposed in a number of independent features, such as [± Author] and [± Participant] (Harbour 2016). Second, for the characterizations ‘+ Topic continuity’ (*ille*) and ‘+ Contrast’ (*ipse*), we refer to the discussion in Section 3. Next, the ‘+ Deixis’ property expresses the fact that *ille* can (but need not be) used deictically: we take it that the structural correlate of this property is a functional layer which in the case of *ille* is always projected, and can optionally be ‘activated’, e.g. by means of an interpretable Deixis-feature.16 For *ipse*, the ‘– Deixis’ setting captures the exclusively anaphoric nature of the element, i.e. the fact that it is referential, but not compatible with a deictic use. Finally, we would like to relate the presence or absence of a Deixis-projection to a difference *qua* categorial status of the two pronouns under discussion: concretely, adopting

---

16 For a recent analysis of the structural representation of (spatial) deixis, see Lander & Haegeman (2018).
the terminology from Déchaine & Wiltschko (2002), we characterize ille as a DP, and ipse as a ϕP.17 Analysing ille as a DP correctly captures the fact that this item (in its bare use) has the same syntactic distribution as DPs projected by a lexical noun: it can only be used in argument and adjunct positions, never as a predicate. In this respect it differs from ipse, which we argued above (cf. (7)) can act as a (secondary) predicate. We therefore consider ipse to be a ϕP, i.e. a nominal category with ϕ-features but no outer D-shell.

When ille and ipse undergo merger by expansion, a new element lui is created with the feature set [+3rd Person, + Emphatic, + Deixis, DP], which comes about in the following way. First, the ‘+3rd Person’ property is simply inherited from both merging elements. Second, lexical merger of an element endowed with a feature ‘+ Topic continuity’ and one with a ‘+ Contrast’ property clearly cannot give rise to an element which (always) expresses both of these notions simultaneously, as this would yield an interpretive clash. Instead, we take it that what arises is an item that is merely emphatic, not narrowly specialized to express topicality, focality, contrast etc. This gives the correct result, as lui can be used to express all of those nuances, whenever pro (or il) is for some reason pragmatically infelicitous. Third, recall that we take the setting ‘+ Deixis’ to be correlated with the presence of a functional projection. It suffices that this category is present in one of the two merging items (in this case ille) in order for it to be inherited by lui. This correctly accounts for the fact that strong personal pronouns can be used anaphorically as well as deictically (cf. Section 1.1). Finally, by the same logic as that applied to the presence of a Deixis-projection, merger-by-expansion of a full DP (which itself contains a ϕP) and a structurally smaller bare ϕP is expected to yield a DP. This accounts for the fact that Romance (strong) personal pronouns can no longer occur in the predicative construal exemplified in (7) (which we assume to be different from cases of subject doubling by means of a strong pronoun in e.g. Italian; see Belletti 2005 for a possible analysis of the latter phenomenon). Finally, the binding properties of lui can be derived as follows. Recall from Section 1.2 that Romance strong personal pronouns are subject to Principle B and not Principle C of the Binding Theory. Cru-

---

17 Déchaine & Wiltschko (2002) propose a three-way topology of pronouns which is highly reminiscent of that of Cardinaletti & Starke (1999b) introduced in Section 1.2. Briefly put, they distinguish DPs, ϕPs and NPs, whereby ϕPs contain an NP-layer, and DPs a ϕP (and an NP). For reasons of space, we cannot attempt a detailed comparison between the two proposals: suffice it to say that despite some similarities, the two theories differ in a number of crucial respects: for one thing, Déchaine & Wiltschko (2002: 409, fn. 1) point out that their categorization cuts across distinctions between phrasal pronouns, clitics and agreement morphemes. See also Déchaine & Wiltschko (2002: 439) for a brief comparison between the two proposals, and for a possibility to analyse Cardinaletti & Starke’s strong, weak and clitic pronouns as three distinct types of their own ϕP.
cially, this property was already present in (Late) Latin *ipse* (cf. (7)): we propose that it was retained by the newly formed pronouns. Specifically, we conceive of Principles C and B as standing in a subset-superset relation, rather than as being opposites. Concretely, the distribution of elements subject to Principle C (‘X should never be c-commanded by a coreferential category’) is more restricted than that of elements obeying Principle B (‘X should not be locally c-commanded by a coreferential category’). Put differently, the range of syntactic domains that an R-expression (a lexical DP, or a demonstrative) can appear in is a proper subset of that in which a personal pronoun can appear. In the case at hand, the newly created strong pronoun *lui* can appear in the set of binding configurations that is the union of the sets of environments that *ille* and *ipse* can appear in, which happens to coincide with the range of environments that *ipse* can appear in.

Turning then to Old French *il*, we can say that this element emerged as a *bona fide*, unmarked personal pronoun, which is devoid of *ille*’s topic feature and of the contrastive semantics typical of *ipse*. Instead, it only expresses the notion ‘third person’, which is arguably the one semantic feature shared by both *ille* and (anaphoric) *ipse*. This development would be an instance of merger by approximation, with the proviso that there was already some functional overlap between the two merging elements. The binding properties of weak pronouns (which just as their strong counterparts obey Principle B) can be derived from a proposal made in Cardinaletti & Starke (1999b: 192), to the effect that weak pronouns per definition lack the structural layer where indexical features are hosted: this effectively disqualifies weak pronouns as R-expressions.

Importantly, we assume that Old French strong *lui* and weak *il* were created around the same time, and by this token that weak pronouns are not derived from strong personal pronouns. The main advantage of this line of analysis is that it can account for one of the more puzzling corpus findings reported in Section 2.3, namely that despite being an innovation, overt third person subject pronouns are initially more frequent than overt first and second person pronouns (in embedded clauses, that is). Concretely, we interpret this last observation to mean that the creation of a paradigm of weak personal pronouns was set in motion in third person contexts. A plausible alter-
How lexical merger can drive grammaticalization

native would be to say that the creation of a full set of weak subject pronouns is a secondary development, which took place after the category strong lui had been innovated. However, this analysis has the drawback that it leaves the ‘Person’ effect present in the Old French corpus unaccounted for. Note also, that it is certainly not the case that weak pronouns are simply phonologically weaker versions of strong pronouns; witness the fact that French weak il is derived from the Latin nominative form ille, and strong lui (indirectly) from the dative illi. If, on the other hand, we assume that the genesis of strong and weak pronouns derived from ille corresponds to two independent but simultaneous processes, the observed time-lag between the spread of overt weak pronouns in third person and non-third person environments is, in fact, predicted. Weak first and second person pronouns come into being later, perhaps through analogy with weak il.

4.3 Lexical merger and grammaticalization

Given the analysis detailed in the previous section, we can say that by resolving morphosyntactic competition, lexical merger effectively acts as a ‘catalyst’ for grammaticalization. When the process involved is merger by expansion, the outcome is a novel functional vocabulary item which is not reduced with respect to its historical sources (in fact it may actually be bigger). On the other hand, merger by approximation gives rise to a more prototypical case of grammaticalization, with the loss of functional features and/or phonological segments. In addition to a number of generally accepted structural processes underlying, or even driving, grammaticalization, such as upward reanalysis (Roberts & Roussou 2003) and economy-based constraints like the Head Preference Principle and Late Merge Principle from van Gelderen (2011: 13–14), lexical merger may thus be an additional mechanism which can give rise to the creation of new functional categories. Importantly, in contrast with the other processes mentioned, the primary locus of change would be the lexicon, rather than phrase structure syntax.

5 Some notes on pro-drop in Old French

Before we conclude, we would like to briefly touch upon two issues related to the loss of pro-drop in Old French (a development which as we pointed out earlier is not to be conflated with the genesis of weak personal pronouns in the language). For recent discussion of this much-debated topic, and in particular the relation between the erosion of agreement morphology and the demise of zero anaphora, we refer to Zimmermann (2014) and especially Simonenko, Crabbé & Prévost (2019), and to the references cited there.
First, there does not seem to be any reason to conclude that Old French was a partial null subject language (in the sense of Holmberg 2005, 2010; Holmberg, Nayudu & Sheehan 2009; see also Barbosa 2019; Roberts 2019: 207–216): the hallmark of such a system, viz. the possibility for a (main clause) third person singular null subject to receive a generic interpretation (corresponding to that of English one and French on), does not seem to be present in Old French (Buridant 2000: 409, §326). In addition, as pointed out in Danckaert & Prévost (submitted), there is no evidence that the presence of a c-commanding coreferential category in the immediately superordinate clause favours the occurrence of embedded null subjects, which again suggests that despite the observed Person effects, Old French was not a partial null subject language (contra Ingham 2018).

Secondly, we would like to offer some speculation about null subjects in Old French main clauses, whose productivity we hypothesized earlier is to be ascribed to an innovation which must have taken place after the Late Latin period. According to the influential Government and Binding analysis of Adams (1987), main clause null subjects in Old French are licensed by V-to-C movement: in particular, the finite verb was argued to license pro under Government. We will here briefly explore an alternative to this idea, which does not refer to the now abandoned notion of Government. With Poletto (2000) and Poletto & Tortora (2016), we can assume that in the high functional field four projections are available which are related to subject agreement (but which are independent of subject agreement morphology on finite verbs). These projections can optionally be lexicalized overtly: this is famously the case in many Northern Italian dialects. The two lowest of the relevant functional heads, which are located just below the lower edge of the CP-domain, encode second person singular and third person (singular and plural) (for an overview, see Poletto 2000: 38). Interestingly, there is evidence that in Old French main clauses, it is subjects of exactly these kinds that are most likely to be left unexpressed. The results of a logistic regression modeling the influence of the factors Person and Number on the expression of pronominal subjects are summarized in Table 6, where the level ‘first person singular’ is set as a baseline. Note that we lumped together singular and plural third persons. In addition, given that for certain Person-Number combinations (in particular first and second person plural) our token counts are unfortunately rather low, estimates are averaged over all texts in our corpus: in other words, we did not take into account the diachronic dimension, nor any random effects associated with individual texts (as a result, the figures in Table 6 may have to be interpreted with some caution).
How lexical merger can drive grammaticalization

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.31415</td>
<td>0.08618</td>
<td>0.000267 ***</td>
</tr>
<tr>
<td>2SG</td>
<td>-0.57005</td>
<td>0.28600</td>
<td>0.046243 *</td>
</tr>
<tr>
<td>3SG+PL</td>
<td>-1.18675</td>
<td>0.09961</td>
<td>&lt;2e-16 ***</td>
</tr>
<tr>
<td>1PL</td>
<td>0.06526</td>
<td>0.21474</td>
<td>0.761218</td>
</tr>
<tr>
<td>2PL</td>
<td>-0.15875</td>
<td>0.16218</td>
<td>0.327629</td>
</tr>
</tbody>
</table>

Table 6  Person and pro-drop in Old French declarative main clauses.

Let us assume that verb movement in Old French main clauses targets a position in the low left periphery, say Fin, which sits above the two lowest subject agreement projections from Poletto (2000), but below the two highest ones (on Fin as a target for verb movement, see Haegeman 1996; Poletto 2002; Wolfe 2019). We can then assume that on its way to Fin, the verb passes through, and thereby ‘activates’ the lower regions of the subject agreement field: this would correctly account for the prevalence of third person and second person singular null subjects in Old French main clauses. Null subjects with different person-number specifications could either be generated by the inherited agreement-based pro-drop grammar (which there is at this point no reason to believe declined faster (or slower) in main than in embedded clauses), or perhaps by a grammar involving V-movement to Force, i.e. to a position much higher than the entire subject agreement field. As pointed out in Poletto (2018), an analysis that links V-to-C movement to the licensing of null subjects predicts that rates of pro-drop should be similar in declarative and interrogative main clauses (which does not in fact seem to be the case in Old Italian; for relevant discussion, see also Cognola & Walkden 2019). In future research it will be interesting to test whether effects similar to those summarized in Table 6 can also be observed in interrogative main clauses in Old French.

6  Conclusion

In this paper we have considered how Romance third person personal pronouns came into being, a question which in the literature has received relatively little attention. On the assumption that this development does not bear the signature of a prototypical case of grammaticalization, we suggested that the two most prominent Late Latin lexical items expressing third person pronominal reference, to wit ille and ipse, were merged together to yield true (tonic) personal pronouns, endowed with the union of the feature sets of both
input items. We extended this analysis to account for the rise of weak subject pronouns in Gallo-Romance: we suggested that this development also involved lexical merger of the same two items, but differently from the previous process, only the intersection of the feature sets of \textsc{ille} and \textsc{ipse} was preserved. Our main theoretical claim is that in addition to other processes, such as upward reanalysis and phrase-to-head reanalysis, lexical mergers may also play a role in driving the creation of new functional material. In future research, it will be interesting to explore whether mergers of lexical items are perhaps more widespread than previously assumed.

References


Cappellaro, Chiara. 2018. Genesis and diachronic persistence of overabundance: Data from Romance languages. In Antje Dammel, Matthias Eitelmann & Mirjam Schmuck (eds.), \textit{Reorganising grammatical variation: Di-
achronic studies in the retention, redistribution and refunctionalisation of linguistic variants, 119–148. Amsterdam: Benjamins.


Cognola, Federica & George Walkden. 2019. *pro*-drop in interrogatives and declaratives: A parallel study of Old High German and Old Italian. Linguistik Online 100. 95–140.


How lexical merger can drive grammaticalization


How lexical merger can drive grammaticalization


Trager, George. 1932. The use of the Latin demonstratives (especially ILLE and IPSE) up to 600 A.D., as the source of the Romance article. New York: Institute of French Studies.


Lieven Danckaert
STL - Savoirs, Textes, Langage, UMR 8163
Rue du Barreau BP 60149
59653 Villeneuve-d'Ascq Cedex
France
lieven.danckaert@univ-lille.fr
https://pro.univ-lille.fr/lieven-danckaert

Liliane Haegeman
Vakgroep Taalkunde (UGent)
Blandijnberg 2
9000 Gent
Belgium
Liliane.Haegeman@UGent.be
https://research.flw.ugent.be/
nl/liliane.haegeman

Sophie Prévost
LATTICE - Langues, Textes, Traitements informatiques et Cognition (UMR 8094)
1 rue Maurice Arnoux
92120 Montrouge
France
sophie.prevost@ens.fr
https://www.lattice.cnrs.fr/amp/membres/direction/sophie-prevost/