The conditioning of overabundance: A probabilistic grammar perspective on French ALLER in periphrastic tenses

Word count: 10,226

author1first author1lastl
author1affiliation
author1@instit1.xx

Author2first author2last
author2affiliation
author2d@instit2.xx

Author3first author3last
author3affiliation
author3d@instit3.xx

Author4first author4last
author4affiliation
author4d@instit4.xx

Abstract In this paper we study overabundance within the paradigm of the French 'aller' ('to go') verb, where GO-forms coexist with BE-forms (coming from the paradigm of the 'être' verb). Through extensive and fine-grained corpus work coupled with robust inferential analyses combining random forests and Bayesian modeling, we show that this overabundance phenomenon is heavily influenced by several linguistic variables such as animacy of the subject, verbal tense and person, negation, type of complement, but also – and mainly – by speaker-level and interaction-level variables such as age, gender, region of origin or the number of participants that are present. Going further than previous works on this alternation, we provide evidence for meaningful interactions between the two sets of factors. We argue for the need to take prescriptive rules into account when studying overabundance, but also their differentiated integration by speakers depending on their social profile.

Keywords: paradigmatic morphology; corpus; random forests; predictive margins; sociolinguistics; compound tenses

1 Introduction

Since Thornton (2011) defined overabundance in morphology as "the general phenomenon of having two or more forms that realize the same cell in an inflectional paradigm", various empirical studies have documented cases of rivalry between overabundant forms (called *cell-mates* by Thornton 2011) for different languages: for example, Italian (Thornton 2012), Spanish (Guzmán Naranjo 2017), Czech (Bermel et al. 2018; Guzmán Naranjo & Bonami 2021) or Estonian (Aigro & Vihman 2024). The general aim of this paper is to document another case of overabundant forms rivalry and to incorporate its study into a probabilistic approach to grammar (Bresnan et al. 2007; Bresnan & Ford 2010; Szmrecsanyi et al. 2017: a.o). According to this approach, speakers' linguistic productions are not only driven by categorical rules, but also by probabilistic choices depending on information of different levels. In this view, the choice between inflectional competitors – here the overabundant forms – that the speaker makes when producing a sentence is the result of a set of factors that together determine the probability of each possible *cell-mate* in the given context. We apply this approach to the case of the French verb ALLER

: it displays overabundance in compound tenses and the probabilistic distribution of its overabundant forms remains understudied (Sammons et al. 2015; Glikman & Patard 2022).

This article is organized as follows. In Section 2, we discuss the paradigm of the ALLER ('to go') verb in French and the incursion of the verb ÊTRE ('to be') into it. In Section 3, we present the dataset we explored and how we annotated it. In Section 4, we describe the relationship between the occurrence of either form of the paradigm and several sentence-level (linguistic), speaker-level (social) and interaction-level variables, before running inferential analyses in Section 5 by using both random forests and Bayesian modeling. Before the conclusion, we discuss our findings in Section 6 and, taking a step back, we compare them to previous studies on the alternation. In particular, we show how social, interaction-level and linguistic factors are intertwined and how it is necessary to study them in combination rather than in isolation to grasp the whole extent of the fine-grained influences over the realization of competing forms from the ALLER paradigm.

2 Overabundance in the paradigm of ALLER

It is very common in productions from French speakers to observe the use of the verb *être* 'to be' in contexts where the intended meaning is *aller* 'to go'. The attested examples in (1)[1] illustrate this alternation: in both sentences, the conveyed meaning is "going somewhere", with a directional prepositional phrase, and in (1a) *aller* is used, whereas in (1b) it is *être*.

(1) a. on **était** **allés** jusqu'à Montmartre (CEFC) 'We'd gone all the
PRO.3SG be.PST.3SG go.PTCP to Montmartre
way to Montmartre.'
b. on **a** **été** jusqu'à Porte_d'Italie (CEFC) 'We went all
PRO.3SG have.PRST.3SG be.PTCP to Porte_d'Italie
the way to Porte d'Italie.'

This alternation is possible not only for the locative meaning, but also for more abstract meanings as in (2), as well as when the verb takes an infinitive complement (3) or for the locution *ça va* ('it's okay') (4).

(2) a. je **suis** **allée** sur Google (CEFC) 'I went on Google.'
PRO.1SG be.PRST.1SG go.PTCP on Google
b. ils **ont** **été** sur internet (CEFC) 'They went online.'
PRO.3PL have.PRST.3PL be.PTCP on internet

(3) a. on **est** beaucoup **allé** voir les films asiatiques (CEFC)
PRO.3SG be.PRST.3SG a.lot go.PTCP see the movies Asian
'We went to see Asian movies a lot.'
b. qui **a** **été** voir le film de Michael Jackson (CEFC)
who have.PRST.3SG be.PST.PTCP see the movie of Michael Jackson
'Who went to see the Michael Jackson movie?'

[1] All examples are extracted from our study corpus: the oral part of *Corpus d'Etude pour le Français Contemporain* (CEFC, https://www.ortolang.fr/market/corpora/cefc-orfeo) . For more details about the corpus, see section 3.2.

(4) a. c' est le seul moment où c' **est** **allé** mieux (CEFC) 'That's
it is the only moment where it be.PRST.3SG go.PTCP better
the only time it got better.'
b. et ça **a** **été** beaucoup mieux (CEFC) 'And it got a lot
and it have.PRST.3SG be.PTCP much better
better.'

This incursion of BE into the domain of GO is not new, nor is it restricted to French. It is attested for example in other Romance languages, such as Ibero-Romance (Juge 1999) or Gallo-Romance (Bach 2022). As formulated by Vandeloise (2007), the interpretation of GO and BE in the past tense "leads to the same result but by opposite paths[2]": GO describes the change of location and anticipates the final location, while BE suggests the change of location and focuses on the final location.

This interpretation is compatible with the 3-step diachronic path proposed by Glikman & Patard (2022) for French. Step 1, termed the *bridging context*, occurred in Old and Middle French. It consisted of the use of the verb past tense *être* with locative prepositional phrases, leading to an ambiguous interpretation: either a static reading, corresponding to the meaning of BE, or a dynamic reading, drawing the parallel between BE and GO. During the period going from the 15th to the 17th centuries, the verb *être* appears in some contexts that are no longer ambiguous, especially with an infinitive complement, as in (3). These contexts are the *switch contexts*, corresponding to Step 2. Then, since the 18th century, the use of *être* has spread to all possible contexts where *aller* can be used in the past tense (Step 3), as in examples (1)-(4). This is consistent with the observations of Damourette & Pichon (1911-1927), two French grammarians, who had identified these different uses of the verb *être* in early 20th century: "An important point is the collaboration of the verb *être* in the conjugation of the verb *aller*. We'll explain: while the verb *aller* has a complete conjugation on its own, […] this conjugation is doubled, in the case of *sûtes*[3] and all compound past tenses, by the concurrent use of the verb *être* with all the same meanings[4]" (Damourette & Pichon 1911-1927: 93).

From a morphological point of view, the paradigm of the French verb ALLER displays well-known cases of suppletion (Boyé 2000), as shown in Table 1 with the suppletive stems of the verb in simple tenses (*all-, ir-, v-, aill-*). In the case of compound tenses[5], we observe that each regular verbal form, consisting of the auxiliary *être* ('be') and the past participle, is in competition with a suppletive form, made up of the auxiliary *avoir* ('have') and the past participle of the verb ÊTRE ('be'), as illustrated in examples (1) to (4)[6]. Numerous morphologists, among which Ackerman & Webelhuth (1998); Ackerman et al. (2011); Bonami & Webelhuth (2012); Bonami (2015), consider these compound tenses as grammaticalized periphrases belonging to the verb inflectional paradigm. According to this analysis, the lexeme ALLER shows overabundance in the periphrastic tenses: the competition between the ALLER forms in compound tenses corresponds to the rivalry

[2] Our translation of "conduit au même résultat mais par des chemins opposés"(Vandeloise 2007: 357)
[3] The word *sûtes* is specific to Damourette & Pichon (1911-1927) and corresponds to indicative perfect.
[4] DeepL translation of "*Un point important, c'est la collaboration du verbe* être *à la conjugaison du verbe* aller. *Nous nous expliquons : certes le verbe* aller *possède par lui seul une conjugaison complète, […] mais cette conjugaison est doublée, en ce qui concerne le* sûtes *et tous les tiroirs passés composés, par l'emploi concurrent du verbe* être *avec toutes les mêmes significations.*"
[5] In French grammars, these compound tenses are usually called *passé composé*, *plus-que-parfait*, *futur antérieur*, *subjonctif passé* and *conditionnel passé*.
[6] Overabundance also exists in the simple past tense, *passé simple*, (*j'allai* vs. *je fus le voir*, Blampain & Hanse 2012). However, we set aside this case of overabundance, because this tense is reserved for formal and literary usage and is almost unattested in spontaneous speech.

between *cell-mates* (Thornton 2011). In this respect, the behavior of French ALLER differs from that of Ibero-Romance and Gallo-Romance counterparts: instead of overabundance, these languages display overlapping suppletion, a situation in which the lexeme GO fills its perfective tense cells by borrowing the forms of BE, which, in addition, do not cease to exist independently (Juge 1999; 2019; Bach 2022). This is illustrated by the Portuguese paradigm for GO (5) and BE (6) in the present and preterite.

	1SG	2SG	3SG	1PL	2PL	3PL
PRST.IND	vais	vas	va	allons	allez	vont
PST.IND.IPFV	allais	allais	allait	allions	alliez	allaient
FUT.IND	irai	iras	ira	irons	irez	iront
PRST.COND	irais	irais	irait	irions	iriez	iraient
PST.SBJV	aille	ailles	aille	allions	alliez	aillent

Table 1: Part of the paradigm of French verb ALLER in simple tenses

(5) Portuguese GO: IR
   a. PRESENT: vou, vais, va, vamos, vão
   b. PRETERITE: **fui**, **foste**, **foi**, **fomos**, **foram**

(6) Portuguese BE: SER
   a. PRESENT: sou, és, é, somos, são
   b. PRETERITE: **fui**, **foste**, **foi**, **fomos**, **foram**

Based on the main dimensions by which this overabundance manifests, as defined by Thornton (2019) and rephrased by Guzmán Naranjo & Bonami (2021: 2), we can characterize further the phenomenon under study by describing: (a) the *lexical prevalence* (i.e. the number of lexemes affected); (b) the paradigmatic prevalence (i.e. the size of the set of paradigm cells affected); (c) the balance (i.e. the statistical distribution of rival forms for each cell); (d) the conditions (i.e. the factors conditioning the preference for one or the other form, be it usage and/or grammatical factors). First, its *lexical prevalence* is the lowest possible, since it affects only a single lexeme (ALLER), but its *paradigmatic prevalence* is much bigger, as all the cells in periphrastic tenses are concerned, as shown in Table 2.

	1SG	2SG	3SG	1PL	2PL	3PL
PRF.IND	suis allé ai été	es allé as été	est allé a été	sommes allés avons été	êtes allés avez été	sont allés ont été
PLPRF.IND	étais allé avais été	étais allé avais été	était allé avait été	étions allés avions été	étiez allés aviez été	étaient allés avaient été
FUT.PRF	serai allé aurai été	seras allé auras été	sera allé aura été	serons allés aurons été	serez allés aurez été	seront allés auront été
PST.COND	serais allé aurais été	serais allé aurais été	serait allé aurait été	serions allés aurions été	seriez allés auriez été	seraient allés auraient été
PST.SBJV	sois allé aie été	sois allé aies été	soit allé ait été	soyons allés ayons été	soyez allés ayez été	soient allés aient été
PST.INF				être allé avoir été		
PST.PTCP				étant allé ayant été		

Table 2: Part of the paradigm of French verb ALLER in periphrastic tenses

Regarding the *balance* of the overabundance phenomenon, previous research did not document the statistical distribution of the rival forms across the cells of the paradigm. They only give the general distribution : 50.1% (570 occ./1137) of *être allé* forms in Sammons et al. (2015)'s study about Ontario French, and 47% (238 occ./507) in Glikman & Patard (2022)'s work about European French. Both studies suggest that the overabundant forms are well attested, but lack in describing the cell by cell rivalry. Regarding the *conditions* that the use of rival forms are subject to, the two studies give insight on the grammatical and extra-grammatical parameters at play. Based on corpus analysis, Sammons et al. (2015) show that the choice between the 'cell-mates' is conditioned by both linguistic (clause polarity, and number of the verb form) and social factors (socioeconomic status, locality of residence of the speaker, speaker gender, category of French speaker in the bilingual context of Ontario). The oral corpus data of Glikman & Patard (2022) suggest that the rivalry may be influenced by a geographical factor, but the number of occurrences is very low. Their questionnaire results point out that a number of speakers consider themselves to be sensitive to prescriptivism in the use of the overabundant forms. They explain the alternation between the 'cell-mates' as a register-based phenomenon: *avoir été* forms are perceived as informal, and *être allé* forms as formal.

3 Corpus constitution and annotation

Our study aims at better describing the *balance* and the *conditions* of the ALLER overabundance, on the basis of corpus data. Unlike previous corpus-based studies of this phenomenon (Glikman & Patard 2022; Sammons et al. 2015) that limited their data to contexts where the verb was followed by a complement (PP or infinitive), we did not restrict our data to specific syntactic contexts. Instead, we kept all sentences in which either form could be used. This approach is crucial to capture the full scope of the rivalry between the cell-mates. As we are studying different varieties of European French, we included all cases where the alternation was acceptable in at least one variety. For example, the sentence (7b) is possible in Swiss French and thus both structures in (7) qualify

for inclusion in the dataset, even though (7a) is strongly preferred in the other European French-speaking regions.

(7) a. ça a été long 'It took a long time' (Standard French)
PRO.3SG have.PRST.3SG be.PTCP long
b. c' est allé long 'It took a long time' (Swiss-French)
PRO.3SG be.PRST.3SG go.PTCP long

In accordance with probabilistic grammar approaches, we focus on spoken data. We used the oral part of *Corpus d'Etude du Français Contemporrain*[7] (CEFC) corpus (Benzitoun et al. 2016; Debaisieux & Benzitoun 2020). CEFC is the result of the Orféo ANR project[8] which aimed to create a large corpus for the French language, with both written (6 million tokens) and oral (4 million tokens), by aggregating many different existing corpora. The oral part we ended up using includes data from different projects (Bérard 2020):

- the CFPP2000 (*Corpus de Français Parlé Parisien des années 2000*, 381,704 tokens) is made of interviews about Paris districts and suburb in the 2000s (Branca-Rosoff et al. 2000);
- the CFPB resource (*Corpus de Français parlé à Bruxelles*, 58,688 tokens, 5 hours) was built with the same protocol as the one used for CFPP2000 and contains interviews about Bruxelles, recorded in 2013-2015 (Dister & Labeau 2017);
- C-Oral-Rom (225,554 tokens, 22 hours) contains monologues, conversations and professional interactions, recorded between 1980 and 2005 (Moneglia & Martin 2008; Deulofeu & Blanche-Benveniste 2006);
- CRFP (*Corpus de référence du français parlé*, 374 789 tokens, 30 hours) mainly consists of private conversations but also includes professional and public speech (Delic 2004);
- the *French Oral Narrative* corpus (131,794 tokens) is a corpus of stories presented in public between 1998 and 2005 (Carruthers 2013);
- the OFROM part of CEFC (258,089 tokens) consists of the transcription of excerpts from guided interviews recorded in 2008-2012 in Switzerland (Avanzi et al. 2016);
- TCOF (*Traitement de Corpus Oraux en Français*, 374,789 tokens, 28 hours) is made of informal conversations between students, interviews, public and professional speech (André & Canut 2010);
- the corpus from Tokyo University of Foreign Studies (TUFS, 663,742 tokens and 52 hours) contains casual conversations recorded by French students between 2005 and 2011 (Kawaguchi 2011);
- the VALIBEL part of CEFC (402,285 tokens, 41 hours) corresponds to 72 interviews about linguistic representations of Belgian French speakers, recorded between 1988 and 2008 (Francard et al. 2002).

We extracted our dataset from this corpus aggregate, which contains approximately 2.8 million tokens representing a variety of speech in a variety of situations.

3.1 Extracting and selecting the data

We relied on the automatic POS and syntactic dependency annotation of the corpus (Nasr et al. 2020) to extract overabundant forms. We searched for ÊTRE and AVOIR lemmas that have an auxiliary relation with respectively the past participle forms *allé* and *été*.

[7] https://repository.ortolang.fr/api/content/cefc-orfeo/4/documentation/site-orfeo/home/index.html
[8] https://anr.fr/Projet-ANR-12-CORP-0005

Given that periphrastic tenses of ÊTRE are not only possible overabundant forms of ALLER ('to go'), but also compound tenses of the verb ÊTRE ('to be'), we refined our query to automatically exclude the *avoir été* occurrences that could not have the GO meaning. In particular, we excluded the cases where *avoir été* had a past participle dependent. This automatic extraction gave us 2,830 sentences, which we manually filtered to get only the *avoir été* occurrences that can alternate with *être allé*. Two different annotators read each occurrence to determine whether alternation with the *être allé* counterpart was possible in the given context. We kept the occurrences positively annotated by both annotators, and the annotation team discussed each non-congruent decision. Our guidelines were (a) to keep ambiguous occurrences, i.e. with both stative and dynamic meanings possible (8), (b) to exclude the aborted sentences if they did not allow the identification of the verb complement (9).

(8) j' ai été à Bourg-en-Bresse dans le matériel (CEFC)
1SG have.PRST.1SG be.PST.PTCP to CITY in the equipment
'I have been in Bourg-en-Bresse in the equipment sector.' or 'I went to Bourg-en-Bresse in the equipment sector.'

(9) j' **aurais été** j' aurais f (CEFC) 'I would
1SG have.PST.COND.1SG be.PST.PTCP 1SG have.PST.COND.1SG f
have I would f[aborted].'

In the end, we discarded 772 sentences and had a final number of 2,058 occurrences with 61.9% of *être allé*.

3.2 Annotation

The study by Sammons et al. (2015) suggests that the conditions of overabundance are related to the sociolinguistic profile of the speaker. Although we do not have access to as much fine-grained social data as in their study, we aimed to extract as much speaker- and interaction-related information as possible on our data, by using the corpus metadata. However, the research teams that made up each of the 9 oral subcorpora of CEFC had different goals and strategies when annotating transcriptions, resulting in a very composite set of metadata (Benzitoun & Etienne 2020). Therefore, we ensured that we had the maximum amount of speaker-related information by filtering out all CEFC files that did not provide information on the following metadata: gender, age, country/region of origin, and number of speakers. This conservative approach left us with 1,157 observations, which form our dataset for the remainder of this paper, and ensures good comparability across observations and thus robust statistical inferences.

In the present study, gender is a binary variable (F, M), as delineated in the metadata. The age variable is categorized into 3 levels (16-20, 21-60, 61 +), with no possibility of refining the splits. As for the region of origin, Switzerland (CH) and Belgium (BE) were both regarded as homogeneous regions in which all speakers were grouped together, thus ignoring potential country-internal sociolectal variation. Meanwhile, speakers from (continental) France were divided into three groups: Northern and Eastern France (F–NE); the area around Paris, Île-de-France (F–IDF); and a large part of metropolitan France which encompasses Southern and Western parts of the country (F–S&W). The three main French 'subregions' were determined by aggregating different administrative subdivision levels available in the metadata ('*département*', '*région*', etc.). The few observations from non-continental France and from places labeled as 'unknown' (N = 33) were discarded, for lack of a good way to put them in.

For the interaction-level variables, although a more precise specification of the contexts and topics under discussion in the interaction would have been very valuable, the information was too sparse and not systematized enough across subcorpora and transcriptions, and we could only keep the number of speakers in the interaction (1, 2 or 2+).

The annotation of linguistic factors was a two-step process. First, we automatically annotated the morphological categories (person, number, tense) on the basis of the form of the auxiliary. The second step corresponds to the annotation of syntactic information, with an automatic pre-annotation, followed by a manual verification.

The first variable is the type of complement. Previous corpus studies agree that the distribution of the overabundant forms is not affected by the distinction between infinitive or PP complement. To ensure that our data are consistent with these previous conclusions, we systematically checked whether the complement was an infinitive (3) or a PP (1) (2). Given that we did not restrict ourselves to these contexts, we also annotated whether the complement was the clitic proform *y* (10)[9], or whether it was absent, because recoverable from the context, as in (11) or (12). The 'Other' category gathers complements that are sparsely represented in the data: adverbs (4), NPs (13) or even adjectives (7).

(10) a. on y est allées avec NAME (CEFC) 'We went there with
PRO.3SG Y be.PRST.3SG go.PTCP with NAME
NAME'

b. on y a été avec ton père avec maman (CEFC)
PRO.3SG Y have.PRST.3SG be.PTCP with you father with mommy
'We went there with you father, with mommy.'

(11) [tu vas sur la Canebière à Marseille tu peux pas marcher]
'You go on the Canebière in Marseille, you can't walk.'

a. tu es déjà allée (CEFC) 'Have you ever been there
PRO.2SG be.PRST.2SG already go.PTCP
?'(litt. 'You already went ?')

(12) [on a de la famille en Floride]
'We have family in Florida.'

a. mon frère a été cette année (CEFC) 'My brother went
my brother have.PRST.3SG be.PTCP this year
this year.'

(13) a. on est allés euh Place Stan (CEFC) 'We went to Place
PRO.3SG be.PRST.3SG go.PTCP DISF square Stan
Stan'

b. on avait été rue Tillemont (CEFC) 'We'd been on
PRO.3SG have.PRST.3SG be.PTCP street Tillemont
Tillemont street.'

We also annotated the presence or absence of negation because it is mentioned in Sammons et al. (2015) as a relevant factor. The negative polarity items found with the verbs are *pas*, *jamais* and *plus*. In most contexts, the expletive *ne* is absent, because the corpus mainly contains spoken colloquial French. However it still appears in 26 sentences (out of 114 negation contexts), as in (14).

[9] In the CEFC corpus, first/last names are anonymized with the label NAME.

(14) a. je n' y suis jamais allée (CEFC) 'I've never been there. '
PRO.1SG NE Y be.PRST.1SG never go.PTCP
b. ça n' a pas été (CEFC) 'It did not work out.'
PRO.3SG NE have.PRST.3SG not be.PTCP

The last annotated variable is the animacy of the subject of the verb. Most subjects refer to animate entities, but we encountered inanimate subjects in the form of the demonstrative pronoun *ça/c'*, not only in the locution *ça va* (4), but also in contexts like (15). There is also an NP subject that is inanimate in our data (16).

(15) c' est allé vite quand_même (CEFC) 'It went fast anyway.'
PRO.3SG be.PRST.3SG go.PTCP fast anyway

(16) la caméra n' a pas été dans les douches (CEFC) 'the camera
the camera NE have.PRST.3SG not be.PTCP in the showers
did not go into the showers'

Table 3 summarizes the variables annotated in the corpus, with the distribution of each level by variable.

Variables	Levels	Occ.
GENDER	F	682
	M	475
AGE	16-20	254
	21-60	733
	61+	170
REGION	F–N&E	142
	F–S&W	341
	F–IDF	242
	CH	194
	BE	238
NBINTERLOC	1	133
	2	626
	2+	398
TENSE	PRF.IND	1037
	PLPRF.IND	105
	OTHER	15
SUBJANIM	Anim	1086
	Inan	71
PERSON	1	516
	2	104
	3	537
CPLTTYPE	PP	578
	Inf.	293
	y	138
	NoCplt	103
	Other	40
NUMBER	SG	1059
	PL	98
POLARITY	Positive	1043
	Negative	114
VERB	*être allé*	664
	avoir été	493

Table 3: List of variables annotated in the data set

4 Results: Descriptive approach

In this section, we present the distribution of the overabundant forms for each annotated variable. First, we document the *balance* of the overabundance in the paradigm of ALLER. Second, we explore the potential factors conditioning the preference for one or the other form, by describing the linguistic variables, then the speaker- and interaction-level variables.

4.1 Distribution in the paradigm of ALLER

The statistical distribution of *être allé* and *avoir été* for each cell of the paradigm is shown in Table 4. As already noticeable in Table 3, the perfect and pluperfect tenses are the most represented, as well as the 1SG and the 3SG. With proportions ranging from 81% to 57%, the *être allé* forms are in the majority for 1SG, 3SG, 1PL, and 3PL of the periphrastic perfect. Cell 2SG displays a more balanced distribution with 51.5% (35/68) of *être allé*, and there is a majority of *avoir été* for 2PL. For the pluperfect tense, the distribution is more balanced with exactly the same number of occurrences for 1SG, 2SG and 2PL, and a majority of *être allé* only for 3SG and 1PL. Despite the sparsely populated cells in the rest of the paradigm, it seems that the past conditional favors *avoir été* (8/10), except for cell 1SG. This is consistent with the data of Sammons et al. (2015) that show exactly the same distribution (8 *avoir été*/10). Given the very unbalanced distribution of the different tenses in our dataset, we henceforth consider only two levels for the variable TENSE: Perfect Indicative vs. Other.

The *balance* presented in Table 4 suggests that the rivalry is not strongly influenced by the structure of the paradigm: each cell appears to be a potential candidate for the rivalry between the two overabundant forms. Therefore, there is room for other conditioning factors, both linguistic and extralinguistic, which are the focus of the following two sections.

		1SG	2SG	3SG	1PL	2PL	3PL
PRF.IND	*être allé*	253	35	261	13	7	35
	avoir été	186	33	180	3	14	17
PLPRF.IND	*être allé*	26	5	20	3	1	0
	avoir été	26	5	17	0	1	1
FUT.PRF	*être allé*	0	0	1	0	0	0
	avoir été	0	0	0	0	0	0
PST.COND	*être allé*	2	0	0	0	0	0
	avoir été	2	1	2	0	2	1
PST.SBJV	*être allé*	0	0	0	0	0	0
	avoir été	0	0	1	0	0	0
PST.INF	*être allé*			1			
	avoir été			1			

Table 4: Corpus distribution of overabundant forms in ALLER paradigm

4.2 Linguistic variables

Table 5 presents the proportion of *être allé* and *avoir été* as a function of the linguistic variables.

As mentioned above, the periphrastic perfect tense seems to favor the use of *être allé*, but the number of observations is not sufficient to have a significant effect (χ^2 =3.0177, df=1,p=0.082). The first and third persons are preferentially associated with *être allé*, and the second with *avoir été*. The direction of the effect of the variable NUMBER is consistent with that found in Sammons et al. (2015): the plural appears to attract the *être allé* forms, but the effect is not statistically significant (χ^2 =0.23244, df=1,p=0.6297). Given the distribution of the rival forms in Table 4 and the marginal significance of the variables

NUMBER and PERSON, we decide to merge the information on person and number into a single variable, called PERSNB, for statistical modeling (Section 5). Regarding person and number, we also annotated the distinction between the two possible pronouns that carry the first-person plural value: formal *nous* and informal *on* (variable SUBJONNOUS). First-person plural pronouns seem to attract *être allé* verb forms (χ^2 =6.968, df=2,p=0.03), with a stronger effect for the *nous* pronoun, although the number of occurrences is very low.

At the level of the sentence, the negation seems to slightly reduce the overall preference for *être allé*. This trend is less strong than in the study of Sammons et al. (2015) where the negative polarity increases significantly the proportion of *avoir été* to 67.7% in Ontario French. The presence of an inanimate subject clearly favors the use of *avoir été*, but this observation relies on a small number of observations (71/1157). Finally, the distribution of the type of verbal complement shows that the proform *y* seems to strongly favor the use of *être allé*, and that the absence of an overt complement favors *avoir été* to a lesser extent. In contrast, there is no significant difference in the choice of the overabundant form with infinitive and PP complements (χ^2 =0.14653, df=1,p=0.7019). This observation is fully consistent with the findings of previous studies (Sammons et al. 2015; Glikman & Patard 2022). Based on these figures, we have merged the infinitive and PP levels within the category "Other" for the rest of the article.

	être allé	*avoir été*
TENSE		
PRF.IND	58.3% (N= 604/1036)	41.7 % (N= 432/1036)
OTHER	49.6% (N= 60/121)	50.4% (N= 61/121)
NUMBER		
SG	57.1% (N= 605/1059)	42.8 % (N= 454/1059)
PL	60.2% (N= 59/98)	39.8% (N= 39/98)
PERSON		
1	57.8% (N= 298/516)	42.2% (N= 218/516)
2	46.2% (N= 48/104)	53.8% (N= 56/104)
3	59.2% (N= 318/537)	40.8% (N= 219/537)
SUBJONNOUS		
nous	84.2% (N= 16/19)	15.8% (N= 3/19)
on	60.5% (N= 124/205)	39.5% (N= 81/205)
Other	56.2% (N= 524/933)	43.8% (N= 409/933)
POLARITY		
Positive	57.7% (N= 602/1043)	42.3 % (N= 441/1043)
Negative	54.4% (N= 62/114)	45.6% (N= 52/114)
SUBJANIMACY		
Animate	59.3% (N= 644/1086)	40.7% (N= 442/1086)
Inanimate	28.2% (N= 20/71)	71.9% (N= 51/71)
COMPLEMENTTYPE		
No Complement	46.6% (N= 48/103)	53.4% (N= 55/103)
Proform *y*	83.3% (N= 115/138)	16.7% (N= 23/138)
PP	56.2% (N= 325/578)	43.8% (N= 253/578)
Infinitive	54.6% (N= 160/293)	45.4% (N= 133/293)
Other	35.6% (N= 16/45)	64.4% (N= 29/45)

Table 5: Linguistic factors and the use of *être allé* vs. *avoir été*

4.3 Speaker-level (sociolinguistic) and interaction-level variables

Table 6 gives an overview of the proportions of *être allé* and *avoir été* uses depending on the different speaker-level variables identified from the metadata. We observe that gender seems to affect the use of the forms: women appear to use significantly more *être allé* than men (χ^2 =12.369, df=1,p<0.001). Regarding age, while younger groups (16-20 and 21.-60 y.o.) seem to favor *être allé* uses, speakers aged 61 and more seem to favor *avoir été* (χ^2 =13.655, df=2,p<0.01). For the last speaker-level variable, country/region, an interesting sociolectal pattern emerges, with speakers from Belgium and Northern and Eastern France using more *avoir été*, speakers from Southern and Western France, as well as from Île-de-France, using more *être allé*, and speakers from Switzerland using a balanced mix of the two alternatives.

	être allé	*avoir été*
GENDER		
F	61.7% (N= 421/682)	38.3 % (N= 261/682)
M	51.2% (N= 243/475)	48.8% (N= 232/475)
AGE		
16-20 y.o.	63.0% (N= 160/254)	37.0% (N= 94/254)
21-60 y.o.	58.3% (N= 427/733)	41.7% (N= 306/733)
61+ y.o.	45.3% (N= 77/170)	54.7% (N= 93/170)
REGION		
Belgium (BE)	33.2 % (N= 79/238)	66.8 % (N= 159/238)
Switzerland (CH)	51.5 % (N= 100/194)	48.5 % (N= 94/194)
F–N&E	46.5 % (N= 66/142)	53.5 % (N= 76/142)
F–IDF	59.1 % (N= 143/242)	40.9 % (N= 99/242)
F–S&W	80.9 % (N= 276/341)	19.1 % (N= 65/341)

Table 6: Speaker-level factors and the use of *être allé* vs. *avoir été*

Finally, at the interaction level, Table 7 provides an overview of how different contexts involving either one person, two people, or more than two people affect the production of the overabundant forms. The more frequent use of *être allé* in our data appears to come from interactions with only 1 and 2 speakers, while interactions with more than 2 speakers show a more balanced pattern (χ^2 =17.672, df=2,p<0.001).

	être allé	*avoir été*
1 person	69.2% (N= 92/133)	30.8% (N= 41/133)
2 people	59.6% (N= 373/626)	40.4% (N= 253/626)
More than 2 p.	50.0% (N= 199/398)	50.0% (N= 199/398)

Table 7: *Être allé* vs. *Avoir été* depending on NBINTERLOCUTORS

4.4 Correlations between variables

Figure 1 summarizes the existing correlations between the variables we studied (numerically recoded for this purpose). Statistically significant correlations are those not crossed off. The size of the dot (upper side of the graph) is proportional to the numeric value of the correlation coefficient that was calculated (found in the lower side of the graph), with positive correlations shown in blue and negative ones in red. The main take-away from this overall graph is that not so many variables are correlated (many crosses all over), and none are correlated to such a high degree that it would be detrimental to further inferential analysis (max coefficient: –0.31, between the PERSNB variable and the SUBJONNOUS one, which is derived from it (all 'on' subjects are 3^{rd} person singular, and all 'nous' are 1^{st} person plural).

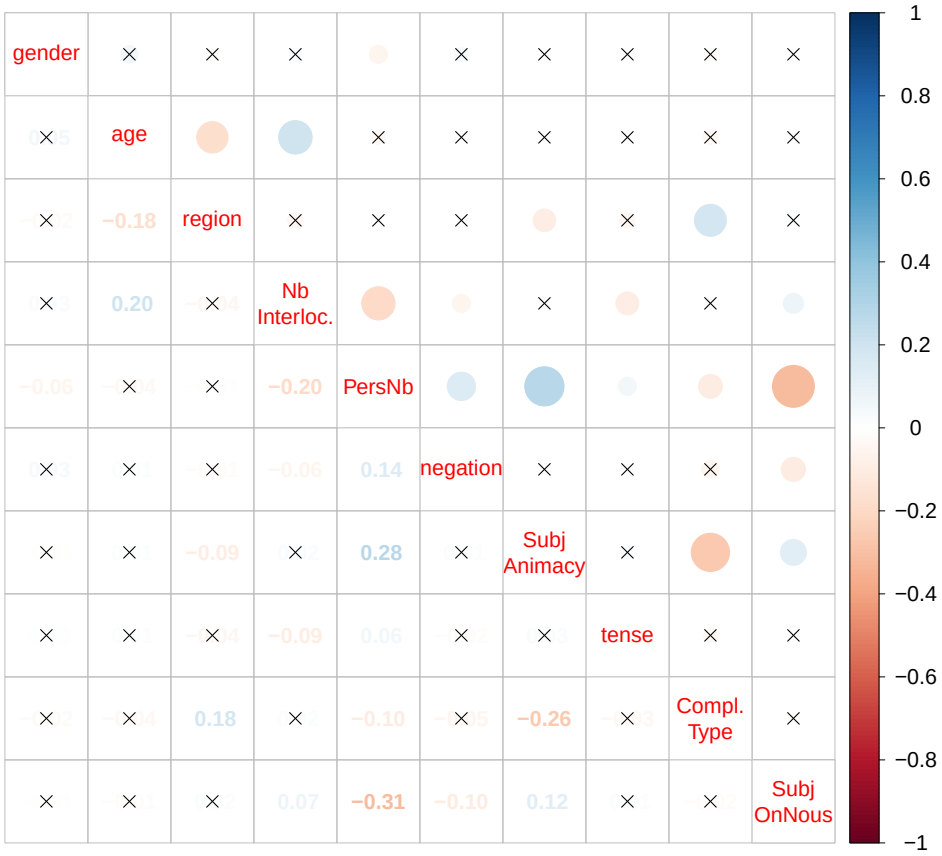

Figure 1: Correlation matrix for all variables of interest (method = Pearson)

5 Generalizing over the results: Conditional inferential trees and conditional random forest

To investigate the robustness of the associations between the use of *être allé* or *avoir été* and variables from our dataset, we used conditional inference trees and random forests (Levshina 2020). This NLP-inspired inferential technique is best suited for a dataset such as ours, where there are many predictors of interest, with high heterogeneity in the number of observations by predictor. On the contrary, classic analyses such as frequentist inferential models, and in particular logistic regressions would be ill-suited here, because of this high-variability and the high number of predictors. First we will present the overall results from a conditional random forest calculated with all the predictors isolated in the previous sections, then we will do a more in-depth analysis of the interactions between predictors.

5.1 Trees, random forest and overall results

Conditional inference trees operate a recursive (here, binary) partition of the data: they recursively split the data according to the predictors they are fed, to maximize the difference between the possible values of the dependent variable. The splits create bins of data (subsets) according to one predictor[10], across which the difference in value for the dependent variable is maximal. In each bin, further binning is then operated with the remaining predictors, until no more meaningful differences can be found. Figure 2 gives a visual representation of a lone tree fit with the *partykit* package (Hothorn et al. 2006; Hothorn & Zeileis 2015), within the R framework (R Core Team 2024) and the RStudio software (Posit Team 2024).

To give an example on how this can be read, in this tree, it is calculated that the split that is most accurately predictive of the dependent variable values is made according to the REGION predictor, with speakers from Southern and Western France producing a higher proportion of *être allé* than speakers from all the other areas. Then, for these speakers from Southern and Western France, the predictor that entails the most meaningful split of the data, as regards predictive accuracy of the dependent variable, is SUBJANIMACY, with inanimate subjects (produced by the speakers from this subset) being linked to less *être allé* observations than animate subjects. For the speakers from other areas however, the most accurate predictor of the dependent variable values is a refined definition of *region*, opposing people from Belgium to others. For these other regions (Île-de-France, Northern and Eastern France and Switzerland), the next most meaningful variable is GENDER. In observations from male speakers of this subset, the most important predictor is then the COMPLEMENTTYPE variable (proform *y*, no complement or any other complement), while for observations from female speakers of this subset, the following most important predictor is NBINTERLOCUTORS. And so on.

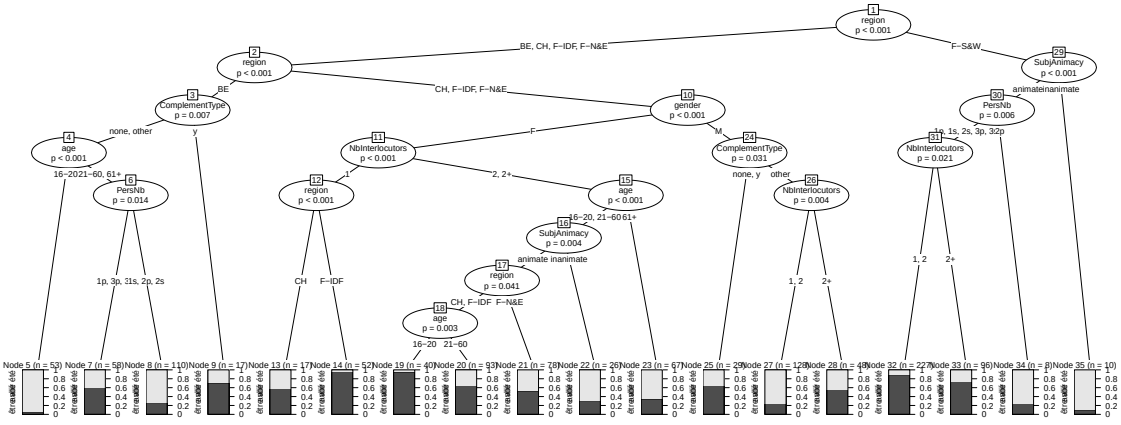

Figure 2: Random tree predicting *être allé* vs. *avoir été* uses

These results from the conditional inference tree are illustrative only, and other trees could give different splits depending on which observation the algorithms would have used as a starting point for the splits, and on many other parameters such as the minimal number of observations per final nodes, the maximal number of nodes, and so on. As an

[10] Details for the algorithm that selects predictors can be found in Levshina (2020), but the main idea is that the values for the dependent variable are reshuffled across all observations, and the chosen predictor is the one for which these permutations entail the biggest difference in association between predictor and dependent variable when comparing the post-permutation to the before-permutation associations.

example, neither the TENSE nor the SUBJONNOUS predictors are present in the random example tree above. To mitigate this variability in the results, conditional random forests are a useful expansion. They operate with the same approach, but generate (*grow*) a specified number of trees, over which predictions for each considered predictor of the dependent variable are averaged. In the end, it allows for a more accurate prediction of the dependent variable values. By averaging over all predictions from the individual conditional trees, random forests also allow to rank the predictors in terms of *how well they predict* the dependent variable values (i.e., *variable importance*).

Here, we used the *ranger* package (Wright & Ziegler 2017), to grow a forest of 2000 conditional inferential trees. The maximal number of predictor permutations per split (*mtry* parameter, which sets how many randomly selected predictors are tested at each split) was set to 3 after fine-tuning to see which number yielded the best overall predictive accuracy. The forest terminal node minimal size was 10 observations. The dependent variable was the form produced, either *avoir été* or *être allé*. Ten predictors were fed to the forest, with three speaker-specific predictors (GENDER, AGE and REGION), one interaction-level predictor (NBINTERLOCUTORS), and six sentence-internal linguistic predictors (NEGATION, SUBJANIMACY, TENSE, COMPLEMENTTYPE, SUBJONNOUS, PERSNB).

Various measures help characterize the predictive power of this forest. The Brier Score or *mean squared error* highlights the proportion of prediction errors by the forest, 0 reflecting perfect accuracy and 1 total inaccuracy. Here, we get an OOB[11] Brier Score of 0.168. Conversely, with a 10-fold cross-validation procedure, we can also estimate the mean proportion of correctly classified observations, which here amounts to 0.835 (calculated with the *rsample* (Frick et al. 2024) and *yardstick* (Kuhn et al. 2025) packages). Finally, the *pROC* package (Robin et al. 2011) helped us calculate the Area Under the *receiver operating characteristic* (AUC), which compares true positives and true negatives in predictions. An AUC of 0.5 would characterize an uninformative classifier, while 1 would mean a perfect one. Here, we get an AUC of 0.912. All in all, all measures point to a reliable predictive capacity for our forest. In this forest, Figure 3 illustrates how the different predictors rank in terms of how important they are in refining the prediction accuracy for *être allé* over *avoir été* uses. The absolute 'importance' numeric values on the x-axis are here less important than the ordering of variables itself, and the order of magnitude of the differences between them.

[11] *Out-Of-Bag* score. Each tree within the forest is trained on a bootstrap subsample of the complete dataset. The OOB score is calculated using the data points not included in this bootstrap sample, yielding an unbiased estimation of the model's prediction error.

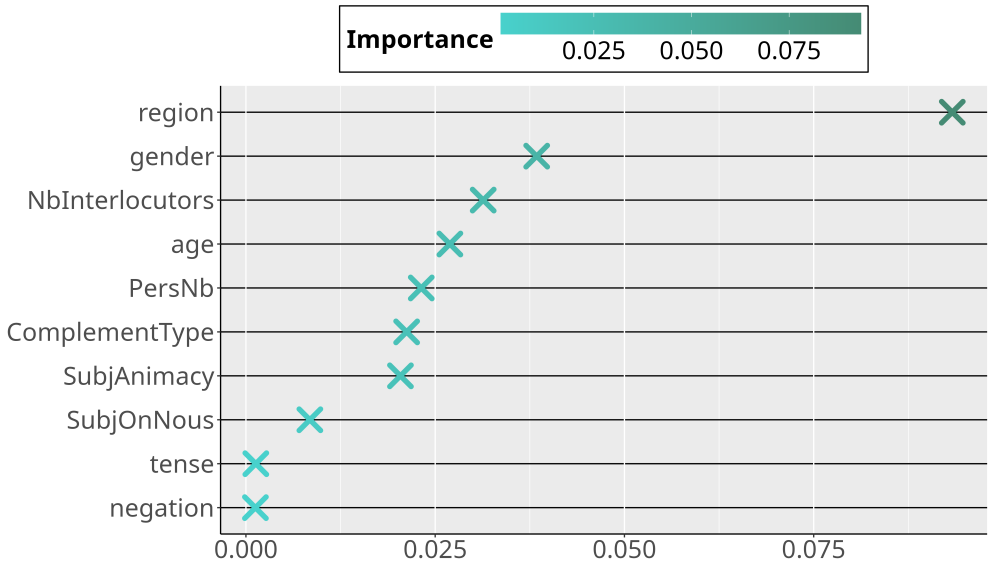

Figure 3: Relative variable importance in predicting *être allé* vs. *avoir été* uses

It appears that across all trees within the forest, the variable most reliably and most accurately predictive of the dependent variable values is REGION. Other speaker-specific predictors such as GENDER and AGE, as well as the interaction-level predictor NBINTERLOCUTORS are next. All sentence-internal predictors appear to be lower in the importance ranking. NEGATION and TENSE appear to have less importance when compared to the others. This analysis may be linked to a difference in nature between the sociointeractional predictors and the more linguistic ones.

The random forest we built here combines 9 independent variables. It is worth noting that this forest has a better accuracy than both a forest which includes only the 5 sentence-level variables (OOB Brier score = 0.229, AUC = 0.68, accuracy = 0.632) and a forest focusing only on the 3 speaker-level and the 1 interaction-level variables (OOB Brier score = 0.191, AUC = 0.812, accuracy = 0.731). The precision metrics from the linguistic-only forest illustrates how insufficient it would be to only focus on the sentence-level parameters at play, while the differences between the socio-interactional only forest and the full one we present here illustrate how important it is, even if social and interaction-level parameters appear to be the most important ones, to study both sets of factors at once when trying to understand the alternation.

That being said, even though the linguistic predictors create important descriptive contrasts in the use of either alternative forms, this variation phenomenon seems to be primarily sociolectal and interaction-driven. Taken together, the descriptive results from the previous sections and the results from this random forest approach call for a more in-depth analysis, where interactions between predictors are also accounted for.

5.2 Disentangling the relative weighs of linguistic and sociolinguistic factors

The *predictiveMargins* R package (Grafmiller & Sönning 2022) allows to explore the forest predictions further. More precisely, it allows an extraction of all predicted values for all trees within the forest, and for all combinations of the predictors that can be found in all these trees. With these predictions, it is possible to calculate the mean prediction for each different level of the target predictor variables, and thus to better understand in what direction a specific predictor influences the overall predictions from the forest. We

used the `avg_predictions()` function of the package to extract these for the predictors identified above and plot them. Importantly, different schemes can be applied to weigh the other predictors with this function, and we chose the default 'isolated' scheme here, which takes into account the overall distributions of these predictors and their levels in the data (Sönning & Grafmiller 2024).

5.2.1 Linguistic factors in isolation

Figure 4 illustrates the output of such a manipulation with the SUBJANIMACY predictor, where inanimate and animate subjects are contrasted. The dotted line represents a .5 predicted probability of *être allé* production. It appears that, all other predictors being weighed according to their distribution in the overall data, the presence of animate subjects in a sentence is linked to a higher predicted probability of *être allé* productions (0.60), compared to inanimate subjects (0.32). Here and in all subsequent graphs of this type, error bars represent 90% intervals around the mean predicted value.

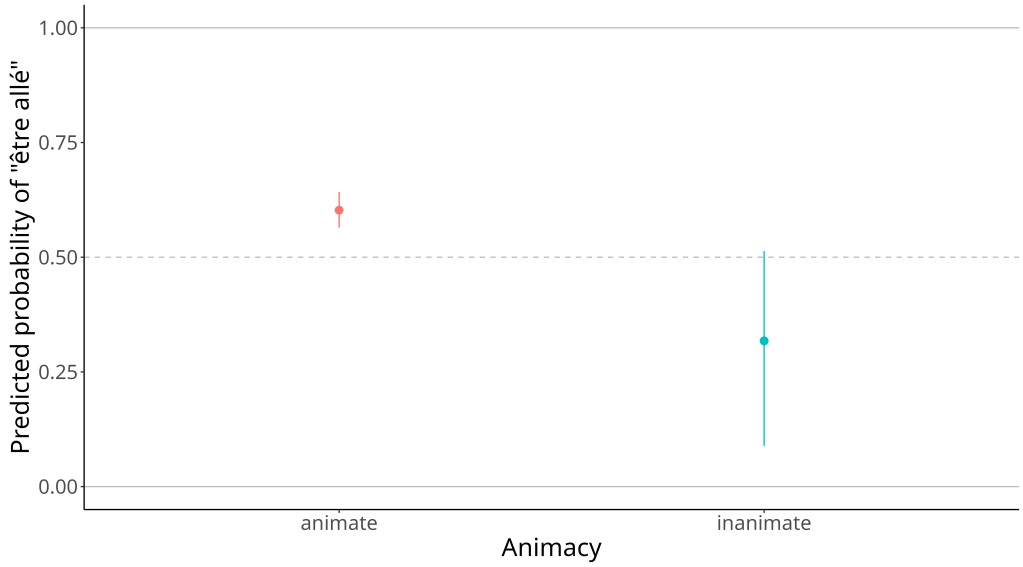

Figure 4: Predictions for "être allé" depending on SUBJANIMACY

When looking at other linguistic predictors, and first NEGATION, average predicted probabilities for *être allé* do not differ much whether in a negative sentence (0.57) or not (0.59). This could explain the relatively low importance of this predictor in the overall forest, compared to the other variables of interest. The TENSE predictor indicate that *passé composé* entails a slightly higher predicted probability (0.59) of *être allé* productions than other verb tenses (0.53). Regarding the SUBJONNOUS variable, there are also slight differences between predicted probabilities for *être allé* productions in sentences where the subject is *nous* (0.64), and sentences where the subject is *on* (0.59) or any other subject (0.58). The type of complement shows more meaningful differences (Figure 5), with the '*y*' proform being associated with a higher predicted probability of *être allé* productions (0.75) than when there is either no complement to the verb (0.59) or any other complement (0.56).

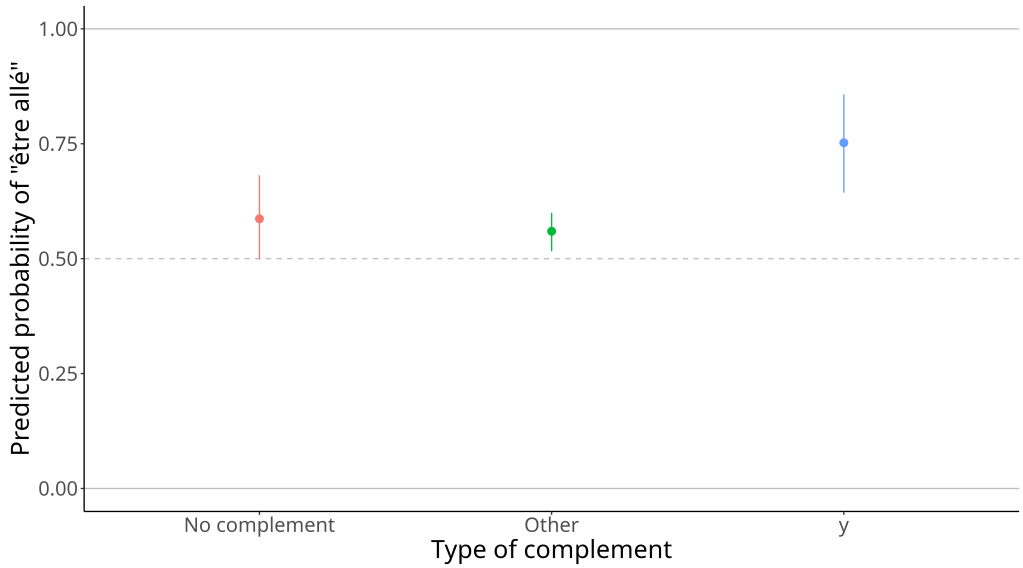

Figure 5: Predictions for “être allé” depending on ComplementType

Also of interest, the predictions for the different levels of the person-number combinations indicate that the 2nd person yields overall lower predicted probabilities of *être allé* productions, both in the singular (0.49) and even more so in the plural (0.41), than all other combinations (1sg: 0.58, 1pl: 0.65, 3sg: 0.61, 3pl: 0.62) (Figure 6).

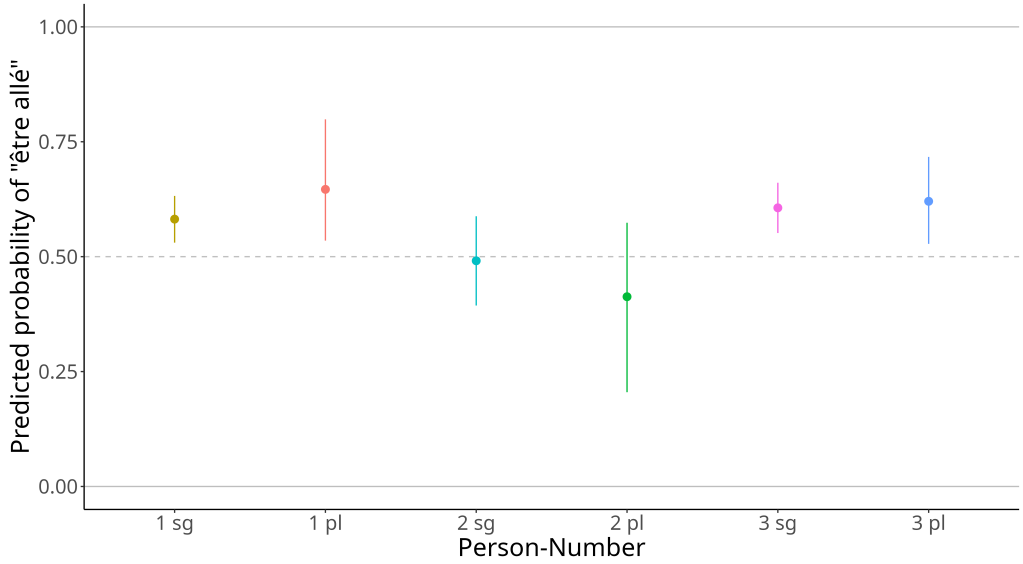

Figure 6: Predictions for “être allé” depending on PersNb

5.2.2 Sociolinguistic and interaction-level factors in isolation

As for speaker-level predictors, all three variables that we included in the forest have a fairly high importance in predicting the production of *être allé* (see Figure 3 above). In more details, average predicted probabilities for the age variable yield nuanced results, with a slightly lower predicted production of *être allé* for speakers older than 60 (0.56) than for speakers aged 21-60 (0.58) and than speakers aged 16-20 (0.61). The gender

variable shows (Figure 7) a similar pattern to that identified in the descriptive section (Table 6), with a higher predicted production of *être allé* for speakers identified as females in the recordings metadata (0.62) than for speakers identified as males (0.54).

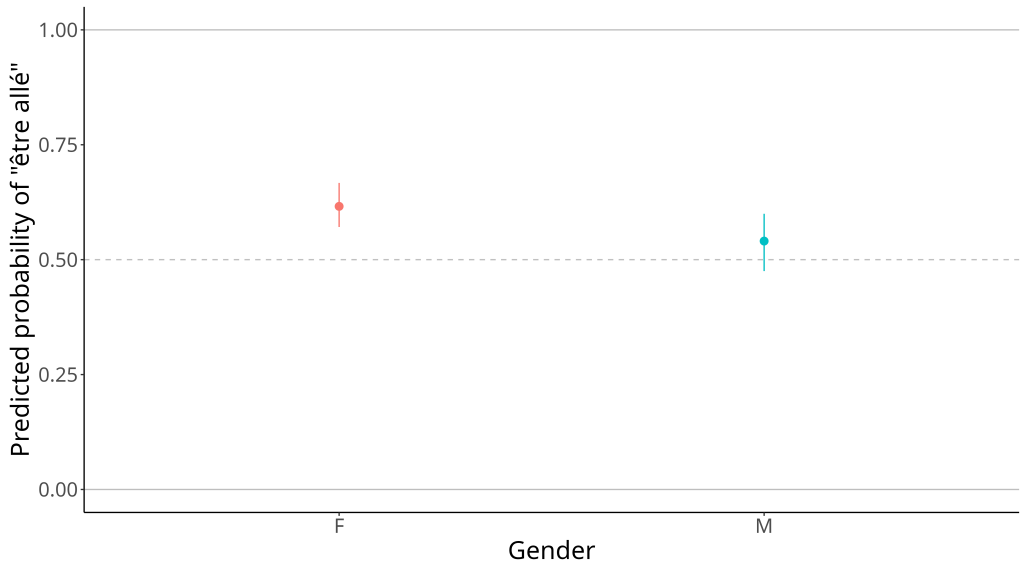

Figure 7: Predictions for "être allé" depending on GENDER

As for the speaker's region of origin, an interesting pattern emerges (Figure 8), again consistent with the descriptive data from Table 6. Here, the lowest predicted probability of *être allé* production is associated with speakers from Belgium (0.41), close to neighboring speakers from Northern and Eastern France (0.48) and Switzerland (0.53), while speakers from Southern and Western France are associated with the highest predicted values (0.78), with speakers from the central metropolitan region of Paris in between (0.60). Moreover, this variable is the one with the highest relative importance in all predictors included in the forest calculation (Figure 3). This makes a strong case for a case of *diatopic* variation, where French speakers from the geographical area comprising Belgium, Switzerland and the Northern and Eastern parts of continental France make use of less *être allé* and more *avoir été*, compared to speakers from the rest of France.

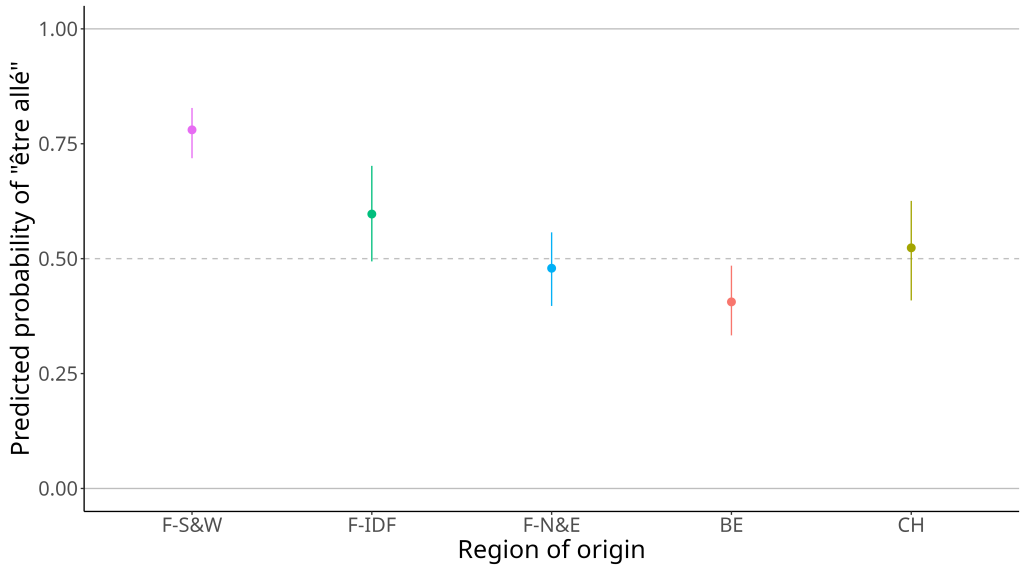

Figure 8: Predictions for “être allé” depending on REGION

Finally, the interaction-level predictor that relates to the number of participants in the interaction is also informative (Figure 9), with situations with only one speaker yielding higher predicted probabilities for the production of *être allé* (0.67) than situations with exactly two interlocutors (0.59), which in turn yield higher predicted values than situations with more than two speakers (0.54). This gradient pattern might be correlated to the type of speech produced depending on the number of participants in the corpus, since, for instance, observations with only one speaker are mostly situations where the speaker is telling a story, with a type of speech that may mimic a written, more normatively-constrained version of the language. But the difference that is also visible between situations with either only 2 or more-than-2 speakers is telling, in that it suggests a different use of the competing forms *être allé* and *avoir été* depending on the interaction setting and on the social dynamics between participants.

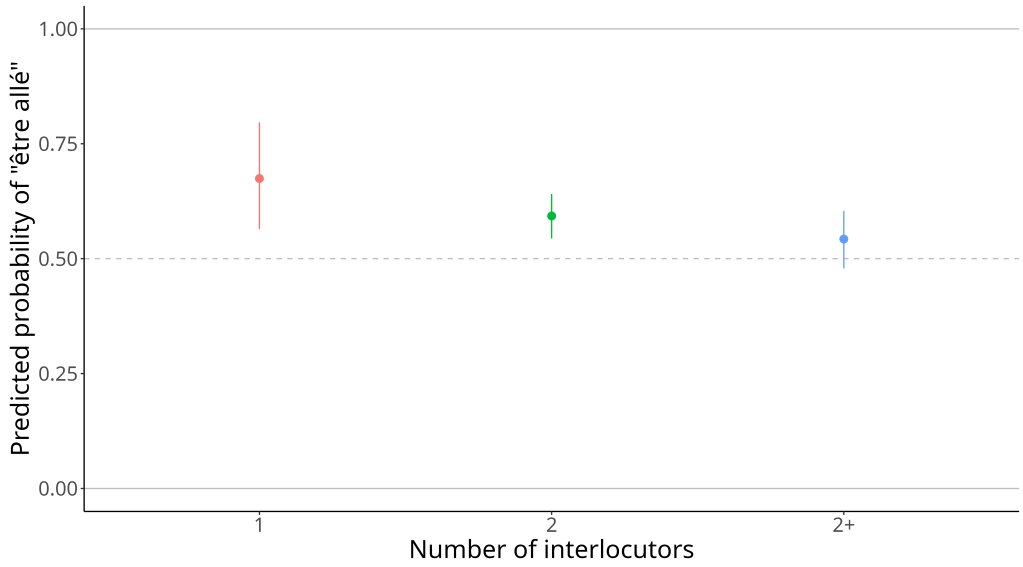

Figure 9: Predictions for “être allé” depending on NBINTERLOCUTORS

5.2.3 Mixing predictors

The results we presented above point to a network of variables influencing the production of either *être allé* or *avoir été*. These variables range from sentence-internal, linguistic variables to speaker-specific variables, with an interaction-level variable complexifying the general picture. To go a bit deeper than the variable relative importance values from Figure 3, we used the *predictiveMargins* package and its `avg_predictions` function to characterize the interactions between some variables. 3-way interactions would be too ambitious with the amount of data we extracted from the corpus, but some 2-way interactions allow for some informative observations over the respective importance of linguistic (SUBJANIMACY, COMPLEMENTTYPE, NEGATION), sociolinguistic (REGION, GENDER, AGE) and interaction-level (NBINTERLOCUTORS) variables.

To ensure these interactions remain meaningful, the calculations of average predictions from the random forest were confronted to Bayesian logistic regressions. Bayesian models are well-suited for analyzing data from small or unbalanced data (Sorensen et al. 2016), and we believe they are a good complement to the NLP-inspired approach with inferential trees and random forests. The combination of both methods of analysis ensures in our view the robustness of the results we present. These Bayesian inferential models are run over the exact same dataset than the random forest (N = 1157 observations), with a binary dependent variable corresponding to the verb that was produced (*être allé*, coded 1, or *avoir été*, recoded 0). All models were run in R and RStudio with the *brms* package (Carpenter et al. 2017; Bürkner et al. 2017; Bürkner 2018). For each model, the independent variables in interaction correspond to those the average predictions from the forest are calculated upon. They are mean-centered coded all the time, to ensure a good interpretation of the results (Brehm & Alday 2022). For clarity, we will only report for each model the estimated coefficient and the probability of having a meaningful effect of the combination of the two variables under consideration, and not the effects from isolated predictors (already described above). The estimated coefficient ($\hat{\beta}$) is a numerical characterization of the effect of the interaction of independent variables on the dependent variable, while the probability associated to it corresponds to the posterior distribution calculated by the model, which characterizes the probability for the real coefficient to be either greater ($P_{(\beta>0)}$) or smaller ($P_{(\beta<0)}$) than 0 (which would mean no meaningful interaction of the variables).[12]

5.2.3.1 Sociolinguistic factors and animacy

We will first present the interaction between the three speaker-level variables we studied and SUBJANIMACY. Figure 10 illustrates the predicted probabilities for *être allé* production depending on region of origin of the speakers but also on animacy of the subject. The left panel is reminiscent of results from Figure 8, with a gradient pattern where speakers from Belgium yield the lowest predicted probability (0.42), close to speakers from Northern and Eastern France (0.49) and speakers from Switzerland (0.53), while speakers from Southern and Western France show the highest predicted probability (0.81) and speakers from the Île-de-France region show intermediate values (0.61). Generally, the animate vs. inanimate opposition also resembles that from Figure 4, with higher predicted probabilities for sentences with animate subjects than for sentences with inanimate subjects. However, for sentences with inanimate subjects, the differences between regions seems

[12] More technical specifications about the models are available in the supplementary materials at https://osf.io/ahu6m/?view_only=2a3b503bffef4ba1af1fa6cc7b45b870.

to be 'neutralized' (granted, with wide margins), with consistent means across all geographical areas (F–S&W: 0.32, F–IDF: 0.32, F–N&E: 0.27, Belgium: 0.24). Speakers from Switzerland seem less affected by this variable (0.44), which is consistent with a sociolectal use of idiomatic '*c'est allé*' constructions in places where speakers from other locations might say '*ça a été*' (cf. examples (7)). A Bayesian logistic regression run for this interaction indeed yields evidence for a difference between animate and inanimate conditions when contrasting for example speakers from Southern and Western France and from Belgium ($\hat{\beta} = 4.26$, $P_{(\beta>0)} = .82$), or speakers from Switerzland and from Belgium ($\hat{\beta} = 8.22$, $P_{(\beta>0)} = 1$).

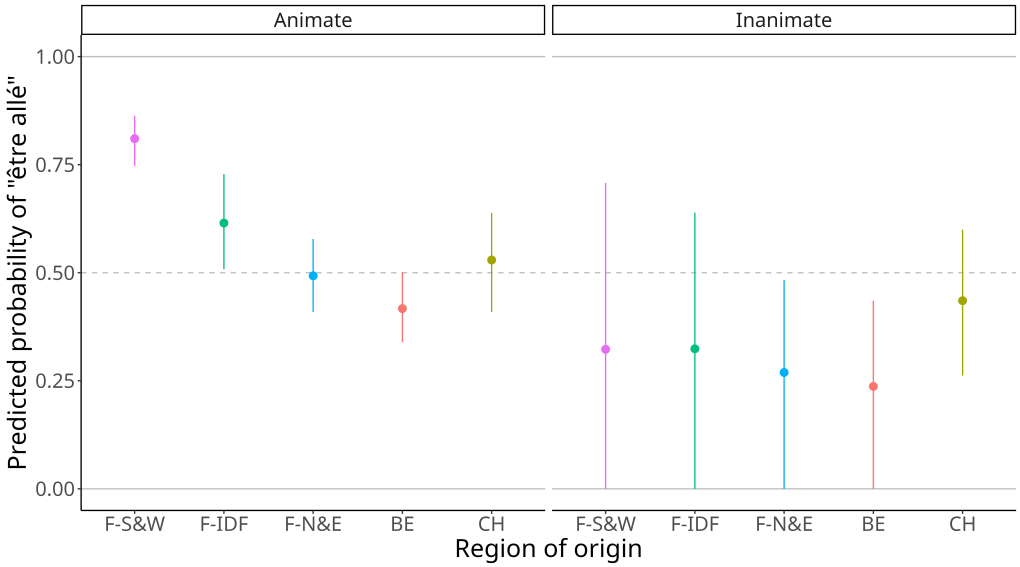

Figure 10: Predictions for "être allé" depending on REGION and SUBJANIMACY

Subject animacy also seems to interact in a noticeable way with gender, and Figure 11 gives an illustration of the combined effect of both variables on the predicted probability of *être allé* productions. The overall difference between the left and right panels is, again, reminiscent of Figure 4, and sentences with an animate subject yield generally higher predicted probabilities than sentences with inanimate subjects, regardless of gender of the speaker. For sentences with an animate subject, speakers identified as female are linked to higher predicted probabilities (0.64) than speakers identified as male (0.55), following the overall pattern for gender effects in Figure 7. In sentences with inanimate subjects however, this difference between genders seems to disappear (albeit with high prediction margins), with roughly equivalent mean predicted probabilities (0.31 for F, 0.33 for M). A Bayesian logistic regression gives further evidence for this interaction ($\hat{\beta} = -0.80$, $P_{(\beta<0)} = .92$).

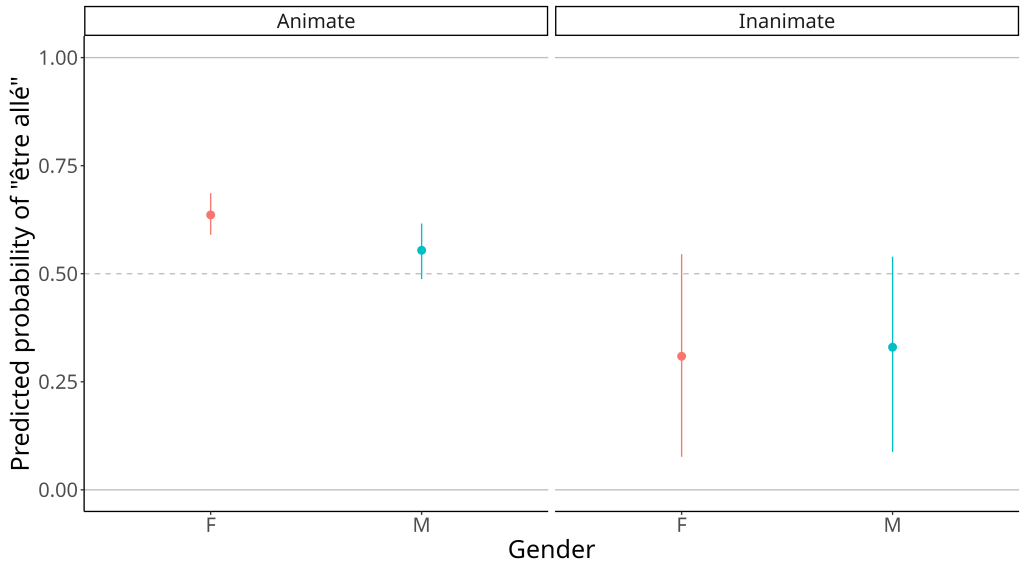

Figure 11: Predictions for “être allé” depending on GENDER and SUBJANIMACY

There does not seem to be much interaction between AGE and SUBJANIMACY, possibly because the productions of inanimate subjects were mostly coming from one age group (21-60 yo., N = 56/71).

5.2.3.2 Sociolinguistic factors and complement type

The complement type seems to interact only with REGION. Figure 12 gives an overview of how the regional differences hold, depending on whether there is no complement (left panel), whether the complement is the proform *y* (right panel), or whether there is any other complement (center). The left and center panels show a pattern similar to that from Figure 8, with predicted probabilities for *être allé* productions lowest for speakers from Belgium, Northern and Easter France, and highest for speakers from Southern and Western France, with speakers from Île-de-France and Switzerland somewhere in between. This contrasted pattern does not totally wear off in the right panel, but it seems to be reduced.

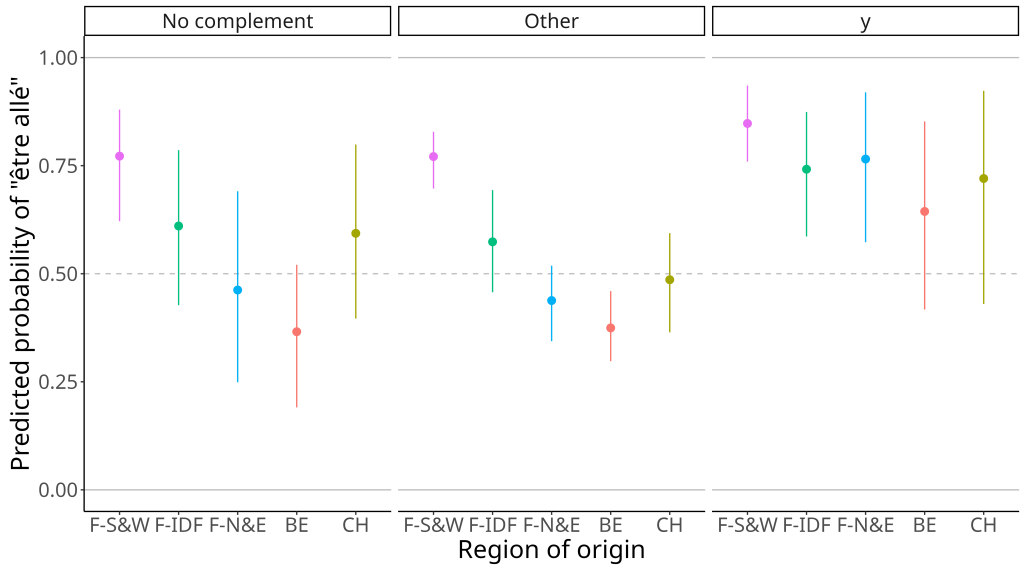

Figure 12: Predictions for "être allé" depending on REGION and COMPLEMENTTYPE

A Bayesian logistic regression with interacting *region* and *complement type* independent variables gives support to this observation. More precisely, it yields evidence for a difference between the 'other complement' and 'y' conditions when contrasting for example speakers from Belgium on one hand and, on the other hand, speakers from Northern and Eastern France ($\hat{\beta} = 0.88$, $P_{(\beta>0)} = .84$) or from Île-de-France ($\hat{\beta} = -1.04$, $P_{(\beta<0)} = .93$).

5.2.3.3 Sociolinguistic factors and negation

The influence of NEGATION over regional differences is less noticeable than the influence of subject animacy, but it is still meaningful (Figure 13). This time, the presence of negation (right panel) seems to widen the differences in predicted probabilities across geographical areas, with lower values in sentences with negation than in sentences without it, especially for speakers from Northern and Eastern France and from Belgium. A Bayesian logistic regression provides evidence for this when comparing the latter group to speakers from Southern and Western France ($\hat{\beta} = 0.71$, $P_{(\beta>0)} = .86$), to speakers from Île-de-France ($\hat{\beta} = 0.94$, $P_{(\beta>0)} = .93$), and also to speakers from Switzerland ($\hat{\beta} = 1.84$, $P_{(\beta>0)} = .96$).

This effect of NEGATION seems to be of a different nature than that of SUBJANIMACY or COMPLEMENTTYPE, in that it does not apply to all groups of people as defined, e.g., by region or gender, with reduced differences between groups. Rather, here, only people from some regions (Northern and Eastern France and Belgium) seem to be sensitive to the effect of negation, which on the contrary widens the difference in linguistic behavior across groups. Given the distribution of our dataset across regions (238/1157 observations from Belgium, and 142 from N&E France) and the relatively small number of negative sentences overall (114/1157), this in fact might be a reason why negation is so low on the variable importance ranking (Figure 3), to be investigated further with better control over the number of observations from all regions.

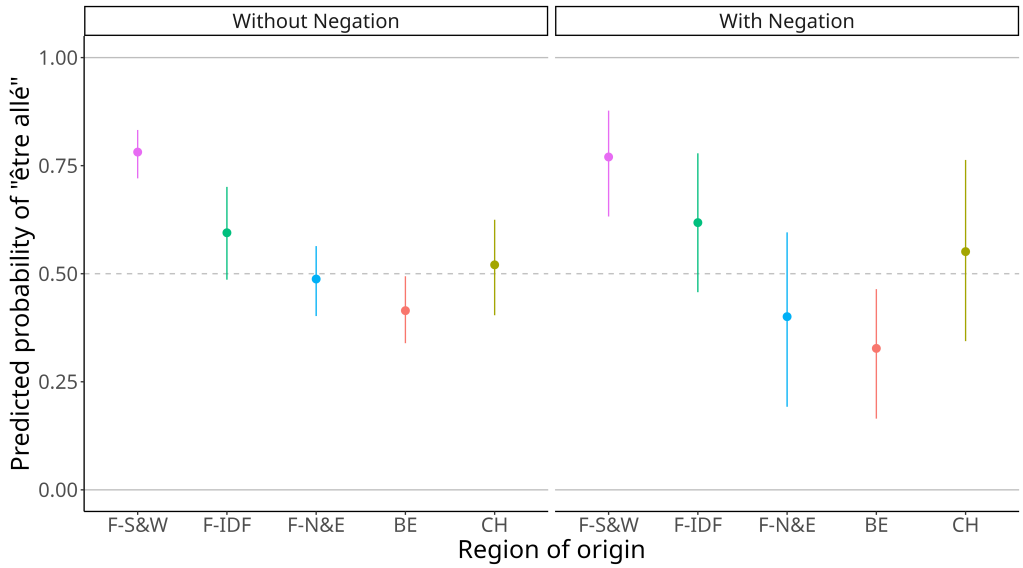

Figure 13: Predictions for “être allé” production depending on REGION and NEGATION

When considering the GENDER and NEGATION variables together (Figure 14), only women appear to be sensitive to the polarity of the sentence. In fact, the predicted probability for women to produce *être allé* is higher in sentences that do not contain negation (0.62) than in those containing negation (0.59), while for men, the predicted probabilities are identical regardless of polarity (0.54). A Bayesian logistic regression provides further support for this interaction ($\hat{\beta} = -0.61$, $P_{(\beta<0)} = .93$).

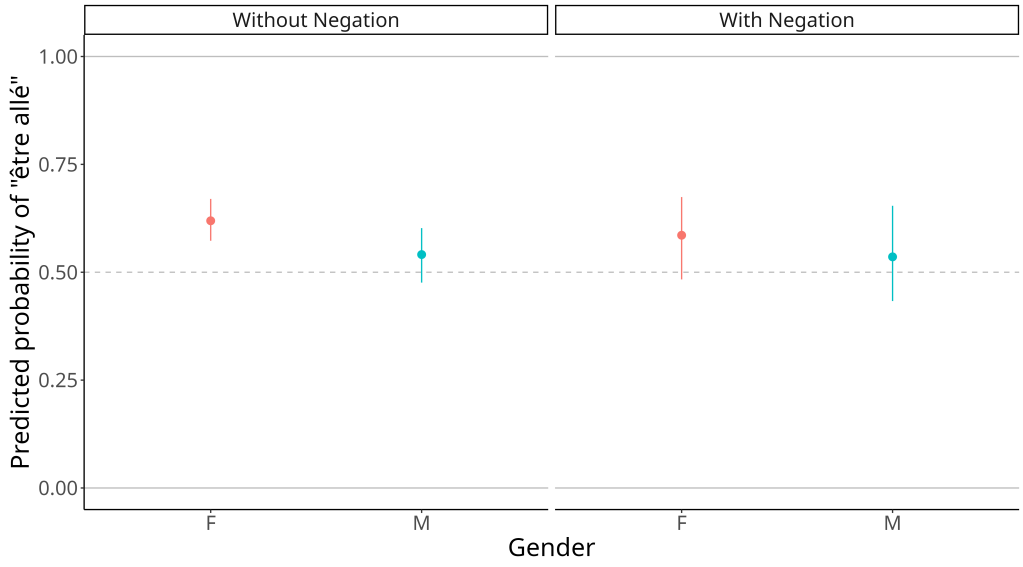

Figure 14: Predictions for “être allé” production depending on GENDER and NEGATION

5.2.3.4 Number of interlocutors

The interaction-level variable NBINTERLOCUTORS is a peculiar one, being both sentence- and speaker- external. A serial investigation of its relationship to all other variables shows that it interacts with SUBJANIMACY (Figure 15): while sentences with animate subjects

follow the gradient overall pattern, sentences with inanimate subjects are associated with roughly similar predicted probabilities whatever the number of interlocutors. The predictive margins are important again, but a Bayesian logistic regression provides evidence for an interaction when contrasting between 'one speaker' and both 'two speakers' ($\hat{\beta} = 1.36$, $P_{(\beta>0)} = .85$) and 'more than two speakers' ($\hat{\beta} = 3.71$, $P_{(\beta>0)} = 1$).

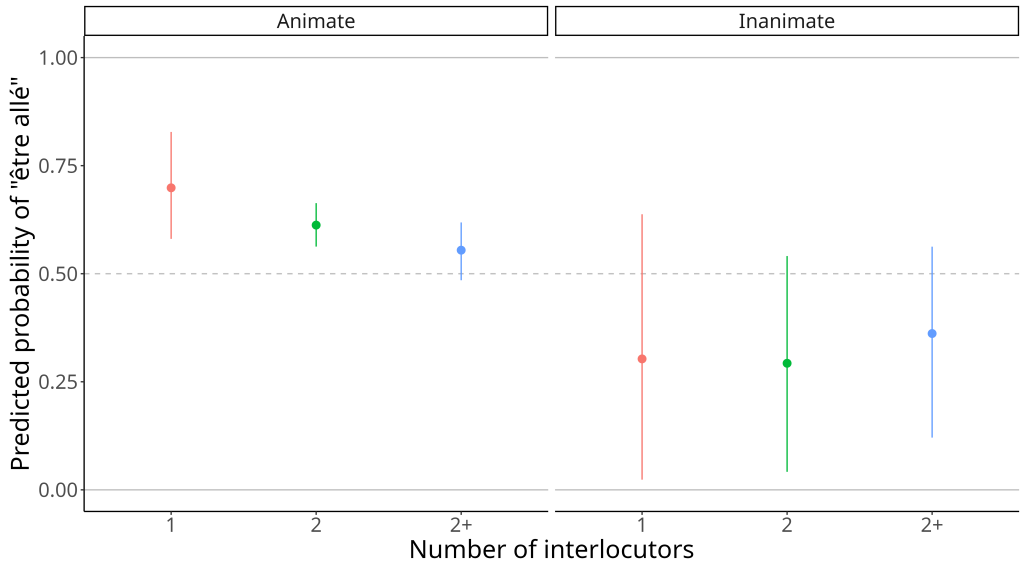

Figure 15: Predictions for "être allé" depending on NBINTERLOCUTORS and SUBJANIMACY

A similar phenomenon is happening with the interaction between NBINTERLOCUTORS and COMPLEMENTTYPE (Figure 16), with an apparent reduction of the differences between number of interlocutors in the predicted probabilities for *être allé* productions when the complement is 'y' and not any other complement (or no complement). A Bayesian logistic regression provides evidence for this, for the contrast between one speaker and more than two speakers ($\hat{\beta} = -5.24$, $P_{(\beta<0)} = .89$).

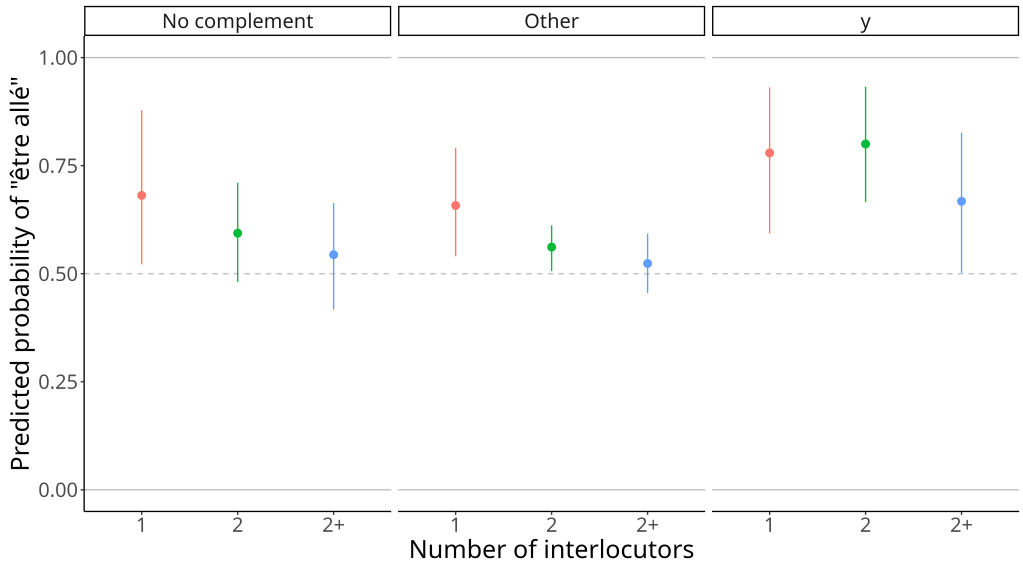

Figure 16: Predictions for "être allé" depending on NBINTERLOCUTORS and COMPLEMENTTYPE

No interaction is found between NEGATION and NBINTERLOCUTORS. However, when switching to speaker-level variables, a clear-cut pattern emerges from the interaction between the latter and GENDER (Figure 17). The gender predictor on its own seemed to have a very binary influence over the production of *être allé* forms with speakers identified as female producing them more (Figure 7). Here, this overall phenomenon is still visible, but the number of speakers has a strong influence over the binary distinctions. The gradient pattern from Figure 9 is only visible for speakers identified as female, while speakers identified as males are not associated with virtually no difference in predicted probabilities for *être allé* productions (respectively 0 .57, 0.54, 0.53 for 1/2/2+ speakers). A Bayesian logistic regression provides evidence for these interactions ($\hat{\beta} = -1.98$, $P_{(\beta<0)} = 1$ for the contrast between 1- and 2-speaker observations when comparing female to male speakers, and $\hat{\beta} = -2.10$, $P_{(\beta<0)} = 1$ for the contrast between 1- and more-than-2-speaker observations).

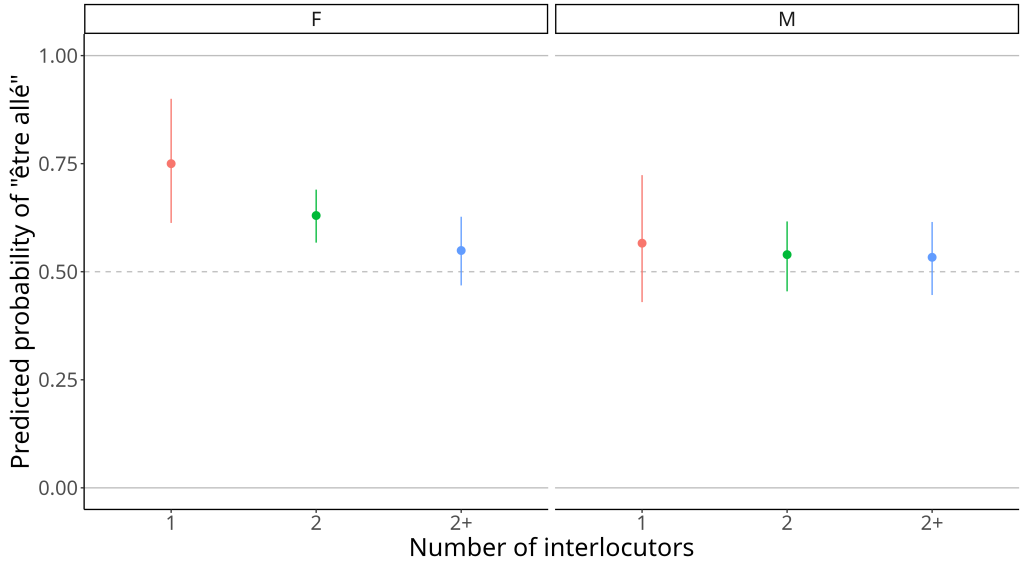

Figure 17: Predictions for "être allé" production depending on NBINTERLOCUTORS and GENDER

The picture is more blurry for the interactions between NBINTERLOCUTORS and both AGE and REGION. For the region variable, Figure 18 illustrates that speakers from Switzerland (rightmost panel) seem relatively more immune to the influence of NBINTERLOCUTORS than speakers from other locations. No easy-to-read pattern emerges from the inferential Bayesian analyses however, with some meaningful interactions (see Supplementary Material) but the high margins there call for a more controlled approach (experimental, for instance) and further work.

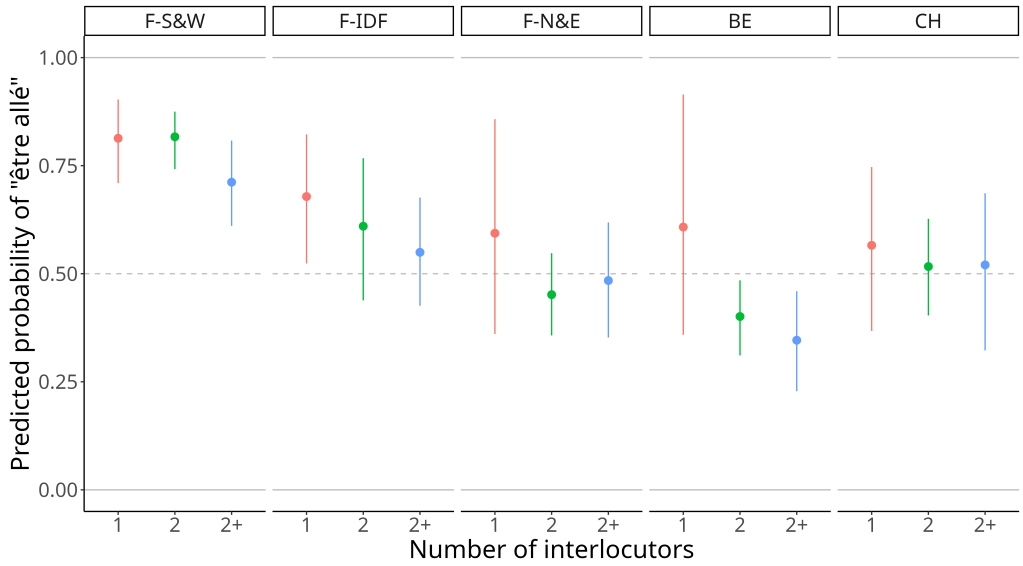

Figure 18: Predictions for "être allé" production depending on NBINTERLOCUTORS and REGION

All in all, these results nonetheless make this predictor a particularly interesting one. On its own, it ranks third on the variable importance continuum (Figure 3), behind REGION and GENDER only. As a variable weighing at the situation level, it interacts meaningfully with both sentence-internal linguistic predictors such as subject animacy AND speaker-level variables such as gender.

6 Discussion

The *conditions* of overabundance that we identified have to do with linguistic properties and extralinguistic properties related to the speaker and the interaction. The latter are the most important, and their effects are summarized below:

- REGION ~ speakers from Belgium and North and Eastern France tend to use less *être allé*, than speakers from other French-speaking regions;
- GENDER ~ women tend to use more *être allé* than men;
- AGE ~ the older speakers (more than 60) tend to use slightly less *être allé* than the younger speakers;
- NBINTERLOCUTORS ~ the more speakers there are in the interaction, the less *être allé* is used.

As for the linguistic properties, the following trends stand out:

- PERSNB and SUBJONNOUS ~ the 2SG and 2PL cells are less likely to combined with *être allé* than other cells, and the pronoun *nous* (1PL) favors *être allé* compared to other types of subject;
- COMPLEMENTTYPE ~ the proform *y* favors the use of ÊTRE ALLÉ compared to other types of complement;
- SUBJANIMACY ~ inanimate subjects are less likely to combine with *être allé* than animate subjects;
- NEGATION ~ the negative polarity of the sentence disfavors the use of *être allé*.

TENSE show a pattern that did not turn out to be statistically meaningful, probably due to the small size of its potential effect, and the lack of observations with tenses other than the periphrastic past (121/1,157).

The importance of geographical variation confirms the trends suggested by the corpus data from Glikman & Patard (2022). It is also consistent with the results from Sammons et al. (2015), who observe statistically significant different distributions between four localities of Ontario. Thus, the broad trends that we document here for European French would probably deserve to be studied in narrower geographical areas, in order to detect more fine-grained distribution differences.

The observed effects of the other two speaker-related factors (GENDER, AGE) do not align entirely with Sammons et al. (2015)'s results. First, it is difficult to compare the observation concerning the age of the speaker with Sammons et al.'s results because they studied a homogeneous age group (adolescents). However, the authors compared their results, which were based on data collected in 2005, with the results of a previous study, based on data collected in 1978 in the same four localities with the same age group (Alexandre 2004). If we compare the productions of the two groups and apply our general observation that the older the people are, the more *avoir été* forms they produce, we might expect that the 1978 group uses more *avoir été* than the 2005 group. This is indeed the case: 69.9% vs. 40.4% of *avoir été* in the 1978 and 2005 groups, respectively. Trends seem to converge in Europe and Canada.

Second, based on their 2005 data, Sammons et al. (2015) indicate that women tend to use the *être allé* variant slightly less than men do (43% vs. 54.8%) and that the gender effect is in flux, given that the opposite trend was observed in data from the 1978 group. According to the authors, this evolution suggests that the *avoir été* forms are not "overly stigmatized" (p. 414). The gender effect in our data appears to be statistically robust and consistent with that found in the 1978 group. However, we were unable to document a potential evolution over time because of the structure of our corpus metadata. As the authors of the Ontario study suggest, and as we will discuss below, if the gender effect is understood, at least partially, with regard to the speaker's relationship to the norm, then we can hypothesize that the difference in direction of the gender effect is due to differences in linguistic norms on both sides of the Atlantic Ocean (Bigot 2021). In other words, the *avoir été* variant may be less stigmatized in Ontario, or more generally in Canada, than in Europe.

The linguistic conditioning of the rivalry between the overabundant forms is less significant than the social conditioning. This is, once more, consistent with previous findings in the literature. Both Glikman & Patard (2022) and Sammons et al. (2015) conclude that the type of postverbal complement (PP or infinitive) does not affect the choice between the two forms of the verb ALLER. Our results go in the same direction: the proportion of *être allé* is roughly the same with both kinds of complement. Nevertheless, since we chose to study the alternation in its entirety (see section 3), we found that certain types of complement affect the verb's form. Specifically, we demonstrate that when the complement is realized as the proform *y*, the *être allé* variant is more likely to occur. In the same vein, expanding the scope of the cases examined revealed that the type of subject influences production choices. *Avoir été* forms are strongly favored with inanimate subjects, which mainly appear in the context of the idiomatic construction *ça va* ('it's okay') and its variants. In the study from Sammons et al. (2015), only two linguistic factors are significant: number (plural favors *être allé*) and polarity (negation favors *avoir été*). Our results confirm the direction of the polarity effect, but with a very low importance in the random forest, compared to the other factors considered. As already mentioned in Section 4, number appears to have a similar effect in our data, though it is not statistically robust. Moreover, rather than considering person and number as two independent dimensions, we treated them as features that define cells of the paradigm. We concluded

that cells with values of 2SG and 2PL favor the *avoir été* forms, whereas 1PL cells more often entail the other variant. Therefore, the lack of number effect in the random forest is crucially related to how we encoded the morphosyntactic features of person and number.

In addition to documenting the conditions of overabundance expressed as factor-by-factor results, we provided evidence for meaningful interactions between linguistic and sociolinguistic properties in two different scenarios. In the first scenario, a linguistic property cancels out the effect of a speaker-related factor. In other words, the linguistic property has the same effect regardless of the speaker's profile. Thus, the inanimate subject *ça* attracts *avoir été* forms regardless of region or gender, while with animate subjects we observe differentiated distributions linked to region and gender. Similarly, the proform *y* uniformly attracts *être allé* variants, as opposed to other types of complement. Similar patterns emerge with the number of interlocutors: the differentiated patterns observed according to this interaction-related factor tend to be flattened by the inanimate subject *ça* or the complement *y*. The second scenario corresponds to cases where a linguistic property affects only a subgroup of speakers. In other words, the effect of the linguistic property depends on the speaker's profile. The polarity effect belongs to this second case scenario: the negative polarity only attracts the *avoir été* forms for women and speakers from Belgium, and Northern and Eastern France. The production of the other subgroups of speakers remains more or less the same regardless of the polarity of the sentence.

In terms of probabilistic grammars, the first scenario corresponds to the common probabilistic grammar shared by French speakers in Europe. Regardless of the speaker's profile, the subject *ça* and the complement *y* trigger a similar probability of using the *être allé* overabundant forms. These two soft constraints are both very "local" in that they concern the verb's morphology or an idiomatic construction based on the verb. In fact, the proform *y* can be analyzed as a verbal prefix (Miller 1992; Miller & Sag 1997; Aguila-Multner 2023), that corresponds to a morphological exponent of the verb ALLER. Additionally, the inanimate subject *ça* is part of the idiom *ça va* ('it's okay'), an ALLER-based construction stored as a unit that can be inflected.

The second scenario resembles other cases of variation documented in probabilistic grammar works, such as those on English varieties (Bresnan & Hay 2008; Bresnan & Ford 2010; Szmrecsanyi et al. 2017). For instance, Bresnan & Hay (2008) examined the dative alternation with the verb *give* in two varieties of English and discovered that the animacy of the recipient exerts a stronger influence in New-Zealand English than in American English. In our case, the attraction of negative polarity for *avoir été* is stronger in the varieties of French spoken in northern and eastern France and Belgium. The fact that only women seem to be sensitive to negative polarity in their use of the overabundant forms would mean that probabilistic grammars are not only geographically modulated but also socially. However, this effect is quite small and must first be confirmed.

As the results of Glikman & Patard (2022)'s questionnaire emphasize, linguistic prescriptivism could influence the choice between the overabundant forms, *avoir été* being informal register and *être allé* being formal. This dimension appears to be helpful for understanding the effects of certain aforementioned variables, as for example the effect of the variable SUBJONNOUS. The morphosyntactic value of the first-person plural can be expressed with the pronouns *nous* and *on*; the former being more formal than the latter. The fact that *nous* attracts the *être allé* forms finds a suitable explanation in the tendency to use words belonging to the same register. Second, the observed trend for GENDER could be interpreted as women tending to stick to the norm-driven use of competing forms. This is consistent with the results of sociolinguistic research showing some

gender-based differences in the use of standard versus non-standard linguistic forms, with 'women' using more standard forms (Trudgill 1972) – or at least being more careful in their use of non-prestigious forms so as to not threaten their (relatively precarious) social position (Eckert 1989). Third, the weight of prescriptivism may help to explain the gradient pattern associated with the number of people in the interaction. In the corpus that we used here, the number of interlocutors is broadly linked to the participants' use of colloquial/informal language: one-speaker interactions are mostly public speaking extracted from the *French Oral Narrative* corpus, two-speaker interactions are often interviews, with two people who do not know each other, and more-than-two-speaker interactions are usually casual or professional conversations between people who know each other. Then, given our results, we could hypothesize that the more casual the interaction, the less prescriptive forms (*être allé*) are used. However, given the interaction between gender and number of interlocutors (see section 5.2.3), our results suggest that women productions are more sensitive to the context of interaction (here the number of interlocutors) than men productions. This is consistent with the work of Wodak & Benke (2017), who offers a more nuanced perspective on the relationship between gender and language use. There, norm-abiding variants are used depending on the situational context of interaction, which the speakers assess differently based on their "*community of practice*" (Eckert & McConnell-Ginet 1992) and on their identity taken as a whole (gender identity being only one dimension of it). If the gradient pattern linked to the participants' level of (in)formality is valid only for women, then our results confirm that women exhibit more norm-driven linguistic behaviors and are more sensitive to the situational context than men.

7 Conclusion

The corpus study presented in this paper documents the rivalry between the *cell-mates* of the periphrastic tenses in the ALLER's paradigm. We describe the statistical distribution of the overabundant forms across the paradigm. The *balance* presented in Table 4 indicates that the overabundance is strongly rooted in the ALLER paradigm, as the rivalry is expressed through all the cells involved. Therefore, overabundance appears to be widespread and no regularizing force leading to its reduction can be detected. We further documented the *conditions* of overabundance, by expanding the scope of the cases examined. This allowed us to discover two new linguistic conditions: inanimate subjects *ça* favor the *avoir été* variant, and the *y* complement the *être allé* variant. Our quantitative study showed that the choice of the form is strongly influenced by speaker- and situation-related variables. There appears to be a geographical distribution and, as discussed in Section 6, the effect of other extralinguistic conditions (gender, number of interlocutors) can be interpreted as the manifestation of prescriptivism-driven behaviors. Linguistic conditions are less likely to affect the probability of one *cell-mate* occurring, but interestingly, their effects are consistent with those found in the literature and interact with speaker-related variables.

We included the conditions of overabundance in a probabilistic grammar perspective: we assumed that the choice of a morphological form is conditioned by multiple constraints, the probabilistic effects of which are part of the grammatical knowledge of speakers. Our results suggest that each speaker's probabilistic grammar is either conditioned by their individual profile (e.g. effect of negation according to the region), or is common to the speaking-community (e.g. effect of the *y* proform).

Finally, it is worth noting that, despite the coarse-grained annotation of extralinguistic properties, we get very specific results, in line with the literature. Nevertheless, a limitation of our work lies in the relatively small number of observations. We had to downsize our sample massively because some metadata (age, region) were lacking (see 3.2). A larger sample size would certainly allow to achieve more robust results. However, big corpus projects are rare and most accessible corpora are still built by research teams working on a specific linguistic phenomenon at first, for a limited period of time. For this reason for instance, we could not take advantage of the diachronic dimension of the CEFC data that we explored. Even though the data span from the 1980s to the 2010s, the REGION variable is highly correlated with the period during which each sub-corpus was constituted. This clearly calls for a more standardized collection of speaker data when creating or making corpora open access. This is necessary to further explore probabilistic grammars.

Abbreviations

ACC = accusative, DAT = dative, DEM = demonstrative, NOM = nominative, PL = plural, SG = singular, PP = prepositional phrase

Data availability/Supplementary files

Our dataset and the script we used for analysis are available at https://osf.io/ahu6m/?view_only=2a3b503bffef4ba1af1fa6cc7b45b870.

Funding information

Will be added in de-anonymized version.

Acknowledgements

Will be added in de-anonymized version.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

Will be added in de-anonymized version.

References

Ackerman, Farrell & Stump, Gregory T. & Webelhuth, Gert. 2011. Lexicalism, periphrasis, and implicative morphology. In Börjars, Kersti & Borsley, Robert D. (eds.), *Non-transformational syntax*, chap. 9, 325–358. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444395037.ch9

Ackerman, Farrell & Webelhuth, Gert. 1998. *A theory of predicates*. Stanford: CSLI.

Aguila-Multner, Gabrielle. 2023. *The morphosyntax of French complex predicates: clitic climbing and periphrasis*: Université Paris Cité dissertation. https://theses.hal.science/tel-04522153/.

Aigro, Mari & Vihman, Virve-Anneli. 2024. Preferences in the use of overabundance: predictors of lexical bias in Estonian. *Cognitive Linguistics* 35(2). 289–312. https://doi.org/doi:10.1515/cog-2023-0035

Alexandre, N. 2004. *Variation in the spoken French of Franco-Ontarians: Preposition 'de' followed by the deictic pro-forms 'ça' and 'la', 'aller' in compound past tenses and prepositions 'à', 'au' and 'en' preceding geographical place names*: York University MA thesis.

André, Virginie & Canut, Emmanuelle. 2010. Mise à disposition de corpus oraux interactifs: le projet TCOF (Traitement de Corpus Oraux en Français). *Pratiques. Linguistique, littérature, didactique* (147-148). 35–51. https://doi.org/10.4000/pratiques.1597

Avanzi, Mathieu & Béguelin, Marie-José & Diémoz, Federica. 2016. De l'archive de parole au corpus de référence: la base de données orales du français de Suisse romande (OFROM). *Corpus* (15). 10.4000/corpus.3060.

Bach, Xavier. 2022. Overlapping suppletion and periphrasis: On have, be, and go in gallo-romance. *Word Structure* 15(2). 115–137. https://doi.org/10.3366/word.2022.0202

Benzitoun, Christophe & Debaisieux, Jeanne-Marie & Deulofeu, Henri-José. 2016. Le projet orfÉo : un corpus d'étude pour le français contemporain. *Corpus* 15. https://doi.org/10.4000/corpus.2936

Benzitoun, Christophe & Etienne, Carole. 2020. Méthodologie d'harmonisation et de traitement des données orales du CÉFC. *Langages* 219. https://doi.org/10.3917/lang.219.0039

Bermel, Neil & Knittl, Luděk & Russell, Jean. 2018. Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. *Corpus Linguistics and Linguistic Theory* 14(2). 197–231. https://doi.org/doi:10.1515/cllt-2016-0032

Bigot, Davy. 2021. *Le bon usage québécois : étude sociolinguistique sur la norme grammaticale du français parlé au Québec*. Québec: Presses de l'Université Laval.

Blampain, Daniel & Hanse, Joseph. 2012. *Dictionnaire des difficultés du français*. De Boeck Sup.

Bonami, Olivier. 2015. Periphrasis as collocation. *Morphology* 25(1). 63–110. https://doi.org/10.1007/s11525-015-9254-3

Bonami, Olivier & Webelhuth, Gert. 2012. The phrase-structural diversity of periphrasis: A lexicalist account. In Chumakina, Marina & Corbett, Greville (eds.), *Periphrasis: The role of syntax and morphology in paradigms*, British Academy. https://doi.org/10.5871/bacad/9780197265253.003.0006

Boyé, Gilles. 2000. *Problèmes de morpho-phonologie verbale en français, en espagnol et en italien*: Université Paris-Diderot – Paris VII dissertation. https://theses.hal.science/tel-00276756.

Branca-Rosoff, Sonia & Fleury, Serge & Lefeuvre, Florence & Pires, Mat. 2000. Discours sur la ville. *Corpus de français parlé parisien des années 2000* http://cfpp2000.univ-paris3.fr/.

Brehm, Laurel & Alday, Phillip M. 2022. Contrast coding choices in a decade of mixed models. *Journal of Memory and Language* 125. 104334. https://doi.org/https://doi.org/10.1016/j.jml.2022.104334

Bresnan, Joan & Cueni, Anna & Nikitina, Tatiana & Baayen., Harald. 2007. Predicting the dative alternation. In Bouma, Gerlof & Kraemer, Irene & Zwarts, Joost (eds.), *Cognitive foundations of interpretation*, Amsterdam: Royal Netherlands Academy of

Science. https://hdl.handle.net/11858/00-001M-0000-0013-1A32-0.
Bresnan, Joan & Ford, Marilyn. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language* 86(1). 168–213. https://doi.org/10.1353/lan.0.0189
Bresnan, Joan & Hay, Jennifer. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 2(1). 245–259. https://doi.org/10.1016/j.lingua.2007.02.007
Bürkner, Paul-Christian et al. 2017. brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software* 80(1). 1–28. https://doi.org/https://doi.org/10.18637/jss.v080.i01
Bérard, Lolita. 2020. La partie orale du Corpus d'Étude pour le Français Contemporain (CÉFC). *Langages* 219. https://doi.org/10.3917/lang.219.0025
Bürkner, Paul-Christian. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10(1). 395–411. https://doi.org/10.32614/RJ-2018-017
Carpenter, Bob & Gelman, Andrew & Hoffman, Matthew D & Lee, Daniel & Goodrich, Ben & Betancourt, Michael & Brubaker, Marcus & Guo, Jiqiang & Li, Peter & Riddell, Allen. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1). https://doi.org/https://doi.org/10.18637/jss.v076.i01
Carruthers, Janice. 2013. French Oral Narrative Corpus. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2555.
Damourette, Jacques & Pichon, Edouard. 1911-1927. *Des mots à la pensée. Essai de grammaire de la langue française.* Paris: D'Artrey.
Debaisieux, Jeanne-Marie & Benzitoun, Christophe. 2020. Orféo : un corpus et une plateforme pour l'étude du français contemporain. *Langages* 219. https://shs.cairn.info/revue-langages-2020-3?lang=fr.
Delic, Équipe. 2004. Présentation du Corpus de Référence du Français Parlé. *Recherches sur le français parlé* 18. 11–42.
Deulofeu, José & Blanche-Benveniste, Claire. 2006. C-Oral-Rom – French Corpus. In Kawaguchi, Yuji & Zaima, Susumu & Takagaki, Toshihiro (eds.), *Spoken language corpus and linguistic informatics*, 181–198. Amsterdam: John Benjamins. https://doi.org/10.1075/ubli.5.14deu
Dister, Anne & Labeau, Emmanuelle. 2017. Le Corpus de Français Parlé à Bruxelles: origines, hypothèses, développements et prédictions. *Cahiers AFLS* 21(1). http://cfpp2000.univ-paris3.fr/cfpb.html.
Eckert, Penelope. 1989. The whole woman: Sex and gender differences in variation. *Language variation and change* 1(3). 245–267. https://doi.org/10.1017/S095439450000017X
Eckert, Penelope & McConnell-Ginet, Sally. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21. 461–490. http://www.jstor.org/stable/2155996.
Francard, Michel & Geron, Geneviève & Wilmet, Régine. 2002. La banque de données VALIBEL: des ressources textuelles orales pour l'étude du français en Wallonie et à Bruxelles. In Pusch, Claus D. & Raible, Wolfgang (eds.), *Korpuslinguistik – korpora und gesprochene sprache / romance corpus linguistics – corpora and spoken language*, 71–80. Tübingen: Gunter Narr.
Frick, Hannah & Chow, Fanny & Kuhn, Max & Mahoney, Michael & Silge, Julia & Wickham, Hadley. 2024. *rsample: General resampling infrastructure*. https://CRAN.R-project.org/package=rsample. R package version 1.2.1.

Glikman, Julie & Patard, Adeline. 2022. Être pour aller en français d'Europe. *SHS Web Conf*. 138. 02001. https://doi.org/10.1051/shsconf/202213802001

Grafmiller, Jason & Sönning, Lukas. 2022. *predictiveMargins: Predictive margins for random forests*. R package version 0.11.0.

Guzmán Naranjo, Matías. 2017. The se-ra alternation in Spanish subjunctive. *Corpus Linguistics and Linguistic Theory* 13(1). 97–134. https://doi.org/doi:10.1515/cllt-2015-0017

Guzmán Naranjo, Matías & Bonami, Olivier. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa: a journal of general linguistics* 6(1). 1–31. https://doi.org/doi:10.5334/gjgl.1626

Hothorn, Torsten & Hornik, Kurt & Zeileis, Achim. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, Torsten & Zeileis, Achim. 2015. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research* 16(118). 3905–3909. http://jmlr.org/papers/v16/hothorn15a.html.

Juge, Matthew L. 1999. On the rise of suppletion in verbal paradigms. In *Proceedings of the annual meeting of the Berkeley Linguistics Society*, vol. 25. 183–194. https://doi.org/10.3765/bls.v25i1.1195

Juge, Matthew L. 2019. The sense that suppletion makes: Towards a semantic typology on diachronic principles. *Transactions of the Philological Society* 117(3). 390–414. https://doi.org/10.1111/1467-968X.12175

Kawaguchi, Yuji. 2011. Corpus TUFS – Tokyo University of Foreign Studies. http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.html?contents_xml=corpus&menulang=en.

Kuhn, Max & Vaughan, Davis & Hvitfeldt, Emil. 2025. *yardstick: Tidy characterizations of model performance*. https://CRAN.R-project.org/package=yardstick. R package version 1.3.2.

Levshina, Natalia. 2020. Conditional inference trees and random forests. In Paquot, Magali & Gries, Stefan Th. (eds.), *A practical handbook of corpus linguistics*, 611–643. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_25

Miller, Philip H. 1992. *Clitics and constituents in phrase structure grammar*. New York: Garland.

Miller, Philip H. & Sag, Ivan A. 1997. French clitic movement without clitics or movement. *Natural Language & Linguistic Theory* 15(3). 573–639. https://doi.org/10.1023/A:1005815413834

Moneglia, Massimo & Martin, Philippe. 2008. The C-Oral-Rom resource. In Cresti, Emanuela & Moneglia, Massimo (eds.), *C-Oral-Rom: Integrated reference corpora for spoken romance languages*, 1–70. John Benjamins Publishing Company. https://doi.org/10.1075/scl.15.03mon

Nasr, Alexis & Dary, Franck & Béchet, Frédéric & Fabre, Benoît. 2020. Annotation syntaxique automatique de la partie orale du ORFÉO. *Langages* 219. https://doi.org/10.3917/lang.219.0087

Posit Team. 2024. *RStudio: Integrated development environment for R*. Posit Software, PBC Boston, MA. http://www.posit.co/.

R Core Team. 2024. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Robin, Xavier & Turck, Natacha & Hainard, Alexandre & Tiberti, Natalia & Lisacek, Frédérique & Sanchez, Jean-Charles & Müller, Markus. 2011. pROC: an open-source

package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12. 77. https://doi.org/10.1186/1471-2105-12-77

Sammons, Olivia N. & Nadasdi, Terry & Mougeon, Raymond. 2015. 'moving' through the past: Thirty years of avoir été in Ontario French. *Journal of French Language Studies* 25(3). 397–422. https://doi.org/10.1017/S0959269514000362

Sorensen, Tanner & Hohenstein, Sven & Vasishth, Shravan. 2016. Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology* 12(3). 175–200. https://doi.org/10.20982/tqmp.12.3.p175

Szmrecsanyi, Benedikt & Grafmiller, Jason & Bresnan, Joan & Rosenbach, Anette & Tagliamonte, Sali & Todd, Simon. 2017. Spoken syntax in a comparative perspective : The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 118. 1–26. https://doi.org/10.5334/gigl.310

Sönning, Lukas & Grafmiller, Jason. 2024. Seeing the wood for the trees: Predictive margins for random forests. *Corpus Linguistics and Linguistic Theory* 20(1). 153–181. https://doi.org/doi:10.1515/cllt-2022-0083

Thornton, Anna M. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In Maiden, Martin & Smith, John Charles & Goldbach, Maria & Hinzelin, Marc-Olivier (eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199589982.003.0017

Thornton, Anna M. 2012. Reduction and maintenance of overabundance: A case study on italian verb paradigms. *Word Structure* 5(2). https://doi.org/10.3366/word.2012.0026

Thornton, Anna M. 2019. Overabundance: A canonical typology. In Rainer, Franz & Gardani, Francesco & Dressler, Wolfgang U. & Luschützky, Hans Christian (eds.), *Competition in inflection and word-formation*, 223–258. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02550-2_9

Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1(2). 179–195. https://doi.org/10.1017/S0047404500000488

Vandeloise, Claude. 2007. Le verbe aller: L'affranchissement du contexte d'énonciation immédiat. *Journal of French Language Studies* 17(3). 343–359. https://doi.org/10.1017/S0959269507003031

Wodak, Ruth & Benke, Gertraud. 2017. Gender as a sociolinguistic variable: New perspectives on variation studies. In Coulmas, Florian (ed.), *The handbook of sociolinguistics*, chap. 8, 127–150. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405166256.ch8

Wright, Marvin N. & Ziegler, Andreas. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1). 1–17. https://doi.org/10.18637/jss.v077.i01