

FUTURE-TIME REFERENCE IN AUSTRALIAN ENGLISH

GIULIANA REGNOLI
UNIVERSITY OF SALERNO

Abstract – The present paper investigates the *be going to* vs. *will/shall* alternation in Australian English. By drawing on quantitative data collected through the Corpus of Global Web-Based English (GloWbE), random forests techniques and conditional decision trees are applied to assess the importance of selected independent variables and to visualize how contexts of specialization are set apart based on the interaction of multiple factors. The results reveal that the predictors traditionally thought to influence the probabilistic grammar of the *be going to* vs. *will/shall* alternation in other L1 varieties of English only partly account for the variation observed in Australian English. The model reports better predictive results for *will/shall* and shows that text-specific preferences play a significant role in this alternation. Additionally, Proximity and Sentence Type emerge as the strongest predictors, indicating that grammaticalization processes affect different grammatical categories and are closely tied with semantics as well as syntactic complexity.

Keywords: Australian English; future-time reference; *will/be going to* alternation; World Englishes

1. Introduction

Australian English (AusE) represents a typical case of an English as a Native Language (ENL) (Strang 1970) or Inner Circle (Kachru 1985) variety of English. To use Schneider's (2007: 118-127) distinction, it is a 'settler' variety emerged from the intense contact of convicts and settlers brought together from the British Isles - predominantly from south-eastern England and Ireland - in 1788. Quickly established as the official language of the Great South Land, a middle-class, educated form of an emergent Australian vernacular had seemingly already stabilized by the mid-19th century (Pawley 2004). AusE has today evolved into a 'major' variety of World English (Svartvik 1997) having reached phase 5 of Schneider's (2007: 118-127) Dynamic Model, i.e. 'differentiation' around the end of the 20th century. This is manifested in the variety's development of internal standards as well as its capacity to support internal diversification and group-specific identification (see e.g. Leitner 2004a; Leitner 2004b).

In contrast with earlier studies which often expressed the view that AusE is a homogeneous variety (see e.g. Bernard 1969), its ongoing internal differentiation has been receiving scholarly attention. Not only is this indexical of the emergence of regional differences and Aboriginal varieties but also of the birth of new dialects (Collins and Blair 2001). Australia's multiethnic society has generated a multitude of varieties spanning from those spoken by native-born non-Indigenous Australians to those spoken by Indigenous as well as non-English migrant communities. This has led to draw a distinction between 'Australian English' - the variety of English spoken by people born in Australia - and 'English in Australia', an umbrella term used to describe the multitude of varieties spoken by immigrant settlers and convicts (Leitner 1984). Though providing a thorough definition of AusE today represents a "vexed question" (Collins 2014: 449), the focus of this study lies on what Leitner (2004: 1) refers to as "mainstream AusE" - a variety rooted in the country's Anglo-Celtic heritage and recognizable as the national language as well as a salient marker of national identity.

Research on the historical development of AusE has traditionally been primarily focused on lexical (see e.g. Murray & Manns 2020; Ramson 1966) and phonological variation (see e.g. Cox 1999; Mitchell and Delbridge 1965a, 1965b). As far as grammar is concerned, while earlier studies examined usage and acceptability through data gathered by means of elicitation tasks (see e.g. Watson 1978; Collins 1979), recent developments have seen an increased usage of corpora of AusE (see e.g. Engel and Ritz 2000; Collins 2008), generally supporting the hypothesis that grammatical differences between AusE and other L1 varieties are seemingly concerned with tendencies rather than categorical differences. Much of the research conducted to date has focused on the verb system (i.e. aspect, voice, mood, and modality), but it has also extended to the intersection between grammar and lexis mainly in relation to the different ways of expressing comparison and alternative case selection (see e.g. Peters et al. 2009 for a comprehensive account).

However, little attention has been given to longitudinal and real-time studies of the development of AusE grammar. Thus, in the attempt to go beyond the dichotomy between 'colonial lag' and 'innovation' (Hundt 2009), there is ample scope for further research on differential grammatical change in AusE in comparison with available evidence for other L1 English varieties. One aspect that has so far been neglected is futurity. The English future-time reference (FTR) system has long been recognized as a variable system undergoing change and its main variants (*will/shall* and *be going to*) have both been argued to have gone through, and, potentially, still be undergoing grammaticalization (Szmrecsanyi 2003; Tagliamonte et al. 2014; Denis and Tagliamonte 2018).

To fill the gap in the literature, the present study investigates FTR in AusE by drawing on quantitative data extracted from the Corpus of Global Web-Based English (GloWbE). More specifically, it focuses on what Bybee et al. (1994) define "primary" future markers, i.e. those constructions consisting of an auxiliary verb with or without an infinitive, viz. *be going to* and *will/shall*. In keeping up with Szmrecsanyi's (2003) novel approach to variable marking in FTR, the aim of the study is to shed light on whether, and to what extent, there are correlations between future marker distributions and their syntactic environment in written discourse. The research questions at the basis of this investigation may be summarized as follows:

- (1) Which predictors guide the semantic differentiation between *will/shall* and *be going to* in AusE?
- (2) How does this contribute to previous theoretical assumptions on grammaticalization and syntactic complexity in the FTR system?

The remainder of the paper is structured as follows: Section 2 addresses the current state of research on grammaticalization and FTR in present-day English (PDE) with a special focus on World Englishes (WEs). Section 3 describes how the data were extracted and analyzed. The results will be presented and interpreted in Sections 4 and 5. Concluding remarks will be given in Section 6.

2. Theoretical background

2.1. Grammaticalization and *will/shall* vs. *be going to*

Language use has long been investigated in terms of form-function asymmetry, i.e. speakers employing different constructions for similar discourse functions (see e.g. Labov 1969; Sankoff and Thibault 1981). In a variationist perspective, this dynamic and ever-

changing nature of language allows for specific mechanisms of change which "[...] occur while language is being used, and [...] create language" (Bybee 2001: 190) itself. This heterogeneity is structured and conditioned by both linguistic and extralinguistic factors (Labov 1994; 2001).

One major phenomenon capturing a set of important patterns of variation is grammaticalization, i.e. "the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions and, once grammaticalized, continue to develop new grammatical functions" (Hopper and Traugott 2003: 18). As far as scholarly treatments are concerned, previous research has tended to examine isolated cases of grammaticalization setting aside the investigation of multiple forms undergoing grammaticalization in the same grammatical domain. Recent trends have started to account for Hopper's (1991: 22-24) concept of 'layering', i.e. the coexistence of different stages of grammaticalization within the same grammatical function.

The FTR system is a case in point. It has long been recognized as a variable system undergoing change and its main variants, *will/shall* and *be going to*, have both been argued to have gone through and to still be undergoing the process of grammaticalization (Bybee et al. 1994; Hopper and Traugott 2003). The development of *will/shall* started in the Old English period, when from a verb of volition and willingness (*willan*) it fully grammaticalized into a general FTR marker during Late Modern English (Warner 1993: 181). To date, it remains unclear when and at which rate FTR uses of *will/shall* increased over time mainly since the meanings of volition, intention, and futurity are difficult to distinguish and prone to subjective interpretation (Bohmann 2023). Traces of its original volitional semantics are nowadays attested only in formulaic expressions, i.e. *if you will* and in the noun *will*. The grammaticalization of *be going to* into a future marker originated as the progressive aspect of the lexical verb *go* with the meaning of physical movement around the 16th century (Tagliamonte et al. 2014). A century later, *be going to* became more frequent and its status of verb of motion, though still retaining its original meaning, gradually receded to one of intention with further connotations of prediction, purpose, determination, and proximity (Binnick 1971; Nicolle 1997; Kuteva 2001). Eventually, *be going to* started to occur more frequently. As is the case with *will/shall*, it is still not clear how or when *be going to* evolved to express pure prediction because of the difficulty to distinguish the meanings of intention and motion in contextualized examples (Pérez 1990). Tagliamonte et al. (2014: 7) outline the following historical developments of the FTR system:

Old English	16th century	17th century	Modern English
<i>shall</i> present obligation	<i>shall</i> simple future	<i>shall</i> 1st p. future	<i>shall</i> restricted; formulaic; infrequent
<i>will</i> volition; willingness	<i>will</i> modal future	<i>will</i> 2nd 3rd p. future; prescriptive rules of usage	<i>will</i> simple future
	<i>be going to</i> actual motion	<i>be going to</i> immediate or impending; correlated with motion verbs	<i>be going to</i> ?? future

Table 1
Historical developments of the FTR system.

be going to has shown layering with *will/shall* since the end of the Middle English period (Tagliamonte et al. 2014). While the grammaticalization of the FTR system has been

considered "completed centuries ago" (Mair 1997a: 1538), the two variants continue to show, both synchronically and diachronically, semantic and pragmatic differences as well as change in frequency and collocation (see e.g. Szmrecsanyi 2003).

will/shall is generally asserted to be the unmarked or simple future with an "irrealis potential" (Klinge 1993: 315) involving the already mentioned nuances of volition and willingness. Quirk et al. (1985: 47), however, define *will* as "the closest approximation to a colorless, neutral future". Conversely, *be going to* seems to suggest immediate futurity (Bickick 1971) or prior intention (Nicolle 1997) indexing the speaker's certainty that some event will occur. Another commonly agreed connotation of *be going to* is the proximity of some future event (Hopper and Traugott 2003), thus retaining its original meaning of motion embedded in its inherent progressive aspect. In this sense, the action expressed by *be going to* is oriented to the present (Quirk et al. 1985). Most of these connotations are supported by empirical studies (see Section 2.2) which show that, overall, *be going to* has been gradually increasing its proportional share of the FTR system with the subsequent decline in frequency for *will/shall* in PDE.

2.2. Futurity and World Englishes

Variable marking of FTR has been predominantly concerned with diachronic and synchronic investigations of British and North American varieties of English. In acknowledging that the body of literature dealing with FTR variation is sizable, for space limitations, I will only address those studies which are, both theoretically and methodologically, in line with the rationale of the present study.

Diachronically, genre (in terms of register) has been one of the first predictor variables, i.e. those factors hypothesized to guide the choice between *be going to* and *will/shall*, to be investigated. Mair's (1997a; 1997b) study of grammatical change and futurity in the Freiburg-LOB Corpus of British English and the Freiburg-Brown corpus of American English show that future marker distributions are sensitive to register. *be going to* has become more frequent because it tends to be chosen more often as the preferred stylistically informal alternative to *will/shall*. Mair (1997a: 1541–1542; 1997b: 203) explains this result as a case of 'colloquialisation' of norms in written English: it is not a matter of grammatical changes but rather of informal options, which tend to be chosen more frequently in PDE. Szmrecsanyi's (2003) variationist study of FTR in two spoken registers of BrE and AmE shed light on the FTR system being sensitive to other predictor variables. He examines the correlations between future marker distributions and their syntactic environment in three major corpora: the British National Corpus (BNC), the Santa Barbara Corpus of Spoken American English (CSAE), and the Corpus of Spoken Professional American English (CSPA). The results demonstrate that future marker distributions are sensitive to four main factors, viz. contexts of negation, contexts of subordination, *if*-clause environments, and sentence length. As far as negation is concerned, *will/shall* is rarely explicitly negated. Moreover, the syntactic embedding of *be going to* is among the most robust factors influencing FTR variation as it appears to be much more frequent in longer, more subordinated and more syntactically complex environments. Szmrecsanyi (2003) explains this in terms of information processing since syntactically dependent contexts indicate environments of higher structural complexity. This has been corroborated by Torres Cacoullos and Walker (2009). Their multivariate analysis of transcripts of sociolinguistic interviews in the Quebec variety of Canadian English (CanE) shed light on the fact that the choice between *be going to* and *will* is not determined by semantic nuances such as proximity, willingness or intention but rather on lexical, syntactic, and pragmatic ones. More specifically, while *will* is favored in indefinite

adverbials and apodoses of *if*-clauses, *be going to* is more frequent in interrogatives and clauses with complement-taking predicates. Moreover, they report that *be going to* is favored for 2nd and 3rd person subjects. According to Tagliamonte (2002) and Tagliamonte et al. (2014), this is mainly because 1st person subjects are more likely to show volition than other grammatical persons, which is "a reading said to be associated with *will*" (Tagliamonte 2002:750). Findings from Denis and Tagliamonte's (2018) study of the FTR system in CanE converge on an effect of both grammatical person and sentence type, with 3rd persons favoring the choice of *be going to* over *will* and interrogatives strongly favoring *be going to* over *will*, the latter having established its own functional/grammatical niche in apodosis clauses. Moreover, *be going to* is comparatively more frequent in the language of younger speakers. Not only does this confirm that younger generations are influential transmitters of language change (in this regard, see Kerswill 1996; Tagliamonte 2016) but also that "a slow but ongoing development consistent with the long-term gradual trajectory of competition between *BGT* [*be going to*] and *will*" (Denis and Tagliamonte 2018: 411) is in place. By recurring to Kroch's (1989) concepts of grammar competition and the subsequent 'Constant Rate Effect', according to which if one form obsolesces the rate of change across all contexts will be constant, Denis and Tagliamonte (2018) confirm the presence of grammar competition and predict that as the English FTR system continues to evolve, *be going to* will continue to (is going to?) expand. However, since *will* has established its own niche, this change may probably never culminate in obsolescing but rather in longitudinal partitioning (see also D'Arcy and Tagliamonte 2015). All the above has been very recently corroborated by Bohmann (2023). His investigation of FTR in selected L1 and L2 varieties of English provides the first large-scale multivariate analysis of this phenomenon. Drawing on both written and spoken data from the International Corpus of English (ICE), Bohmann examines the distribution of *be going to* and *will* across registers and grammatical contexts, uncovering significant insights into the dynamics of FTR variation. Notably, his findings emphasize the critical role of register in shaping linguistic choices, demonstrating that formal and informal contexts distinctly influence the use of futurized verb phrases. Moreover, he explores the interaction between the grammatical complexity of verb phrases - such as bare infinitives versus constructions like passives, progressives, and perfects - and their association with specific FTR markers. These findings align with broader trends in WEs research, highlighting the nuanced interplay of grammatical structure, register, and the sociolinguistic environment in shaping variation patterns in English worldwide. By addressing these complexities, Bohmann not only advances the current understanding of FTR but also reinforces the importance of multivariate approaches in capturing the subtleties of linguistic variation across WEs.

To summarize the above, the grammaticalization trajectories outlined have shown to impact the systematicity of FTR variation in PDE. *be going to* is the historically more recent form of the system and, as such, is a clear example of a sociolinguistic innovation. Research has suggested that *be going to* is comparatively more frequent in informal rather than formal registers (Mair, 1997b), in spoken rather than in written English (Mair 1997a) as well as in the language of younger speakers (Denis and Tagliamonte 2018). As Mair (1997a: 1541) predicted *be going to* will "become the most frequent, semantically neutral and syntactically least constrained expression of futurity".

Though there is a lot of research about the FTR system in English, to our knowledge, studies of other settler varieties such as AusE have yet to be documented. Such studies might shed new light on different degrees of specialization as well as different projections of grammaticalization.

3. Data and methodology

3.1. Corpus

The data for the present study has been extracted from the online version of the Corpus of Global Web-Based English (GloWbE). The corpus is based on 1.9 billion words of text from twenty English-speaking countries and consists of informal blogs (accounting for about 60% of the corpus) and other web-based materials including newspapers, magazines, and company websites (Davis and Fuchs 2015: 4). In keeping up with the study's research questions, the Australian sub-corpus of GloWbE (AusGloWbE) was selected. AusGloWbE consists of 28,881 websites and 129,244 webpages, for a total of 148,208,169 words (Davis and Fuchs 2015: 8).

The choice of GloWbE was largely driven by its open-access availability, making it the only open-source corpus of standard AusE. One of the major advantages of GloWbE is its large, diverse dataset, which facilitates comparisons among different standard varieties of English (see e.g. Levshina 2015; Szmrecsanyi et al. 2016), thus supporting replicability in research. However, there are some limitations to consider. A notable drawback of the corpus is the automated inclusion of web-sourced material which offers minimal control over data accuracy and introduces high levels of noise. Additionally, GloWbE provides limited information about text-type and register variation, which may reflect a specific set of language practices that do not encompass all social contexts or registers of English (see e.g., Liao & Fidalgo 2017). Nevertheless, GloWbE remains a valuable tool for conducting cross-linguistic analyses due to its extensive coverage of varieties of English.

3.2. Variable extraction and predictor variables

All variant forms of *will*, *shall* and *be going to*, i.e. *will*, *'ll*, *won't*, *shall*, *shan't*, *(not) going to*, and *(not) gonna* were extracted by means of the regex 'trigger.[v*]' (e.g. *will*.[v*]). A sample of 500 random tokens per variant was selected and extracted to keep the dataset manageable to code. *shan't* was the only exception as it included 97 hits in the corpus. This procedure yielded a total of 2,597 tokens. These, however, had to be further processed manually both to ensure that *will* and *be going to* were reliable instances of FTR and to code some of the predictor variables outlined below. To this purpose, utterances containing each target variant were extracted and imported into an Excel spreadsheet for manual processing. Unsuitable forms which had to be excluded consisted in:

- Duplicates, e.g. strings 43 and 451.
- Spatial meaning of *be going to*, e.g. *I am going to Paris*
- Tag questions, e.g. *You'll become a professor, will you?*
- Future-in-the-past contexts, e.g. *I was hoping you were gonna call me.*
- Fixed expressions, e.g. *We'll see/we shall see; I'll bet...*

Among the excluded tokens, *will/shall* instances are found in fixed expressions, predominantly, and in tag questions, exclusively. *Be going to* dominates in utterances indexing future-in-the-past contexts and spatial meaning.

A total of 194 tokens had to be excluded according to the criteria outlined above resulting in a final count of 2,403 observations divided in 1279 instances of *will/shall* and 1124 instances of *be going to*, accounting for the 53% and 47% of the final dataset, respectively. Tokens were first manually coded according to Type (*be going to* vs. *will/shall*), Verb (all variants of *be going to* and *will/shall*), and File (the file texts of AusGloWbE). The data were then semi-automatically coded according to a selection of

predictor variables which were selected on the basis of the previous literature (see Sections 2.1 and 2.2) and the study's research aims (see Section 1). These include Genre, Clause Type, Sentence Type, Proximity, and Grammatical Person. Table 2 summarizes the predictors selected and their levels, with the reference level set to the option expected to disprefer *be going to*, as indicated by the underlined levels.

Predictor	Levels
Genre	<i>blog</i> vs. <u><i>general</i></u>
Clause type	<u><i>main clause</i></u> vs. <i>subordinate clause</i>
Sentence type	<i>interrogative</i> (yes/no + wh- questions) vs. <i>negative</i> vs. <u><i>declarative</i></u>
Proximity	<i>near future</i> (within the same day or so) vs. <u><i>far future</i></u>
Pronominal person	<u><i>3rd person</i></u> vs. <i>non-3rd person</i>

Table 2
Predictor variables and their levels.

Genre, with the levels *blog* and *general*, reflects the composition of GloWbE and accounts for (an approximation of) informal and formal language, respectively. Since future marker distribution is sensitive to register and *be going to* is the preferred stylistically informal alternative to *will/shall* (Mair 1997a; 1997b), the hypothesis this predictor tests is that *blogs* favor the choice of *be going to*, as in (1a):

- (1) (a) the next time the blind man comes in he's going to test him (string 1804)
(b) The Parliament shall, subject to this constitution, have power to make laws (string 7)

Clause Type considers the levels *main* and *subordinate*. Following the evidence in the literature (Szmrecsanyi 2003; Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014; Denis and Tagliamonte 2018; Bohmann 2023), the hypothesis in relation to this predictor is that subordinate clauses favor *be going to* compared to main clauses, as in (2b):

- (2) (a) Occasionally, I will make guacamole or put them in a garden salad (string 204)
(b) If that person thinks that the experiment is going to show that priming reduces walking speed (string 1832)

Sentence Type, with the levels *interrogative*, *negative*, and *declarative*, hypothesizes that interrogatives and declaratives favor *be going to* over *will/shall* as in (3a) and (3c). Conversely, negative sentences strongly favor *will/shall* as in (3b) (Szmrecsanyi 2003; Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014; Denis and Tagliamonte 2018; Bohmann 2023):

- (3) (a) if you want to access every part of the game, you're gonna need at least 8? (string 1560)
(b) It needs to go hand in hand otherwise it won't work or will just be a gimmick (string 898)
(c) It's gonna be hard, but this time I'm ready for it (strong 1564)

Proximity considers the levels near future and far future. Overall, conditional phrases, generalizations, verbs of duration and those indicating a future state of something index far future actions (Tagliamonte 2002). This predictor then tests the hypothesis that near future favors *be going to*, as in (4a) while far future favors *will/shall*, as in (4b) (Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014):

- (4) (a) I think I'm gonna cry (string 1329)
 (b) Private sessions also available but will incur extra charges (string 856)

Pronominal Person considers the levels 3rd person and non-3rd person. The literature on this predictor is far from unanimous, revealing several complexities and occasional disagreements. For instance, Torres Cacoullos and Walker (2009) find that 1st person subjects favor *will*, while 2nd person subjects are identified as the most favorable environment for *be going to*. In contrast, Fehringer and Corrigan (2015) report no significant difference based on pronominal person. Denis and Tagliamonte (2018) present a more nuanced perspective, noting that animate non-1st person subjects are most inclined toward *be going to*. Bohmann (2024) finds that the effect of the subject is minimal and significantly influenced by modality. Consequently, the literature does not converge on a single, clear hypothesis regarding this predictor. However, given that 1st person subjects are more likely to show volition and are thus more associated with *will/shall*, as in (5b), the hypothesis in relation to this predictor is that *be going to* is favored for 3rd person subjects, as in (5a):

- (5) (a) the industry is probably going to face foreign competition from companies (string 1816)
 (b) I won't support you because you choose not to work (string 686)

As mentioned above, the data were semi-automatically coded for each constraint. Genre and Proximity were manually coded. Proximity is particularly cost-intensive to code as it often requires attention to the wider context beyond the individual sentence. A Perl script (Wall et al. 2000) was written to automatically code Clause Type, Sentence Type, and Grammatical Person. With respect to Sentence Type, while Perl identified negative and interrogative clauses, a manual check was necessary to assign each instance to the right category. For coding accuracy, a second researcher independently coded the predictor variables genre and proximity and double-checked the final version of the dataset. Potential diverging classifications were discussed and agreed upon accordingly.

3.3. Statistical modeling

The analysis below adheres to standard variationist methods, which primarily focus on describing, comparing, and measuring linguistic variation in probabilistic terms (see e.g. Levshina 2015). Specifically, conditional random forests (CRFs) and conditional inference trees (CITs) are employed to assess variable importance rankings and to explore the interactions among the most significant predictor variables, respectively. They are performed using the `party` package implemented in R (Hothorn et al. 2006; Strobl et al. 2007).

CRFs are used for nonlinear multiple regression and recursively split the data into a set of split points for each factor tested, ranking and establishing their importance, while

addressing the problem of data over-fitting. The CRF algorithm grows 'forests' of conditional decision trees (i.e. CITs) in which each tree provides a classification regarding the impact of each independent variable (i.e. our predictor variables) on the response variable (i.e. *be going to* and *will/shall*), and finally selects the variable(s) with the best classification, implementing a binary split based on such variable(s) (Levshina 2020). CITs represent randomly sampled subsets of the data, wherein each CIT considers "the likelihood of the value of the response variable [...] based on a series of binary questions about the values of predictor variables" (Tagliamonte and Baayen 2012: 159) and uses permutation tests to measure the importance of the predictor variables.

The CRFs and subsequent CITs built for the present analysis are based on the predictor variables introduced in Section 3.2. Following Tagliamonte and Baayen (2012: 161), a conditional variable importance measure was incorporated to ensure that it "correctly reports spurious predictors [...] to have a very low variable importance". To enhance the accuracy of the analysis, the following steps were taken: (i) the reference level for each predictor was set to the option that was supposed to disprefer *be going to*, as indicated by the levels underlined in Table 2, and (ii) the bias in variable selection was reduced to provide more reliable variable importance scores. This was achieved by employing the `cforest_unbiased` function.

4. Results

4.1. Distribution of predictor variables

Figure 1 shows the distribution of *be going to* and *will/shall* according to Proximity and Genre. Proximity favors *be going to* in near future (N=585) and *will/shall* in far future (N=1818) structures. This might corroborate the hypothesis discussed in Section 3.2 according to which near future favors *be going to* and far future favors *will/shall* (Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014). Exploration of Genre suggests that *be going to* is the preferred stylistically informal alternative to *will/shall* as it occurs more in blogs (N=492) than in general webpages (N=1921) (Mair 1997a; 1997b). This aligns nicely with the hypothesis according to which blogs favor the choice of *be going to*, thus confirming that future marker distribution is sensitive to register.

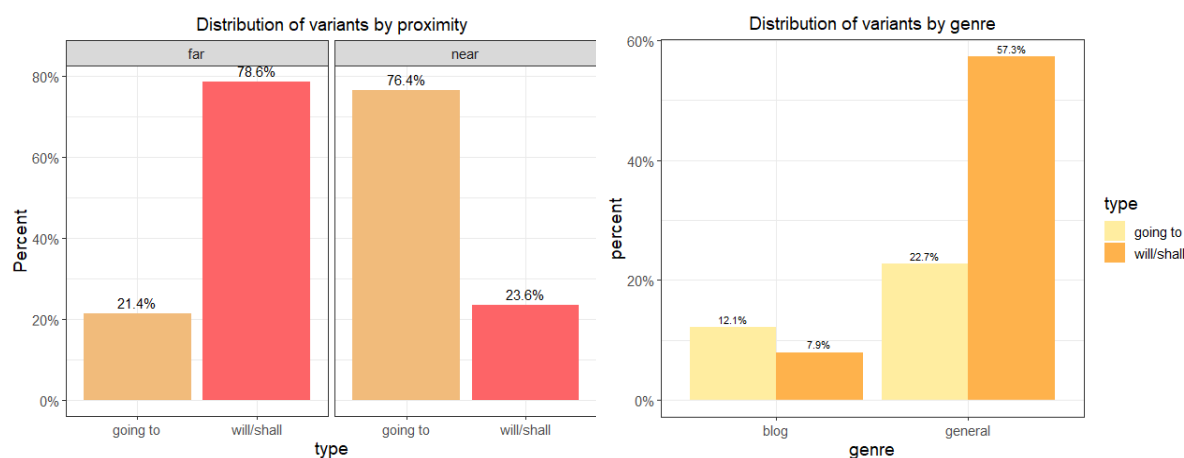


Figure 1
Distribution of *be going to* and *will/shall* according to Proximity and Genre.

The *be going to* vs. *will/shall* distribution according to Sentence Type and Clause Type is shown in Figure 2. As far as Sentence Type is concerned, section 3.2 hypothesized that while declaratives and interrogatives favor *be going to*, negative sentences strongly favor *will/shall*. Figure 2 shows that in affirmative sentences (N=1496), *will/shall* is slightly more prevalent compared to *be going to*. The preference for *will/shall* becomes even more pronounced in negative sentences (N=804), thus confirming previous research (Szmrecsanyi 2003; Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014; Denis and Tagliamonte 2018). As far as interrogative sentences are concerned, *be going to* is the preferred variant in AusE (N=103). This effect is explained by the fact that *be going to* retains a semantic component which can be used to enquire about already-decided intentions or projected plans (Torres Cacoullos and Walker 2009: 343). However, it is important to take this finding with a grain of salt, considering the paucity of data related to interrogatives. The hypothesis in relation to Clause Type is that subordinate clauses favor *be going to* compared to main clauses (Szmrecsanyi 2003; Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014; Denis and Tagliamonte 2018; Bohmann 2023). Frequency distribution shows that *will/shall* is favored both in main (N=1365) and subordinate (N=1038) clauses, with *be going to* occurring only slightly more frequently in subordinate rather than main clauses. Specifically, out of all subordinate clauses, 37% occur with *be going to*, while only 33.1% of main clauses do so, thus corroborating, though only impressionistically, the hypothesis outlined above.

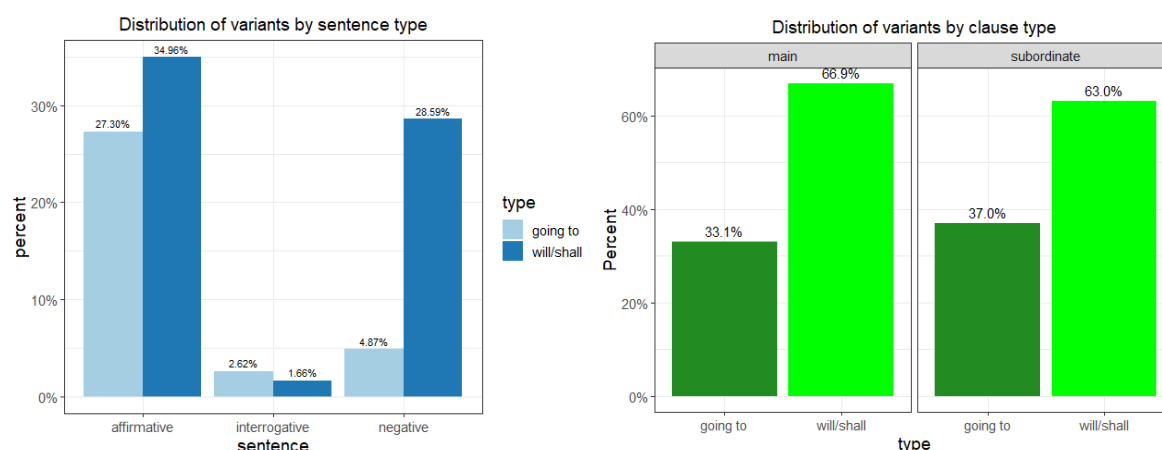


Figure 2
Distribution of *be going to* and *will/shall* according to Sentence Type and Clause Type.

Figure 3 shows the frequency distribution of *be going to* and *will/shall* according to Pronominal Person. The hypothesis in relation to Pronominal Person posits that *be going to* is favored for 3rd person subjects. Frequency distribution shows that *will/shall* is the predominant variant in AusE, while *be going to* occurs only slightly more frequently with non-3rd person subjects (N=1366). Considering the several complexities and occasional disagreements in the literature (see Section 3.2), this result should be interpreted with caution.

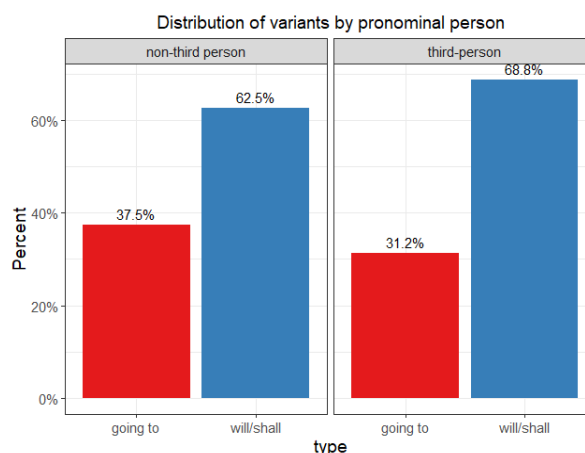


Figure 3
Distribution of *be going to* and *will/shall* according to Pronominal Person.

4.2. The CRF model

To assess the effectiveness of the CRF model in predicting the usage of *be going to* and *will*, concordance (C) and discrimination indices (Dxy) were analyzed. These metrics provide insight into the model's ability to distinguish between the two tense forms. Harrell's C-index (C) measures the degree to which the model's predictions align with the observed outcomes, with higher values indicating better predictive accuracy. The Dxy index evaluates the model's ability to discriminate between classes, with positive values suggesting good discrimination, and negative values indicating poor discrimination (Harrell et al. 1996). The results are summarized in Table 3.

Future tense form	C	Dxy	n	Missing
<i>Be going to</i>	0.1375	-0.7251	2403	0
<i>Will/Shall</i>	0.8625	0.7251	2403	0

Table 3
Model performance metrics for future tense form prediction.

The model shows varying performance levels between *be going to* and *will/shall*. It exhibits a low C-index of 0.1375 (accounting for 13.75% of the observations in the dataset) and a negative Dxy of -0.7251 for *be going to*, and a high C-index of 0.8625 (accounting for the remaining 86.25% of the total observations) and a positive Dxy of 0.7251 for *will/shall*. The model's output highlights the dominance of *will/shall* in the dataset and suggests that it has stronger predictive performance and better discriminatory power for *will/shall* compared to *be going to*. This may reflect both the frequency bias of *will/shall* and possibly stronger associations with the predictors included in the model.

Figure 4 shows the factors contributing most significantly to the model's predictions.

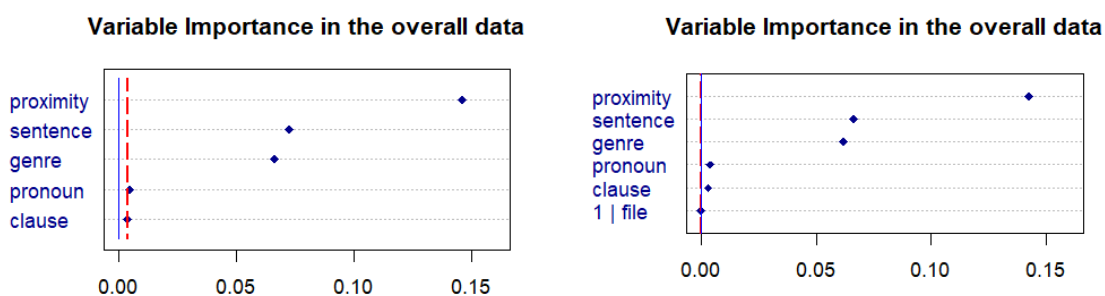


Figure 4
Conditional permutation variable importance for all predictor variables.

Predictors on the rightmost side of the red-dotted line are the most significant ones. Those situated on its left side have little to no impact on the choice between *be going to* and *will/shall*. Specifically, the graphic on the left shows variable importance according to the predictor variables discussed in Section 3.2. The graphic on the right accounts for the same with the inclusion of File as random effect. As explained in 3.2, file is not a fundamental predictor in the analysis as it only accounts for the text files in AusGloWbE. However, since each text file may express divergent lexical choices behind the realization of *be going to* and *will/shall*, its choice as random effect may explain idiosyncratic and random variation in the corpus and help "[...] avoid possibly misleading generalisations" (Gries and Bernaisch 2016: 21).

Both graphics show that the ranking of variable importance is similar, with Pronominal Person and Clause Type resulting more significant when considering random variation. Specifically, Proximity emerges as the most significant constraint guiding FTR variation, followed by Sentence Type and Genre. Pronominal Person and Clause Type have less influence but still contribute to the model's overall predictions. This result is in line with the descriptive statistics discussed in section 4.1.

4.3. The CIT

Figure 5 presents the CIT resulting from the CRF model described in Section 4.2. It provides a more fine-grained overview of the way in which these constraints interact to drive the model's predictions. To read the tree, please note that each ellipse represents a decision node where a specific variable is used to split the data. The branches leading from each node represent the different levels of the variables being tested. Finally, the leaves (i.e. terminal nodes) at the bottom of the tree show the distribution of the dependent variables, i.e. *be going to* and *will/shall*, in the corpus (Levishina 2020).

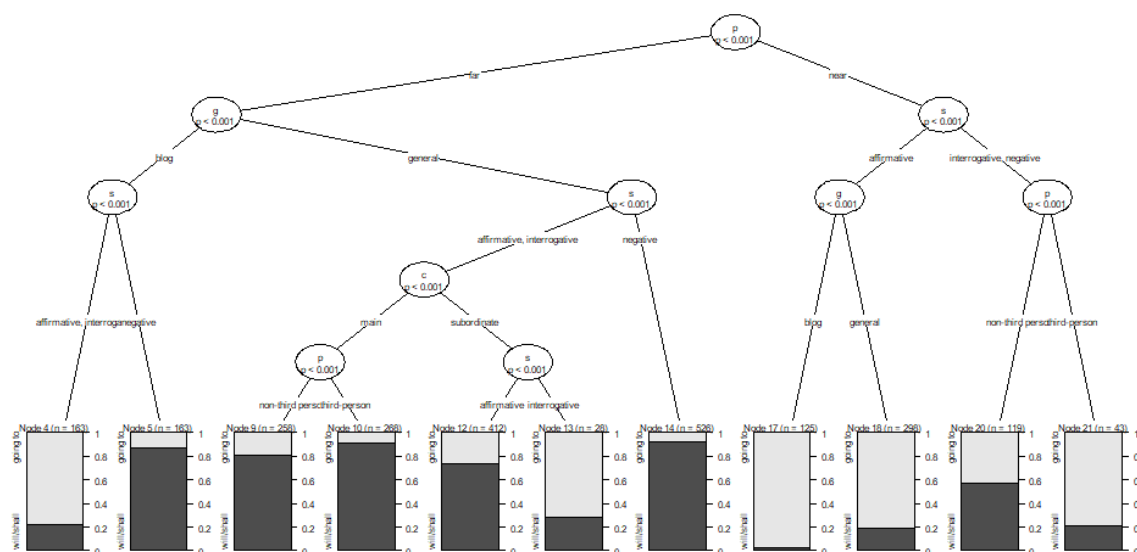


Figure 5

Conditional inference tree. Nodes from top to bottom: p=proximity; s=sentence; g=genre; p=pronoun; c=clause.

The CIT shows that Proximity is the variable having the most significant influence on the prediction of every other factor tested. Starting from the left side of the tree, this first split sets apart far futures from near futures. As a matter of fact, far futures predict *will/shall* better, especially when correlated with texts from blogs and interrogative sentences (node 5), or with general texts and negative sentences (node 14). When general texts are correlated with affirmative and interrogative sentences, *will/shall* is the preferred variant in main clauses with either 3rd (node 9) or non-3rd (node 10) pronominal subjects, or in subordinate affirmative clauses (node 12). On the contrary, far futures have a strong predictive effect for *be going to* when they are correlated with affirmative sentences in texts from blog (node 4) and a moderate predictive effect with subordinate interrogative clauses (node 13) in general texts. Looking at the right branch of the tree, the first partition is for near futures, which have a strong predictive effect for *be going to* in affirmative sentences occurring in blogs (node 17) and in interrogative and negative sentences with 3rd person pronominal subjects (node 21). *Be going to* is still the preferred variant in affirmative sentences in general texts (node 18) and in interrogative and negative sentences with non-3rd person pronouns (node 20).

5. Discussion

Futurity in AusE is a neglected area of investigation within the theoretical framework of WEs. While much research has been devoted to British and American Englishes, and, more recently, to L2 varieties of English (see e.g. Bohmann 2023), differential grammatical change in AusE is still relatively under-researched.

The analysis above shows that several key factors influence the semantic differentiation between *be going to* and *will/shall* in the AusE FTR system. The output of the CRF model discussed in Section 4.2 reveals that *will/shall* exhibits higher predictive power compared to *be going to*. It also highlights a significant imbalance in the data set, which may have introduced bias toward the majority class (*will/shall*).

A closer examination of the variables that contribute most significantly to the model's predictions suggests that factors traditionally thought to influence the probabilistic grammar of the *be going to* vs. *will/shall* alternation in other varieties of English may only partly account for the same patterns in AusE. As a matter of fact, future marker distributions in AusE are particularly sensitive to Proximity, Sentence Type, Genre, and Clause Type. The findings for Sentence Type Genre and Clause Type dovetail nicely with what previous research about other L1 varieties of English has established (Mair 1997a; 1997b; Torres Cacoullos and Walker 2009; Tagliamonte et al. 2014; Denis and Tagliamonte 2018). However, the emergence of Proximity as the most important predictor stands in stark contrast to the existing literature. Denis & Tagliamonte (2018) and Torres Cacoullos & Walker (2009) investigate Proximity but do not find a significant effect. Similarly, Tagliamonte et al. (2014: 99) show that the effect of Proximity is limited to communities with very low rates of *be going to*, i.e. at early stages of the grammaticalization process, thus suggesting that its influence diminishes rapidly over time. This may be due to a different projection of grammaticalization in AusE but more data is needed to confirm this. Regarding the effect of Pronominal Person, Section 3.2 emphasized that the literature on this predictor is far from unanimous and does not converge on a single hypothesis. However, since 1st person subjects are more likely to express volition and are thus more strongly associated with *will/shall*, the present study hypothesized that *be going to* would be favored for 3rd person subjects. The CIT in Section 4.3, however, reveals a more nuanced picture. It shows that *will/shall* is the preferred variant for both 3rd and non-3rd person subjects in main and affirmative clauses referring to near futures, while *be going to* is favored in affirmative sentences in general texts and in interrogative and negative sentences with non-3rd person pronouns in far futures. These findings open new avenues for research on the interplay between grammaticalization and semantics in future-time reference.

In addition, the data show that *be going to* and *will/shall* do not have a degree of specialization with respect to Clause Type and Pronominal Person. This may be due to a different projection of grammaticalization in AusE. As a matter of fact, grammaticalization seemingly affects different grammatical categories since Proximity is, to some degree, related to semantics and Sentence Type to both semantics and syntactic complexity. Whether this confirms specific variety effects is a speculation that, undoubtedly, warrants further exploration. Should this be the case, it would support previous assumptions according to which grammatical differences between AusE and other L1 varieties of English are seemingly concerned with tendencies rather than categorical differences (see e.g. Engel and Ritz 2000; Collins 2008). At present, however, there is not enough evidence to sustain any concrete discussion. It is possible that, given more and more diversified data, specific variety effects would emerge.

On a final note, as far as methodological implications are concerned, the analysis shows that the CRF model performed significantly better when adding File as a random effect in the formula. GloWbE's lack of metadata makes it impossible to investigate speaker-specific preferences, hence using the corpus file's ID may help partially account for file-specific variation in order to take into consideration idiosyncratic and random variation in the corpus.

6. Conclusion

This paper has presented the first empirical small-scale study of FTR variation in AusE. Based on 2,403 observations retrieved from the Australian component of GloWbE, the

study aimed to shed light on the correlations between future marker distributions and their syntactic environment in written discourse. To do so, five factors, i.e. Genre, Clause Type, Sentence Type, Proximity, and Pronominal Person, were tested to predict FTR variation in AusE. The reported results show that Proximity and Sentence Type are the most influential variables guiding the choice between *be going to* and *will/shall*, soon followed by Genre, Pronominal Person, and Clause Type. Though more data is needed to draw more accurate conclusions, this may shed light on the relationship among grammaticalization, semantics, and syntactic complexity.

A next step in this study could involve further elaboration on the CRF model outlined above. Refinements such as addressing class imbalance, incorporating additional predictors, and validating the model could significantly improve its performance, particularly for *be going to*. Future research may also focus on the semantics of the *will/be going to* alternation with a closer inspection of register and genre variation, and test for language-internal predictors such as Animacy, Verb Type, and Clause Length as further conditioning factors predicting FTR variation (see e.g. Tagliamonte et al. 2014). Further steps may expand the scope of the research by making cross-varietal comparisons with typologically similar varieties of English, e.g. New Zealand English.

Bionote: Giuliana Regnoli is Assistant Professor of English Language and Linguistics at the University of Salerno and a Postdoctoral Research Fellow at the University of Regensburg. Her research interests include variationist sociolinguistics, sociophonetics, language attitudes, perceptual dialectology, and World Englishes. She is currently working on children's English in Cameroon and on Italian university students' attitudes toward English(es) worldwide.

Author's address: gregnoli@unisa.it

Acknowledgements: I am grateful to Roberta La Peruta for her invaluable contribution to this project. We conceptualized the project together and collaborated on the stages of data collection and analysis. Her support and feedback have been greatly appreciated. All errors are mine alone.

References

- Bernard J. 1969, *On the uniformity of Australian English*, in "Orbis" 18, pp. 62-73.
- Binnick R.I. 1971, *Will and be going to*, in *Papers from the Seventh Regional Meeting of the Chicago Linguistics Society*, Chicago Linguistic Society, Chicago, pp. 40-53.
- Bohmann A. 2023, *Future-time reference in World Englishes*, in "World Englishes" 43 [1], pp. 2-22.
- Bybee J. 2001, *Phonology and Language Use*, Cambridge University Press, Cambridge.
- Bybee J., Perkins R. and Pagliuca W. 1994, *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*, The University of Chicago Press, Chicago.
- Collins P. 1979, *Elicitation experiments on acceptability in Australian English*, in *Macquarie University Speech and Language Research Centre (SLRC) Working Papers 2*, pp. 1-49.
- Collins P. 2008, *The progressive in World Englishes: A corpus-based study*, in "Australian Journal of Linguistics" 28, pp. 225-249.
- Collins P. 2012, *Australian English*, in Koch, H. and Nordlinger, R. (eds.), *The Languages and Linguistics of Australia: A Comprehensive Guide*. Mouton de Gruyter, Berlin, pp. 449-484.
- Collins P. and Blair D. 2001, *Language and Identity in Australia*, in Blair, D. and Collins, P. (eds.), *English in Australia*, John Benjamins, Amsterdam, pp. 1-13.
- Cox F. 1999, *Vowel change in Australian English*, in "Phonetica" 56, pp. 1-27.
- D'Arcy A. and Tagliamonte S.A. 2015, Not always variable: Probing beneath the vernacular grammar, in "Language Variation and Change" 27, pp. 255-285.
- Davies M. and Fuchs, R. 2015, *Expanding horizons in the study of World Englishes with the 1.9 billion word global web-based English corpus (GloWbE)*, in "English World-Wide" 3 [1], pp. 1-28.
- Denis D. and Tagliamonte S.A. 2018, *The changing future: Competition, specialization and reorganization in the contemporary English future temporal reference system*, in "English Language and Linguistics" 22 [3], pp. 403-430.
- Engel D. and Ritz M. 2000, *The use of the present perfect in Australian English*, in "Australian Journal of Linguistics" 20, pp. 119-140.
- Fehrer C. and Corrigan K.P. 2015, *The rise of the going to future in Tyneside English: Evidence for further grammaticalisation*, in "English World-Wide" 36 [2], pp. 198-227.
- Gries S.Th. and Bernaisch T. 2016, *Exploring epicentres empirically: Focus on South Asian Englishes*, in "English World-Wide" 37 [1], pp. 1-25.
- Harrell F.E., Lee K.L. and Mark D.B. 1996, *Multivariable prognostic models: Issues in developing models, evaluating assumptions, and validating predictions*, in "Statistics in Medicine" 15 [4], pp. 361-387.
- Hopper P. 1991, *On Some Principles of Grammaticalization*, in E.C. Traugott and B. Heine (eds.), *Approaches to Grammaticalization*, John Benjamins, Amsterdam, pp. 17-35.
- Hopper P. and Traugott E.C. 2003, *Grammaticalization* (2nd ed.) Cambridge University Press, Cambridge.
- Hothorn T., Hornik K., van de Wiel M.A. and Zeileis A. 2006, *A lego system for conditional inference*, in "The American Statistician" 60 [3], pp. 257-263.
- Hundt M. 2009, *Colonial Lag, Colonial Innovation or Simply Language Change?*, in G. Rohdenburg and J. Schlüter (eds.) *One Language, Two grammars? Differences between British and American English*. Cambridge University Press, Cambridge, pp. 13-37.
- Kachru B.B. 1985, *Standards, Codification and Sociolinguistic Realism: The English Language in the Outer Circle*. In R. Quirk and H.G. Widdowson (eds.), *English in the World: Teaching and Learning the Language and Literatures*, Cambridge University Press for The British Council, Cambridge, pp. 11-30.
- Kerswill P. 1996, *Children, adolescents and language change*, in "Language Variation and Change" 8, pp. 177-202.
- Klinge A. 1993, *The English modal auxiliaries: From lexical semantics to utterance interpretation*, in "Journal of Linguistics" 29, pp. 315-357.
- Kuteva T. 2001, *Auxiliation: An Enquiry into the Nature of Grammaticalization*. Oxford University Press, Oxford.
- Labov W. 1969, *Contraction, deletion, and inherent variability of the English copula*, in "Language" 45, pp. 715-762.
- Labov W. 1994, *Principles of Linguistic Change. Vol. I: Internal Factors*, Blackwell, Oxford.
- Labov W. 2001, *Principles of Linguistic Change. Vol II: Social Factors*, Blackwell, Oxford.
- Leitner G. 1984, *'Australian English' or 'English in Australia': Linguistic identity or dependence in broadcast language*, in "English World-Wide" 5, pp. 55-85.
- Leitner G. 2004a, *Australia's Many Voices: Australian English – the National Language*, Mouton de Gruyter, Berlin.

- Leitner G. 2004b, *Australia's Many Voices: Ethnic Englishes, Indigenous and Migrant Languages – Policy and Education*, Mouton de Gruyter, Berlin.
- Levshina N. 2015, *How to do Linguistics with R: Data Exploration and Statistical Analysis*, John Benjamins, Amsterdam.
- Levshina N. 2020, *Conditional Inference Trees and Random Forests*, in M. Paquot and S. Th. Gries (eds.) *A Practical Handbook of Corpus Linguistics*, Springer, Switzerland, pp. 611-643.
- Liao Y., and Fidalgo R. 2017, *The GloWbE corpus and its potential for linguistic research*, in “International Journal of English Linguistics” 7 [3], pp. 107-117.
- Mair C. 1997a, *The Spread of the Going-to-future in Written English: A Corpus-based Investigation into Language Change in Progress*, in R. Hickey and S. Puppel (eds.), *Language History and Linguistic Modelling. Vol. II*, Mouton de Gruyter, Berlin, pp. 1537-1543.
- Mair C. 1997b, *Parallel Corpora: A Real-time Approach to the Study of Language Change in Progress*, in M. Ljung (ed.), *Corpus-based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora*, Rodopi, Amsterdam, pp. 195-209.
- Mitchell A.G. and Delbridge A. 1965a, *The Speech of Australian Adolescents*, Angus and Robertson, Sydney.
- Mitchell A.G. and Delbridge A. 1965b, *The Pronunciation of English in Australia*, Angus and Robertson, Sydney.
- Murray L. and Manns H. 2020, *Lexical and Morphosyntactic Variation in Australian English*, in L. Willoughby and H. Manns (eds.), *Australian English Reimagined: Structure, Features and Developments*, Routledge, London, pp. 120-133.
- Nicolle S. 1997, *A relevance-theoretic account of be going to*, in “Linguistics” 33, pp. 355-377.
- Pawley A. 2004, *Australian Vernacular English: Some Grammatical Characteristics*, in B. Kortmann, K. Burridge, R. Mesthrie, E.W. Schneider and C. Upton (eds), *A Handbook of Varieties of English. Vol. II: Morphology and Syntax*, Mouton de Gruyter, Berlin, pp. 611-642.
- Pérez A. 1990, *Time in motion: Grammaticalisation of the be going to construction in English*, in *La Trobe University Working Papers in Linguistics* 3, pp. 49-64.
- Peters P., Collins P. and Smith A. 2009, *Comparative Studies in Australian and New Zealand English Grammar*, John Benjamins, Amsterdam.
- Quirk R., Greenbaum S., Leech G. and Svartvik J. 1985, *A Comprehensive Grammar of Contemporary English*, Longman, New York.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <<https://www.R-project.org/>>.
- Ramson W. 1966, *Australian English: An Historical Study of the Vocabulary, 1788-1898*, Australian National University Press, Canberra.
- Sankoff D. and Thibault P. 1981, *Weak Complementarity: Tense and Aspect in Montreal French*, in B. B. Johns and D. R. Strong (eds.), *Syntactic Change*, Ann Arbor: University of Michigan Press, Ann Arbor, pp. 205-216.
- Strang B.M.H. 1970, *A History of English*, Methuen, London.
- Schneider E.W. 2007, *Postcolonial English: Varieties Around the World*, Cambridge University Press, Cambridge.
- Strobl C., Boulesteix A., Zeileis A. and Hothorn T. 2007, *Bias in random forest variable importance measures: Illustrations, sources, and a solution*, in “BMC Bioinformatics” 8, 25.
- Szmrecsanyi B. 2003, *Be going to versus will/shall. Does syntax matter?*, in “Journal of English Linguistics” 31 [4], pp. 295-323.
- Szmrecsanyi B., J. Grafmiller, Heller B. and Röthlisberger M. 2016, *Around the world in three alternations: Modelling syntactic variation in varieties of English*, in “English World-Wide” 37 [2], pp. 109-137.
- Tagliamonte S.A. 2002, *Comparative Sociolinguistics*, in J. K. Chambers, P. Trudgill and N. Schilling-Estes (eds.), *Handbook of Language Variation and Change*, Blackwell., Oxford, pp. 729-763.
- Tagliamonte S.A. 2016, *Teen Talk: The Language of Adolescents*, Cambridge University Press, Cambridge.
- Tagliamonte S.A. and Baayen R.H. 2012, *Models, forests and trees of York English: Was/were variation as a case study for statistical practice*, in “Language Variation and Change” 24 [2], pp. 135-178.
- Tagliamonte S.A., Durham M. and Smith J. 2014, *Grammaticalization at an early stage: Future be going to in conservative British dialects*, in “English Language and Linguistics” 18 [1], 75-108.
- Torres Cacoullos R. and Walker J. 2009, *The present of the English future: Grammatical variation and collocations in discourse*, “Language” 85 [2], pp. 321-354.
- Wall L., Christiansen T. and Orwant J. 2000, *Programming Perl*, O'Reilly Media, Sebastopol.
- Warner A. 1993, *English Auxiliaries: Structure and History*, Cambridge University Press, Cambridge.

Watson K. 1978, *Teachers' Attitudes to Usage*, in *Applied Linguistics Association of Australia (ALAA) Occasional Papers* 2, pp. 32-40.