Comparing Explanatory Principles of Complement Selection Statistically: a Case Study Based on Canadian English

By Juho Ruohonen and Juhani Rudanko

Abstract

Several factors have been identified in the recent literature to explain variation in the selection of sentential complements in recent English, and the article begins with a survey of such factors. The article then offers a case study of the impact of such factors on non-finite complements of the adjective *afraid* on the basis of the Strathy Corpus of Canadian English. Attention is paid for instance to the Extraction and Choice Principles, passive lower predicates, and text type. Multivariate analysis is applied to compare and to shed light on such different explanatory principles. The Choice Principle proves to be by far the most significant predictor of the alternation, while the heavily correlated syntactic feature of Voice appears non-significant. Fiction, as opposed to the informative registers, shows a notable preference for *to* infinitives, though this finding needs to be replicated in datasets where controlling for author idiolect is possible. Theoretically plausible odds ratios are observed on the Extraction Principle and negation of the predicate, but they are not statistically significant. In the former case, this may well be due the variable's collinearity with the Choice Principle and its low overall frequency, resulting in a low effective sample size.

1. Introduction

Consider (1a-b), from the Strathy Corpus of Canadian English:

(1a) I was afraid to hang up. (1992, NEWS)
(1b) …Quebeckers are not afraid of going it alone … (1991, NEWS)

In (1a) the adjective *afraid* selects a *to* infinitive as its complement. In (1b) the adjective selects what may be called an *of -ing* complement, consisting of the preposition *of* and a following *-ing* clause, which is a gerund. It is assumed here that each type of complement is sentential and has its own understood or covert subject. The postulation of an understood subject, which is found in traditional grammar (for instance, see Jespersen [1940] 1961: 140) and in much current work, makes it possible to represent the argument structure of the lower verb in (1a-b) in a straightforward fashion. Another property shared by the sentences in (1a-b) is that in both the constructions are control structures, and that they do not involve NP Movement. This follows from the fact that in both sentences the higher subject receives a theta role from the higher predicate. Since the constructions are control structures, the lower subject may be represented by the symbol PRO, which is an abstract pronominal element lacking phonological realization, in accordance with current work in syntax. A further property shared by the sentences in (1a-b) may then be stated by saying that both sentences display subject control. In other words, PRO is controlled by the higher subject in each sentence. The two sentences may be bracketed in their essential aspects as in (1a´) and (1b´).

(1a´) $[[I]_{NP}$ was $[[afraid]_{Adj}$ $[[[PRO]_{NP}$ $[to]_{Aux}$ $[hang\ up]_{VP}]_{S2}]_{AdjP}]_{S1}$
(1b´) $[[Quebeckers]_{NP}$ are not $[[afraid]_{Adj}$ $[[[of]_{Prep}$ $[[[PRO]_{NP}$ [going it alone$]_{VP}]_{S2}]_{NP}]_{PP}]_{AdjP}]_{S1}]$

The bracketing of (1b´) also makes use of the traditional notion of nominal clause, with an NP node dominating the lower clause. This is motivated because while gerundial complements are sentential, they are at the nominal end of the cline of nouniness that characterizes sentential complements (see Ross 2004).

The two types of sentential complements of *afraid* have often been treated under the same sense of the adjective in standard dictionaries. For instance, in the *Shorter Oxford English Dictionary*, they are given under the sense 'frightened, alarmed, in a state of fear' of the adjective. (For a fuller treatment of the sense of the adjective with the two complements, see Rudanko 2014: 225-227.) At the same time, Bolinger's Generalization, according to which a "difference in syntactic form always spells a difference in meaning" (Bolinger 1968: 127), constitutes an invitation to inquire into the meanings and uses of the two sentential complementation patterns and specifically into the factors that bear, or may bear, on the variation in question.

The data of the present article are from the Strathy Corpus of Canadian English (henceforth Strathy), which has been made available by Mark Davies on his Brigham Young University website. One purpose of this study is then to give information on the complementation of the adjective *afraid* in that core variety of

English. However, the main purpose of this study is methodological. It is to examine and to compare the role of different factors bearing on the variation between infinitival and gerundial complements selected by one and the same predicate, using *to* infinitive and *of -ing* complements of *afraid* as a case study, in order to shed light on the salience of the factors in complement selection with the help of statistical analysis.[1] The factors to be examined are introduced in the remainder of this introductory section with illustrations from earlier work, and the statistical analysis on the basis of the corpus considered here is carried out in section 3.

A syntactic factor that has come to be widely accepted in the literature on complementation is the Extraction Principle. The essence of it was formulated by Günter Rohdenburg and Uwe Vosberg in their pioneering work in the late 1990s and 2000s. Vosberg offers a concise definition of the principle as follows:

> In the case of infinitival or gerundial complement options, the infinitive will tend to be favoured in environments where a complement of the subordinate clause is extracted (by topicalization, relativization, comparativization, or interrogation etc.) from its original position and crosses clause boundaries. (Vosberg 2000a: 308; see also Vosberg 2000b: 202)

Two examples from Vosberg's (2003b: 204) work may serve to illustrate the Extraction Principle.

(2a) …protesting that he was only taking me to his brother's farm, which I remember to hear spoken of frequently. (1752, Lennox, *The Female Quixote*)

(2b) …he had moved his free hand to a side pocket, in which he remembered to have some bread and meat. (1854, Dickens, *Hard Times*)

Both *to* infinitives and gerundial complements are selected by the matrix verb *remember*, and in accordance with the Extraction Principle the *to* infinitives in sentences (2a-b) are favored by Relativization, which applies in both (2a) and (2b). In sentence (2a) the gap (or extraction site) linked to the relative pronoun is between *of* and *frequently*, and the example also shows how, in the case of a prepositional complement, the preposition may be left behind in the case of extraction. As for sentence (2b), the gap is at the end of the sentence, and the sentence also shows that it is appropriate to relax the definition of extractions to include the extraction of adjuncts that are part of the predicate (or the VP) of the relevant sentence. (For discussion and illustration of adjuncts in connection of the Extraction Principle, see Vosberg 2006: 69 and Rudanko 2006: 43). In a further important contribution to the study of extractions, Rohdenburg (2016) does not use the term Extraction Principle, but this very recent study confirms its essence on the basis of a detailed discussion of different types of extraction contexts, for the author observes in the conclusion that the *to* infinitive, for which he uses the term "marked infinitive," "enjoys a privileged status in extraction contexts" and that the *to* infinitive outranks "all kinds of gerunds" in such contexts (Rohdenburg 2016: 481).

A second generalization that has been proposed in the recent literature as a

[1] In addition to the two non-finite complements investigated in this article, the adjective *afraid* selects other types of sentential complements, including *that* clauses. However, *that* clauses differ grammatically from the two patterns studied here, because they are finite with expressed subjects, and they deserve a treatment of their own.

factor influencing the selection of *to* infinitival and gerundial complements is the Choice Principle. The principle was defined by Rudanko (2017) as follows:

> In the case of infinitival and gerundial complement options at a time of considerable variation between the two patterns, the infinitive tends to be associated with [+Choice] contexts and the gerund with [–Choice] contexts. (Rudanko 2017: 20)

A [+Choice] context is then defined on the basis of the semantic role of the lower subject. If that subject has the semantic role of Agent, the context is [+Choice], and if the lower subject does not have the Agent role, the context is [–Choice].

The Choice Principle makes crucial use of the theory of semantic roles. As far as the definition of the Agent is concerned, it is probably unrealistic to expect all linguists to agree on any one definition of the concept of Agent, but there is a sufficient degree of consensus for the concept to be used. In broad terms, language may uncontroversially be viewed as the "communicative resource for the definition and enactment of (past, present, and future) realities" (Duranti 2004: 451), and the Agent role is one aspect of the resource.

In more narrowly linguistic terms, Gruber commented on agentive verbs, writing that an "[a]gentive verb is one whose subject refers to an animate object which is thought of as the willful source or agent of the activity described in the sentence" (Gruber 1967: 943). In later work it has been recognized that not only the verb of the sentence but the larger predicate needs to be taken into account when considering the agentivity or otherwise of a subject. Thus Marantz (1984: 24) pointed to predicates such as *throw a baseball* and *throw a fit*, and noted that the direct object of a verb can affect the role of the subject (see also Chomsky 1986: 59-60). The larger predicate is therefore taken into account in the present treatment.

In current work agentivity is also generally viewed as a cluster of features. Lakoff (1977) was probably the first to propose an approach based on features. He proposed as many as 14 features in his discussion of what he called the "prototypical uses" of "prototypical agent-patient sentences." Some of them are less central, including no. 14 on the list ("the agent is looking at the patient"), but three of the features stand out as salient. These are volitionality, control and responsibility. As regards the first, Lakoff's formulation is the "agent's action is volitional," but Dowty in his discussion of what he terms the Agent Proto-Role (Dowty 1991: 572) speaks of "volitional involvement in the event or state," which suggests itself as a more fully developed definition of the feature in question. As regards control, Lakoff writes that the "agent is in control of what he does" (Lakoff 1977: 244), which seems adequate (except for the need for gender neutrality). And as for responsibility, Lakoff writes that the "agent is primarily responsible for what happens (his action and the resulting change)" (Lakoff 1977: 244), which again seems appropriate (except for the need for gender neutrality).

The three features singled out here are also prominent in Hundt's (2004) discussion of the notion of agentivity. They may be supplemented by the consideration of imperatives: an imperative is more natural and more likely with an agentive predicate than with a non-agentive one. To illustrate the distinction, consider (3a-b) from the present dataset:

(3a) … he is afraid to sing it for her. (1988, FIC)
(3b) I was simply afraid to lose my job. (1988, NF)

In sentence (3a) the lower predicate *sing it for her* is agentive, with the predicate encoding an event as volitional on the part of the referent of the subject, under his control and as something that he would be responsible for. Further, an imperative of the form *Sing it for her!* seems entirely natural. The lower predicate of (3b) *lose my job* represents the event in question as lacking these properties.

A third distinction that should be made concerns time period. The Strathy Corpus represents fairly recent English, but it is still helpful to make some chronological division to take at least some account of diachronic change. A fourth possible factor concerns the possible effect of the *horror aequi* principle. This principle has been formulated by Rohdenburg. He writes:

> Very briefly, the *horror aequi* principle involves the widespread (and presumably universal) tendency to avoid the use of formally (near-) identical and (near-)adjacent (non-coordinate) grammatical elements or structures. (Rohdenburg 2003: 236)

In the case of an adjective pattern, the structures of interest from the point of the *horror aequi* principle would concern the form of the verb preceding the adjective, and whether the verb consists of a *to* infinitive or an *-ing* form.

The present authors also wanted to probe the potential influence of insertions, passive subordinate clauses and of negation on complement selection. With respect to negation, negations in the subordinate clause were very rare and had to be excluded from consideration, but negations of the higher predicate were more numerous. Regarding the latter feature, the authors discriminated between *no*-negation, i.e. the use of an adverb, pronoun, or determiner incorporating the negative *n*-element (Tottie 1991), and the more numerous *not*-negation, in order to study their possible differential effects. Lastly, Strathy's ready-made classification of texts into seven different registers suggested itself as another factor worth investigation.

With the different factors identified that may potentially have an influence on complement selection,[2] section 3 is devoted to their statistical comparison.

2. Overview of the Corpus

Compiled at Queen's University and with online access provided by Mark Davies at Brigham Young University, Strathy covers a 90-year timespan from 1921 to 2011. Its texts are classified into the seven registers of Academic, Newspaper, Magazine, Spoken, Fiction, Non-Fiction, and Miscellaneous writing. The Academic and Newspaper registers together make up over half the data, while magazines account for one-fourth. Strathy may therefore be considered a collection of fairly formal texts. Table 1 shows the distribution of the registers in

[2] An anonymous reviewer suggests that transitivity, as discussed by Hopper and Thompson (1980), might also be considered as a factor impacting the choice between the two types of complement. The two authors define transitivity as a cluster of ten features, some of which (such as volitionality and agency) overlap with the Choice Principle, while others (such as the necessary presence of two participants) do not. The reviewer's suggestion would presumably include the idea that higher transitivity correlates with higher selection rates for the *to* infinitive. Putting this hypothesis to the test would have to begin with a thorough discussion of how the concept of transitivity is to be operationalized for statistical analysis. Satisfactory treatment of this question would merit a study of its own, and the present authors defer such undertakings to future work.

the six time periods constituting the corpus.

Along the diachronic dimension, over 80% of Strathy's data are from 1990 or later, while over 90% are no older than 1980. Owing to this heavy concentration of the data within a fairly short time window and the relative sparseness of material from earlier decades, caution must be exercised when interpreting results on diachronic effects in this corpus.

	1920s-1940s	1950s-1970s	1980s	1990s	2000s	2010s	Total
Spoken	0	0	0	94,527	5,592,381	187,689	5,874,597
Fiction	1,739,983	329,263	506,611	860,022	452,736	12,766	3,901,381
Magazines	0	0	1,388,416	2,185,009	6,359,030	55,358	9,987,813
Newspaper	0	0	835,569	1,805,388	9,948,930	510,807	13,100,694
Nonfiction	761,739	172,617	735,567	822,731	2,728	0	2,495,382
Academic	125,134	193,961	1,996,289	2,523,454	9,575,865	230,650	14,645,353
Misc	0	0	49,437	0	26,620	0	76,057
Total	2,626,856	695,841	5,511,889	8,291,131	31,958,290	997,270	**50,081,277**

Table 1: Strathy's registers and time periods cross-classified.

3. Analysis of the Data

3.1. Descriptive Statistics

To collect *to* infinitive complements of the adjective *afraid*, the present authors used the basic search string "afraid to *," and to collect *of -ing* complements, they used the basic search string "afraid of *ing." The former search retrieves 392 tokens. Additional searches were conducted for strings with one or two words between *afraid* and *to.* These supplementary searches retrieved another ten tokens in all. The total at this point was 402. One of these was dropped as a duplicate of another token. Among the remaining tokens there are 8 tokens that the present investigators have classified as indirect complements or degree complements, to use a designator from Baltin (2006). An example is given in (4).

(4) The problem traditionally with trying to prosecute a pimp is that prostitutes are too afraid to testify. (1993, MAG)

The *to* infinitive in (4) is licensed by the degree modifier *too* in front of the adjective, not by the adjective.[3] With these exclusions, the remaining total is 393

[3] The present investigators did not regard all *to* infinitives in strings of the form "*too afraid to* Verb …" as degree complements. Consider the example in (i), given with some context:

(i) Those who dare to take liberties will go highest in the dance.... Tradition is not enough.... Here there is this bright country but people are too afraid to try, too afraid to seem foolish. (1996, MAG)

The present investigators view the *to* infinitive in (i), and in three other analogous sentences, as a complement of *afraid*, rather than as a complement of the degree modifier. The presence of a degree word in front of an adjective does not make an overt degree complement obligatory, and in (i) a degree complement is covert and can be understood from the context, being of the type "too afraid of trying or seeming foolish to take liberties." Analogous considerations apply in the other three sentences.

tokens.

As regards *of -ing* complements, the number of tokens retrieved is 121, including five with insertion(s) between *afraid* and the preposition. Among this total, there are 13 tokens where the complement is clearly nonsentential, of the type *afraid of something*, and these can be dropped without further discussion. This leaves us with 108 tokens.

Figure 1 shows the normalized frequencies of the two variants for each time period listed in Table 1. It appears that the overall frequency of non-finite complements of *afraid* is declining. This downward trend may conceivably be due to increasing competition from near-synonymous predicates such as *scared* and *terrified* (Rickman and Rudanko 2018: 15-52).[4]

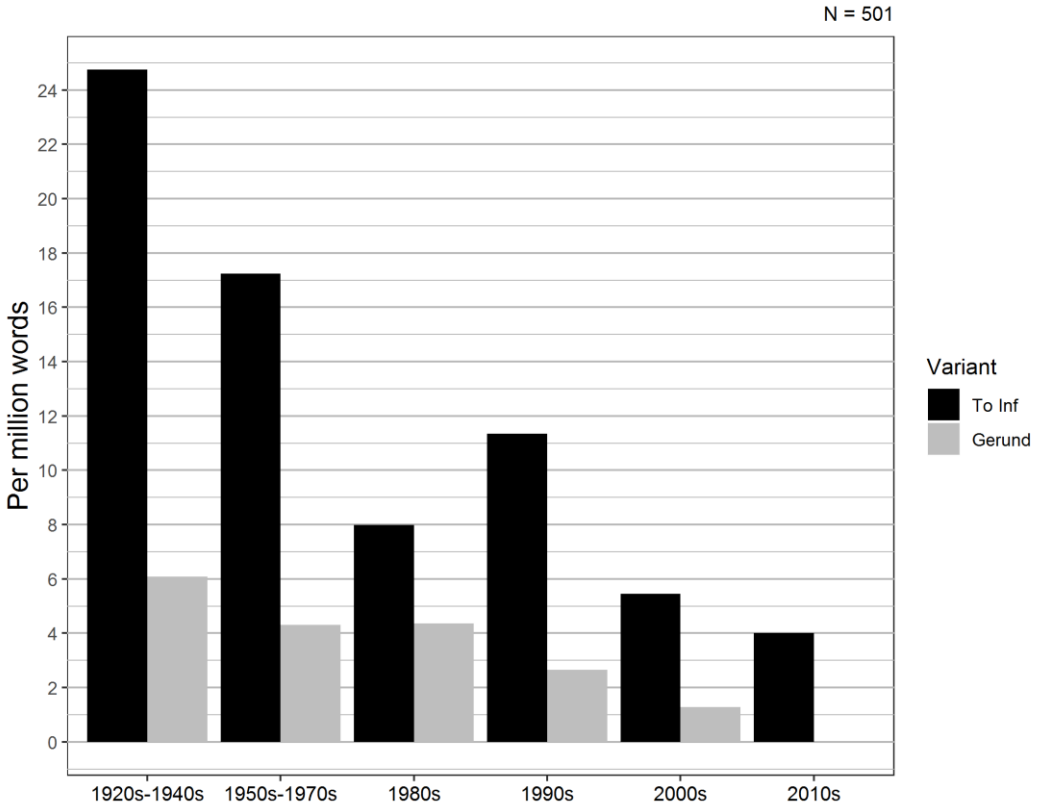

Figure 1: Normalized frequencies of the two variants in Strathy by time period.

3.2. Choice, Voice, and the Limitations of Univariate Analyses

To illustrate the very strong correlation between [±Choice] and the type of non-finite complement selected after *afraid*, we begin with a traditional analysis using Pearson's χ^2 test of independence,[5] seen in Table 2. Consistent with previous

	[+Choice]	[-Choice]	Total
to infinitives	369	24	393
of -ing	27	81	108

Table 2: Contingency table of Choice and type of non-finite complement after *afraid* in Strathy.

[4] An anonymous reviewer suggests that another potential explanation is an increase in the use of finite complements with overt subjects coreferential with the higher subject.

[5] We use Yates' correction in Pearson's χ^2 test throughout this article.

studies, the correlation is dramatic with χ^2 = 238.6 (df = 1) and p < .001. This correlation, however, is not yet proof that [±Choice] exactly is the causal factor behind the selection pattern. Previous work (Rudanko 2015: 41-48) has indicated that Voice is also strongly associated with the same variation. To illustrate, consider examples (5a-c), where (5b-c) involve [-Choice] contexts:

(5a) He walloped me for fair last week, and I was afraid to hit back. (1925, FIC)
(5b) He was afraid to get whipped. (1966, FIC)
(5c) I haven't told anybody because I was afraid of being rooked. (1936, FIC)

The strong correlation between the passive and *of -ing* is also true of our dataset, as seen in Table 3 below. χ^2 equals 53.35 (df = 1) with p-value < .001.[6]

	Active	Passive	Total
to infinitives	390	90	480
of -ing	3	18	21

Table 3: Contingency table of Voice and type of non-finite complement after *afraid* in Strathy.

The similarity of the respective correlations of Choice and Voice with complement selection is unsurprising, since the *raison d'etre* of the passive is to topicalize the patient (that is, the prototypically unagentive participant) of a transitive clause by promoting it from object to subject. This three-way correlation poses an insoluble problem for univariate analyses. Since [±Choice] and Voice are both strongly associated with complement selection and since they are also strongly associated with each other, a univariate analysis cannot disambiguate whether it is in fact [±Choice] or Voice that bears more significantly on complement selection. This is where multivariate analysis can help.

3.3. Multivariate Analysis – Preliminaries

We used the *lme4* library (Bates *et al*) in R (v.3.4.4) to fit mixed-effects logistic regression models (Hedeker and Gibbons 2006: 149-162) to the data. Broadly speaking, regression enables us to answer the following question: what is the value of each explanatory variable in predicting the outcome when we already know the value of every other explanatory variable (McElreath 2016: 123)? In our case, the outcome of interest is the choice between *of -ing* and the *to* infinitive.

Since we are modeling a probability, which is necessarily constrained to the [0,1] interval, we cannot model it directly. This is because all regression models fall short of predicting the outcome perfectly – their predictions necessarily have varying amounts of error in them. With many observations and explanatory variables, using the probability scale to describe effects would have the consequence that sooner or later, the model would inevitably predict a value outside the possible range. This is circumvented by first converting the probability into odds, then taking the natural logarithm of those odds. This *log odds* ranges between -∞ and ∞, yet transforming it back into a probability always

6 Pearson's χ^2 test is arguably not ideal for these data due to the low number of Active *of -ing* complements. A more suitable approach is arguably Fisher's exact test, which also yields p < .001.

yields a value in the [0,1] range. This enables the effect coefficients to be unrestricted while ensuring that the model cannot predict impossible values. Figure 2 illustrates how probability maps onto the log odds scale and vice versa. Log odds also happens to equal the quantile of the corresponding probability in the logistic cumulative distribution function, motivating the name of the method.

Coefficients in logistic regression describe effects on the probability of "success" on the log odds scale. Exponentiating the coefficient of an explanatory variable yields an odds ratio, i.e. the estimated multiplicative change in the odds of "success" corresponding to a one-unit change in the value of the explanatory variable, *ceteris paribus*. Some find odds ratios easier to understand than the log odds scale. Odds ratios between 0 and 1 reduce the estimated probability of the outcome, while ones greater than 1 increase it.

Lastly, the term 'mixed' refers to the possibility of including both fixed and random effects in the model. A fixed effect is any explanatory or confounding variable for which sufficient data are available to perform unbiased maximum likelihood estimation (Agresti 2015: 138-143).

By contrast, a random effect is typically a nuisance variable with a large number of discrete categories. The best linguistics example is idiolect, which is usually known or suspected to cause variability in the outcome but cannot be

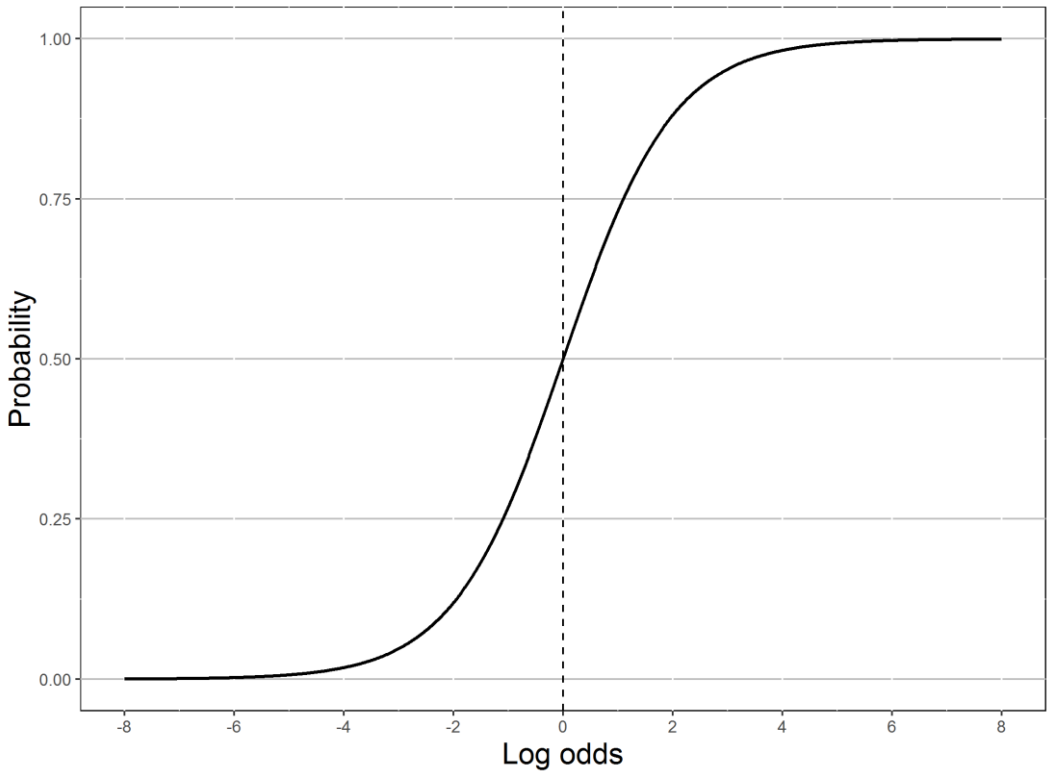

Figure 2: Probability as a function of log odds.

quantified for inclusion in the model as a fixed, numeric predictor. Including all its discrete categories as fixed effects may increase the number of parameters in the model to such a degree that the regularity conditions of maximum likelihood estimation are no longer met, resulting in inflated estimates (Breslow and Day 1980: 249-250). Treating it instead as a random effect reduces this bias by assuming a common (typically normal) distribution for the random effects, shrinking their estimates towards their shared overall mean in inverse proportion

to their respective sample sizes (Agresti 2007: 302-304).

Strathy unfortunately lacks speaker/writer information for the most part, so idiolect cannot be controlled for. Another potential source of variation is the identity of the subordinate verb (Baayen 2008: 295-300; Hämäläinen 2002: 351; Levshina 2016: 252-253). It is conceivable that some verbs may be likelier than others to occur in one or the other construction after *afraid*, possibly as its collexemes (Stefanowitsch and Gries 2003).

3.4. Insertions, Horror Aequi, No-Negation, and Register

Only six insertions were found.[7] Two examples are shown in (6a-b):

(6a) I saw too how ravenously she ate, how afraid she was to accept kindness, how distrustful of coaxing. (1944, NF)
(6b) Our Lord was afraid of this, not afraid merely of dying. (2001, SPOK)

It just so happens that all four *to* infinitives with insertions are [+Choice] and the two *of -ing* examples [-Choice]. In other words, all six tokens are perfectly predicted by the Choice principle, thus containing little information on the influence or lack thereof of insertions. Consequently, insertion was excluded from consideration as an explanatory variable after verifying that ignoring it did not cause confounding with any of the remaining variables (Hosmer *et al* 2013: 92).

Potential *horror aequi* contexts numbered 13. Two are shown below:

(7a) In terms of getting ahead in academia, you have to publish and write things. That means not to be afraid to write and put things down. (2004, ACAD)
(7b) How much were you distressed by feeling afraid to go out of your house alone? (2004, ACAD)

In (7a) *afraid* is immediately preceded by a *to* infinitive, yet still governs a *to* infinitive, contrary to what *horror aequi* would predict. By contrast, example (7b) does accord with *horror aequi*, given that the adjective is preceded by an *-ing* form and proceeds to select a *to* infinitive. However, (7a-b) and all other observations with a preceding *to* infinitive or verbal *-ing* form were perfectly predicted by [±Choice], apparently overriding what little effect prior context might otherwise have had. *Horror aequi* was thus also excluded from the ensuing multivariate analysis after checking that this had no appreciable effect on the other coefficients.

There were 15 instances of *no*-negation of the matrix predicate. They all combined with an infinitive, however, 14 of these infinitives were also predicted by the Choice Principle. Consequently, the *to* infinitive seen in (8) below was the only observation containing substantive statistical information on the effect of *no*-negation.

(8) They should never be afraid to find themselves alone because they have said what they believed to be true[.] (2000, SPOK)

[7] Constructions with one word or two words between *afraid* and the following *to/of* were regarded by the present authors as insertions.

Since the only statistically informative *no*-negation token was an infinitive, the resulting effect estimate was an infinite coefficient favoring infinitives and causing severe model convergence problems. One might therefore be tempted to regard *no*-negations as "categorical contexts" which should be discarded from the dataset (Tagliamonte 2006: 86-87). Based on work conducted with a larger corpus (Ruohonen and Rudanko, under review), however, we do not believe that *no*-negated predicates constitute a categorical context. We believe them to be simply another factor with some probabilistic bias for the *to* infinitive. We therefore pooled the 15 *no*-negation cases together with unnegated tokens after confirming that doing so did not cause confounding for the remaining variables.

Complement negation, of which the only example encountered is seen in (9a) below, was also excluded from statistical consideration. After these exclusions, the only type of negation included in the multivariate analysis was *not*-negation of the higher predicate, exemplified by (9b) below:

(9a) I witnessed a brutal beating being inflicted by one schoolboy on another, so savage we were afraid not to stop and intervene. (1993, NEWS)
(9b) I'm not afraid to die, but I want to live a while longer to help Tim. (1921, FIC)

Register was initially entered as a nominal-scale variable retaining the 6 distinct categories that occurred in the dataset, i.e. all except Miscellaneous. In preliminary model fitting, however, most of the category contrasts turned out to be of negligible predictive value. Retaining all of them was weakening inference on the other variables by introducing what seemed to be statistically superfluous distinctions. The one category distinction that showed promise, however, was that between fiction and everything else. We therefore simplified the six-way division into a dichotomy between Informative and Imaginative texts, the former constituting a conflation of all non-fiction categories and the latter representing fiction, now relabeled. A likelihood ratio test comparing a model with the full set of categories to one using only the dichotomy returned a χ^2 statistic of 1.06 (df = 4) and a p-value of .9, strongly indicating that the additional category distinctions did not significantly improve the fit.[8] Thus, the model described in the ensuing sections utilizes the binary classification for register.

3.5. Multivariate Analysis – Results

Since it is unmistakably the more marked alternative with *afraid*, we treated *of -ing* as the "success" outcome from the perspective of the regression. We fit the following model:

Fixed effects:

1. [±Choice] (dichotomous)
2. Voice (dichotomous)
3. Extraction (dichotomous)
4. *Not*-negation of Predicate (dichotomous)
5. Register (dichotomous, Informative or Imaginative)
6. Year (continuous, centered around its mean of 1987 and divided by 10)

[8] All models were fit using the default Nelder-Mead optimization algorithm with the number of adaptive Gauss-Hermite quadrature points set to 20, unless explicitly stated otherwise.

Random effects:

1. Subordinate verb

A simple and intuitive measure of model fit with binary outcomes, the *concordance index* results from first forming every possible pair of two observations where one has a "success" and the other a "failure", then calculating the proportion of these pairs that are concordant, i.e. with the "success" observation having the higher estimated probability (Agresti 2015: 172). This statistic equals .937 for our model, constituting "outstanding" discrimination of successes and failures (Hosmer *et al* 2013: 177).

When interpreting regression results, it is imperative to distinguish between effect size and statistical significance. Effect size quantifies a variable's estimated effect on the outcome and is therefore the parameter of primary interest. Statistical significance, by contrast, quantifies our degree of confidence that the observed effect is not due to mere chance. This confidence (or lack thereof) depends to a large extent on effective sample size. Even minuscule effects are statistically significant if backed by enough data, while even strong effects will fail to reach statistical significance if backed by too little data. Figure 3 displays the fixed effects' estimated effect sizes on the log odds scale. The 95% confidence intervals around the point estimates are directly analogous to p-values.[9] An interval spanning zero implies p > .05. The farther the confidence interval lies from zero, the smaller the p-value.

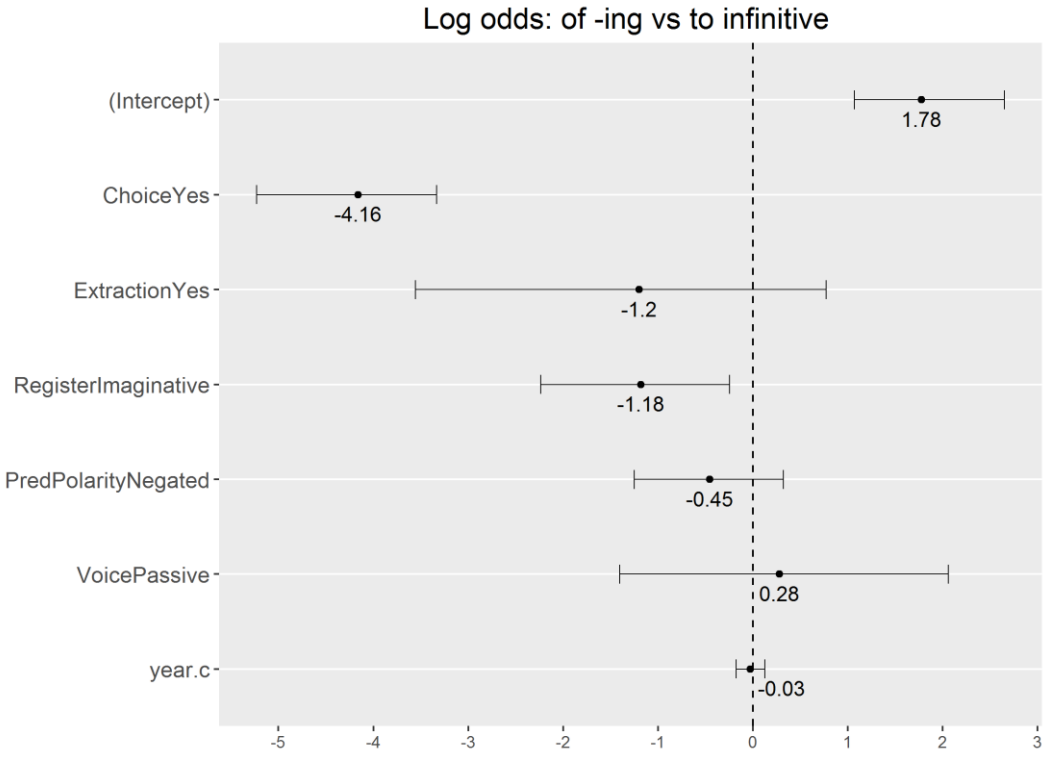

Figure 3: Estimated effects on the log odds of *afraid* selecting *of -ing.*

[9] These are profile likelihood confidence intervals, which are more accurate than standard Wald intervals but much more CPU-intensive to calculate. See Hosmer *et al* (2013: 15-20) for details. Likewise, all p-values reported in this section are based on the likelihood ratio test, which is known to be more reliable than the Wald test (Agresti 2013: 174-175).

The Intercept reflects that with all the variables at their baseline, i.e. [-Choice], no extraction, an informative register, an unnegated predicate, active voice, and year at its mean of 1987, the estimated odds of *of -ing* relative to the *to* infinitive are $e^{1.78} = 5.93$. This corresponds to a probability of about 85%. With everything else equal, [+Choice] contexts are estimated to have only $e^{-4.16} = .015$ times the odds of selecting *of -ing* that [-Choice] contexts do. For the example case above, this translates to a reduction of the probability of the gerundive complement to just 8%. This is by far the largest and most statistically significant effect in the model ($p < .001$). By contrast, when we already know the value of [±Choice] and the other variables, Voice is estimated to exert virtually no influence on variant selection. It appears that [±Choice], not Voice, is the decisive factor in determining which variant is used.

Consistent with the indications of previous literature, extraction contexts are estimated to improve the odds of the *to* infinitive by a factor of $e^{1.2} = 3.3$.[10] The effect is not statistically significant ($p = .25$), which requires explanation. There are a total of 14 extraction contexts in the dataset. 11 of them have a *to* infinitive. However, all of these 11 are also [+Choice], thus containing little information about the independent effects of either principle. In only 3 tokens, seen in (10a-c), do the two principles clash:

(10a) What she had been afraid of witnessing did not occur. (1930, FIC)
(10b) [D]ependency on others and physical and/or mental disabilities that we as individuals are afraid of having risk being translated as signifying a lack of worth[.] (2004, ACAD)
(10c) "Where is the life you are so afraid to lose?" (1982, ACAD)

In (10a) and (10b), [-Choice] overrides the Extraction Principle, yielding *of -ing*.[11] In (10c), the Extraction Principle overrides [-Choice], yielding an infinitive. This may be an indication that both principles are valid but [±Choice] tends to take precedence where they conflict. Though such a hypothesis seems plausible enough, the observations are far too few to constitute statistical evidence. The high p-value reflects this uncertainty. Due to the overall rarity of extraction contexts and the observed collinearity of the two variables, we would simply need much more data to reliably disentangle the effects of extraction from those of [±Choice].

With all else equal, the odds of *afraid* selecting a *to* infinitive in the imaginative register are estimated $e^{1.18} = 3.2$ times the odds in the informative register. Our first suspicion was that perhaps some specific author's idiosyncratic style was simply overrepresented among the fiction tokens, causing a spurious association between that register and the *to* infinitive. Indeed, the fiction data do contain clusters of observations from one and the same novel. However, based on what we could ascertain using the limited author information available, only two fiction examples with an unagentive *to* infinitive could be plausibly attributed to idiolect:

(11a) "He was afraid to get whipped." (FIC, 1966, Robert Kroetsch)

[10] Changing the sign of the log odds yields an interpretation of the effect with the respective roles of the "success" and "failure" outcome reversed.
[11] The predicate *witness* (*something*) can be agentive, but in the present example it is unagentive, with the sense of 'be a witness of' or 'see'.

(11b) "He was afraid to be a fool. So he was a coward instead." (FIC, 1966, Robert Kroetsch)

Absent a better explanation, we acknowledge that the register effect may be legitimate. However, it must be replicated in other datasets before we can conclude that it is not just a statistical fluke of this specific sample.

There appears to be no appreciable diachronic trend. Before modeling year as a continuous variable, we algorithmically fitted models with every possible diachronic cutoff point to see whether the dataset could be diachronically split in a way that made theoretical sense and improved the fit. It turned out that modeling diachrony as a dichotomy between pre-2005 and newer data did indeed improve the fit, and the new variable had a rather impressive coefficient of -1.3. However, the effect is due entirely to four unexpected *to* infinitives[12] occurring in 2005:

(12a) Loath to be left alone, he may have argued that his bones were accustomed to being in motion, and that he was afraid to be left in the sedentary silence of the grave. (ACAD, 2005)
(12b) Patients are afraid to die alone and families may feel as if they have abandoned their loved ones at the time of their death. (ACAD, 2005)
(12c) In Canada we witnessed the situation when politicians are not afraid to be made fun of. (MAG, 2005)
(12d) The one – how do you say it, don't be afraid to die but don't do something stupid to bring it on faster. (2005, SPOK)

In addition, we ran a similar algorithm to fit a separate model for every one of the 1,325 possible ternary diachronic divisions.[13] Some of these models achieved further statistically significant improvements in fit over the model with the binary split at 2005, but the implied diachronic effects were even less plausible. These models suggested either significant back-and-forth developments in the early decades where data were sparse, or they pointed to a major spike in *to* infinitive rates from 1995 to 2000 followed by a partial reversal afterwards. We are skeptical of any true non-linear diachronic effects, ternary or binary. We are aware of no documented or anecdotal shifts in recent usage towards an increased preponderance of *to* infinitives. We suspect such effects to be random fluctuations peculiar to this dataset, whose inclusion as meaningful predictors would constitute overfitting, i.e. "capitalizing on chance" (Fox 2016: 690).

Lastly, *not*-negation of the predicate appears to have fairly little effect on the choice between the *to* infinitive and *of -ing*. Its odds ratio of $e^{0.45} = 1.57$ is favorable to the infinitive, but the effect is not statistically significant in this dataset.

[12] Unexpected from the standpoint of the Choice Principle. While the examples in (12a-d) – and those in (11a-b) – are unexpected from the standpoint of the Choice Principle, this does not of course mean that they are ungrammatical or unidiomatic in any way. (The Choice Principle expresses a tendency and it is not a categorical rule.) As regards the semantic interpretation of sentences of the type of (12a) and (12c), see the comments in Rickman and Rudanko (2018: 64).
[13] Getting all 1,325 models to converge required Bayesian methods. We used the *brms* package (Bürkner 2017), setting a uniform prior between -5 and 5 for the fixed effects and an exponential prior with a rate parameter of 1 for the random-effect standard deviation. We used four Markov chains with 1,000 warmup iterations and 1,000 sampling iterations each.

3.6. Verb-Specific Effects?

Thankfully, the identity of the lower verb seems to exert only a very minor influence on the choice of non-finite complement. The model's Intraclass Correlation (Hosmer *et al* 2013: 327), which estimates the proportion of the model's total explanatory power that is due to verb-specific idiosyncrasies, is only 8.6%. The p-value is .26.[14]

Get exhibits what is perhaps the most notable verb-specific effect, favoring *of -ing* at an odds ratio of $e^{0.74} = 2.1$. Examples (13a-b) show *get* defying the Choice Principle:

(13a) I'm afraid of getting off HRT because of the headaches and I worry about my bones. (2005, MAG)
(13b) Afraid of getting ahead and hoping foolish hopes, of getting close enough to think they might have a chance. (1983, NF)

Another verb appearing to favor *of -ing* is *fly*, whose estimated odds ratio in favor of the gerundial complement is $e^{0.65} = 1.9$. This is largely due to the three occurrences (out of a total of five) of the phrase *afraid of flying*. Lastly, something vaguely reminiscent of a fixed phrase is seen in *afraid to die.* The semantics of *die* would lead us to expect *of -ing* to occur almost categorically. There are many counterexamples in our data, two of which are shown below:

(14a) They think they'll win because they're not afraid to die. (2001, NEWS)
(14b) [S]he was not fit to live but was afraid to die. (1935, NF)

Die is estimated to favor the infinitive at an odds ratio of $e^{0.59} = 1.8$.

4. Concluding Remarks

This study has explored the effects of six variables on the choice between two non-finite complement types after the adjective *afraid* in the Strathy Corpus of Canadian English. We began by demonstrating the strong univariate association between [±Choice] and complement type. We then proceeded to carry out a multivariate analysis of the influence of [±Choice] and five other variables on this binary outcome, using mixed-effects logistic regression.

Arguably the most important result has been the disentanglement of the effects of [±Choice] and Voice. As a syntactic operation that promotes transitive patients to subjects, the passive is almost perfectly correlated with [-Choice]. It was therefore far from clear, a priori, what roles the two variables played in complement selection. Our analysis provides strong evidence that the Choice Principle is indeed the main operative factor between the two, while Voice alone is relatively inconsequential in the variation concerned.

We found little evidence of a linear diachronic effect. Regarding non-linear diachronic effects, exhaustive experimentation with different diachronic cut-points enabled us to create the appearance of statistically significant diachronic threshold(s) around the turn of the millennium – with newer material appearing to favor *to* infinitives. However, we felt it unwise to read much into this

[14] Since the variance of a random effect cannot be negative, we divided the raw p-value of this likelihood ratio test by two. See Agresti (2018: 278) for a more detailed explanation.

phenomenon. As we already alluded in Section 2, the sparseness of data from the decades leading up to 1980 make Strathy less than ideal for rigorous diachronic analysis.

Our dataset displayed a mysterious tendency for imaginative texts to favor the *to* infinitive. Though the effect seemed fairly robust in terms of statistical significance, we urge caution in its interpretation. It is particularly important to note that the design of the corpus did not enable us to control for idiolectal variation in complement choice. Fiction texts are characteristically long, and it is well within the realm of possibility that the effect could be a consequence of a small number of overrepresented idiolects in the corpus. Future studies seeking to replicate the finding must take account of this possibility.

The nature, if not magnitude, of our findings on negated predicates and extraction contexts was broadly in line with what previous research has suggested, but neither factor was statistically significant. Negated predicates showed a subtle and statistically non-significant tendency to favor *to* infinitives. Extraction contexts showed signs of a stronger association, but their low overall frequency and their collinearity with [+Choice] in our dataset prevented statistically substantive inferences from being made.

Causation is harder to prove than correlation. Disentangling the independent contributions of several different variables requires much more data than obtaining a significant χ^2 statistic in a univariate crosstabulation. Multicollinearity among explanatory variables is the rule rather than the exception in observational datasets, which is what corpora essentially are. This greatly increases the amount of data required. Furthermore, the quantitative bottleneck to obtaining statistically significant results in multivariate analyses of a binary outcome is often the frequency of the rarest outcome rather than the overall sample size (Hosmer *et al* 2013: 408). This is the case for our study. Despite the reasonable size of the dataset, the low overall frequency of *of -ing* severely limits the amount of statistical inference that can be conducted. Fortunately, larger and more up-to-date contrastive corpora have recently become available (Davies 2013), so many options exist for those seeking to either replicate our results, disprove them, or confirm the ones that had to be declared speculative and preliminary due to the limited data size.

Juho Ruohonen
University of Helsinki

Juhani Rudanko
University of Tampere

Bibliography

Agresti, Alan. 2007. *An introduction to categorical data analysis* (2nd ed.). Hoboken: Wiley-Interscience.

Agresti, Alan. 2018. *An introduction to categorical data analysis* (Third edition.). Hoboken, NJ: John Wiley & Sons.

Agresti, Alan. 2013. *Categorical data analysis* (3rd ed.). Hoboken, N.J. : Chichester: Wiley ; John Wiley [distributor].

Agresti, Alan. 2015. *Foundations of linear and generalized linear models*. Hoboken, New Jersey: John Wiley & Sons.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baltin, Mark. 2006. Extraposition. In Martin Everaert & Henk van Riemsdijk (eds.). *The Blackwell companion to syntax*, volume II, 237–271. Malden, MA.: Blackwell.

Bolinger, Dwight. 1968. Entailment and the meaning of structures. *Glossa* 2:119–127.

Breslow, N E. & N. E. Day. 1980. *Statistical Methods in Cancer Research. Vol 1: The Analysis of Case-Control Studies.* International Agency on Cancer, Lyon, France.

Chomsky, Noam. 1986. *Knowledge of language. Its nature, origin, and use*. New York: Praeger.

Davies, Mark. 2013 *Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day*. Available online at https://corpus.byu.edu/now/

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67 (3): 547–619.

Duranti, Alessandro. 2004 Agency in language. In Alessandro Duranti (ed.) *A Companion to Linguistic Anthropology*. Malden, Mass.: Wiley-Blackwell, 451–473.

Fox, John. 2016. *Applied regression analysis and generalized linear models* (Third edition.). Los Angeles: SAGE Publications, Inc.

Gruber, Geoffrey. 1967. Look and see. *Language* 43 (4): 937–948.

Hämäläinen, Taina. 2008. *Espanjan kielioppi*. Helsinki: Finn Lectura.

Hopper, Paul J. & Sandra Thompson. 1980. Transitivity in grammar and discourse. *Language* 56 (2): 251–299.

Hosmer, David W., Stanley Lemeshow, & Rodney X. Sturdivant. 2013. *Applied logistic regression* (3rd ed.). Hoboken, N.J.: Wiley.

Hundt, Marianne. 2004. Animacy, agentivity, and the spread of the progressive in Modern English. *English Language and Linguistics* 8 (1): 47–69.

Jespersen, Otto. 1940. *A modern English grammar on historical principles. Part V Syntax*, Volume IV Reprinted 1961. London and Copenhagen: George Allen & Unwin/Ejnar Munksgaard.

Lakoff, George. 1977. Linguistic gestalts. In A. B Woodford, Samuel Fox & Shulamith Philosoph (eds.). *Papers from the thirteenth regional meeting of the Chicago Linguistic Sociery*. Chicago: Chicago Linguistic Society, 236–287.

Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2): 235–268.

McElreath, Richard. 2016. *Statistical rethinking: a Bayesian course with examples in R and Stan*. Boca Raton: CRC Press/Taylor & Francis Group.

Marantz, Alec. 1984. *On the nature of grammatical relations.* Cambridge, Mass.: The MIT Press.

Rickman, Paul. & Juhani Rudanko, J. 2018. *Corpus-based studies on non-finite complements in recent English*. Houndmills, Basingstoke: Palgrave Macmillan UK.

Rohdenburg, Günter. 2003. Cognitive complexity and *horror aequi* as factors determining the use of interrogative clause linkers in English. In Günter Rohdenburg & Britta Mondorf (eds.). *Determinants of grammatical variation in English*. Berlin: Mouton de Gruyter, 205–249.

Rohdenburg, Günter. 2016. Testing two processing principles with respect to the extraction of elements out of complement clauses in English. *English Language and Linguistics* 20: 463–486.

Ross, John Robert. 2004. Nouniness. In Bas Aarts, David Denison, Evelyn Keizer & Gergana Popova (eds.). *Fuzzy grammar: a reader*. Oxford: Oxford University Press, 351–422. Originally published as Ross, John Robert. 1973. Nouniness. In Osamu Fujimura (ed.). *Three Dimensions of Linguistic Research*. Tokyo: TEC Company, 137–257.

Rudanko, Juhani. 2006. Watching English grammar change. *English Language and Linguistics* 10 (3), 31–48.

Rudanko, Juhani. 2014. A new angle on infinitival and *of -ing* complements of *afraid* with evidence from the TIME Corpus. In Kristin Davidse, Caroline Gentens, Lobke Ghesquière & Lieven Vandelotte (eds.). *Corpus Interrogation and Grammatical Patterns*. Amsterdam: John Benjamins, 23–238.

Rudanko, Juhani. 2015. *Linking form and meaning: studies on selected control patterns in recent English*. Basingstoke: Palgrave Macmillan.

Rudanko, Juhani. 2017. *Infinitives and gerunds in recent English*. London: Palgrave Macmillan Springer.

Ruohonen, Juho & Juhani Rudanko. Semantics, syntax, and *horror aequi* as

predictors of non-finite alternation: a multivariate analysis of clausal complements of *afraid* based on the NOW Corpus. Manuscript submitted for publication.

Shorter OED = The New Shorter Oxford English Dictionary on Historical Principles. 1993. Edited by Lesley Brown. Oxford: Clarendon Press.

Stefanowitsch, Anatoli & Stefan Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Tottie, Günnel. 1991. Lexical diffusion in syntactic change: frequency as a determinant in the development of negation in English. In: Dieter Kastovsky (ed.). *Historical English Syntax*. Berlin: Mouton/de Gruyter, 439–467.

Vosberg, Uwe. 2003a. The role of extractions and horror aequi in the evolution of *-ing* complements in Modern English. In Günter Rohdenburg & Brita Mondorf (eds.). *Determinants of grammatical variation in English*. Berlin: Mouton de Gruyter: 305–327.

Vosberg, Uwe. 2003b. Cognitive complexity and the establishment of *-ing* constructions with retrospective verbs in Modern English. In Marina Dossena & Charles Jones (eds). *Insights into Late Modern English*. Bern: Peter Lang, 197–220.

Vosberg, Uwe. 2006. *Die grosse Komplementverschiebung*. Tübingen: Narr.

Software Used

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. *Fitting Linear Mixed-Effects Models Using lme4*. Journal of Statistical Software, 67(1), 1–48. doi:10.18637/jss.v067.i01.

Bürkner, Paul-Christian. 2017. brms: an R package for advanced Bayesian Multilevel Models using Stan. *Journal of Statistical Software*, 80(1), 1–28. doi:10.18637/jss.v080.i01

Harrell, Frank E, with contributions from Charles Dupont & many others. 2018. *Hmisc: Harrell Miscellaneous*. R package version 4.1-1. https://CRAN.R-project.org/package=Hmisc

R Core Team 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

Corpora Consulted

Strathy Corpus of Canadian English. 2013. Created by the Strathy Language Unit at Queen's University. Available at: https://corpus.byu.edu/can/