Future constructions in English and Norwegian

A contrastive corpus study

Stefan Hartmann[1] & Olaf Mikkelsen[2,3]
[1]Heinrich Heine University Düsseldorf; [2]University of Poznan; [3]University Paris 8

The choice between the future constructions *will/shall* and *BE going to* is among the most well-investigated topics in English linguistics. A host of semantic, pragmatic, and syntactic factors has been suggested to drive the alternation between these constructions. Recent research has taken a contrastive perspective and investigated whether similar factors also apply to Norwegian, which shows a very similar alternation (*skal/vil* vs. *kommer til å*). This paper follows up on this line of research, taking new data into account. Drawing on the Open American National Corpus (OANC) and the Spoken BNC2014 for English on the one hand and the NoTa corpus as well as the Big Brother corpus for Norwegian, we carve out commonalities and differences between the alternation patterns in English and Norwegian, and we argue that in both languages, it may actually be semantic, rather than structural, aspects that play the most crucial role in language users' choice between competing future constructions.

Keywords: Future constructions; English; Norwegian; alternations

1. Introduction

Alternations present an interesting challenge for linguists, and arguably they also provide a unique window to the cognitive processes at play in everyday language use: Why and how do variants that fulfill roughly the same functions develop, and how do language users choose between them? As Pijpops (2020) has pointed out, the term *alternation* has been defined quite differently in the linguistic literature. In the framework of Construction Grammar (CxG), which focuses on modeling language users' linguistic knowledge and has hence developed a keen interest in individual differences (see e.g. Ungerer & Hartmann, 2023: 49–51), variants that "present a choice point for an individual language user" (Pijpops, 2020: 284) are of particular interest. For example, users of English can choose between different constructions that encode 'transfer', cf. *I gave her the book* vs. *I gave the book to her* (the so-called dative alternation; see e.g. Zehentner 2019). Such alternations often show what Goldberg (e.g. 1995, 2019) calls partial

productivity, i.e. they cannot be freely extended but show more or less systematic constraints that in turn can point to subtle semantic or pragmatic differences. A number of such alternations have been investigated quite thoroughly, especially for English, one of them being the future alternation between *will/shall* and *BE going to* (e.g. Szmrecsanyi, 2003; Denis & Tagliamonte, 2018). A number of syntactic and semantic as well as cognitive factors have been proposed that can potentially account for this variation. Especially when it comes to proposed cognitive factors underlying patterns of variation, the question arises how such claims can be substantiated. This is where a contrastive perspective can offer promising insights: if a putative determinant of variation has a cognitive basis, then, all other things being equal, it can be expected to apply cross-linguistically to languages that show the same alternation pattern (also see Tamminga et al.'s 2016 concept of 'psychophysiological conditioning'). This is of course complicated by the fact that all other things are hardly ever fully equal between two languages (even within one language, it can often be challenging to tease apart the constructions that are involved in an alternation), and that it is very rare that two or more languages show patterns of variation that are sufficiently similar to speak of 'the same' alternation. Despite such challenges, there are some cases in which languages show alternations that are similar enough to make a systematic comparison seem highly promising. The future alternation in English and Norwegian is one such case. Like English, Norwegian has a future construction grammaticalized from modal verbs (*skal/vil*) on the one hand, and a future construction grammaticalized from a verb of movement on the other (*kommer til å*).

The present paper follows up on previous contrastive research on the future alternation in English and Norwegian (Hasselgård, 2015; Mikkelsen & Hartmann, 2022), taking new data into account and extending the scope from the 'complexity principle', which Szmrecsanyi (2003) offered as an explanation for the variation in English, to a more semantically-oriented account that can potentially explain the patterns of variation in both languages. Furthermore, we depart from large parts of the previous literature in operationalizing the alternation in question not as a binary choice, which allows for more fine-grained distinctions between the patterns in question and takes seriously previous findings that have shown that e.g. *going to* and *gonna* differ considerably in their collocational preferences, and arguably also in their semantics (see Berglund 2000 and especially Lorenz 2013). Treating all the different forms as independent constructions allows for more fine-grained hypothesis testing. On the one hand, if the complexity principle is the underlying explanation for the observed variation, we would expect a correlation between syntactically complex environments, such as negation, interrogation or subordination, and the longer form *BE going to*, not the shorter *BE gonna*. On the other hand, if our own account in terms of semantic principles is correct, we can expect to see the intention-based English futures *BE gonna/going to* behave in the same way as the intention-based Norwegian *skal*.

The remainder of the paper is structured as follows: We first give an overview of previous research, focusing on empirical studies that tried to tackle the future alternation patterns in English and Norwegian (Section 2). Then we turn to our own corpus study, emphasizing the

aspects that the current study adds to our previous analysis presented in Mikkelsen & Hartmann (2022) (Section 3). In Section 4, we summarize the main results and discuss potential avenues for future research.

2. Previous research

The English future alternation has been investigated extensively (e.g. Brisard, 1997; Szmrecsanyi, 2003; Hilpert, 2008; Torres Cacoullos, 2009; Bergs, 2010; Denis & Tagliamonte, 2018; Engel & Szmrecsanyi 2022, to mention only some fairly recent work). As with other alternation phenomena, the overarching question is which factors determine whether language users choose one of the alternating constructions, and to what extent the constructions differ in terms of their semantics and/or pragmatics. A number of potentially influential factors have been proposed, including differences related to function – *BE going to* being described as being more associated with intentions and *will* with predictions – and semantics – *BE going to* expressing a higher degree of certainty and temporal proximity than *will*, in addition to being related to predictions deriving from present circumstances, whereas *will* is seen as more contingent on future conditions. According to Szmrecsanyi (2003), the distribution of *going to* vs. *will* can partly be attributed to the complexity principle proposed by Rohdenburg (1996: 151), according to which "in the case of more or less explicit grammatical options, the more explicit one(s) will tend to be favoured in cognitively more complex environments." In his study of *BE going to* vs. *will*, he operationalizes (syntactic) complexity as contexts of interrogation, negation and subordination. As the literature on the English future alternation has repeatedly found strong effects of sentence type, we will adopt these variables in the present study, following Torres Cacoullos & Walker (2009) and Deis & Tagliamonte (2018) in adding interrogatives as a further sentence type. In their study of the future alternation in spoken Québec English, Torres Cacoullos & Walker (2009: 343) find that interrogatives strongly disfavor *will*, potentially for semantic reasons as questions with *going to* seem to "retain a nuance of intention", while "questions with *will* concern desire or willingness" (Torres Cacoullos & Walker 2009: 343). Apart from such semantic differences, other factors that are mentioned in the literature are regional variation – *BE going to* has been shown to be more frequent in North American than in British varieties of English (Denis & Tagliamonte, 2018: 407) – and register variation, with *BE going to* being more strongly associated with informal settings (Szmrecsanyi, 2003: 296). Finally, it has been shown that these factors interact in intricate ways, as semantic factors vary across different registers, particularly when comparing informal conversation and writing (Engel & Szmrecsanyi 2022).

Hilpert (2008: 39–45) has used collostructional analysis (Stefanowitsch & Gries 2003, 2005, Gries & Stefanowitsch 2004) to check which lexemes the different English future constructions combine with. Collostructional analysis, explained in more detail in Section 3.1, uses association measures to check which lexical items combine with a given syntactic construction with above-

or below-chance frequency, or to compare different constructions in terms of the lexical items they combine with. Gries & Stefanowitsch (2004: 114) had already compared *BE going to* and *will* in the British component of the International Corpus of English (ICE-GB), showing that *will* combines with stative verbs like *remain, depend, become,* or *know*, while *BE going to* combines with verbs that are high in agency and transitivity (in the sense of Hopper & Thompson, 1980), such as *do, say, put* or *kill* (Hilpert, 2008: 40). These collocational patterns are compatible with the typically intentional readings of *BE going to* and prediction-based uses of *will*.

The Norwegian future alternation, contrary to its English counterpart, has received relatively little attention thus far (Mac Donald 1982, Vannebo 1985), but a recent corpus study (Mikkelsen & Glynn, forthcoming) shows that *skal* is an intention-based construction associated with near temporal reference in contrast with *vil* and *kommer til å*, which are both used to express predictions. Furthermore, *vil* is specialized in contexts expressing predictions dependent on future contingencies, while *kommer til å* expresses predictions that are higher in certainty and typically exhibits an open temporal reference. Finally, *kommer til å* has been considered a more informal variant than the somewhat stylistically marked *vil*. A collostructional analysis of Norwegian future constructions has yet to be published, but Mikkelsen & Horbowicz (2022) find that even for a small sample of 600 items, 'lexical verb' is the strongest predictor in a three-way choice between *vil, skal* and *kommer til å*. This suggests that the Norwegian constructions may be more lexically restricted than their English counterparts, and this is indeed corroborated by the present study (Section 3.2).

Examples (1-5) below illustrate the above-mentioned semantic and formal differences. English *will* (1) expresses a prediction dependent on a future contingency, similar to its Norwegian cognate *vil* (3). English *BE going to* (2) is used to express an intention that is to be carried out in the near future, much like Norwegian *skal* (4). Finally, Norwegian *kommer til å* (5) expresses a strong prediction about the future.

(1) *if you retire on seventy grand a year they WILL continue to pay you seventy grand a year* (BNC2014, S2C9)
(2) *well we 're GOING TO go to bed now* (BNC2014, SK3S)
(3) *hvis du kommer og sier at du er fra Norge så VIL de foretrekke å snakke engelsk med deg?* (NoTa, 117)
'if you come saying you're from Norway they will prefer to speak English with you?'
(4) *jeg SKAL kjøpe noen rundstykker og greier fordi F2 kommer på lunsj i_morgen* (NoTa, 5) 'I'm going to buy some bread rolls because F2 is coming for lunch tomorrow'
(5) *jeg KOMMER aldri TIL Å få lappen for noen ting* (NoTa, 20)
'I'm never going to get my drivers licence'

These examples illustrate that while English and Norwegian *will*/*vil* share many similar uses, *skal* is more semantically similar to *BE going to*, while *kommer til å* bear some similarity to both *will* and *BE going to* (see also Hasselgård 2015). It should be pointed out that this comparison

only holds for contexts with future time reference, as the two modals *skal* and *vil* are frequently employed with deontic or epistemic meanings (Mikkelsen & Hartmann 2022). Another way of thinking about this difference is to say that the English constructions are more semantically bleached. A systematic comparison between the systems of future reference in English and Norwegian can therefore contribute to a better understanding of each of the individual systems, but also of the ways languages tend to encode future time reference in general.

3. Our study

3.1 Data and methods

We draw on four different corpora, two for English, two for Norwegian. In line with previous studies, we focus on spoken data. For English, we use the SpokenBNC2014 (Love et al. 2017) for British English and the Open American National National Corpus (OANC, Ide et al. 2002); for Norwegian, we use the Norwegian Speech Corpus NoTa and the Big Brother Corpus. BNC2014 contains 1,251 conversations from 672 speakers of British English; the spoken component of OANC contains the well-known Switchboard corpus, which contains 2,307 two-sided telephone conversations, as well as 93 additional transcripts of face-to-face conversations[1]. All corpora were queried for the future constructions in question: the Norwegian ones for *kommer til å, skal,* and *vil*, the English ones for *going to, gonna, shall, will, 'll,* and *won't*. We took random samples of 5,000 attestations from each corpus. After manually deleting false hits, 6,582 hits remained in the Norwegian data, and 8,770 in the English data (see Table 1).
Note that we fully rely on the transcriptions provided in the corpora, the accuracy of which may differ to some extent between or even within corpora. For example, the transcription guidelines of the SpokenBNC2014 state that contractions that are acceptable in Standard English, such as *she'll*, should not be corrected, while non-standard contractions such as *gonna* should only be retained if it is very clear that the speaker used the contracted rather than the full form; transcribers were asked to use the standard form if they were unsure[2]. By contrast, the annotation guidelines of the Switchboard corpus, which makes up a large proportion of the spoken OANC data[3], state that the transcribers should use dictionary forms, rather than imitating pronunciation[4]. Hence, we have to be very careful in interpreting the differences between the more fine-grained sub-variants.

Table 1: Number of hits (true positives) in the four corpora.

[1] See https://www.anc.org/OANC/index.html (last checked 05/02/2024).
[2] See http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf (last checked 20/05/2024).
[3] See https://web.archive.org/web/20240222112935/https://anc.org/data/oanc/contents/ (last checked 02/06/2024).
[4] See https://catalog.ldc.upenn.edu/docs/LDC97S62/swb1_manual.txt (last checked 02/06/2024).

Construction	Corpus	
	NoTa-Oslo	**Big Brother**
kommer til å	175	273
skal	2639	2581
vil	495	419
Sum	**3309**	**3273**
	SpokenBNC2014	**OANC-Spoken**
will	518	865
'll	1900	1402
won't	233	210
shall	113	20
going to	391	1626
gonna	1051	441
Sum	**4206**	**4564**

The data were annotated for a number of variables that have been found to be influential in previous studies (see Mikkelsen & Hartmann, 2022: 20). A sample of the data was coded by both authors, yielding high inter-annotator agreement rates between .8 and 1 (Cohen's kappa; see Mikkelsen & Hartmann, 2022: 23 for details). The variables that we take into account in the present study are the following:

- Does the construction occur in a main clause or a subordinate clause?
- Does the construction occur in an *if*-clause and if so, does it occur in the protasis (*if X*...) or the apodosis (...*then Y*)?
- Does the construction occur in a negated context?
- Does the construction occur in a question?

For English, it has often been observed that *BE going to* is favored in subordinate clauses and *will* in main clauses, with the protasis and apodosis of *if*-clauses behaving differently (Denis & Tagliamonte 2018: 414). Also, some researchers have argued that *BE going to* is favored in

negated contexts and in interrogatives (Denis & Tagliamonte 2018: 413), although it should be mentioned that Denis & Tagliamonte's (2018: 413) study did not find an effect of negation.
A number of other variables have not been taken into account here for practical reasons. For instance, Szmrecsanyi (2006) also includes sentence length (as a proxy to syntactic complexity, as he assumes that the distribution of *will* vs. *BE going to* can partly be explained by the complexity principle, as mentioned in Section 2). However, sentence boundaries are notoriously difficult to operationalize in spoken language corpora, and they are not pre-annotated in some of the corpora that we are using. Also, previous studies have found a strong effect of priming (see e.g. Gries 2016: 117): When one of the variants has been used in the immediately preceding discourse, language users tend to prefer it. Although priming can be assessed semi-automatically, it requires some manual checking. This is why we can only report results for priming for parts of the data (the SpokenBNC2014 data) at the moment.
As in Mikkelsen & Hartmann (2023), we largely followed Szmrecsanyi (2003) and Denis & Tagliamonte (2018) in the operationalization of the variables. As for the distinction between main and subordinate clause, which is probably the most challenging variable to operationalize among those four, we used a simplified version of Szmrecsanyi's (2003) coding protocol, annotating a clause as subordinate if there was an overt subordinating conjunction or, in the case of null complementizers (*I think [that] there could be a complementizer*), if an overt complementizer could be inserted without changing the meaning of the clause.
These annotations were used as input for a conditional inference tree and random forest analysis as well as a Boruta analysis. In the English data, we use the variety (British vs. American English) as an additional predictor, with AmE corresponding to the OANC data and BrE corresponding to the BNC data. In the Boruta analysis, we additionally include the individual lexemes that the constructions combine with as a predictor.

Conditional inference trees and random forests. Since Tagliamonte & Baayen's (2012) pioneering work, conditional inference trees and random forests have become a widely-used tool for the analysis of linguistic variation (see e.g. Levshina, 2020). These techniques are particularly well-suited for what Tagliamonte & Baayen (2012: 161) call "'small *n* large *p*' problems", i.e. situations in which we are dealing with relatively small datasets but at the same time want to take a comparatively large number of variables into account. This is the case here, as we work with samples of 5,000 attestations for each of the corpora under investigation (before excluding false hits, which means that the actual number of datapoints will be lower).
Conditional inference trees use binary partitioning to gauge the importance of predictors in a given dataset. More specifically, the conditional inference tree algorithm tests the null hypothesis that the distribution of a response variable is equal to the conditional distribution of the response variable given a predictor (see Levshina, 2020: 615). To test this hypothesis, the algorithm reshuffles the labels of the response variable, thus breaking the association between response and predictor variable; this allows for computing a statistic that quantifies the difference in the association between response and predictor before and after the permutation (Levshina, 2020:

615). The algorithm makes a split in the variable that best distinguishes between different values in the response variable, thus partitioning the data in two sets (Levshina, 2020: 612). This is repeated until no further splits can be made (Levshina, 2020: 612). The result is a "tree" in which the dataset is repeatedly split along the variables that have been identified as influential.

A random forest consists of many conditional inference trees. The random forest algorithm draws a random sample of candidate predictors for each individual tree, which means that each variable has the chance to appear in different contexts with different covariates, which leads to more reliable results especially in situations of multicollinearity (Levshina, 2020: 616–617). Random forests allow for calculating conditional variable importance scores, which "show how important each variable is, taking into account all others and their interactions" (Levshina, 2020: 617). We use the package *party* (Hothorn et al., 2006; Strobl et al., 2008) for R (R Core Team 2023) for fitting conditional inference trees and random forests to our data, and for calculating variable importance scores.

Boruta analysis. A potential disadvantage of random forests is that they do not provide definitive cut-off points for variables that do not have a meaningful relationship with the dependent variable (Schweinberger, 2023). This is where another extension of random forests can prove helpful, namely Boruta analysis (Kursa & Rudnicki, 2010), which is a variable selection procedure that reports which variables have some non-random relationship with the dependent variable (Schweinberger, 2021: 246). Building on ideas for feature ranking developed by Stoppiglia et al. (2003), the Boruta algorithm aims at finding all relevant variables in a given information system by adding randomness to the system and collecting results from the randomized samples, thus reducing the impact of random fluctuations and correlations (Kursa & Rudnicki, 2010: 3). Following Kursa & Rudnicki (2010: 3–4) and Schweinberger (2023), the inner workings of the Boruta algorithm can be described as follows: In a first step, the algorithm adds copies of all attributes to the dataset and reshuffles the datapoints. These copies are called shadow attributes. The dataset is always extended by at least five shadow attributes, i.e. if there are less than five variables, additional shadow attributes are added by copying one or more of the attributes more than once. Then a random forest classifier is trained on the extended dataset, and the z-scores are computed to assess the relative importance of both the actual predictors and the shadow variables. The algorithm then finds the maximum z-score among the shadow attributes, and checks which of the "real" attributes scored better. Predictors that did not outperform the best shadow variable are deemed 'unimportant' and permanently removed from the model; predictors that did outperform the best shadow variable are deemed 'important'. The procedure is repeated until the algorithm has either reached a previously set number of random forest runs or when the importance is assigned for all attributes. A further advantage of Boruta analysis is that it is fairly robust and computationally efficient, which allows for including variables of the type that would typically be included as random effects in mixed-effects models, which is not possible for tree-based models; these variables typically have many variable levels, which can prove problematic for conditional inference trees and random forests. This is why we included

lexeme as a predictor only in the Boruta analysis but not in the conditional inference tree and random forest analysis.

Multiple distinctive collexeme analysis. To check which lexemes co-occur with which constructions, we use multiple distinctive collexeme analysis (Stefanowitsch & Gries, 2005). Distinctive collexeme analysis is used to compare two or more constructions in terms of their collocational preferences. The method is a subtype of collostructional analysis (see Stefanowitsch, 2013 for an overview), which uses simple association measures such as the p-values of a Fisher-Yates Exact Test or a Chi-squared test to assess which items co-occur at above-chance or below-chance frequency. We use Flach's (2021) R package *collostructions* for our analysis, and we use the log-likelihood ratio G^2 as a measure of collostruction strength.

3.2 Results and discussion

3.2.1 Conditional inference trees and random forests

For our tree-based analyses, we use the different variants of the constructions under investigation as outcome variables. While previous research, including our own (Mikkelsen & Hartmann, 2022), has often lumped together variants of the same construction such as English *will, 'll, won't* or *going to* and *gonna*, it can prove insightful to disentangle these variants, especially in the light of research that shows that these variants often show difference not only in their collocational behavior but also in their semantics (Lorenz, 2013).

Figure 1 below shows the conditional inference tree for the Norwegian data. The first split here occurs in the category of negation, associated with *vil*, followed by the context of interrogation, associated with *skal* (particularly independent interrogative clauses). *Skal* is also the preferred choice in the context of subordination, and is near-categorically used in protasis (*i.e.*, the subordinate clause expressing the condition in a conditional sentence). To the contrary, in main clauses, and particularly apodosis (*i.e.*, the main clause dependent on the condition in a conditional sentence), *skal* is less strongly preferred.

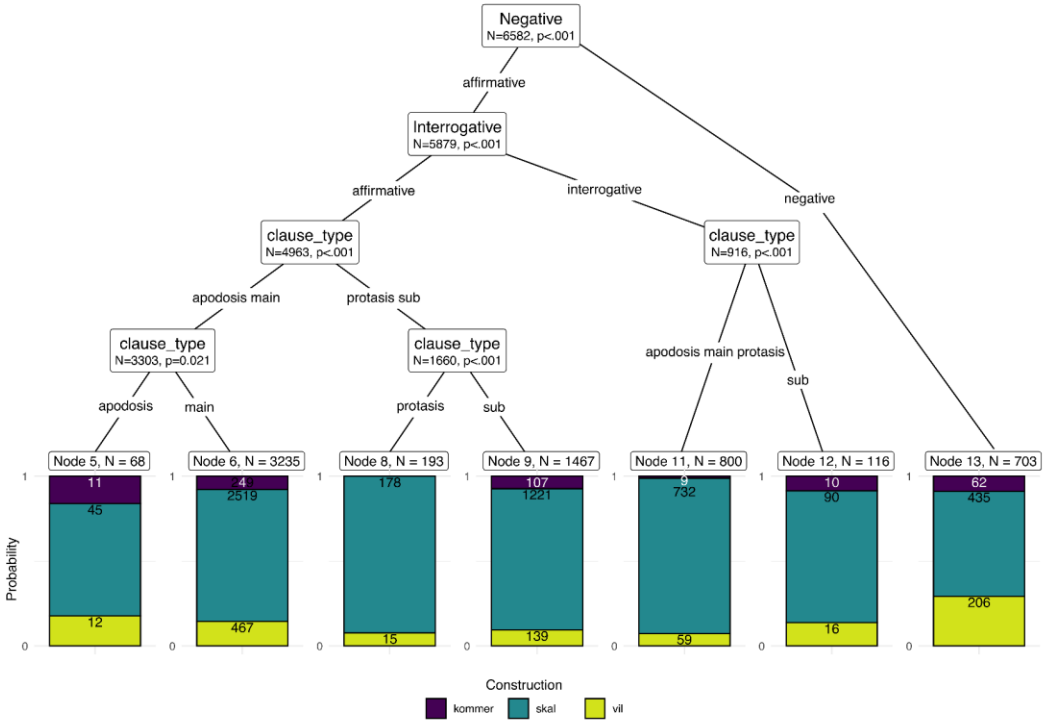

Fig. 1: Conditional inference tree for the Norwegian data.

Moving on to the English data, the conditional inference tree presented in Figure 2 shows that here as well, the first split occurs along the category of negation. Both in British English and in American English, contexts of negation are more associated with *BE going to* and *BE gonna*, with a stronger tendency towards the latter in the British English data. As mentioned in Section 3.1, the higher proportion of *gonna* in the BrE data is probably due to the transcription conventions in the AmE part of the data, which avoided *gonna* in favor of dictionary forms. The second split in the tree corresponds to the varieties themselves, as clause type, according to this particular tree, is more important than interrogation in American English, while the opposite is true for British English. However, the conditional variable importance in Figure 3 indicates that overall – aggregated over many trees –, the factor of clause type is more important than regional variation. When we compare the syntactic environments, we see that *BE going to* and *gonna* are preferred in subordinate clauses and in protases of *if*-clauses, especially so in AmE; in BrE, we find a higher proportion of *'ll* in subordinate clauses as well as in non-interrogative main clauses and apodoses of *if*-clauses. In AmE, the highest proportion of *'ll* is found in non-interrogative main clauses and in apodoses of *if*-clauses as well, whereas in subordinate clauses that do not belong to an *if*-clause, the share of *'ll* is much lower than in BrE. Across both varieties, interrogatives are the main niche of *shall*, which in BrE tends to occur in apodoses of *if*-clauses.

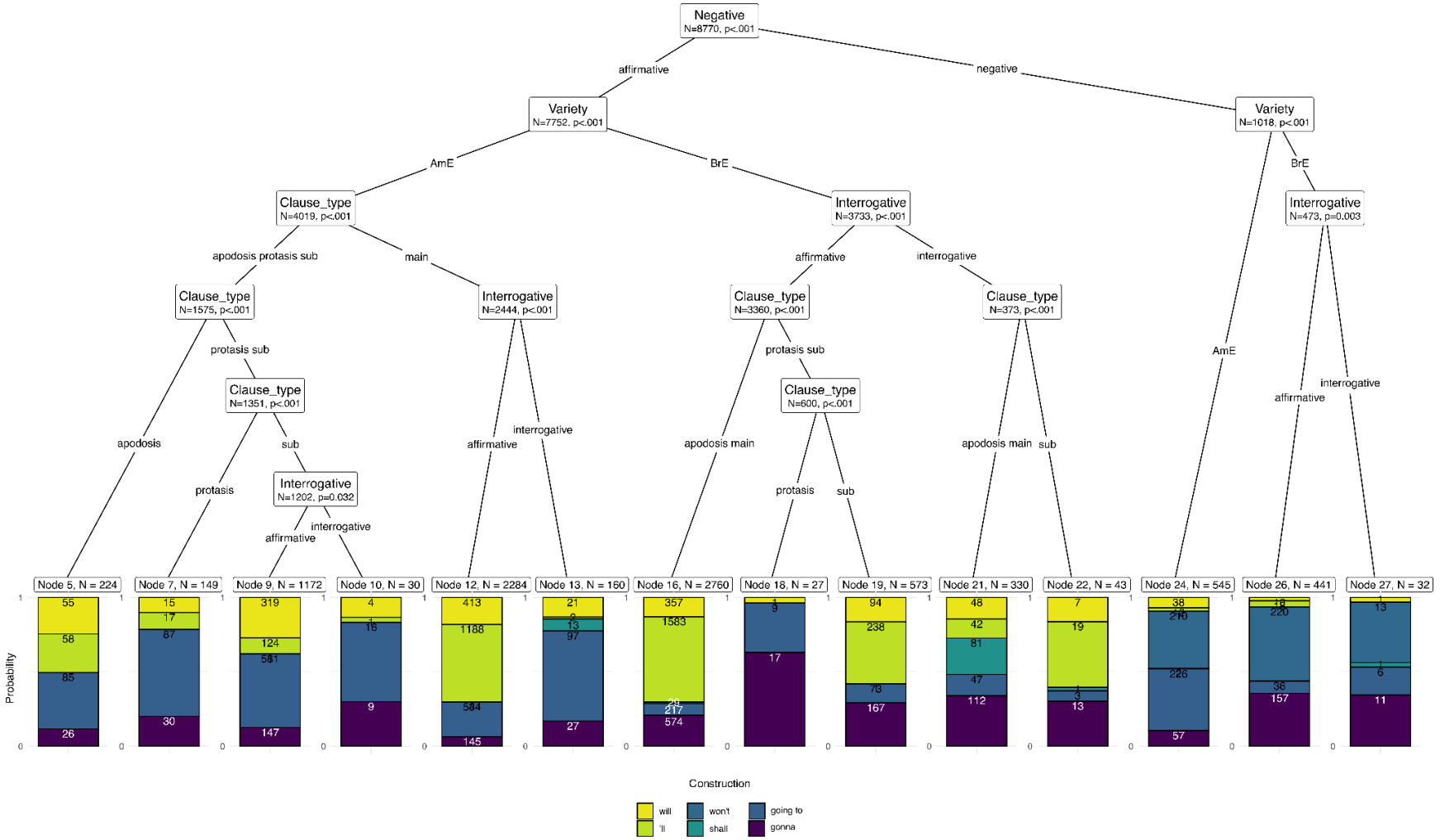

Fig. 2: Conditional inference tree for the English data.

One of the clearest patterns that can be discerned from Fig. 2 is that *BE going to* and *gonna* are strongly preferred in protasis (near categorically so in British English). To a lesser degree, *will/'ll* are preferred in apodosis. In Mikkelsen & Hartmann (2022), we suggested that this was due to semantic factors. We will return to this idea in Section 4. The random forest analysis largely lends support to the tendencies that we observe in the conditional inference trees in Fig. 1 and Fig. 2. For the random forest analysis, we grew 2,000 trees for each of the two languages, with two predictors being randomly selected for each individual tree in the case of the Norwegian data and (because of the higher overall number of predictors) three in the case of the English data. The resulting conditional variable importance scores are shown in Fig. 3. To assess the fit of the model, we measured the accuracy of the random forest by dividing the number of correct predictions by the number of observations (see Levshina, 2020: 632). The Norwegian forest fares much better with an accuracy of 0.79 than the English forest, which only reaches an accuracy value of 0.51. This might partly be due to the higher number of response variable levels in the English data, but it could also indicate differences between the alternations in the two languages due to lexeme-specific factors. This is why we took lexemes into account as an additional predictor in the Boruta analysis.

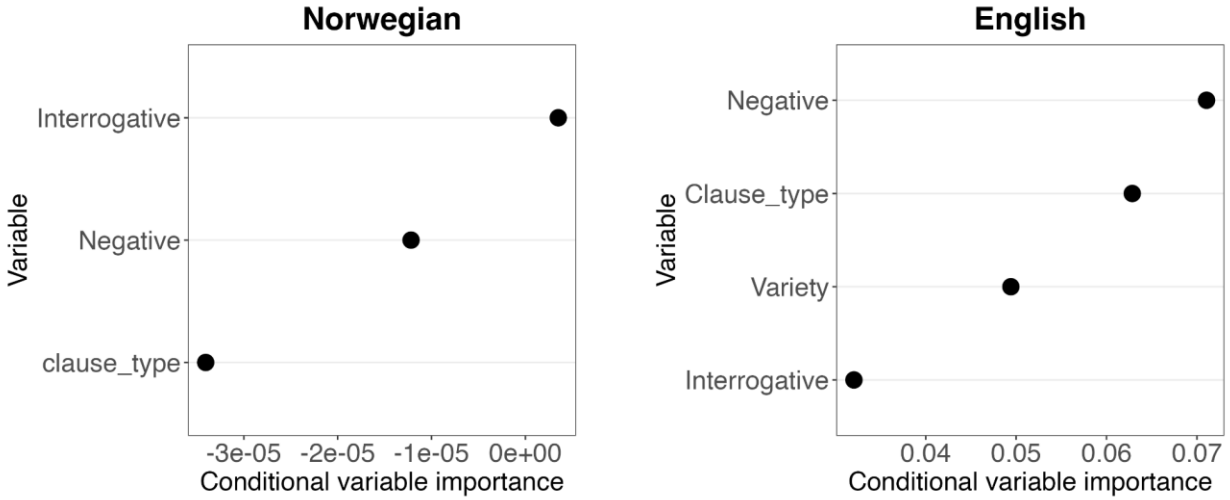

Fig. 3: Conditional variable importance of the variables taken into account in the random forests for the Norwegian (left) and the English data (right).

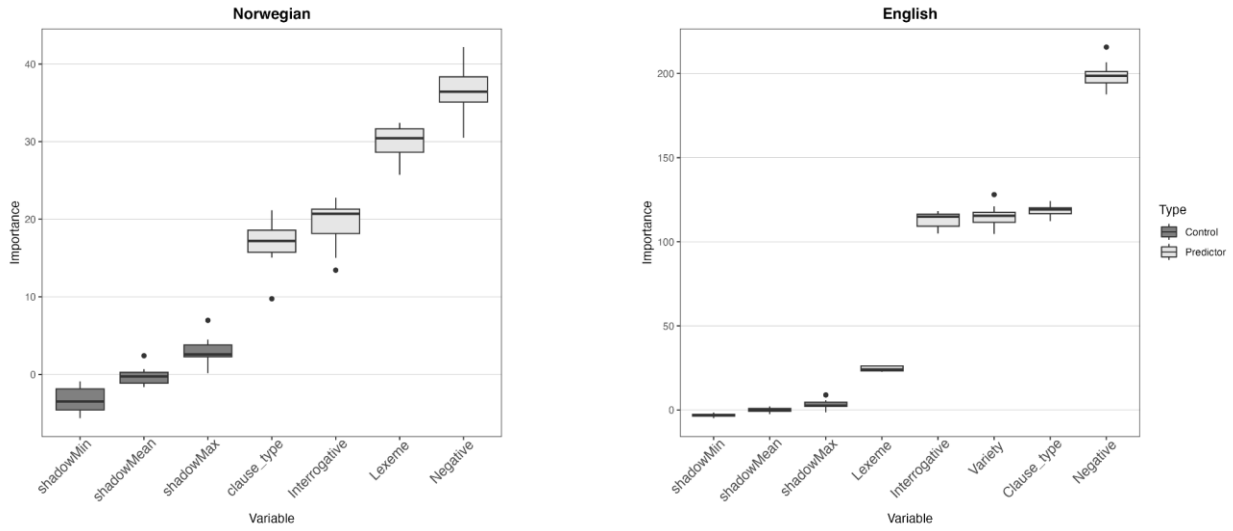

Fig. 4: Results of a Boruta analysis of the Norwegian (left) and English data (right).

When we compare the relative importance of the predictors across both languages, the Boruta analysis (Figure 4) shows that the context of negation is the most important predictor across both languages. We also see that clause type is a more important predictor for English than for Norwegian and that the Norwegian constructions are substantially more lexically specific than their English counterparts. This will be addressed in Section 3.2.2, which reports on the results of the collostructional analysis. As for the results of the random forest analysis and the Boruta analysis in Figures 3 and 4, it should be added that the underlying datasets differ in many respects including their size, which is why a direct comparison between the English and Norwegian results would be problematic – as a reviewer correctly points out, the variable 'Lexeme' obtains a similar score for the English and Norwegian data in Figure 4; however, the importance score scales with the sample size, which is why we cannot compare the raw values of the importance scores across the two datasets.

As mentioned above, we also annotated the SpokenBNC2014 data for the variable "priming": a form was considered primed if a variant of it occurred in the preceding 150 words; in this case, we operationalized *will* and *going to* in a binary way, i.e. *'ll* and *won't* were accepted as primes for *will,* and *gonna* as prime for *going to* (and vice versa). A Boruta analysis of these data (reported on in more detail in the supplementary material) shows that priming emerges as the second most important variable after negation. Future follow-up analyses should double-check whether priming plays a similarly important role in Norwegian.

3.2.2 Collostructional analysis

The top attracted collexemes of the Norwegian constructions in Table 2 show that *skal* is attracted to both verbs denoting situations in their inceptive stage, such as *ta* ('take'), *begynne* ('begin'), *prøve* ('try') and verbs referring to concrete actions such as *lage* ('make'), *prate* ('talk') and *skrive* ('write'). This would be compatible with the intentional readings of the construction. *Kommer til å,* on the other hand, prefers verbs that denote a change, such as the inchoatives *bli* ('become'), *skje* ('happen') and *få* ('get'), or verbs expressing feelings such as *savne* ('miss'), *angre* ('regret') and *elske* ('love'). These verbs are all used to express predictions. The collexemes of *vil* reveal a set of idiosyncratic chunks with highly specialized discursive functions that might be considered constructions in their own right. These include the elucidation construction *det vil si* ('that is to say'), the hedging constructions *jeg vil si*/*påstå* ('I would say') or the speech act construction *jeg vil anbefale* ('I would like to recommend'). Finally, the observed attraction between *vil* and *ha* seems to correspond to the deontic use of *vil ha* ('want').

Table 2: Results of a multiple distinctive collexeme analysis showing the top 10 attracted collexemes for *vil, skal,* and *kommer til å,* respectively. Significance thresholds: ***** = p < .00001, **** = p < .0001, *** p < .001, ** p < .01.

Auxiliary	Collexeme	observed freq.	expected freq.	G^2	Significance
vil	si	164	65.3	144.875	*****
vil	tro	30	5.1	85.4514	*****
vil	ha	193	106.2	80.0677	*****
vil	påstå	5	0.8	14.657	***
vil	innrømme	3	0.4	11.8541	***
vil	stokke	3	0.4	11.8541	***
vil	satse	7	1.8	11.5233	***
vil	anbefale	4	0.7	11.103	***
vil	snuble	4	0.7	11.103	***
vil	beholde	2	0.3	7.90083	**
skal	ta	208	178.4	30.3419	*****
skal	lage	61	48.4	28.4332	*****
skal	begynne	82	69	15.6624	****
skal	prøve	71	59.5	14.4162	***
skal	prate	31	24.6	14.4124	***
skal	skrive	31	24.6	14.4124	***
skal	stå	69	57.9	13.7024	***
skal	spille	43	34.9	13.6048	***
skal	gjøre	288	265.7	10.4223	**
skal	slutte	22	17.4	10.2203	**
kommer	bli	74	16.5	132.006	*****
kommer	merke	5	0.3	26.9252	*****
kommer	skje	10	1.6	23.2878	*****
kommer	savne	4	0.3	21.5318	*****
kommer	angre	3	0.2	16.1426	****
kommer	elske	3	0.2	16.1426	****
kommer	få	32	15.7	15.0172	***

kommer	klare	7	1.4	13.6338	***
kommer	vinne	7	1.4	12.9372	***
kommer	gråte	2	0.1	10.7575	**

While measures of association such as G^2 can only be interpreted relatively and not in absolute terms, it is immediately obvious that the association scores are lower in both variants of English, when compared to the Norwegian data. Again, as we saw in the preceding section, this points to a higher degree of lexical specificity in the case of the Norwegian constructions and/or semantic bleaching in the case of the English ones.

For British English (Table 3), we see that, other than the strong association between *will* and *be*, and *'ll* and *have*, the most strongly attracted collexemes reveal some highly specific recurrent combinations such as *that's not gonna happen*, *we'll see*, *I'll tell you*, *they won't know*, *what are you gonna do?*, *shall I put this on?* and *I'm not gonna lie*. For American English, among the similarly associated patterns we find *I'll bet*, *it's going to cost*, *I will admit*, *they won't let them*, and *shall we say?* Outside of these combinations, it is difficult to discern any particular groups of verbs that can inform us more about the semantics of each construction. Again, it seems that the English constructions are less restricted than their Norwegian counterparts in what concerns the type of verbs that can occur in them.

Table 3: Results of a multiple distinctive collexeme analysis showing the top 10 attracted collexemes for *going to, gonna, will, won't, shall*, respectively, in the SpokenBNC2014 corpus. Significance thresholds: ***** = p < .00001, **** = p < .0001, *** p < .001, ** p < .01, * p < .05.

Auxiliary	Collexeme	observed freq.	expected freq.	G^2	Significance
going to	sit	6	1.2	11.99307	***
going to	write	4	0.7	10.06096	**
going to	guess	2	0.3	5.88667	*
going to	eat	6	2.1	5.46419	*
going to	aid	1	0.1	4.75344	*
going to	arrange	1	0.1	4.75344	*
going to	bite	1	0.1	4.75344	*
going to	cap	1	0.1	4.75344	*
going to	care	1	0.1	4.75344	*
going to	chuck	1	0.1	4.75344	*
gonna	happen	25	11	20.34037	*****
gonna	do	100	70.5	16.46342	****
gonna	lie	4	1	11.1056	***
gonna	listen	3	0.7	8.32705	**
gonna	keep	14	7	8.1267	**
gonna	win	4	1.2	6.6748	**
gonna	make	23	14.7	5.67766	*

gonna	grow	2	0.5	5.54994	*
gonna	paint	2	0.5	5.54994	*
gonna	teach	2	0.5	5.54994	*
will	be	163	116.9	25.05678	*****
will	choose	2	0.2	8.38395	**
will	help	6	2.1	5.98228	*
will	like	4	1.1	5.72178	*
will	cancel	2	0.4	4.8268	*
will	guarantee	2	0.4	4.8268	*
will	mean	2	0.4	4.8268	*
will	apologise	1	0.1	4.19028	*
will	approach	1	0.1	4.19028	*
will	arrive	1	0.1	4.19028	*
'll	have	232	187.5	21.29	*****
'll	tell	31	17.6	19.48	****
'll	find	21	11.3	16.29	****
'll	sort	10	4.5	15.92	****
'll	see	49	33.9	12.55	***
'll	pick	13	7.2	8.86	**
'll	ring	4	1.8	6.36	*
'll	book	3	1.4	4.77	*
'll	explain	3	1.4	4.77	*
'll	push	3	1.4	4.77	*
won't	know	12	2.4	22.48298	*****
won't	be	75	52.6	12.08005	***
won't	download	2	0.1	11.58918	***
won't	bother	3	0.3	10.89107	***
won't	affect	1	0.1	5.79052	*
won't	beep	1	0.1	5.79052	*
won't	bind	1	0.1	5.79052	*
won't	budge	1	0.1	5.79052	*
won't	calibrate	1	0.1	5.79052	*
won't	compromise	1	0.1	5.79052	*
shall	put	10	2	17.81925	****
shall	google	1	0	7.2424	**
shall	overrule	1	0	7.2424	**
shall	photocopy	1	0	7.2424	**
shall	repick	1	0	7.2424	**
shall	return	1	0	7.2424	**
shall	roll	1	0	7.2424	**
shall	stir	1	0	7.2424	**
shall	whistle	1	0	7.2424	**
shall	call	2	0.3	4.92267	*

Table 4: Results of a multiple distinctive collexeme analysis showing the top 10 attracted collexemes for *going to, gonna, will, won't, shall*, respectively, in the spoken part of the OANC corpus. Significance thresholds: ***** = p < .00001, **** = p < .0001, *** p < .001, ** p < .01, * p < .05, ns = not significant.

Auxiliary	Collexeme	observed freq.	expected freq.	G^2	Significance
going to	cost	18	7.8	19.91562	*****
going to	do	125	93.3	17.06768	****
going to	be	378	327.4	14.97858	***
going to	run	15	7.1	12.93314	***
going to	lose	10	4.6	9.26667	**
going to	beat	4	1.4	8.26295	**
going to	happen	31	21	7.11282	**
going to	try	22	13.9	7.02803	**
going to	produce	3	1.1	6.19602	*
going to	rain	3	1.1	6.19602	*
gonna	hope	3	0.3	14.03994	***
gonna	does	2	0.2	9.35585	**
gonna	has	2	0.2	9.35585	**
gonna	have	56	38.5	8.73254	**
gonna	reach	3	0.5	7.71376	**
gonna	hurt	4	0.9	7.37085	**
gonna	think	5	1.5	5.76006	*
gonna	did	2	0.3	5.73915	*
gonna	achieve	1	0.1	4.67587	*
gonna	approve	1	0.1	4.67587	*
will	come	32	16.9	14.47434	***
will	admit	4	0.8	13.32082	***
will	continue	7	1.9	12.366	***
will	last	6	1.5	11.83065	***
will	increase	3	0.6	9.9878	**
will	pop	3	0.6	9.9878	**
will	fit	4	0.9	8.73537	**
will	grow	6	1.9	8.20368	**
will	consider	2	0.4	6.65665	**
will	evolve	2	0.4	6.65665	**
'll	tell	49	22.1	43.101	*****
'll	bet	20	6.5	40.0955	*****
'll	read	11	4.9	9.80006	**
'll	flip	4	1.2	9.45031	**
'll	go	85	64.5	9.39802	**
'll	enjoy	7	2.8	8.47598	**

'll	see	33	22.1	7.33681	**
'll	meet	5	1.8	7.14053	**
'll	remember	5	1.8	7.14053	**
'll	hold	3	0.9	7.08625	**
won't	let	11	2.4	17.93353	****
won't	notice	3	0.2	14.1084	***
won't	bother	3	0.2	11.96989	***
won't	know	7	2.2	7.43035	**
won't	appoint	1	0	6.16224	*
won't	develop	1	0	6.16224	*
won't	erase	1	0	6.16224	*
won't	firm	1	0	6.16224	*
won't	left	1	0	6.16224	*
won't	lift	1	0	6.16224	*
shall	say	12	0.4	70.38724	*****
shall	push	1	0	5.94045	*
shall	kill	1	0	4.69892	*
shall	wait	1	0	4.29263	*
shall	see	2	0.3	4.20682	*
shall	tell	1	0.3	0.96888	ns
shall	be	2	4	1.50733	ns

When we look at the tree-based analyses and the collostructional analyses together, we see that there are syntactic idiosyncrasies both across languages and across different varieties of the same language, and that these seem to be associated with different lexical preferences, at least to some degree. Starting with the observation that the most important predictor in both English and Norwegian is the context of negation, we can observe that it is associated with different constructions in all three languages/varieties compared here: *BE going to* in American English, *BE gonna* in British English and *vil* in Norwegian. When we examine this context more closely – keeping in mind that the data for the alternation between *going to* and *gonna* are unreliable in large parts of the AmE data –, we find many occurrences of strong negative predictions used with specific verbs in both American (6) and British (7) English, whereas the deontic sense of *vil ha* ('want') is frequent in the Norwegian data (8).

(6) *soon they're not GOING TO be able to handle you know all of the crime* (OANC, sw2531)
(7) *that's not GONNA happen* (BNC2014, S0084)
(8) *jeg VIL ikke ha deg mer du er gammel* (NoTa, 35)
'I don't want you anymore, you're old'

The intersection of semantics and syntax may also prove to be a fruitful way to interpret the results. In Section 1, we saw that, semantically, English *BE going to*/*gonna* is similar to

Norwegian *skal*. Both constructions are preferred in the context of interrogation, but while in English it typically expresses inquiries about intentions (9) or state-of-affairs (10); in Norwegian it is most commonly used to make offers (11), similarly to English *shall* (12).

(9) *what are you GONNA do?* (BNC2014, S0556)
(10) *when are we GONNA get paid?* (BNC2014, S0440)
(11) *SKAL vi grille i dag?* (BB, Lars_Joakim)
'Shall we have a barbecue today?'
(12) *SHALL we go for a coffee?* (S0607)

Having separated the full and reduced forms as separate constructions (or 'allostructions' in the sense of Cappelle 2006) may allow us to draw some conclusions that almost necessarily remained undiscovered as long as these constructions were treated together. When we examine the observed preference for *BE going to*/*gonna* in subordinate clauses, we can see that this is related to the strong preference for *'ll* in main clauses. For phonological reasons, *'ll* is nearly exclusively used with personal pronouns (in particular *I'll*), and we might thus hypothesize that the prototypical context of *'ll* is that of intentions within the matrix of a simple main clause (13). The collocational preferences of *'ll* in both British and American English seem to lend some support to this, when compared with the verbs attracted to *will*, which is often used with an expletive pronoun to express a prediction (14).

(13) *I'll do it tomorrow* (BNC2014, S0255)
(14) *I think it WILL probably have a place in a classroom* (BNC2014, S0008)

Overall, then, these results provide evidence against both a purely processing-based account, in which the length of the future marker plays a crucial role, and for an account based on shared semantic factors across different languages and varieties. First, while American English prefers the longer form *BE going to* in contexts of negation, interrogation and subordination, British English favors the shorter *BE gonna* in the same contexts. Second, cross-linguistic semantic features such as intentionality are not enough to explain the variation alone, as they can be overridden by specialized uses of specific constructions, such as the retention of willingness in *vil* (especially in negated contexts this form is closer to English 'want' than to 'will') or suggestion in *skal* (which in interrogative contexts is closer to 'shall' than to 'going to'). Instead, our results suggest that one cannot assume that semantically similar forms should behave in the same way across related languages or even dialects. Indeed, the collocational preferences show that each form (including the contracted ones) has the potential to develop contextual specializations (such as the intentional uses of *'ll* in main clauses) that may in turn affect paradigmatic speaker choice.

4. Conclusion and outlook

Following up on previous work comparing future constructions in English and Norwegian (Hasselgård 2015, Mikkelsen & Hartmann 2022), the present paper has set out to take a closer look at the factors driving the variation between (variants of) *going to* and *will/shall* in English on the one hand, and *skal/vil* and *kommer til å* in Norwegian on the other using a variety of quantitative methods. While we have analyzed the data separately for each of the two languages, the results partly confirm the intuitive impression that the use of the alternants in both languages follows similar patterns. Factors like negation, interrogation, and clause type play a significant role in language users' choice between the different variants in both languages. A number of explanations have been offered for this finding in the previous literature. For instance, Szmrecsanyi (2003) invokes the complexity principle, according to which less complex variants are preferred in cognitively more complex environments (operationalized in terms of syntactic complexity). However, the results of Mikkelsen & Hartmann (2022) indicate that this account seems to work better for the English data than for the Norwegian ones. As an alternative approach, we argue that the influence of these contextual factors can partly be explained semantically. For example, *BE going to* is sometimes interpreted as more subjective in the sense that it relates to the agent's willingness of doing something, whereas *will* is more strongly associated with prediction-based future (see Mikkelsen & Hartmann 2022: 32 and the references cited therein). Especially in casual conversation, we can expect negation to refer to a lack of intention (*I'm not going to do this*) more often than to the negation of a prediction (*Trump's not going to win the election this year*), although we have to leave an empirical validation of this assumption to follow-up studies. Similarly, it can be expected that interrogative sentences typically relate to intentions (*Are you going to come to Mannheim next week?*) rather than to predictions (*Will Trump win the election this year?*). Finally, when it comes to *if*-clauses, the observed distribution is coherent with the hypothesized semantic difference between the constructions: protasis is used with the intention-based future constructions *skal* (Norwegian) and *BE going to* (English), apodosis with prediction-based and contingent future constructions *vil*/*will*. The remaining contexts differ in language-specific preferences where no overarching semantic or other (social, processing) factors are immediately discernible. The different collocational patterns observed, not only across different languages, but equally across regional varieties, suggest that these idiosyncrasies are due to different contextual specializations related to particular lexical elements that have developed over time. It should also be stressed, however, that the account proposed here does not rule out the possibility that the factors captured by the complexity principle also play a role in the selection of constructions encoding future reference. This also shows that there is still a lot of potential for further exploring the commonalities and differences between the future alternation patterns in English and Norwegian. For one thing, the relevance of intention- vs. prediction-based semantics could be assessed more systematically by using animacy and grammatical person as proxies for intention and volition. For another, the impact of fixed phrasal patterns could be studied in more detail by going beyond the lexeme-

based perspective that collostructional analysis offers, and taking the broader context into account, e.g. by using n-gram-based pattern detection methods that allow for identifying recurrent multi-word units. Also, the analysis should be extended to datasets that offer a more reliable transcription of the different constructional variants than large parts of the AmE data analyzed here (see Section 3.1). These steps could help us to discern to what extent simple usage factors may offer a more adequate account of the distribution of the two constructions than the semantic ones discussed here. Almost certainly, both play a key role and are closely intertwined, but more research is required to discern the relative importance of both types of factors in more detail, and to assess whether and to what extent the variation between similar future constructions is more semantically-driven in some languages than in others.

Acknowledgments

We are grateful to two anonymous reviewers for helpful feedback on this paper, as well as to Benedikt Szmrecsanyi for sending us the codebook that was used for annotating the data underlying his 2003 paper. We are of course solely responsible for any remaining errors and shortcomings.

Data availablility

The datasets and scripts used for the present study can be found at https://osf.io/nt32a/.

References

Berglund, Y. (2000). Gonna and going to in the spoken component of the British National Corpus. In Christian Mair & Marianne Hundt (eds.), Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999, 35–50. Amsterdam: Rodopi. https://doi.org/10.1163/9789004490758.

Bergs, A. (2010). Expressions of futurity in contemporary English: A Construction Grammar perspective. English Language & Linguistics 14(2): 217–238. https://doi.org/10.1017/S1360674310000067

Brisard, F. (1997). The English tense-system as an epistemic category: the case of futurity. In Verspoor, M.H. & Sweetser, E. (Eds.) Lexical and syntactical constructions and the construction of meaning (pp. 271–285).

Cappelle, B. (2006). Particle placement and the case for “allostructions.” Constructions. Constructions. https://doi.org/10.24338/CONS-381.

Denis, D., & Tagliamonte, S. A. (2018). The changing future: Competition, specialization and reorganization in the contemporary English future temporal reference system. English Language and Linguistics, 22(3), 403–430. https://doi.org/10.1017/S1360674316000551

Engel, A. & B. Szmrecsanyi. (2022). Variable grammars are variable across registers: future temporal reference in English. Language Variation and Change, 34, 355–378. doi:10.1017/S0954394522000163

Flach, S. (2021). *collostructions*: An R Implementation for the Family of Collostructional Methods. Retrieved from www.sfla.ch/collostructions (last checked 23/02/2024)

Gries, S. Th. (2016). Variationist analysis. In Paul Baker & Jesse Egbert (eds.), Triangulating methodological approaches in corpus-linguistic research, 108–123. New York: Routledge.

Gries, S. Th., & Stefanowitsch, A. (2004). Extending Collostructional Analysis: A Corpus-Based Perspective on "Alternations." International Journal of Corpus Linguistics, 9(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Hasselgård, H. (2015). Coming and going to the future: Future-referring expressions in English and Norwegian. In S. O. Ebeling & H. Hasselgård (Eds.), Cross-linguistic perspectives on verb constructions (pp. 88–115). Newcastle upon Tyne: Cambridge Scholars Publishing.

Hilpert, M. (2008). Germanic Future Constructions: A Usage-Based Approach to Language Change. Amsterdam, Philadelphia: John Benjamins.

Hopper, P. J., & Thompson, S. A. (1980). Transitivity in Grammar and Discourse. Language, 56(2), 251–299.

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

Ide, N., Reppen, R., & Suderman, K. (2002). The American national corpus: More than the web can provide. In M. González Rodríguez & C. P. Suarez Araujo (Eds.), Proceedings of the third international conference on language resources and evaluation (LREC'02). Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2002/pdf/303.pdf

Kursa, M. B. & Rudnicki, W.R. (2010). Feature selection with the Boruta package. Journal of Statistical Software 36(11). 1–13. https://doi.org/10.18637/jss.v036.i11.

Levshina, N. (2020). Conditional Inference Trees and Random Forests. In M. Paquot & S. Th. Gries (Eds.), A Practical Handbook of Corpus Linguistics (pp. 611–643). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_25

Lorenz, D. (2013). On-Going Change in English Modality: Emancipation Through Frequency. Zeitschrift Für Literaturwissenschaft und Linguistik, 43(1), 33–48. https://doi.org/10.1007/BF03379871

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). "The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations." International Journal of Corpus Linguistics 22(3):319–44. doi: 10.1075/ijcl.22.3.02lov.

Mac Donald, K. (1982). Uttrykk for ramtid i norsk. Norskrift 39: 74–87.

Mikkelsen, O. & D. Glynn. (forthcoming). The future that may still be: the spread of *blir å INF* in contemporary Norwegian. In Hartmann, S. & L. Schnee (Eds). *Futures of the past*. Berlin: Language Science Press.

Mikkelsen, O. & Horbowicz, P. (2022): Modelling Semantics in constructional near-synonymy: A usage-based perspective on Norwgian future constructions. Presentation at the conference "Constructions in the Nordics" (CxgN3), Kiel, Germany, September, 2022.

Pijpops, D. (2020). What is an alternation? Six answers. Belgian Journal of Linguistics, 34, 283–294. https://doi.org/10.1075/bjl.00052.pij

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, 7(2), 149–182.

Schweinberger, M. (2023). Tree-based models in R. Brisbane: University of Queensland. https://slcladal.github.io/tree.html. (22 February, 2024).

Schweinberger, M. (2021). On the waning of forms – A corpus-based analysis of decline and loss in adjective amplification. In S. Kranich & T. Breban (Eds.), Lost in change. Causes and processes in the loss of grammatical elements and constructions (pp. 235–260). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/slcs.218.08sch

Stefanowitsch, A. (2013). Collostructional Analysis. In T. Hoffmann & G. Trousdale (Eds.), The Oxford Handbook of Construction Grammar (pp. 290–306). Oxford: Oxford University Press.

Stefanowitsch, A., & Gries, S. T. (2005). Covarying Collexemes. Corpus Linguistics and Linguistic Theory, 1(1), 1–43.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the Interaction of Words and Constructions. International Journal of Corpus Linguistics, 8(2), 209–243.

Stoppiglia, H., Dreyfus, G., Dubois, R., & Oussar, Y. (2003). Ranking a Random Feature For Variable And Feature Selection. Journal of Machine Learning Research, 3, 1399–1414. https://doi.org/10.1162/153244303322753733

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25). Retrieved from http://www.biomedcentral.com/1471-2105/8/25

Szmrecsanyi, B. (2003). Be going to versus will/shall: Does syntax matter? Journal of English Linguistics, 31(4), 295–323.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. Language Variation and Change, 24(02), 135–178. https://doi.org/10.1017/S0954394512000129

Tamminga, Meredith, Laurel MacKenzie & David Embick. 2016. The dynamics of variation in individuals. Linguistic Variation 16(2). 300–336. https://doi.org/10.1075/lv.16.2.06tam.

Torres Cacoullos, R., & Walker, J. A. (2009). The Present of the English Future: Grammatical Variation and Collocations in Discourse. Language, 85(2), 321–354. JSTOR. Retrieved from JSTOR.

Ungerer, T., & Hartmann, S. (2023). Constructionist approaches: Past, present, future. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009308717

Vannebo, K. I. (1985). Tempussystemet i norsk. Norskrift 46: 1–60.

Zehentner, E. (2019). Competition in Language Change: The Rise of the English Dative Alternation. Berlin, Boston: De Gruyter.