Taraldsen's Generalization in Diachrony: Evidence from a Diachronic Corpus

Alexandra Simonenko, Benoît Crabbé, Sophie Prévost

Taraldsen's Generalization in Diachrony: Evidence from a Diachronic Corpus

Alexandra Simonenko, Benoît Crabbé, and Sophie Prévost

1. Introduction

This paper examines the possibility of a dependency between the availability of null subjects and the "richness" of verbal subject agreement, known as Taraldsen's Generalisation (Taraldsen 1980, Rizzi 1986, Adams 1987), from the point of view of grammar change in Medieval French based on corpus data. The original generalisation states that a language having rich (i.e. non-syncretic) subject agreement implies the possibility of non-expression of subjects. Medieval French texts abound in examples of null subjects that would be ungrammatical in Modern French, such as the following one.

(1)   de lui ∅   firent segnor et mestre. Puis ∅   ont gardé devers senestre
        of him (they) made sir   and lord.   Then (they) have looked towards left
        "They made him their sir and lord. Then they looked to the left" (Eneas, v. 76-79, XII c.)

It has been noted that with time such cases become more and more rare (Foulet 1928), an intuition that has been confirmed quantitatively (Fontaine 1985, Hirschbühler 1992, Kaiser 2009, Prévost to appear a.o.). It is likewise commonly assumed that French went from a language characterised by a "rich" or non-syncretic agreement inherited from Late Latin to a language with an impoverished, or syncretic, agreement paradigm (Foulet 1935, Dees et al. 1980, Marchello-Nizia 1992, Buridant 2000, Morin 2001, De Jong 2006, Bettens 2015 a.o.). Some considered the loss of non-syncretic agreement to be a trigger for the null subject loss in Medieval French (Ewert 1943, Vennemann 1975). Such a scenario, however, was questioned for Medieval French on empirical grounds by Schøsler (2002) and Roberts (2014) who reject a direct connection between the two changes because of an apparent temporal lag between them. Noteworthily, these authors make opposite assumptions about the order of changes.

We present the first large scale quantitative investigation of the syncretisation of verbal subject agreement in Medieval French and test a classic analysis which relates non-syncretic agreement and null subjects as parts of the same grammar (e.g. Rizzi 1986, Adams 1987, Alexiadou & Anagnostopoulou 1998, Roberts 2010, Sheehan to appear). We show that agreement syncretisation and the emergence of overt pronominal subjects proceeded at the same rate. On the Constant Rate Hypothesis of Kroch (1989), which states that a grammatical change has the same rate in different contexts, these results are compatible with the traditional analysis. However, we show that this analysis also generates a number of predictions which are not borne out by the quantitative data. We conclude that a more complex model of interaction of subject and inflection parameters is needed.

We will first introduce corpus-based empirical observations concerning subject agreement syncretisation in Medieval French. In section 3 we formalize diachronic predictions made by a family of approaches tying together the two phenomena as parts of the same grammar. In section 4 we present logistic regression models of the two changes in the light of the standard Constant Rate effect assumptions about grammatical changes. We then present some additional aspects of the data which falsify the hypothesis and conclude suggesting an acquisitional view of the connection.

2. Agreement syncretisation

One of the major factors evoked in relation to the disappearance of null subjects in the history of French is syncretisation of subject agreement inflection on the verb. To date, however, there has been no studies exploring the possible diachronic connection from a quantitative point of view. Unlike relatively well-documented appearance of overt pronominal subjects, existing observations concerning inflection syncretisation are based on much more sporadic data. Below we present the first large-scale quantitative investigation of the observable cases of syncretisation with the goal of exploring the extent to which Taraldsen's generalisation holds in diachrony and putting to test synchronic analyses of pro-drop.

Descriptively, the main changes in the verbal agreement that were reflected in the spelling are as in the table 1. For the verbs of the I group (-*er* infinitives) 1P and 3P syncretise in present tense indicative and subjunctive, taking an ending -*e*. For the II group, in which we grouped all other infinitival types with the exception of highly idiosyncratic *être* "be" and *avoir* "have", 1P and 2P syncretise in present and past tense indicative, taking the ending -*s*.

	I group pres. ind.	I group pres. subj.	II group pres. ind.	II group pst. ind.
1P	aim > aim**e** "love"	aim > aim**e**	voi > voi**s** "see"	vi > vi**s**
2P	aime**s**	ains > aime**s**	voi**s**	vi**s**
3P	amet > aim**e**	aint > aim**e**	voit	vit

Table 1: Subject agreement syncretisation in Medieval French

In Modern French 1P, 2P, and 3P singular, as well as 3P plural for many verbs, are phonologically indistinguishable, with the exception of a handful of verbs which also function as auxiliaries.

In order to establish a temporal profile of the surface changes, we calculated the ratio of the new endings (-*e* with the verbs of the I group in the context of 1P and 3P subjects and -*s* with verbs of the II group in the context of 1P subjects) to the sum of the new and "old" endings (zero and -*t* with the verbs of the I group in the context of 1P and 3P subjects respectively and zero with the verbs of the II group with 1P subjects) for each text in the corpus. In other words, our goal was to track the "spread" of innovative endings with 1P and 3P subjects.[1] Figure 1 shows the raise of the new endings. For instance, P(ENDING = NEW | DATE = D, GROUP = I, PERS = 1P) stands for the probability, estimated by relative frequency, of a verb from the I group to have a new ending, -*e*, in the context of a 1P subject.

It must be mentioned here that the extent to which the spelling was synchronically faithful to the spoken language remains largely a matter of speculation, although recent works involving phonological reconstructions of the Medieval French seem to agree that during the early Middle Ages spelling variations stemmed from variations in pronunciation. De Jong (2006:174), based on an analysis of grapheme variation in the documents written in the Paris region during XIII–XIV cc., concludes that prior to the mid-XIV c., graphemes likely reflected pronunciation very closely. Similarly, Bettens (2015) in his discussion of the French rime tradition concludes that non-pronunciation of certain graphemes starts manifesting itself at the beginning of the XIII c. According to him, the strict rime system of the first (attested) versified texts indicates that the final consonants, with very few exceptions, were pronounced in the spoken language. In the rimed texts examined by De Jong (2006:176), first signs of the fall of the final consonants appear in the XIV c. Very similar observations were made earlier by Foulet (1935). Concerning specifically final consonants making part of the verbal endings several grammarians of the XVI c. mention in their work that those were still pronounced in their dialects (Bonin 1992:56). For the purposes of our study we will adopt the hypothesis that at least until the XIV c. the spelling reflects to a large extent a living grammatical system, and that the graphic spread of the "new" endings mirrors the syncretisation of the verbal agreement paradigm in the spoken language.

The MCVF (2010) corpus together with Penn Supplement to MCVF (2010) includes 34 syntactically parsed texts of approximately 1 million words. On the assumption that null subjects correspond to phonologically null pronominal elements, observations about the disappearance of null subjects are

---

[1] We had to limit ourselves to clauses with overt pronominal subjects in order to be able to identify subject's person.
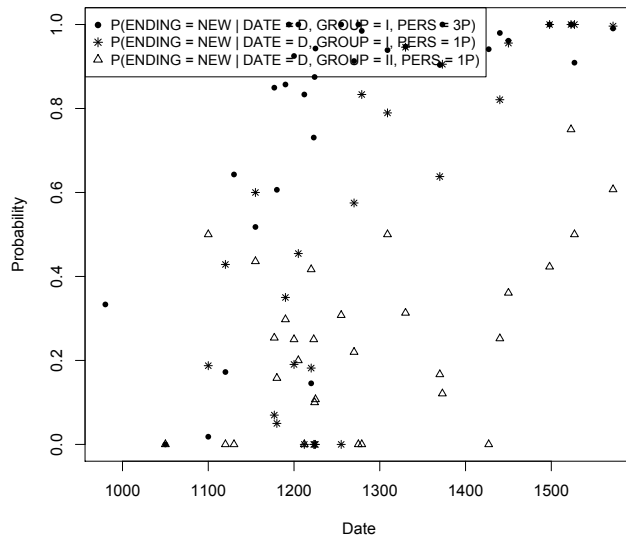
Figure 1: Emergence of new verbal endings

given here as the estimated probability of overt pronominal subjects against null subjects, with nominal and other kinds of overt subjects being excluded from consideration. Specifically, our dataset includes all finite clauses with either an overt pronominal or null subject (total of 104,485), excluding imperatives, subject relatives and wh-questions targeting subjects because of their idiosyncratic subject syntax. We also excluded the cases of subject ellipsis under coordination, still allowed in French.

3. Null subjects Two changes as the loss of Agr

In search for a relation between syncretisation and null subject loss, we will quantitatively explore the standard hypothesis that null subjects and non-syncretic agreement are related via a certain structural property giving rise to both, often associated with a person feature-specified functional head.[2] On this approach the two surface changes can thus be viewed as consequences of one underlying structural change: the disappearance of the relevant functional head, let us call it Agr.

The approach whereby a null subject is a spellout of *pro*, which, to be merged, needs to be identified by a verbal head specified with ϕ-features, is schematically illustrated in figure 2. Here we assume that a head being specified for person implies spellout rules which map every person feature to a unique exponent, to use the Distributed Morphology vocabulary.[3]

Concerning the nature of the licensing relation between Agr and *pro*, we assume that person features introduce conditions on the meaning of a pronoun, specifically, they provide constraints on the domain of the variable it denotes (Heim & Kratzer 1998, Kratzer 2009, Heim 2008, Sauerland 2008 among others). In addition to that, we will assume that a pronoun needs to be supplemented with an element triggering a presupposition about its meaning, whether it comes as part of a pronominal form or as a verbal ending.

---

[2] Depending on a particular implementation of this approach, this head either enters into a feature identification relations with a *pro* or, if no *pro* is postulated, itself serves to fill the verb's subject argument slot, which allows for null subjects (see Sheehan (to appear) for an overview). We will assume that there is a *pro*, although this choice does not have any substantial consequences for our conclusions.

[3] It could be argued that some pro-drop theories deal with abstract morphosyntactic features as correlates of rich agreement and may say nothing about the actual morphological exponents. In this respect it can be recalled that the original Taraldsen's generalization was made based on surface observations, that is, it is a correlation between two sets of surface phenomena. Also, trivially, a theory making a straightforward statement about surface observations is falsifiable to a much higher degree than one that does not, and is therefore to be preferred.
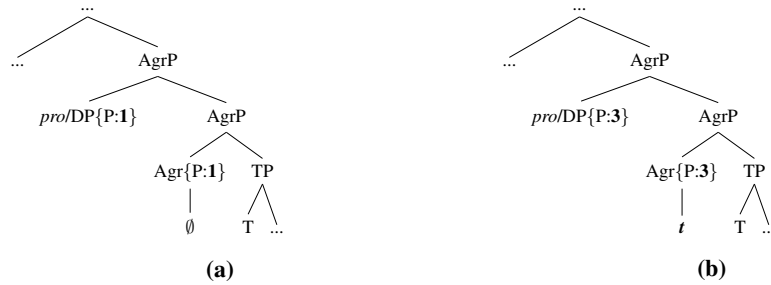
Figure 2: Agr-grammar

On this approach the loss of null subjects can be naturally modelled as a passage from the Agr-grammar to an alternative grammar without Agr, in which verbal endings correspond to the spellout of some other head, let us say T, not specified for person.[4] Since T is not specified for person, it cannot introduce presuppositions necessary for a felicitous use of a featureless pronoun, *pro*. An overt pronoun, which itself carries the information about its interpretation domain, has therefore to be merged.
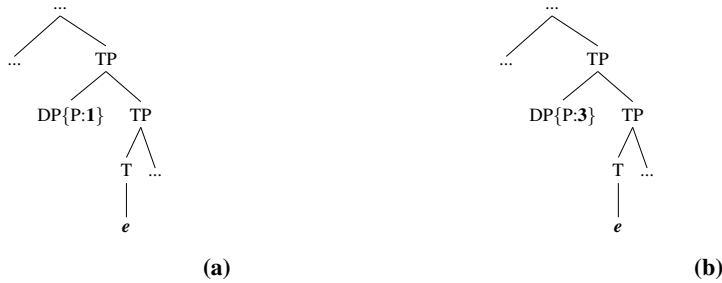

Figure 3: T-grammar

On the view just laid out, the rates of the emergence of overt pronominal subjects and of the new endings are expected to be the same, given the Constant Rather hypothesis of Kroch (1989) that a grammatical change has the same rate in all of its surface manifestations. One caveat of this prediction is that even stable null subject grammars allow for overt subjects, usually under some pragmatically restricted conditions (Bates 1976, Otheguy et al. 2007). This makes it impossible to classify a given overt pronominal subject as an instance of Agr- or T-grammar. The only context which sets the two apart clearly are expletive subjects, which are consistently null in null subject languages.[5] We therefore will compare the spread of overt *expletive* subjects with the spread of the new endings.

There are at least three other immediate predictions this approach generates. Most importantly, it predicts that syncretisation should have the same rate in different contexts: if syncretisation reflects the disappearance of Agr, we do not expect this change to proceed differently depending on the verb type or the person of the subject. Another prediction is that there should be no null subjects in the context of new syncretic endings. This is so because Agr-grammar which, by hypothesis, is the only grammar that licenses null subjects, is associated with spellout rules which produce old, non-syncretic endings. Finally, there is a third prediction that there should be no raise in pronominal subject expression with old, non-syncretic endings: although Agr-grammar, associated with non-syncretic endings, is compatible with overt subjects, their distribution is supposed to be governed by the pragmatic rules which result in the same rate of subject expression during the course of existence of this grammar.[6]

---

[4] We assume that 4P and 5P endings *-ez* and *-ons*, distinct from each other and the rest of the paradigm, are exponents of the feature [plural] in the context of 4P and 5P subjects. In other words, verbal endings from presupposition triggers become mere agreement markers.

[5] From this perspective, a purely pro-drop stage was never attested in Old French: from the earliest texts on we find overt expletives (cf. discussion in Zimmermann (2014)).

[6] For Italian and Spanish pronominal subject expression is reported to be at 30% (Bates 1976, Otheguy et al. 2007).

4. Results
4.1. Testing the main prediction: same rate of syncretisation and subject emergence

To test the prediction that the changes in question proceeded with the same rate we compared two logistic models fitting the data on the appearance of the new endings and overt expletives. First, the model ENDING is a model that predicts whether the verbal ending, Y, is *new* (or syncretic) by contrast with an *old* or non syncretic verbal ending as a function of time. This model has the following form:

(2) $\quad P(Y = new | Time = t) = \frac{e^{\alpha + \beta t}}{1 + e^{\alpha + \beta t}}$

where α is the intercept and β the slope. The intercept is interpreted as an abstract indicator of when the change takes place in time and the slope is interpreted as the rate of change. The more important the slope, the faster the change takes place.

We compare this model with a EXPL SBJ model which predicts whether the subject realisation, Y, is *new* (or overt) by contrast with an *old* realisation where the pronominal subject is dropped. This second model has the exact same form:

(3) $\quad P(Y = new | Time = t) = \frac{e^{\alpha + \beta t}}{1 + e^{\alpha + \beta t}}$

but this time Y represents subject realisation instead of verbal syncretism. In order to illustrate both models we first fitted them separately to the data. The logistic curves and the corresponding parameter estimates are provided in figure 4.
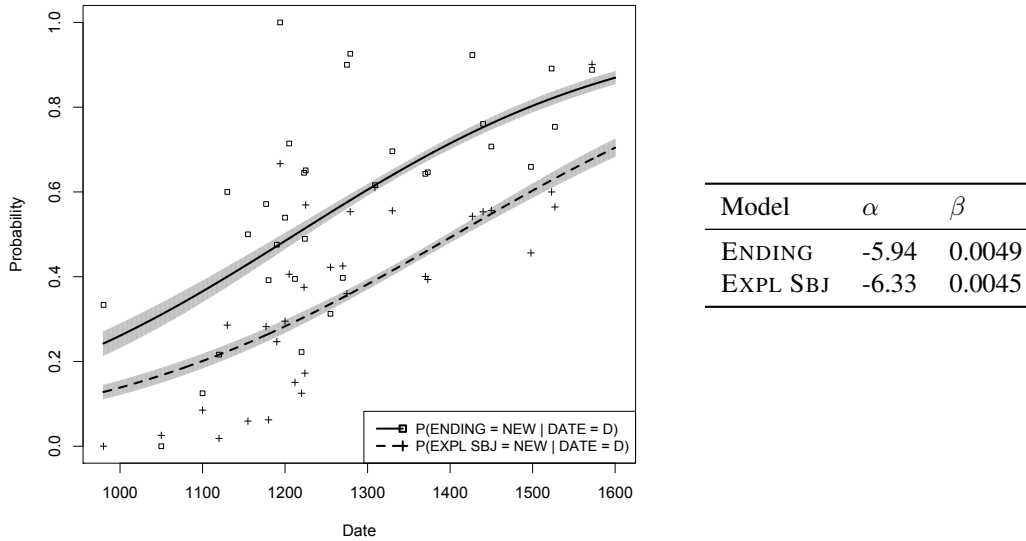

Model	α	β
ENDING	-5.94	0.0049
EXPL SBJ	-6.33	0.0045

Figure 4: Spread of new endings and overt expletive subjects

In our case, a smaller α for the ENDING model indicates that the change starts earlier than the change for the PRO-DROP model. The slope parameters are almost equal. Although this first observation suggests that we cannot reject the Constant Rate hypothesis with the data, we further perform a statistical test of the hypothesis. To do that we test for the contribution of the slope by comparison of two mixed effect models (Gelman & Hill 2006). The first model amounts to predicting the new form Y, whether it is an overt subject or a syncretic verbal ending, by contrast with an old form, that is a covert subject or a non syncretic verbal ending. The prediction is still a function of time but we also add a random intercept α_c for each context c: either an ending context or a subject context, yielding the following model:

(4) $\quad P(Y = new | Time = t, Context = c) = \frac{e^{\alpha + \alpha_c + \beta t}}{1 + e^{\alpha + \alpha_c + \beta t}}$

Informally. this model means that the global model intercept may be further parametrized for each specific context but the slope is constrained to be identical for both contexts. We compare this model to an extended version where this time we add a random slope β_c, thus allowing the slope to vary for each context:

(5) $\quad P(Y = new|Time = t, Context = c) = \frac{e^{\alpha+\alpha_c+(\beta+\beta_c)t}}{1+e^{\alpha+\alpha_c+(\beta+\beta_c)t}}$

Since the slope models the rate of change, this second model allows the rate of change to differ for each context. We test whether the slope introduces a significant difference between the two models (4) and (5) with a log likelihood ratio test which is χ^2 distributed ($df = 2$). The test fails to reject the null hypothesis since $p = 0.37$. As we cannot reject the null hypothesis, we must conclude that the introduction of the slope does not contribute to better predicting the data and we cannot conclude that the Constant Rate Hypothesis does not hold in this case.

On the Constant Rate Hypothesis, that the rate of overt pronominal subject emergence and ending syncretisation is the same is compatible with the analysis of the two diachronic phenomena as stemming from the same grammatical change, which we identified as the passage from a grammar with Agr head to a grammar without.

Based on the results obtained, we can also tentatively formulate the following correlate of Taraldsen's Generalization for the diachronic data, where the probability of null subjects equals ($1-$ P), P being the probability of overt pronominal subjects:

(6)   EXTENSION OF TARALDSEN'S GENERALIZATION ONTO DIACHRONIC DATA
       The lower the probability of syncretic endings the higher the probability of null subjects.

Below we will explore three other predictions made by the current approach.

4.2. Different rate of syncretisation in different contexts

The model for agreement syncretisation defined in (2) merges syncretisation developments in three different environments described in section 2, namely, the passage from unambiguous to ambiguous endings for the verbs of the I group in the context of 3P subjects (present indicative and subjunctive), and 1P subjects (present indicative and subjunctive), as well as for the verbs of the II group in the context of 1P subjects (present and past indicative). According to the approach that maintains that syncretisation is a consequence of the T-grammar winning over the old Agr-grammar, these developments are expected to have the same rate. In order to test this, we modelled them separately. As table 2 shows, the rate of the spread of (-*s*) is much lower than the rate of the spread of -*e*, contrary to what was predicted.

Verb group	Subject person	α	β
I	3P (-*e*)	-18.533	0.016
I	1P (-*e*)	-19.539	0.015
II	1P (-*s*)	-4.602	0.003

Table 2: Logistic regression parameters for ENDING in different environments

4.3. Increase of syncretic endings in the context of null subjects

Another prediction made by the approach we are exploring is that there should be no increase in the frequency of syncretic endings in the context of null subjects. We say increase because of a number of verbal stems ending in -*e* etymologically, so the mere presence of such endings with null subjects does not indicate that it is a T-grammar ending cooccuring with an Agr-grammar subject type. However, we found that -*e* is spreading in the context of 1P and 3P null subjects. For instance, a logistic regression fitting the data on endings in the context of 1P null subjects has parameters $\alpha=-16.56$ and $\beta=0.012$. Since the slope parameter is significantly different from zero (p = 4.27×10^{-12}), it is unlikely that the occurrences of -*e* with 1P null subjects are simply due to a limited number of verbs whose stems had always ended in -*e*.

4.4. Increase of overt subjects in the context of non-syncretic endings

Finally, we predicted no increase in overt pronominal subjects in the context of verbs with old or non-syncretic endings, because this is the context of the Agr-grammar that expresses subjects at a pragmatically determined stable rate. In order to test this, we compared the rate of subject expression in the context of verbs with the non-syncretic endings *-t* and zero on the one hand and syncretic endings *-e*, *-s* on the other. We fitted the four datasets to the following logistic regression model: $P(Pron\ Sbj = overt|Date = d) = \frac{e^{\alpha+\beta*d}}{1+e^{\alpha+\beta*d}}$. The parameters of the models for different datasets are given in table 3.

Verb group	Ending	α	β	Type
I group	zero	−3.446	0.0037	old
II group	zero	−4.792	0.0046	old
I group	*-t*	−12.545	0.0113	old
I group	*-e*	−3.478	0.0045	new
II group	*-s*	−6.557	0.0066	new

Table 3: Logistic regression parameters for PRO-DROP in the context of different endings

Subject expression rate grows over time for all types of endings, both syncretic and non-syncretic, a result not predicted by an approach which sees non-syncretic endings as part of the null subject licensing grammar which has a constant pragmatically determined rate of subject expression.[7]

5. Conclusions

We have shown that some of the predictions of the approach which postulates an utterance-level relation between subject expression and ending type are not borne out. However, given that the overall ending syncretisation has the same rate as the emergence of overt expletive subjects, it would be counterintuitive to conclude that the two processes are entirely independent. We need therefore a model which would dissociate subject expression and the choice of ending at the clause level, but would relate them tightly in the process of language evolution. Such a model may for instance be one where the type of an ending (non-syncretic vs. syncretic), presumably dependent on some unrelated phonological mechanism, presents a parsing difficulty for a null subject-licensing grammar and thus lowers its probability to be chosen by the speaker, which eventually drives it to extinction, similarly to the grammar competition model proposed in Yang (2010).

References

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

Alexiadou, Artemis & Elena Anagnostopoulou (1998). Parametrizing AGR: Word order, V-movement and EPP-checking. *Natural Language & Linguistic Theory* 16:3, 491–539.

Bates, Elizabeth (1976). *Language and context: The acquisition of pragmatics*, vol. 13. Academic Press, New York.

Bettens, Olivier (2015). Chantez-vous français ? remarques curieuses sur le français chanté de moyen Âge à la période baroque. URL `http://virga.org/cvf/`.

Bonin, Michèle (1992). Le système verbal français de Luis Meigret, XVIe siècle. MA thesis.

Buridant, Claude (2000). *Nouvelle grammaire de l'ancien français*. Sedes, Paris.

De Jong, Thera (2006). *La prononciation des consonnes dans le français de Paris aux 13ème et 14ème siècles*. Netherlands Graduate School of Linguistics.

Dees, Anthonij, S. Meilink, Karin van Reenen-Stein & Pieter van Reenen (1980). Un cas d'analogie: l'introduction de *-e* à la première personne du singulier de l'indicatif présent des verbes en -er en ancien français. *Rapport/Het Franse Boek* 50, 105–110.

Ewert, Alfred (1943). *The French Language*. Faber & Faber, London.

[7] With regard to an outstandingly low estimated coefficient in the context of *-t*, it needs to be mentioned that this ending becomes extremely rare past XIII, and the standard error of the estimated coefficient is more important in this case than for the other endings.

Fontaine, Carmen (1985). Application de méthodes quantitatives en diachronie: L'inversion du sujet en français. Université du Québec à Montéal.

Foulet, Lucien (1928). *Petite syntaxe de l'ancien français*. Champion, troisième édition revue. Réédition 1982, Paris.

Foulet, Lucien (1935). L'extension de la forme oblique du pronom personnel en ancien français. *Romania* 61–62, 257–315.

Gelman, Andrew & Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Heim, Irene (2008). Features on bound pronouns. Harbour, David, David Adger & Susana Bejar (eds.), *Phi Theory: Phi-features across Modules and Interfaces*, Oxford University Press, Oxford, 35–52.

Heim, Irene & Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell, Oxford.

Hirschbühler, Paul (1992). L'omission du sujet dans les subordonnées V1: les CNN de Vigneulles et les CNN anonymes. *Travaux de linguistique* 24, 25–46.

Kaiser, Georg A. (2009). Losing the null subject. A contrastive study of (Brazilian) Portuguese and (Medieval) French. *Proceedings of the Workshop "Null-subjects, expletives, and locatives in Romance"*, 131–156.

Kratzer, Angelika (2009). Making a pronoun: Fake indexicals as windows into the properties of pronouns. *Linguistic Inquiry* 40:2, 187–237.

Kroch, Anthony (1989). Reflexes of grammar in patterns of language change. *Language variation and change* 1:03, 199–244.

Marchello-Nizia, Christiane (1992). *Histoire de la langue française aux XIVe et XVe siècles*. Dunod, Paris.

MCVF (2010). Corpus MCVF annoté syntaxiquement, dirigé par France Martineau, avec Paul Hirschbühler, Anthony Kroch et Yves Charles Morin. URL `http://www.voies.uottawa.ca/corpus_pg_en.html`.

Morin, Yves Charles (2001). La troncation des radicaux verbaux en français depuis le moyen age. études diachroniques. *Recherches linguistiques de Vincennes* :30, 63–85.

Otheguy, Ricardo, Ana Celia Zentella & David Livert (2007). Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language* 770–802.

Penn Supplement to MCVF (2010). Penn Supplement to the MCVF Corpus by Anthony Kroch and Beatrice Santorini.

Prévost, Sophie (to appear). Emergence and development of personal pronoun subjects: study of a bilingual Latin–Old French corpus.

Rizzi, Luigi (1986). Null objects in Italian and the Theory of *pro*. *Linguistic Inquiry* 17, 501–557.

Roberts, Ian (2010). A deletion analysis of null subjects. *Parametric variation: Null subjects in minimalist theory*, Cambridge University Press, 58–87.

Roberts, Ian (2014). Taraldsen's Generalization and Language Change: Two Ways to Lose Null Subjects. Svenonius, Peter (ed.), *Functional Structure from Top to Toe: The Cartography of Syntactic Structures*, Oxford University Press, vol. 9.

Sauerland, Uli (2008). On the semantic markedness of phi-features. Harbour, David, David Adger & Susana Bejar (eds.), *Phi Theory: Phi-features across Modules and Interfaces*, Oxford University Press, Oxford, 527–82.

Schøsler, Lene (2002). La variation linguistique: le cas de l'expression du sujet. Sampson, Rodney & Wendy Ayres-Bennett (eds.), *Interpreting the History of French, A Festschrift for Peter Rickard on the occasion of his eightieth birthday*, Rodopi, Amsterdam - New York, NY, 187–208.

Sheehan, Michelle (to appear). Subjects, null-subjects and expletives in romance. URL `http://ling.auf.net/lingbuzz/002466`. To appear in Manuals of Romance Linguistics (MRL): Grammatical Interfaces.

Taraldsen, Tarald (1980). *On the NIC, vacuous application and the that-trace filter*. Indiana University Linguistics Club.

Vennemann, Theo (1975). An explanation of drift. Li, Charles N. (ed.), *Word Order and Word Order Change*, University of Texas Press, Austin, 269–305.

Yang, Charles (2010). Three factors in language variation. *Lingua* 120:5, 1160–1177.

Zimmermann, Michael (2014). *Expletive and Referential Subject Pronouns in Medieval French*, vol. 556. Walter de Gruyter GmbH.