



Article

Ahead of Print

Aleksandrs Berdicevskis*, Evie Coussé, Alexander Koplenig and Yvonne Adesam

To drop or not to drop? Predicting the omission of the infinitival marker in a Swedish future construction

<https://doi.org/10.1515/cllt-2022-0101>

Received November 18, 2022; accepted April 5, 2023; published online May 5, 2023

Abstract: We investigate the optional omission of the infinitival marker in a Swedish future tense construction. During the last two decades the frequency of omission has been rapidly increasing, and this process has received considerable attention in the literature. We test whether the knowledge which has been accumulated can yield accurate predictions of language variation and change. We extracted all occurrences of the construction from a very large collection of corpora. The dataset was automatically annotated with language-internal predictors which have previously been shown or hypothesized to affect the variation. We trained several models in order to make two kinds of predictions: whether the marker will be omitted in a specific utterance and how large the proportion of omissions will be for a given time period. For most of the approaches we tried, we were not able to achieve a better-than-baseline performance. The only exception was predicting the proportion of omissions using autoregressive integrated moving average models for one-step-ahead forecast, and in this case time was the only predictor that mattered. Our data suggest that most of the language-internal predictors do have some effect on the variation, but the effect is not strong enough to yield reliable predictions.

***Corresponding author: Aleksandrs Berdicevskis**, Språkbanken Text, Department of Swedish, Multilingualism, Language Technology, University of Gothenburg, Gothenburg, Sweden, E-mail: aleksandrs.berdicevskis@gu.se

Evie Coussé, Department of Languages and Literatures, University of Gothenburg, Gothenburg, Sweden, E-mail: evie.cousse@gu.se

Alexander Koplenig, Department of Lexical Studies, Leibniz-Institute for the German Language (IDS), Mannheim, Germany, E-mail: koplenig@ids-mannheim.de

Yvonne Adesam, Språkbanken Text, Department of Swedish, Multilingualism, Language technology, University of Gothenburg, Gothenburg, Sweden, E-mail: yvonne.adesam@gu.se

Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

Keywords: corpus; language change; language variation; logistic regression; predictive approach; Swedish; time-series analysis

1 Introduction

Linguists attempt not only to describe language variation and change, but also to explain it, assuming that we understand at least some of the mechanisms behind it (Croft 2000; Labov 1994, 2001, 2011; McMahon 1994). However, our explanations of change are necessarily post-hoc, vulnerable to the “I knew it would happen” fallacy, also known as “hindsight bias” (Fischhoff and Beyth 1975). Measuring the *predictive* accuracy of theories of language change would constitute a more rigorous test of their explanatory power.

The same is true about explaining synchronic variation, that is, looking for factors that can potentially affect it. A common way of performing a quantitative test of the role of certain factors is to fit a statistical model and then to estimate how large the observed effects are and whether they are significant. There is, however, yet another important property of a model: how well it can predict the variation, given the factors (Bresnan et al. 2007; Koplenig 2019; Theijssen et al. 2013).

In this study, we adopt a predictive approach to explore the role of different factors in the usage of one Swedish future tense construction (consisting of the auxiliary *kommer* ‘come’, the optional infinitival marker *att* and an infinitive). Two variants of the construction have co-existed for a long time, and recent studies suggest that approximately during the last two decades, the frequency of the innovative variant has substantially increased. A number of language-internal factors which are assumed to affect the usage of both variants have been proposed in the literature.

We perform a very-large-scale corpus study (drawing on a range of Swedish corpora of different genres with a total size of more than eleven billion tokens) in order to address two main questions:

Q1) How well can we predict variation and change at the micro-level, i.e. which variant will be used in a given utterance?

Q2) How well can we predict variation and change at the macro-level, i.e. how frequent will the variants be in the community as a whole during a given period of time?

In our view, the most rigorous way to evaluate the predictive power of a theory or model of language change is to measure how well it can predict the future. In this article, however, we test our models on a presumably easier task: to determine the value of the dependent variable (that is, one of the two competing variants), given

the values of all the predictors. To approximate the “true” future-prediction task, we split our data into a training set representing the “past” and a test set representing the “future”. We do, however, let the models know the values of the predictors for the test set, which would not be possible if we were predicting the actual future.

Our results show that even for this relatively simple task we do not manage to achieve a substantial and robust increase over the baseline performance.

The rest of the article is structured as follows. In Section 2, we describe what is known about the usage of the future tense construction in contemporary Swedish and identify a number of factors which are supposed to affect the variation. In Section 3, we describe the corpora, the methods of extracting and processing the data, the operationalization of the predictors and the statistical analysis. We present the results in Section 4, discuss them in Section 5 and conclude with Section 6.

We would like to highlight that in Sections 2 and 3, we discuss at length the factors that may affect variation in order to motivate our choice of predictors and our operationalizations. In Section 4, however, we do not attempt to evaluate the relative importance of individual predictors. Since all the predictors, taken together, do not yield any substantial improvement over baseline, their individual contributions are obviously marginal and thus of little interest.

2 Background

2.1 Omission of the infinitival marker

Swedish has a future tense construction consisting of the auxiliary *kommer* (present tense of the verb *komma* with the literal meaning ‘come’) and an infinitive. The infinitive is preceded by the optional infinitival marker *att*. (1a) illustrates the construction with *att* present, (1b) shows *att*-omission, as used in the Swedish online discussion forum Familjeliv (see Section 3.1).

(1a) *Ja, vi tror att väntetiden kommer att vara 1-1,5 år framöver.*

‘Yes, we think the waiting time **is going to be** 1–1.5 years in the future.’

(1b) *Så kommer det nog vara några veckor till.*

‘It probably **is going to be** like that for some more weeks.’

The *kommer (att)* construction is not the only way to refer to the future in Swedish. Future tense may also be expressed by present tense inflection or a construction consisting of the future tense auxiliary *skola* ‘shall’ and an infinitive (Teleman et al. 1999 IV: 243–250).

The *kommer (att)* construction is part of a larger family of constructions consisting of an auxiliary or auxiliary-like verb plus an infinitive with optional *att* (Bylin 2013; Lagervall 1999; Mjöberg 1950; Teleman et al. 1999), as illustrated in the constructed examples (2a) to (2c). The parentheses indicate that the infinitival marker is optional.

(2a) *Julen börjar (att) närma sig.*

‘Christmas is starting to approach.’

(2b) *Ibland vägrar hon (att) sitta kvar i vår famn.*

‘Sometimes, she refuses to remain sitting in our arms.’

(2c) *Jag försöker (att) inte visa min oro inför barnen.*

‘I try not to show my worry to the kids.’

Optional infinitival markers are not restricted to Swedish but have also been observed in other Germanic languages. In English, the infinitival marker *to* may be left out after the verbs *help* (Kjellmer 1985; Levshina 2018; Lind 1983; Lohmann 2011; Mair 2002; McEnery and Xiao 2005) and *try* (Kjellmer 2000).

(3a) *Sarah helped us (to) edit the book.* (Quirk et al. 1985: 1205).

(3b) *I have never seen a fish try (to) get warmed by a fire.* (Kjellmer 2000: 116).

In Dutch, the infinitival marker *te* is optional after some aspectual verbs (4a) and the modal verb *hoeven* ‘have to’ (4b) (Van de Velde 2015, 2017; Coussé forthcoming).

(4a) *Volgens mij is het de bedoeling dat we niet keiveel herrie zitten (te) maken.*

‘I think the idea is that we should not be making a lot of noise.’ (Coussé forthcoming: §18.2.4)

(4b) *Ze heeft er niet veel moeite voor hoeven (te) doen.*

‘She did not have to make a lot of effort.’ (Coussé forthcoming: §18.2.4)

In this section, we review the factors that, according to the literature, affect the omission of the infinitival marker in the future *kommer (att)* construction. We also

complement them with related factors suggested for optional infinitival markers in other verb constructions in Swedish, English and Dutch. In this study, we consciously focus mostly on language-internal, structural predictors (but also time and genre), leaving potential language-external predictors for future studies.

2.2 Language-external predictors

2.2.1 Time

The distribution of the *att* marker in the *kommer (att)* construction has varied in the last decades. Delsing (1993), Olofsson (2008) and Malmgren (2017) observe in Swedish newspaper texts that *att*-omission only occurs sporadically in the period 1965–1998. Newspaper texts from 2004 show a marked increase of *att*-omissions. The literature reviews of Persson (2005) and Olofsson (2008) show that this trend has been commented upon in the normative literature from the early 1990s onwards. Olofsson (2007, 2008) also conducts a small apparent-time study of *att*-omission revealing that younger university students (born 1984–1987) have a higher preference for omission than older students (born before 1980). Similarly, Persson (2005) finds that *att*-omission is higher in a discussion forum with young writers (up till 17 years old) compared to one with adult writers. Most recently, Adesam et al. (forthcoming) performed a large corpus study where they showed that there was indeed a strong recent increase of *att*-omission. The proportion of *att*-omission increased from early 2000s to 2021 in newspaper texts from ca 0.10 to ca 0.40 and in social-media texts from ca 0.35 to ca 0.65.

The ongoing loss of *att* has been related to the grammaticalization of *kommer* as a future tense auxiliary. Several authors suggest that *att*-omission can be considered a case of reduction which is typical for grammaticalization (Christensen 1997: 46–47; Falk 2002; Hilpert 2008: 127). Others argue that the newly grammaticalized *kommer* joins the class of auxiliaries without an infinitival marker (Delsing 1993; Olofsson 2007, 2008; Persson 2005). The absence of an infinitival marker has been considered a feature of prototypical auxiliaries in a gradient approach to auxiliaries (Lagervall 2015; Sundman 1983; Telemann et al. 1999). This feature was initially restricted to modal verbs but is assumed to have spread in the course of history to other verbs that were semantically close (Lagervall 1999: 132; Mjöberg 1950: 72). The ongoing loss of *att* after the future auxiliary *kommer* fits into this historical development. The analogical pull of auxiliaries without an infinitival marker has also been suggested as a motivation for the ongoing loss of infinitival markers in English (Kjellmer 1985, 2000; Lind 1983; Mair 2002) and Dutch (Van de Velde 2015, 2017). The increase of *att*-omissions has also been related to an overall

increase in frequency of the future *kommer* (*att*) construction (Teleman et al. 1999 IV: 246; Svenska Språknämnden 2005: 357).

We expect that the rate of *att*-omission is likely to increase with time in our data.

2.2.2 Genre

Persson (2005), Malmgren (2017) and Blensenius and Rogström (2020) report that *att*-omission after *kommer* is more frequent in social media (blogs, microblogs and discussion forums) than newspaper texts and academic writing. Mjöberg (1950) points out that *att*-omission in general is more often found in newspaper texts than in books. He relates this tendency to the fact that newspapers are produced more rapidly than books and therefore are not edited as carefully. Mjöberg assumes that short function words like the infinitival marker are also easily dropped in the telegram-like style of headlines to save space. The effect of text genre is also noted for English infinitival markers (Kjellmer 1985; Levshina 2018; Lind 1983; Lohmann 2011; McEnery and Xiao 2005). Omission of the infinitival marker is more common in informal settings as opposed to formal genres. Adesam et al. (submitted) show that the *att*-omission rate is indeed much higher in social media than in newspaper texts.

We do not add genre as a predictor. Instead, given that large differences are known to exist between our corpora and that different social-media corpora can be argued to represent different language communities, we opt for running a separate analysis for each of the corpora.

2.3 Language-internal predictors

In this subsection, we describe hypotheses about language-internal predictors and their cognitive motivation.

2.3.1 Predictability

Levshina (2018), studying the omission of the infinitival marker *to* after *help* in English, shows that the predictability of the construction affects the probability of *to*-omission. In general, omission is more likely in more predictable contexts, though Levshina (2018) observes broad variation and rather interesting non-linear patterns.

If explicit marking is optional, then it is more likely to be omitted when the utterance in context is more predictable (Gibson et al. 2019; Haspelmath 2008). If the construction's predictability is low, it is more difficult for the addressee

to process it correctly, and the speaker is more likely to help the addressee by preserving the explicit marker.

One way of measuring predictability is **attraction** (Levshina 2018; Schmid 2000). In our case, attraction is the conditional probability of the infinitive given the *kommer (att)* construction; that is, how often a verb is used after *kommer (att)* in comparison with other verbs. Attraction is correlated with frequency (frequent verbs will occur in the construction more often than infrequent ones), but not equivalent to it.

We add attraction as a predictor and expect high attraction values to facilitate *att*-omission.

2.3.2 Distance

Delsing (1993) points out that early *att*-omission was more common in sentences where the subject intervenes between the auxiliary *kommer* and the infinitive. Olofsson (2007, 2008) finds a higher preference for *att*-omission in test sentence (5a), where the auxiliary and infinitive are separated by a heavy subject and adverbial (underlined), as opposed to (5b), where there are no intervening elements between the auxiliary and infinitive.

(5a) *Kommer John och hans tvillingsyster verkligen (att) börja skolan i år?*

‘Are John and his twin sister really **going to start** school this year?’ (Olofsson 2007: 4, our translation)

(5b) *Jag är övertygad om att kriget kommer (att) vara slut om två veckor.*

‘I am convinced that the war **is going to be** over in two weeks.’ (Olofsson 2007: 4, our translation)

Persson (2005) finds more generally that the separation of the auxiliary and infinitive by one or more intervening elements facilitates *att*-omission (although the trend only shows in his discussion forum texts, not in newspaper texts). *Att*-omission is also more frequent in his material when the intervening element is a sentence adverbial (especially a negating adverb like *inte* ‘not’) as opposed to a time adverbial.

Interestingly, the *kommer (att)* construction seems to be rather exceptional in this respect: for other infinitival constructions, the effect supposedly works the other way round, that is, *att*-omission is stimulated by the adjacency of the auxiliary and infinitive. Teleman et al. (1999 III: 593, fn. 3) state that the omission of *att* in other Swedish infinitival constructions is stimulated by adjacency of the finite verb and the infinitive. In English, omission of the infinitival marker *to* is more frequent when *help* and the infinitive are adjacent (Kjellmer 1985; Levshina 2018; Lind 1983;

Lohmann 2011; McEnery and Xiao 2005). In Dutch, the omission of the infinitival markers only occurs when the auxiliary and infinitive are juxtaposed in a so-called verb cluster (Coussé forthcoming).

A decrease in the probability of the omission of the infinitival marker as the distance increases fits well with the assumption that omission is more likely in more predictable contexts (see Section 2.3.1: “Predictability”). This idea can also be framed as the principle of minimizing cognitive complexity, defined by Rohdenburg (1996: 151): “In the case of more or less explicit grammatical options, the more explicit one(s) will tend to be favoured in cognitively more complex environments.”

Levshina (2018: 4) follows the same logic: “The more words between *help* and the infinitive, the more difficult it is to recognize the latter as part of the construction”. Adding the infinitival marker thus helps to explicitly mark the structure of the construction.

This line of reasoning has also been proposed for Swedish by Mjöberg (1950: 77) and Delsing (1993: 4) who calls it the “clarity hypothesis”. While this hypothesis is cognitively plausible, for the *kommer (att)* construction it does not seem to be borne out by the Swedish facts. Delsing (1993: 4) proposes an alternative cognitive explanation, which he calls the “forgetfulness hypothesis”. When linguistic elements are placed in between the auxiliary and infinitive, speakers might forget what auxiliary they are using, and leave out the *att* infinitival marker. Persson (2005: 41) also follows this line of reasoning.

We use the distance (number of intervening words) between the auxiliary *kommer* and the infinitival phrase as a predictor. We expect that larger distance will stimulate *att*-omission. We also use two predictors related to distance, see Sections 2.3.3 and 2.3.4.

2.3.3 Length of the infinitive chain

The infinitive which has *kommer* as a syntactic head may have another infinitive as a dependent, which, in turn, may have another, and so on. Consider, for instance, the following example with a chain of three infinitives:

(6) *Det verkar också som att Academedia kommer att kunna fortsätta skratta*

‘It also seems that Academedia is going to be able to continue laughing’ (Familjeliv)

Svenska Svenska Språknämnden (2005: 357) suggests that *att*-omission in general is stimulated when several verbs are combined in a verb chain. This effect can be explained by the “forgetfulness hypothesis”, which also featured in Section 2.3.2 to explain the effect of distance.

We use the number of infinitives in the verb chain as a predictor and expect that longer infinitive chains will stimulate *att*-omission.

2.3.4 Presence of a syntactic subject

An additional exploratory predictor is whether a nominal subject of *kommer* is present or not, see (7), where the subject is omitted:

(7) *Kommer att kosta massor.*

‘Going to cost a lot.’ (Familjeliv)

This predictor is related to the distance between *kommer* and the infinitive: if the subject is omitted, it can never occur between the two verbs and thus increase the distance. Nonetheless, we use it as a separate predictor, assuming that there may exist other relevant syntactico-semantic differences between the constructions with and without subject.

2.3.5 Presence of other *att*

Mjöberg (1950) points out that the presence of another *att* in the immediate context triggers a stylistic response among some writers to leave out the infinitival marker *att* (also Teleman et al. 1999 III: 598 fn. 1; Svenska Språknämnden 2005: 357). This tendency is not restricted to the *kommer (att)* construction. In examples (8a, 8b), the presence of an additional *att* presumably increases the probability of omitting the *att* after *kommer*.

(8a) *Han kommer (att) bli tvingad att vända sig till andra kraftkällor.*

‘He is going to be forced to turn to other power sources.’

(Mjöberg 1950: 79)

(8b) *Men jag tror aldrig att vi kommer (att) förstå hur impulser i dessa nätverk ger upphov till medvetande.*

‘But I think that we never are going to understand how impulses in these networks give rise to consciousness.’

(Svenska Språknämnden 2005: 357)

The other *att* may occur both before and after the infinitival marker under scrutiny. Note that the other *att* is not necessarily an infinitival marker but also may be a subordinator, as in (8b).

The presence of an identical marker has a similar effect on infinitival markers in English (Kjellmer 1985, 2000; Levshina 2018; Lind 1983; Lohmann 2011; McEnery and Xiao 2005; Rohdenburg 2009) and Dutch (Van de Velde 2015, 2017; Coussé forthcoming). The underlying motivation for this effect has been argued to be avoidance of identity, also known as “horror aequi”, defined by Rohdenburg (2003: 236) as “the widespread (and presumably universal) tendency to avoid the use of formally (near-) identical and (near-)adjacent (non-coordinate) grammatical elements or structures”.

We add two predictors:

- the presence of another *att* before the *kommer (att)* construction
- the presence of another *att* after the *kommer (att)* construction.

In both cases, we expect the presence of *att* to make the omission of *att* in the future tense construction more likely. We assume that the position of another *att* (or the presence of an *att* both before and after the construction) may also play a role, but we do not have any specific expectations of its direction or size.

2.3.6 Voice of the infinitive

The infinitive in the *kommer (att)* construction can be either in the active or in passive voice. (9) illustrates the passive voice consisting of the so-called *s*-form.

(9) *Det kommer att krävas mer från oss alla*

‘More is going to be required from all of us.’

(Election manifesto of the Swedish Social Democratic Party, 2022)

Persson (2005) reports that *att*-omission is less frequent with passives. McEnery and Xiao (2005) likewise indicate that the passive infinitives after *help* are always marked with *to* in English. Levshina (2018) shows that this is not correct, but does not make any quantitative analysis of this factor due to its infrequency in her data.

Following the logic outlined in Section 2.3.1: “Predictability”, it can be assumed that since passive is a less frequent form, the verb would be less predictable and thus more likely to require explicit marking.

We add voice as a predictor and expect *att*-omission to be less probable with passive voice.

3 Materials and methods

3.1 Corpora

In order to test to what extent the factors outlined in Section 2 enable us to predict which variant of the *kommer (att)* construction will be used (in a specific utterance or during a given period), we need large, deeply annotated, time-stamped corpora. Since the change has been particularly active during approximately the last two decades, we need corpora that contain recent texts. We use seven Swedish corpora created and maintained by Språkbanken Text (SBX) that satisfy these criteria. The basic information about the corpora is presented in Table 1, the data extraction and annotation processes are described in the rest of this section, more details about the corpora are provided in Appendix A. In addition to those seven corpora, we generated an “All” version for which we aggregated all available observations. All data were collected in January–September 2022.

To generate balanced samples per year, we manually inspected each raw dataset and decided on a corpus-specific time span and sample size (denoted as N_{sample}), shown for each corpus in columns 2 and 3 in Table 2. Per corpus, we then randomly drew N_{sample} observations per year (the samples are available in the Supplementary material). Depending on the chosen time span, we used the first N_{train} years to train our models and last N_{test} years to evaluate our models, shown in columns 4 and 5 in

Table 1: Basic information about the corpora used in the study. Observations are instances of *kommer + (att) + infinitive* constructions. One sentence may contain more than one observation.

	Description	#Observations	Years	#Tokens
Familjeliv	“Family life”, a discussion forum with a focus on pregnancy, children and parenting; predominantly female users	5.04 M	2003–2021	4.4 G
Flashback	A discussion forum with a broad variety of topics; predominantly male users	5.31 M	2001–2021	3.7 G
Twitter	Tweets by Swedish users	1.3 M	2006–2019	2 G
Sveriges Television (SVT) News	Short news articles from the SVT webpage	412 K	2004–2021	214 M
Göteborgs-Posten (GP)	Articles from the GP newspaper	205 K	2001–2013	250 M
Dagens Arena (DA)	Articles from the DA online newspaper	15 K	2007–2021	9.5 M
Bloggmix	Texts from Swedish blogs	643 K	1998–2017	616 M
In total		11.6 M	2001–2021	11.0 G

Table 2: Time span, sample size and training/testing split for each corpus.

Corpus	Time span	N_{sample} (observations)	N_{train} (years)	N_{test} (years)
All	2004–2021	100,000	13	5
Bloggmix	2005–2016	5,000	9	3
DA	2013–2021	1,000	7	2
Familjeliv	2004–2020	60,000	13	4
Flashback	2005–2021	80,000	13	4
GP	2001–2013	13,000	10	3
SVT	2007–2021	8,000	12	3
Twitter	2009–2018	6,000	8	2

Table 2. Note that while we measure time in years for the purpose of generating balanced samples, we will measure it in months when training and testing the models. Thus, for DA, for instance, the test set contains not 2, but $2 \cdot 12 = 24$ periods (and 2,000 observations).

3.2 Data extraction and annotation

The data were downloaded using the API (<https://ws.spraakbanken.gu.se/docs/korp>) for the SBX search engine Korp (Borin et al. 2012). All sentences containing at least one relevant observation, that is, the verb *komma* in the present tense (normally *kommer*) followed by an infinitive, were extracted, if the distance between *kommer* and the infinitive was not larger than six tokens. (The distance between two immediately adjacent tokens was taken to be 1; tokens include both words and punctuation marks).

All SBX corpora are annotated by the Sparv pipeline (Hammarstedt et al. 2022), which provides information about part of speech, morphological features, lemma, dependency structure and more. Using this annotation, all observations were automatically labelled either as “noise” (there is no syntactic link between *kommer* and the infinitive; the example should be discarded), “att” (the infinitive’s head is the word *att*, the head of which is *kommer*) or “omission” (the infinitive’s syntactic head is *kommer*, there is no *att*). The values of the predictors (see Section 2 for a general description, Section 3.4 for operationalizations) were also automatically calculated.

The accuracy of the part-of-speech and morphological annotation of the SBX corpora is estimated to vary between 0.91 and 0.96, depending on the genre and the version of Sparv (Adesam and Berdicevskis 2021), while the quality of the syntactic

annotation is lower: the labeled attachment score (LAS) is estimated to vary between 0.78 and 0.84 (Berdicevskis 2020). In social-media corpora, which often contain unedited non-standard texts, the annotation quality is likely to be lower for both morphology and syntax. The Flashback, Familjeliv, SVT, and DA corpora were recently added or updated, and were annotated with a newer version of Sparv which typically yields higher accuracy.

It is thus not a given that the available annotation, especially the dependency tree structure, is good enough for the current study. In order to test the quality of the data extraction and annotation, we performed a manual spot check, see Section 3.3. The extraction and annotation scripts are provided in Supplementary materials.

3.3 Manual spot check of the annotation quality

We randomly selected 300 observations from the downloaded data for the Familjeliv forum. The classification and all feature values were manually checked by AB, EC and YA. The annotated sample is provided in Supplementary materials.

Out of 300 observations, eight (2.7%) were misclassified. Table 3 shows a confusion matrix of the values generated by our automatic annotation and the values of our manual evaluation.

If the example was classified correctly, most of the feature values in the checked examples were also correct. The only features which contained errors were those related to the syntactic subject of the construction and the length of the infinitive chain (see Section 3.4).

Two systematic limitations were made clear by the analysis of the sample. First, if there are several coordinated infinitives, we focus only on the first one, consider (10):

(10) *Just den här eftermiddagen kommer jag att dra ut telefonjacket och låta bli att sätta mig framför datorn*

‘But this afternoon, I **am going to pull** the phone out of its jack and **avoid sitting down** in front of the computer’ (Familjeliv)

Table 3: Confusion matrix for the spot check of the annotation quality (300 observations).

Manual	Automatic		
	att	omission	noise
att	115	0	4
omission	0	157	0
noise	2	2	20

It is difficult to estimate how good the annotation of coordinated structures in the corpora is and whether we can reliably extract relevant structures. Besides, adding non-first conjuncts would arguably introduce another factor which can potentially affect the omission of *att*.

Second, it is possible that the data contain constructions with *kommer* which have the same syntactic structure, but a different, non-future tense, meaning (Teleman et al. 1999 II: 511–512). One example where another interpretation is possible is (11), where *kommer att tänka på* ‘come to think of’ does not signal future, but rather chance.

(11) *Såg ett tidigare inlägg och bör nog precisera mig: Jag behöver inga tips på vad jag ska trycka i mig för jag är inte sugen på något av alla saker som jag **kommer att** tänka på*

‘Saw an earlier post and should probably specify: I don’t need tips for what I should chow down on because I’m not keen on any of the things I **can** think of’ (Familjeliv)

There is no reliable method which could have filtered out these constructions, especially since in many cases they can be argued to be ambiguous. However, they are most probably infrequent (one ambiguous example in the 300-observation sample). Johansson (2006: 139–140) reports a higher proportion of the non-futural constructions (35 out of 576 in her sample, extracted from a corpus of newspaper texts published in 1995–1997), but even with this estimate, their frequency is still rather low.

Overall, with a classification accuracy of 97.3%, we deem the annotation to be reliable enough. No major systematic biases were identified. In the subsequent data processing, we also exclude examples where more than one *att* occurs between *kommer* and the infinitive (which is extremely unlikely), as well as those where the syntactic subject of *kommer* is placed after *att* (which is ungrammatical). The column “#observations” in Table 1 refers to the number of observations after all these filters were applied.

3.4 Operationalization of the predictors

All examples which were labelled as noise were discarded, that is, we only kept those where there was a syntactic link between *kommer* and the infinitive (via *att*, if it was present). We then operationalized the notions outlined in Section 2 by calculating the values of the following predictors (see Appendix A for more details).

Language-external predictors

year, month (see Section 2.2.1: “Time”): the year and the month when the text was written. The distribution of observations over time is very uneven, but we deal with that by drawing balanced samples (see Section 3.1).

genre (see Section 2.2.2: “Genre”): we perform our analysis separately for each corpus.

Language-internal predictors

attraction (see Section 2.3.1: “Predictability”): the conditional probability of a verb given the *kommer (att)* construction, calculated as the ratio of the frequency of the verb with *kommer (att)* to the frequency of *kommer (att)*. Note that this measure, unlike all others, is not a property of an observation, but of a corpus. It is calculated using only the samples included in the training set (see Sections 3.1 and 3.6).

distance to att (see Section 2.3.2: “Distance”): the distance from *kommer* to the infinitival group, that is, to *att* if present, else to the infinitive. This predictor is considerably skewed: for instance, in the “All” corpus, over 57% of all utterances have a distance of one word. Given this skewness, we made the decision to transform these predictors into binary variables (distance of one word vs. distance of two or more words) to ensure that we have adequate data for model fitting, as suggested by Bresnan et al. (2007).

length of the infinitive chain (see Section 2.3.3: “Length of the infinitive chain”): the infinitive which has *kommer* as a syntactic head may have another infinitive as a dependent, which, in turn, may have another, and so on. This feature gauges the complexity of the infinitival group by measuring the number of infinitives in it. This predictor is considerably skewed: for instance, in the “All” corpus, nearly 90% of all utterances have a length of one word. Given this skewness, we made the decision to transform these predictors into binary variables (length of one word vs. distance of two or more words).

subject (see Section 2.3.4: “Presence of a syntactic subject”): whether *kommer* has a syntactic non-clausal subject. If *kommer* is coordinated with another verb, only the first verb is treated as having a subject (i.e. if *kommer* is not the first one, it is treated as subjectless). Due to the limitations of the syntactic annotation, this variable is slightly less reliable than others (approximately 2% of the observations in the sample were misannotated).

att before (see Section 2.3.5: “Presence of another att”): a binary variable: whether there is another *att* in the same sentence before *kommer*. We do not make any attempt to distinguish between *att* as an infinitive marker and as a subordinator.

att after: same as **att before**: whether there is another *att* in the same sentence after *kommer*.

voice (see Section 2.3.6: “Voice of the infinitive”): active or s-form. Most s-forms are passive voice (e.g. *att påverkas* ‘to be influenced’ vs. *att påverka* ‘to influence’),

but some are so-called deponent verbs where the passive form is used with an active meaning (and an active form is missing), e.g. *att hoppas* ‘to hope’. No attempt was made to distinguish deponent verbs from true passives.

The dataset (available in the Supplementary materials) also contains values of other potentially relevant features as well as detailed metadata.

3.5 Visualization of the effect of the predictors

We perform a simple correlation test, in order to preliminarily check whether our expectations about the relation between the individual predictors and the dependent variable are correct. Two predictors are continuous (time, attraction) and six are binary (or have been simplified to be binary). We calculate the phi coefficient (also known as the Matthews correlation coefficient, see e.g. Chicco and Jurman 2020) between each of the binary predictors and the outcome. For binary variables, phi is equal to the Pearson correlation coefficient. To have a comparable measure for continuous predictors, we calculate the Pearson correlation coefficient between each of them and the outcome (since the outcome is a binary variable, the Pearson coefficient is equivalent to the point-biserial correlation coefficient, for which the assumption of linearity does not apply). We use the training set for each corpus (see Section 3.1). The predictors are coded in a way which should yield a positive correlation if the direction of the effect is in agreement with our expectations and a negative correlation if it is opposite.

The results are visualized in Figure 1.

Most predictors do have a certain relation to the outcome, and its direction is usually consistent across different corpora, but its size is always small. Interestingly, for *att before* and *subject* the direction of potential effect is almost always opposite to the one we expected. For *att after*, the direction is inconsistent (which, together with the small values, suggests that there is no effect). Note that while negative values are not what we expected, they are not per se a problem for the predictive approach: if there is a relation, it can be exploited, regardless of its direction.

3.6 Statistical analysis

We focus on the predictive accuracy of statistical models by training them on part of the data (training set) and then measuring their performance on a held-out portion of the data (test set).

Note that at no point do we attempt to estimate statistical significance. This is a conscious choice. First, the very notion of statistical significance has recently been

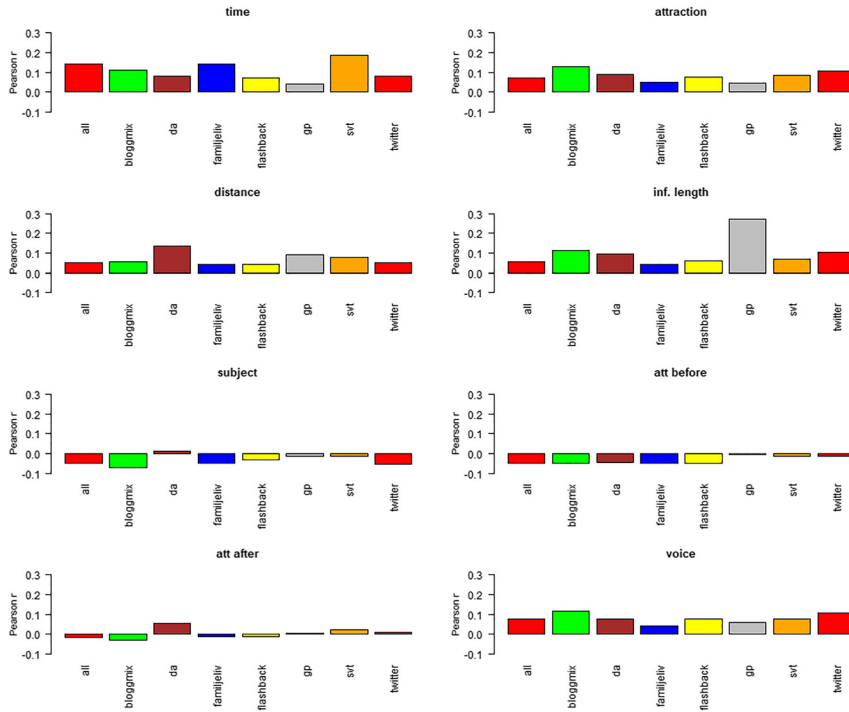


Figure 1: The Pearson correlation coefficients between the individual predictors and the outcome (equivalent to the phi coefficient for binary variables). Negative values indicate effects that go against our expectations.

strongly criticized, see Wasserstein et al. (2019) about the limitations and pitfalls of this approach in general and Koplein (2019) about corpus linguistics in particular.

Second, the predictive approach addresses a question which is somewhat different from questions that are central in many other corpus studies. The question is not “how large is the effect?” (effect size) or “how probable is it that the effect of the same or larger size could have arisen in a sample of a given size if there is no effect in the population?” (statistical significance), but “if we have a theory which claims that the predictors affect the variable in a specific way and know the values of the predictors, can we find out what the value of the variable is?”, which we find equivalent to “how good is our explanatory theory actually?” and thus of utmost importance.

To find out how well we can predict variation and change, we use two different approaches: *Approach A* that is based on a logistic regression model and *Approach B* that is based on an autoregressive integrated moving average (ARIMA) model.

Logistic regression is a statistical technique that models the relationship between a binary outcome (taking on values of either 0 or 1) and one or more predictor variables. The model uses a logistic function to estimate the probability that an observation belongs to either outcome and a threshold is used to make a final prediction. Logistic regression plays an important role in a wide range of fields and disciplines, such as epidemiology, psychology and sociology or (corpus) linguistics, where researchers try to understand the relationship between an outcome and various predictors. In addition, logistic regression is also widely used in machine learning, where models are trained on large datasets to make predictions about unseen data. This approach has numerous applications, including image or text classification, spam detection, and the prediction of medical outcomes, such as the risk of hospitalization or mortality in patients with chronic diseases. For a comprehensive and accessible introduction to logistic regression, see Hosmer et al. (2013).

ARIMA is a technique that is used to describe and analyse time-series data. The resulting model can then be used to make predictions about future values of the time-series. Due to the temporal nature of the data, most of the methods that are used to analyse cross-sectional data cannot be used to analyse time-series datasets. One of the most important issues arising in this context is that many time series exhibit autocorrelation, i.e. the value of a variable at one moment in time is correlated with values of its own past and its own future. Not accounting for this aspect of time-series will lead to incorrect statistical inference and inaccurate predictions. ARIMA models that are widely used to predict time-dependent values (e.g. food prices, inflation rates, unemployment rates, weather patterns, spread of diseases or language change) solve this problem by incorporating a number of past values of the time-series in order to predict its value at a given time step, which is called an autoregression model (the AR part). In addition, ARIMA models also include a weighted moving average of past forecast errors as predictors (the MA part). Finally, the 'I' in ARIMA refers to the process of making a time-series stationary, meaning transforming it to a series whose statistical properties do not change over time. For a comprehensive and accessible introduction to time series analysis and ARIMA models, see Hyndman and Athanasopoulos (2018).

In what follows, both approaches will be explained more formally.

3.6.1 Approach A (logistic regression)

To estimate the probability that *att* is omitted in some utterance indexed by *i*, i.e. $\Pr(y_i = 1)$, depending on a vector of predictors, denoted as x_i , we fit a logistic regression to the training data that can be written as (Bresnan et al. 2007; Rabe-Hesketh and Skrondal 2012):

$$\text{logit}\{\Pr(y_i = 1|x_i)\} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_v x_{vi} \quad (1)$$

where $\text{logit}()$ denotes the logistic function, $y_i = 1$ denotes omission of *att*, $y_i = 0$ denotes that *att* is not omitted, β_0 denotes an intercept, x_{1i}, \dots, x_{vi} represent v predictors and β_1, \dots, β_v represent the corresponding coefficients that are estimated by maximum likelihood.

To test whether the selected predictors help to predict if *att* is omitted, we fit the following three models to the training data of each corpus:

- M^I : a null model without predictors, i.e. $\text{logit}\{\Pr(y_i = 1|x_i)\} = \beta_0$. This model simply always predicts the value of y_i that is more frequent in the training data and thus serves as our baseline. For example, in the Familjeliv training data, *att* is omitted in 297,558 of all 780,000 occurrences. Thus, the null model will always guess that $y_i = 1$ in the test data. With a percentage of $167,317/240,000 = 69.72\%$ this prediction is correct.
- M^{II} : a model with time as a predictor. To parameterize the effect of time x_1 (measured in months), we fit four sub-models: a linear parameterization, a degree-1 fractional polynomial, a degree-2 fractional polynomial and degree-3 fractional polynomial (see Appendix B for formal definitions). Fractional polynomials provide a flexible way to accurately model non-linear relationships by providing a much wider range of shapes than regular polynomials (Royston and Altman 1994; Royston and Sauerbrei 2008; StataCorp 2022).
- M^{III} : as M^{II} this model includes time as a predictor and is parameterized accordingly. Additionally, M^{III} contains the following predictors:
 - x_2 is *subject* (0 – no; 1 – yes).
 - x_3 is *voice* (0 – s-form; 1 – active).
 - x_4 is *att before* (0 – no; 1 – yes).
 - x_5 is *att after* (0 – no; 1 – yes). In addition, we included the interaction between x_4 and x_5 .
 - x_6 is *distance to att* (0 – distance of one; 1 – distance of two or more).
 - x_7 is *length of the infinitive chain* (0 – length of one; 1 – length of two or more).
 - x_8 is *attraction*. Let N_ω denote the number of utterances where a particular verb ω is used after *kommer (att)*, *attraction* is then calculated as $N_\omega / (N_{\text{sample}} \cdot N_{\text{train}})$. As written above, *attraction* is computed based on the training data only. If a verb only occurs in the test data, but not in the training data, we assume an occurrence frequency of 1, i.e. $N_\omega = 1$, and impute a corresponding value from the training data. To parameterize the effect of attraction, we fit separate degree-2 fractional polynomial models as described for x_1 . Note that these models were fitted with all other covariates except for time (x_1). The generated fractional polynomial variables for *attraction* were then used as a covariate in addition to $x_2 - x_7$ to parameterize the effect of time.

- M^{IV} : as main effects, this model includes the same covariates as M^{III} . In addition, it contains all first-order interactions between the generated fractional polynomial variables for time (x_1) and all other covariates.

Due to different parameterizations, models M^{II} to M^{IV} each contain four different sub-models. To select the best-fitting (sub-)model, we first dropped non-converged models and then used the following two different goodness-of-fit measures for selection: $G_{\text{train}}^{\text{A}}$ (related to Q1: how well can a model predict *att*-omission for a given utterance) and $G_{\text{train}}^{\text{B}}$ (related to Q2: how well can a model predict the proportion of omissions for a given period). Both measures are based on the predicted probability $\hat{\rho}_i$ of $y_i = 1$. It is common to transform this probability into a predicted outcome by setting a threshold of 0.5 and classifying the prediction as 1 if the predicted probability is higher than 0.5.

For Q1, however, it is not a given that 0.5 is always the optimal threshold, which is why we try different thresholds and for every threshold calculate the proportion of utterances where the prediction is correct.

For Q2, we do not need to transform the probability into the outcome at all, since we are not interested in predictions at the utterance levels. Instead, we sum the utterance-level probabilities for a given month and use the resulting value as the expected proportion of omissions, which can then be compared to the actual value.

More formally, the measures are defined as follows:

- $G_{\text{train}}^{\text{A}}(M^\pi)$: for each utterance i in the training data, where π denotes one model from the set {II, III, IV}, we compute the predicted probability $\hat{\rho}_i$ of $y_i = 1$. To transform probabilities into actual predicted outcomes, denoted as \hat{y}_i , we define a threshold value δ : a prediction is classified as 1, if $\hat{\rho}_i \geq \delta$. For each possible value of c in the interval of [0.25, 0.75] with steps of 0.01, we compute the percentage of correct classifications and extract the value of c where the classification accuracy is highest, denoted as δ^* . Out of the all sub-models, we select the sub-model with the highest overall classification accuracy. In case there are several models with the same classification accuracy, we select the model that has the smaller AIC (Akaike 1974) on the training data. This sub-model and the corresponding value of δ^* are then used to predict the test data.
- $G_{\text{train}}^{\text{B}}(M^\pi)$: based on each model, we compute the expected proportion of *att*-omissions in a given month m , i.e. average predicted probabilities denoted as $\hat{\text{P}}_m$. On this basis, we calculate the model fit for model M^π for the training data as a measure of prediction accuracy (Koplenig et al. 2022; Tofallis 2015):

$$G_{\text{train}}^{\text{B}}(M^\pi) = \frac{1}{M} \sqrt{\sum_{m=1}^M \ln \left(\hat{\text{P}}_m / \text{P}_m \right)^2} \quad (2)$$

where $m = 1, 2, \dots, M$ index the available months in the training data and P_m represents the observed proportion of *att*-omissions in m . G^B -values are reported as percentages by multiplying the above equation by 100. Note that as long as the difference between \hat{P}_m and P_m is relatively small, $\log(\hat{P}_m/P_m) \approx (\hat{P}_m - P_m)/P_m$. Thus, we can interpret G^B as measuring the approximate (absolute) average percentage difference between P_m and \hat{P}_m .

To test how well each chosen model helps to predict the test data, we compute corresponding measures for the test data for each corpus, denoted as $G_{\text{test}}^A(M^T)$ and $G_{\text{test}}^B(M^T)$.

Higher values are indicative of higher prediction accuracy for G^A and lower values are indicative of higher prediction accuracy for G^B . Therefore, if our predictors help to predict the omission of *att*, we should observe the following patterns across corpus samples:

- (i) G_{test}^A computed for M^I should be **lower** compared to G_{test}^A computed for M^{II} , M^{III} and M^{IV} .
- (ii) G_{test}^B computed for M^I should be **higher** compared to G_{test}^B computed for M^{II} , M^{III} and M^{IV} .
- (iii) If the chosen language-internal factors help to predict the outcome, we additionally expect that M^{III} or M^{IV} achieve a better model fit than M^{II} for both G_{test}^A and G_{test}^B .

3.6.2 Approach B (ARIMA)

As an alternative way to model the temporal nature of our data, we try to predict the observed proportion of *att*-omissions P_m in a given month by fitting non-seasonal ARIMA(p, d, q) models that can be written as (Beckett 2013; Hyndman and Athanasopoulos 2018):

$$P'_m = \beta_0 + \beta_1 x_{1i} + \dots + \beta_v x_{vi} + \eta_m \quad (3)$$

where β_0 denotes an intercept, x_{1m}, \dots, x_{vm} represent v predictors and β_1, \dots, β_v represent the corresponding coefficients. P'_m is a time series that has been differenced d times to transform the original time series P_m into a stationary time series. Differencing refers to computing the difference between consecutive observations. For example the first-order difference is computed as $P_m - P_{m-1}$.

η_m is written as:

$$\eta_m = \Phi_1 P'_{m-1} + \dots + \Phi_p P'_{m-p} + \Theta_1 \varepsilon_{m-1} + \dots + \Theta_q \varepsilon_{m-q} + \varepsilon_m \quad (4)$$

where Φ_p is the p th order autocorrelation parameter, Θ_1 is the q th order moving-

average parameter and ε_m is an independent and identically distributed (i.i.d.) error term. Estimates are derived by maximum likelihood.

To test whether the selected predictors help to predict the proportion of *att* in a given month, we fit the following two models to the training data of each corpus:

- M^I : a null model without predictors, i.e. a pure ARIMA(p, d, q) where P'_m is predicted by a combination of lagged values of P'_m and past errors of P'_m . To select the best-fitting values for p, d and q , we use the Hyndman-Khandakar algorithm (Hyndman and Athanasopoulos 2018; Hyndman and Khandakar 2008) as implemented in the Python pmdarima package (Smith and Taylor 2017).
- M^{II} : a so called ARMAX model that includes the following predictors:
 - x_1 is *subject*.
 - x_2 is *voice*.
 - x_3 is *att before*.
 - x_4 is *att after*. As above, we included the interaction between x_3 and x_4
 - x_5 is *distance to att*.
 - x_6 is *length of the infinitive chain*.
 - x_7 is *attraction*.

All predictors are modelled as described above. In addition, all predictors are averaged by month prior to estimation. This implies that x_1 – x_6 represent proportions. For example, a value of 0.75 for x_2 implies that 75% of the utterances in the corresponding month are in active *voice*. x_7 represents average monthly values. As for the pure ARIMA model, we use the Hyndman-Khandakar algorithm to select the best-fitting values for p, d and q .

To test how well each chosen model helps to predict the test data, we use each selected model for each corpus to compute two types of forecasts:

- (i) A full forecast, where we predict the whole test data set at once.
- (ii) A one-step-ahead forecast, where, in each month m , we use the estimated model parameters to predict the value of P in the subsequent month, i.e. \hat{P}_{m+1} . After prediction, the model is updated with the observed value of P , i.e. \hat{P}_{m+1} and new model parameters are estimated. This process is continued until the end of the test data set is reached

In each case, we compute G_{test}^B as a measure of forecast accuracy for each corpus. If the chosen language-internal factors help to predict the outcome, we expect in each corpus for both the full and the one-step-ahead forecast that $G_{\text{test}}^B(M^I) > G_{\text{test}}^B(M^{II})$.

3.6.3 Summary

To sum up, we use Approach A (based on logistic regression) to address both Q1 (predictions for a given utterance, goodness-of-fit measured as G^A) and Q2 (prediction of proportion for a given period, goodness-of-fit measured as G^B). We use Approach B (based on ARIMA) to address Q2.

The data analysis was done in Stata 17.0 and Python 3.6.8. The commented code is available in Supplementary material.

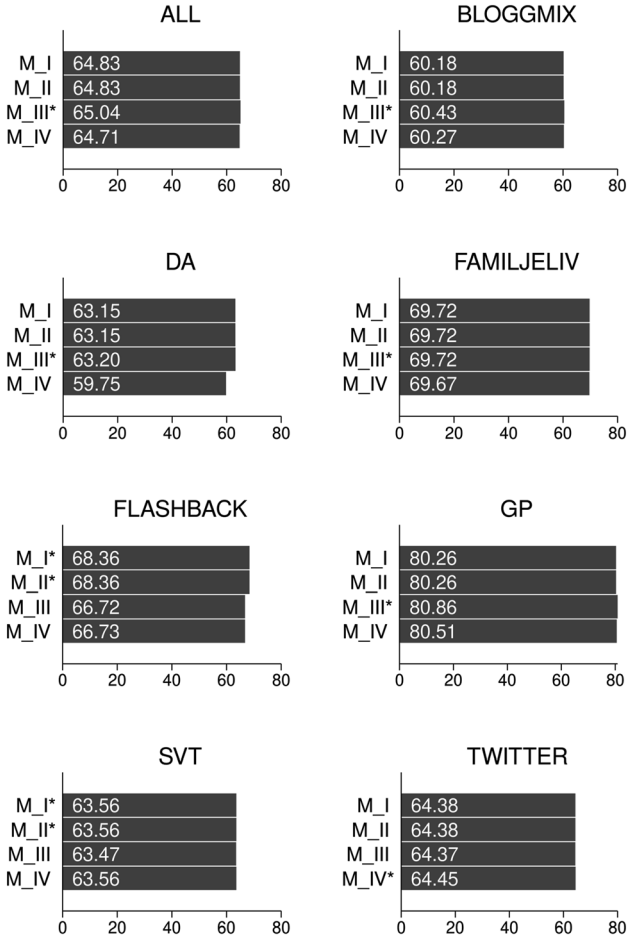
4 Results

4.1 Approach A (logistic regression)

Figure 2 visualizes the results of the logistic classification (measure G_{test}^A , addressing Q1). For each corpus, the percentage of correctly predicted test utterances is visualized per model (also see Appendix C for a table with confusion matrix information for all corpora and all four models). In 2 out of our 8 corpus samples (Flashback and SVT), the prediction accuracies of the models that include predictors (M^{II} , M^{III} and M^{IV}) do not top the baseline prediction accuracy, i.e. $G_{\text{test}}^A(M^{\text{I}})$. For the other 6 corpus samples, M^{III} is the best-fitting model for All, Bloggmix, DA, Familjeliv and GP. For Twitter, M^{IV} is the best-fitting model. However, the difference between $G_{\text{test}}^A(M^{\text{III}})$ or $G_{\text{test}}^A(M^{\text{IV}})$ and $G_{\text{test}}^A(M^{\text{I}})$ as a measure of prediction improvement is below 1 percentage points in all such cases. This indicates that including predictors hardly helps in predicting the outcome.

In Appendix C, we show that if we use a different performance measure (balanced accuracy) which punishes the models that always predict only one variant (as M^{I} does), then one of the linguistically-informed models (M^{III} or M^{IV}) is always the best, but the improvement over baseline (that is, the difference between the best model and the baseline model) is still small (from 0.25 percentage points for Twitter to 4.12 percentage points for DA).

In Appendix D, we show that the obtained results are somewhat better if we split the data into training/test in a non-consecutive manner, i.e. if utterances are randomly split into 80% training and 20% test data. If we use accuracy, one of the linguistically-informed models is the best for all corpora except for DA (the improvement ranges from 0.17 percentage points for GP to 5.15 for Bloggmix). If we use balanced accuracy, one of the linguistically-informed models is always the best (the improvement ranges from 1.42 for SVT to 9.33 for Bloggmix).



Correctly predicted test data (%)

Figure 2: G_{test}^A for each model (M^I , M^{II} , M^{III} and M^{IV}). Per corpus, an asterisk indicates the best-fitting model(s). For Familjeliv, the difference is not obvious from the figure, since the values are rounded to two decimal places.

Figure 3 shows time-series line plots for all corpus samples where the monthly proportion of *att*-omissions is the outcome. In each plot, the training data is depicted in black colour and the test data is coloured in mint green. In 2 out of our 8 corpus samples (Bloggmix and Twitter), the prediction accuracies of the models that include predictors do not top the baseline prediction accuracy, i.e. $G_{\text{test}}^B(M^I)$.

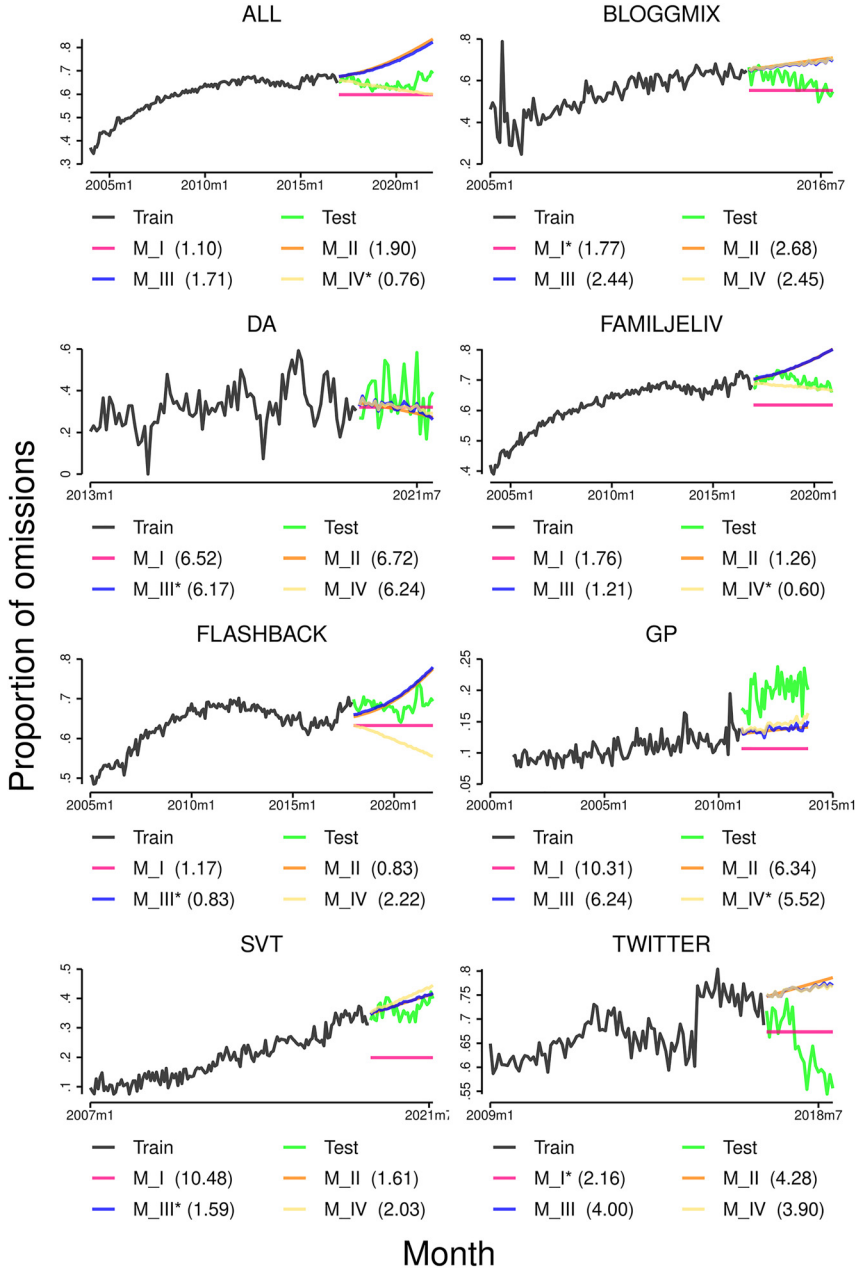


Figure 3: Proportion of *att*-omissions as a function of time and corpus. In each plot, G_{test}^B for each logistic model is given in brackets. An asterisk indicates the best-fitting model per corpus.

While in the remaining 6 cases, the models with the lowest G_{test}^B all include language-internal predictors, a visual inspection shows that the fit between the test data and the predicted proportions is rather poor in all cases.

4.2 Approach B (ARIMA)

Figure 4 visualizes the results of the full forecast based on the ARIMA models with P_m as the outcome. As above, training data is depicted in black colour and test data is coloured in mint green. In 5 out of our 8 corpus samples (All, DA, Familjeliv, Flashback and SVT), the prediction accuracies of the models that include predictors (M^{II}) do not top the baseline prediction accuracy, i.e. $G_{\text{test}}^B(M^{\text{I}})$.

Figure 5 shows the corresponding results of the one-step-ahead forecast. A visual inspection shows that for all corpora, calculated forecasts based on both M^{I} or M^{II} closely match the observed test data. The comparison between forecast accuracies of models without predictors (M^{I}) and models with predictors (M^{II}), show that only in half of our corpus samples (All, Flashback, GP and Twitter), M^{II} achieves a better prediction accuracy than M^{I} , i.e. $G_{\text{test}}^B(M^{\text{II}}) < G_{\text{test}}^B(M^{\text{I}})$.

Taken together and in line with the results presented for *Approach A*, the results for the ARIMA model imply that including the chosen language-internal predictors does not really help to predict the proportion of *att*-omissions for both the full and the one-step-ahead forecast. In Appendix E, we demonstrate that the lack of prediction improvement has to do with the fact that statistical associations between (i) the outcome and (ii) the predictors are not strong enough.

5 Discussion

As is obvious from Figures 3, 4, and 5, the proportion of *att*-omission does indeed change with time (Q2). An ARIMA model is able to predict these changes successfully, but only in the one-step-ahead mode (Figure 5). Even in this successful case, adding language-internal predictors to time does not result in a substantial and consistent performance improvement.

Other approaches, such as ARIMA in the full-forecast mode, (Q2, Figure 4), logistic-regression approaches both at proportion level (Q2, Figure 3) and utterance level (Q1, Figure 2) do not result in accurate predictions. Again, adding language-internal predictors either does not improve the performance or does it only marginally. Table 4 summarizes in which cases language-internal predictors do increase the performance of the model.

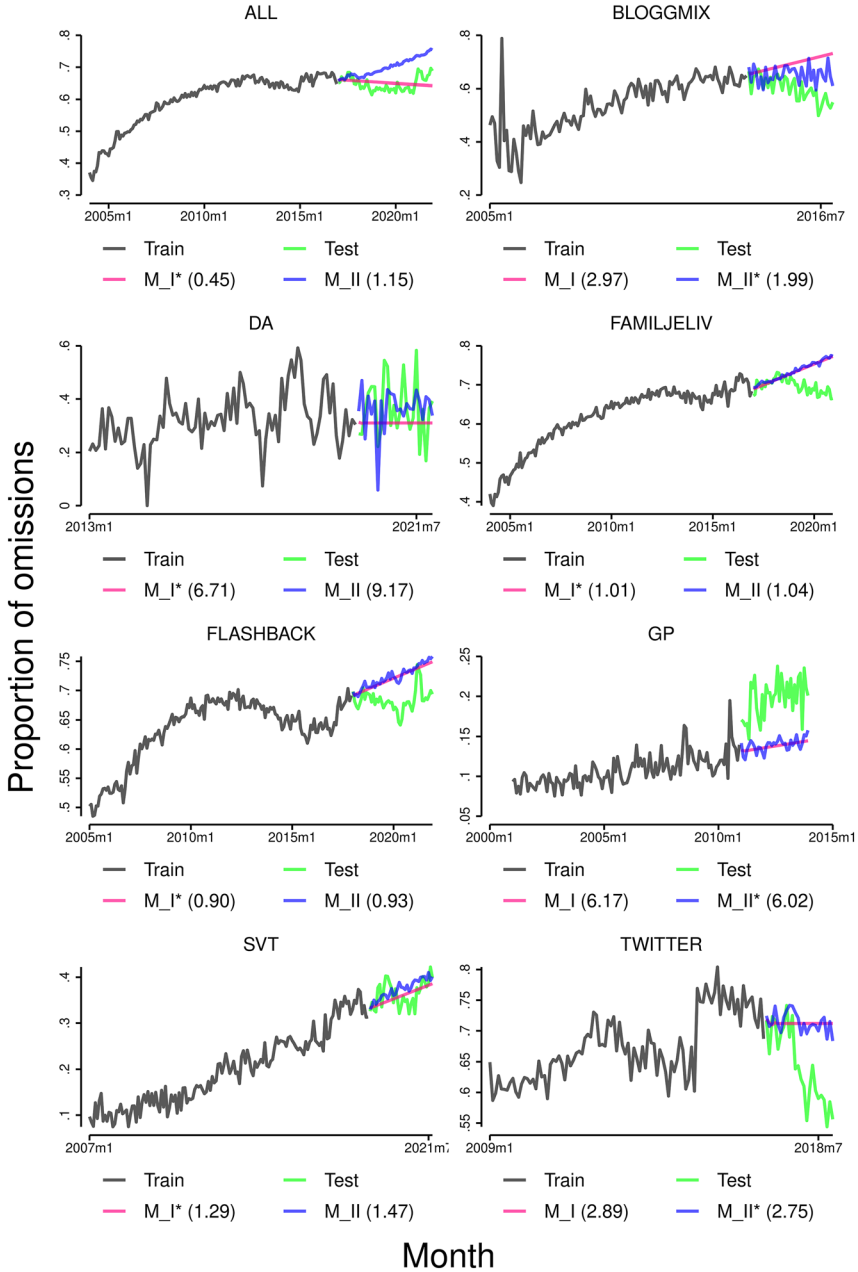


Figure 4: Proportion of *att*-omissions as a function of time and corpus. In each plot, G_{test}^B for each ARIMA model is given in brackets. Model predictions are based on a full forecast, where the whole test data set is predicted at once. An asterisk indicates the best-fitting model per corpus.

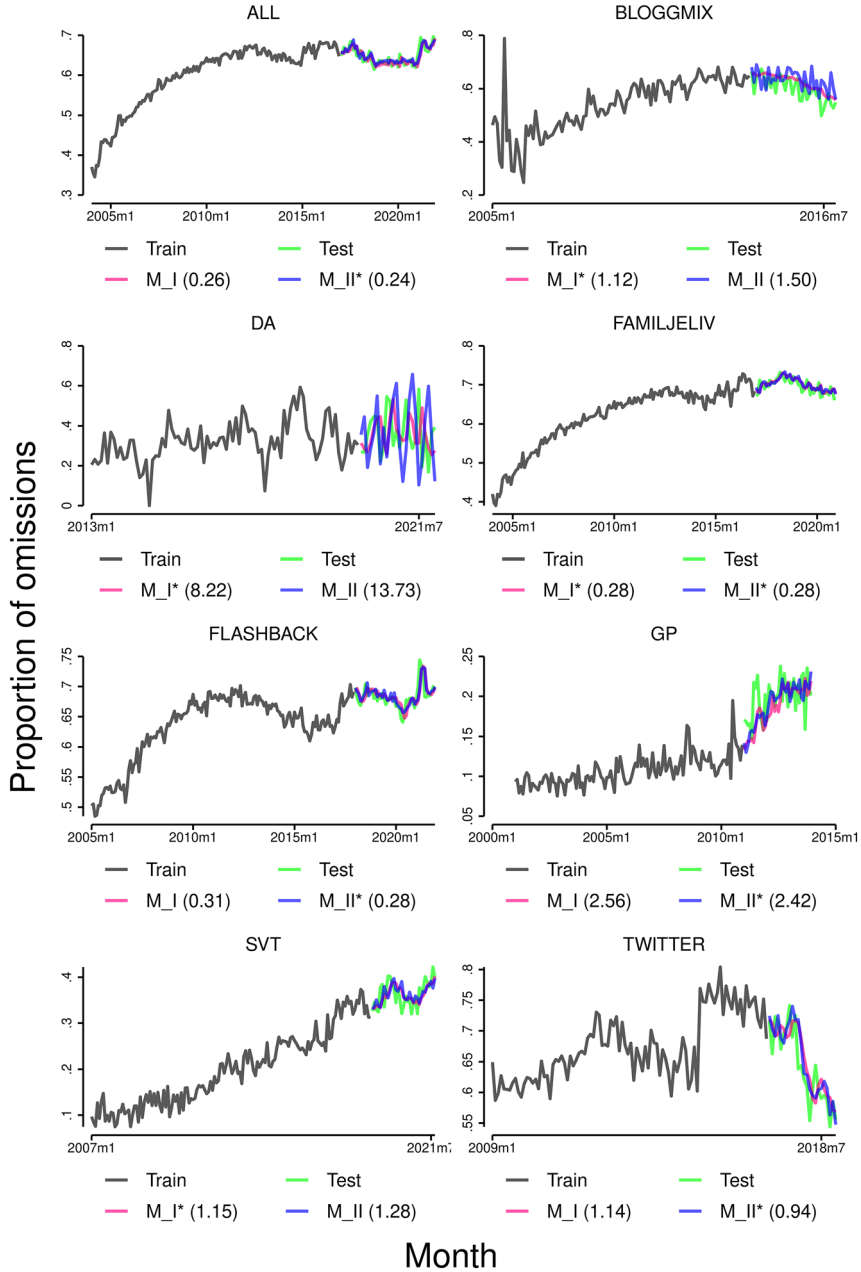


Figure 5: Proportion of *att*-omissions as a function of time and corpus. In each plot, \hat{G}_{test}^B for each ARIMA model is given in brackets. Model predictions are based on a one-step-ahead forecast where, in each month m , the estimated model parameters are used to predict the value of P in the subsequent month, i.e. \hat{P}_{m+1} . An asterisk indicates the best-fitting model per corpus.

Table 4: Cases when the best-fitting model includes language-internal predictors. Ac: utterance-level logistic regression with (Figure 2); A: proportion-level logistic regression (Figure 3); B: full-forecast ARIMA (Figure 4); B': one-step-ahead ARIMA (Figure 5).

	Q1:Ac	Q2:A	Q2:B	Q2:B'	In total
ALL	+	+		+	3
Bloggmix	+		+		2
DA	+	+			2
Familjeliv	+	+			2
Flashback		+		+	2
GP	+	+	+	+	4
SVT	+	+			2
Twitter	+		+	+	3
In total	6	6	3	4	

Including language-internal predictors helps for 3–6 corpora out of 8, depending on the approach. Note, however, that the improvement is always rather modest, and in many cases negligible.

The positive impact of language-internal predictors also varies across corpora. There are large differences between the corpora, some known in advance (size, time span, genre), some transpiring from the time-series visualizations. It does not, however, seem to be possible to explain the data in Table 4 through those differences.

If we apply a performance measure which takes the skewness of the datasets (one outcome is usually more frequent than another) into account (see Appendix C), the linguistically-informed models always perform better than baseline, but the improvement is still small.

Somewhat better results may be achieved by using a non-consecutive (random) split (see Appendix D), but this split is not compatible with the task of predicting the future, and the improvement is still modest.

As demonstrated in Appendix E, if there were strong statistical associations between the outcome and the predictors, our methods would have captured them.

We would like to point out, however, that it would be wrong to conclude that the predictors used do not have any effect on the variation. First, this is not the question we are trying to answer. Second, Figure 1 in fact suggests that they do have an effect (with the exception of *att after*), although usually a very small one.

For two predictors, the direction of the effect interestingly was opposite to our expectations. We expected that when the subject or another *att* in the sentence is present, the omission of *att* in the *kommer (att)* construction is more likely, but the effect seems to be the opposed (with the exception of *subject* in DA).

One tentative explanation can be the degree of formality that the speaker wants to maintain (which is related to their intent of complying with the norm and how

conscious about their style they are). Omitted subjects are more common in colloquial Swedish, and so is *att*-omission; thus omission of the subject might actually be an indication of register. The same may be true for the presence of another *att*: it may indicate a more formal register.

Another potential explanation for *att before* is that it has a priming effect and thus makes a speaker more prone to keeping the *att* in the future tense construction, despite the potential *horror aequi*.

6 Conclusion

The *kommer (att)* future tense construction is clearly in a state of change in Swedish. The language-internal predictors seem to affect the variation, that is, the probability of the *att*-omission (though not always in the expected direction). The influence of those predictors that we selected for this study, however, is not strong enough to create a successful predictive model that would perform considerably better than the majority baseline.

Note also that, as was mentioned in Section 1, we are letting the models know the predictor values in the test set, something which would not be possible if we were truly attempting to predict the unseen future (this limitation does not apply to ARIMA-based M^I and logistic-regression-based M^{II} : these models do not know anything but time). It is reasonable to assume that the true predictions of the future changes would be even less successful.

The answers to both Q1 and Q2 are thus pessimistic: we cannot reliably predict the presence or absence of *att* in an individual utterance and the proportion of *att*-omission in a given corpus in a given period of time. The only exception is predicting the proportion by an ARIMA model in a one-step-ahead mode, which is rather successful, but for this success we do not need language-internal predictors.

These findings may seem unexpected. Previous research on the variation in the *kommer att*-construction has established certain factors that affect *att*-omission. Explanations have been suggested of how and why the association between these factors and the outcome emerges, and quantitative evidence in favour of the association has been provided (and moderately supported by our data, see Figure 1). It is an interesting and not entirely intuitive conclusion that despite all that, predictive approaches may still fail. We are, however, not aware of any studies that explicitly attempted to predict the distribution of variants, so there is no direct contradiction.

Predicting the future is notoriously hard (The Forecasting Collaborative 2023). Still, we believe that it is important to systematically test the ability of linguistic theories and models to do that. If they cannot predict the future, can we really claim that they are able to explain the past and present? And even if we do, how useful are these explanations if they do not generalize well to new data?

In the particular case of the *att*-omission, it may be too early to despair and claim that theories about language are hopeless at predictions. First of all, we have consciously abstained from including any sociolinguistic predictors, such as the age and gender of the individual speakers; the structure of the interactions with each other; the degree of dissemination of a certain variant through the community (Würschinger 2021). Such sociolinguistic predictors can potentially capture certain trends in language change that structural predictors cannot. Second, it is possible to enrich our pool of language-internal predictors by going onto the lexical level and analyzing the behaviour of individual verbs in the *kommer (att)* construction (Persson 2005) and individual subjects. Third, it remains to be seen whether a powerful language model pretrained on a large number of texts (e.g. BERT or GPT) can predict *att*-omission better than a model trained on a number of preselected features.

Supplementary materials

https://github.com/AleksandrsBerdicevskis/kommer_att.

Acknowledgments: This work has been carried out within the Cassandra project (funded by Marcus and Amalia Wallenberg Foundation, donation letter 2020.0060) and supported by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (20182024, contract 2017-00626) and the 10 participating partner institutions. We thank two anonymous reviewers for their insightful comments.

Research funding: This work was supported by the Swedish Research Council (2017-00626) and Marcus and Amalia Wallenberg Foundation (2020.0060).

Appendix A: Additional information about corpora and data extraction

Flashback: The Flashback corpus also contains a very small amount of texts from early 2000, but they were not included in the study, since there is a discontinuity: the forum was not functioning in late 2000 and early 2001. The observations would have been filtered out by the sampling procedure anyway.

Duplicate texts: From the forum corpora (Flashback and Familjeliv), quotes (parts of messages where users cite one or more previous messages) were removed, but retweets were not removed from the Twitter corpus, which means it contains a certain amount of duplicate texts.

Twitter: Twitter doubled the available character space in November 2017 from 140 to 280 characters, which could affect linguistic production of its users (Boot et al. 2019). It can be hypothesized that limited space stimulates *att*-omission and thus that omission would become less frequent after the limit was changed. Visual inspection of the Twitter data (Figure A1), however, does not support this hypothesis: there is indeed a downward trend, but it seems to start earlier than November 2017.

Distance: As mentioned in the main text, all observations with distance = 6 were excluded, since they can only contain constructions without *att* as a result of how the original search query was defined. The original query does not say anything about *att*, it only picks up sentences where the distance between *kommer* and the infinitive is not larger than 6 (including punctuation marks). Hence, the maximum possible distance between *kommer* and *att* in the dataset is 5; all the examples with distance = 6 will be labelled as “omission”. In the corpora, there are (very rare) examples where *att* is present and the distance between it and *kommer* is 6, but they cannot get included in the dataset.

Note also that we count the punctuation marks when extracting the data, but not when determining the distance. Since punctuation marks seldom occur between *kommer* and the infinitive group, this inconsistency is unlikely to affect the results.

Length of the infinitive chain:

Some of the longer chains contain errors. In the sample we checked manually (see Section 3.3), we had only one such instance:

(A1) *Vet man att man exempelvis pga ett funktionshinder alltid kommer att ha svårt att få jobb ...*

‘If one knows that one, because of a disability, is always **going to have** difficulties **getting** a job’ (Familjeliv)

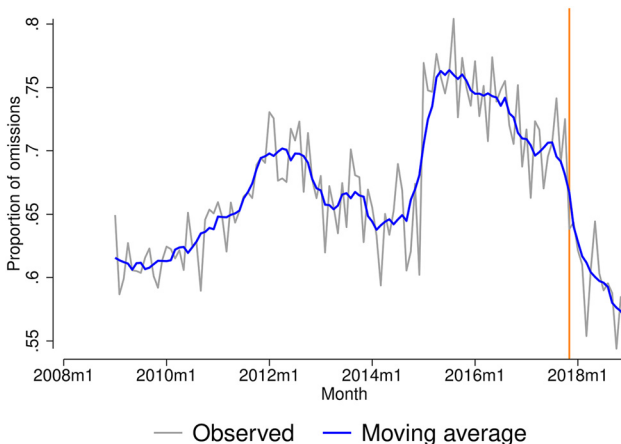


Figure A1: A: frequency of *att*-omission per month. B: same data with moving average (window of seven months). The vertical red line marks November 2017, when the character limit was changed.

According to the syntactic annotation in the corpus, *få* ‘get’ is dependent on *ha* ‘have’, and then the chain contains two infinitives. *få*, however, should rather depend on *svårt* ‘difficult’, and the length of the chain should equal one. It seems that for large chain lengths such errors are relatively common, but there are very few such chains. Besides, we later convert this predictor into a binary factor (length of one word vs. length of two words or more), and thus consider the potential influence of mis-annotated chains negligible.

Other verb forms (participles, supines) are not counted as parts of the infinitive chain. The chain is not necessarily continuous: if some other words occur between infinitives (participles that belong to phrasal verbs, reflexive pronouns that belong to reflexive verbs, arguments of verbs etc.), they are ignored.

Appendix B: Detailed description of the statistical analysis

Definitions of sub-models used to parameterize the effect of time x_1 in model M^I , Approach A:

- A linear parameterization: $\text{logit}\{\text{Pr}(y_i = 1|x_i)\} = \beta_0 + \beta_1 x_{1i}$
- A degree-1 fractional polynomial (Royston and Altman 1994; Royston and Sauerbrei 2008; StataCorp 2022) written as: $\text{logit}\{\text{Pr}(y_i = 1|x_i)\} = \beta_0 + \beta_1 x_{1i}^{(\varphi_1)}$, where φ_1 is chosen from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. For each possible value of φ_1 , a separate model (8 in total) is fitted. The value of φ_1 with best-fit (in terms of deviance) is chosen. Note that $x^{(0)}$ represents $\ln(x)$.
- A degree-2 fractional polynomial written as $\text{logit}\{\text{Pr}(y_i = 1|x_i)\} = \beta_0 + \beta_1 x_{1i}^{(\varphi_1)} + \beta_2 x_{1i}^{(\varphi_2)}$. Both φ_1 and φ_2 are chosen by finding the best-fitting model out of a total of 44 models. Note that if powers are repeated, i.e. $\varphi_1 = \varphi_2$, the second term is multiplied by the natural logarithm of x , i.e. $\ln(x)$.
- A degree-3 fractional polynomial written as $\text{logit}\{\text{Pr}(y_i = 1|x_i)\} = \beta_0 + \beta_1 x_{1i}^{(\varphi_1)} + \beta_2 x_{1i}^{(\varphi_2)} + \beta_3 x_{1i}^{(\varphi_3)}$. As for the degree-1 and the degree-2 model, the powers φ_1 , φ_2 and φ_3 are chosen by selecting the best-fitting model out of a total of 164 models.

Appendix C: Confusion matrix

Obviously, M^I always predicts the more frequent variant. Table C1 shows that M^{II} does the same in almost all cases, i.e. both these models always err on the same

side. It can be argued that the models which are able to predict both variants (and thus better match the true probability distribution) have an advantage, which is not necessarily taken into account when measuring accuracy (percentage of correct predictions), as we do in the main text. To correct for that, we reproduce the results using balanced accuracy as a measure. Balanced accuracy is defined as an arithmetic mean of sensitivity and specificity (Brodersen et al. 2010), where sensitivity is the true positive rate ($TP/(TP + FN)$) and specificity is the true negative rate ($TN/(TN + FP)$). Balanced accuracy can be argued to be better suited for evaluating performance on skewed datasets: always choosing one of two variants will yield 0.50.

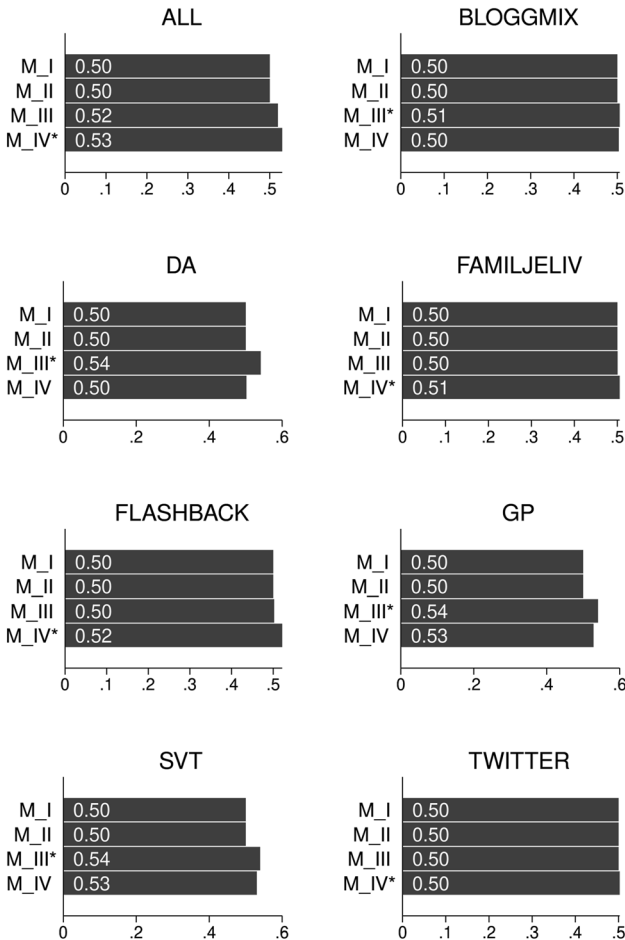
Table C1: Confusion information per split (consecutive/non-consecutive, 1st column, also see Appendix A) and per corpus (2nd column) for all selected models (3rd column). TP – True positives (4th column). TN – True negatives (5th column). FP – False positives (6th column). FN – False negatives (7th column). We treat omission as a positive outcome and “att” as a negative outcome.

Type of split	Corpus	Model	TP	TN	FP	FN
Consecutive	All	I	324,130	0	175,870	0
		II	324,130	0	175,870	0
		III	311,220	13,964	161,906	12,910
		IV	299,634	23,935	151,935	24,496
	Bloggmix	I	9,027	0	5,973	0
		II	9,027	0	5,973	0
		III	8,945	120	5,853	82
		IV	8,986	55	5,918	41
	Da	I	0	1,263	0	737
		II	0	1,263	0	737
		III	47	1,217	46	690
		IV	133	1,062	201	604
	Familjeliv	I	167,317	0	72,683	0
		II	167,317	0	72,683	0
		III	167,316	4	72,679	1
		IV	165,722	1,480	71,203	1,595
	Flashback	I	218,755	0	101,245	0
		II	218,755	0	101,245	0
		III	200,035	13,470	87,775	18,720
		IV	200,916	12,623	88,622	17,839
Gp	I	0	31,300	0	7,700	
	II	0	31,300	0	7,700	
	III	1,074	30,461	839	6,626	
	IV	496	30,904	396	7,204	
Svt	I	0	15,255	0	8,745	
	II	0	15,255	0	8,745	

Table C1: (continued)

Type of split	Corpus	Model	TP	TN	FP	FN	
Non-consecutive	Twitter	III	2,170	13,063	2,192	6,575	
		IV	2,007	13,247	2008	6,738	
		I	7,726	0	4,274	0	
		II	7,726	0	4,274	0	
		III	7,716	8	4,266	10	
	All	IV	7,696	38	4,236	30	
		I	219,859	0	139,336	0	
		II	200,051	24,266	115,070	19,808	
		III	197,941	28,779	110,557	21,918	
	Bloggmix	IV	194,011	32,706	106,630	25,848	
		I	6,837	0	5,279	0	
		II	5,137	1,987	3,292	1,700	
		III	5,768	1,693	3,586	1,069	
	Da	IV	5,652	1,744	3,535	1,185	
		I	0	1,246	0	590	
		II	0	1,246	0	590	
		III	75	1,171	75	515	
	Familjeliv	IV	64	1,176	70	526	
		I	129,633	0	74,069	0	
		II	122,295	8,875	65,194	7,338	
		III	120,897	10,949	63,120	8,736	
	Flashback	IV	120,730	11,104	62,965	8,903	
		I	175,337	0	96,392	0	
		II	172,625	2,645	93,747	2,712	
		III	169,310	7,005	89,387	6,027	
	Gp	IV	167,893	8,488	87,904	7,444	
		I	0	29,290	0	4,380	
		II	0	29,290	0	4,380	
		III	412	28,905	385	3,968	
	Svt	IV	361	28,987	303	4,019	
		I	0	18,349	0	5,611	
		II	0	18,349	0	5,611	
		III	283	18,113	236	5,328	
	Twitter	IV	185	18,218	131	5,426	
		I	8,071	0	4,045	0	
		II	8,071	0	4,045	0	
		III	7,917	198	3,847	154	
			IV	7,956	147	3,898	115

We followed exactly the same steps as described in the main text for Q1 and Approach A with the only difference that we did not fine-tune the threshold value for the logistic regression, opting instead always for 0.5. The results are represented in Figure C1. According to balanced accuracy, one of the linguistically-informed models (M^{III} or M^{IV}) is always the best, but the differences are still small.

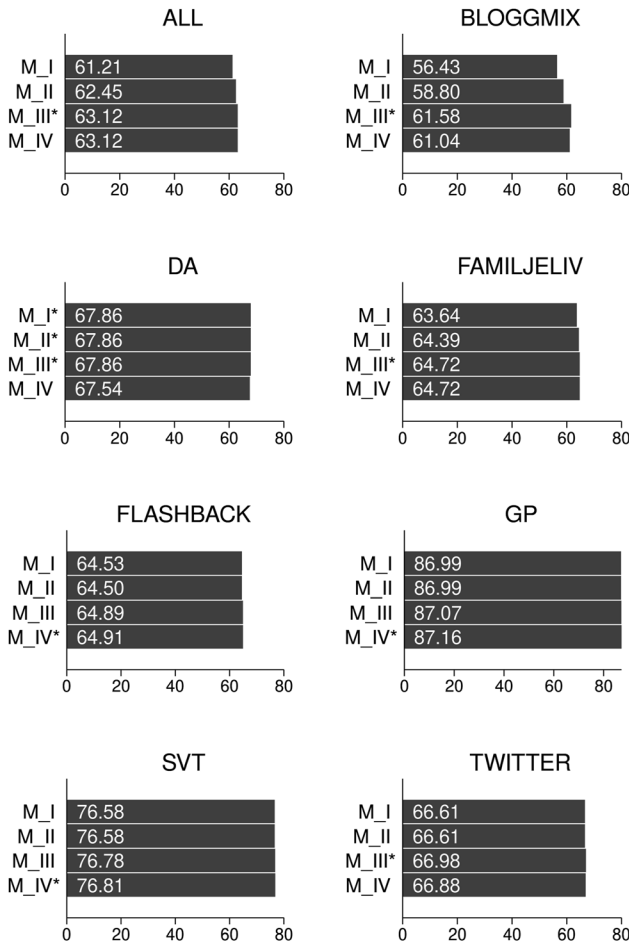


Balanced accuracy

Figure C1: G_{test}^A for each model (M^I , M^{II} , M^{III} and M^{IV}) with balanced accuracy as the performance measure. Per corpus, an asterisk indicates the best-fitting model(s). For Twitter, the difference is not obvious from the figure, since the values are rounded to two decimal places.

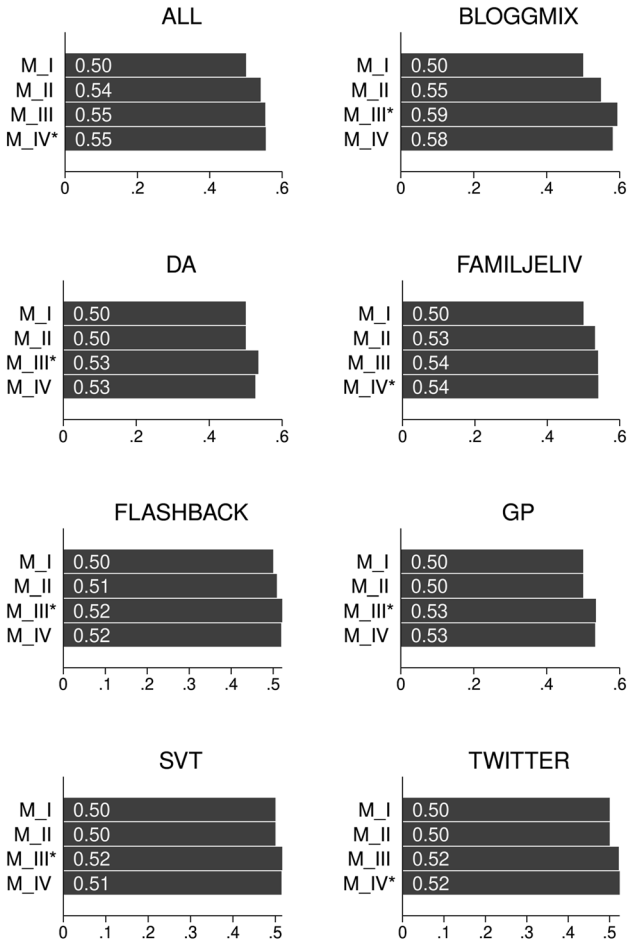
Appendix D: Results for Approach A based on a non-consecutive train/test split

See Figure D1 and D2.



Correctly predicted test data (%)

Figure D1: G_{test}^A for each model (M^I , M^{II} , M^{III} and M^{IV}). Per corpus, an asterisk indicates the best-fitting model(s). Here, analyses are based on a non-consecutive split where utterances are randomly split into 80% training and 20% test data.



Balanced accuracy

Figure D2: G^A_{test} for each model (M^I , M^{II} , M^{III} and M^{IV}) with balanced accuracy as the performance measure. Per corpus, an asterisk indicates the best-fitting model(s). Here, analyses are based on a non-consecutive split where utterances are randomly split into 80% training and 20% test data.

Appendix E: Results for a synthetic corpus where the effect of one covariate is artificially inflated

In the original version of the All corpus, the mean value of x_7 (*attraction*) is 0.026 for $y_i = 0$ (standard deviation $SD = 0.035$) and 0.031 for $y_i = 1$ ($SD = 0.037$). To artificially inflate that statistical association, we first randomly select $\sim 50\%$ of all cases for which $y_i = 0$. For the selected cases, we increase each value of x_7 by 0.200 and thus increase

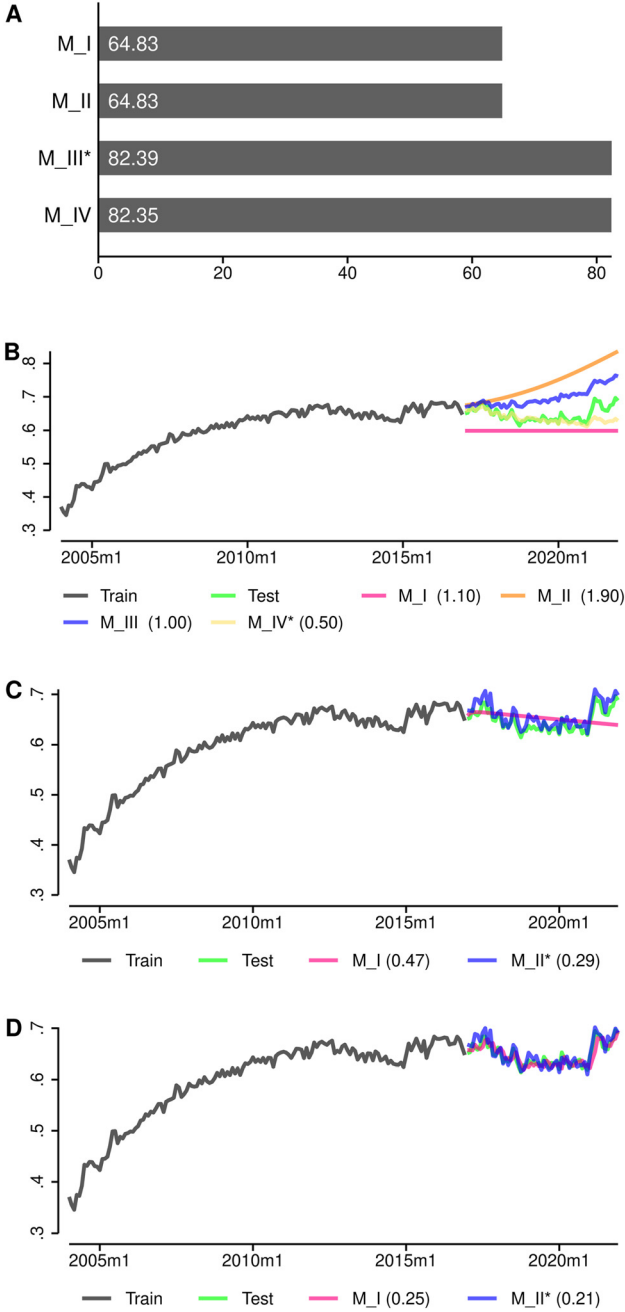


Figure E1: Results for the synthetic dataset where the influence of *attraction* is artificially inflated. (A) Results of the logistic classification analysis (see Figure 2 for details). (B) Results of the logistic proportion analysis (see Figure 3 for details). (C) Results of the ARIMA full forecast (see Figure 4 for details). (D) Results of the ARIMA one-step-ahead forecast (see Figure 5 for details).

the corresponding conditional mean value of x_7 to 0.126 ($SD = 0.106$). If our quantitative approaches work, then we should expect that for the synthetic dataset, including covariates should sharply increase model-fits. Figure E1 shows that this is indeed the case for both approaches.

References

- Adesam, Yvonne, Aleksandrs Berdicevskis & Evie Coussé. Forthcoming. Språkförändring på bar gärning: En storskalig korpusstudie av pågående förändringar i stavning, lexikon och grammatik [Language change in the act: A large scale corpus study of ongoing changes in spelling, lexicon and grammar]. *Svenskans beskrivning* 38, Submitted for publication.
- Adesam, Yvonne & Aleksandrs Berdicevskis. 2021. Part-of-speech tagging of Swedish texts in the neural era. In *Proceedings of the 23rd Nordic conference on computational linguistics, NoDaLiDa*. Available at: <https://aclanthology.org/2021.nodalida-main.20/>.
- Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6). 716–723.
- Beckett, Sean. 2013. *Introduction to time series using Stata*, 1st edn. College Station, Tex: Stata Press.
- Berdicevskis, Aleksandrs. 2020. *Choosing a new dependency parser for Sparv*. Technical report. Available at: https://github.com/spraakbanken/golddatools/blob/master/report_parsing_20200603.pdf.
- Blensenius, Kristian & Lena Rogström. 2020. Att hantera grammatisk förändring i en deskriptiv ordbok [Handling grammatical change in a descriptive dictionary]. *Nordiska studier i lexikografi* 15. 81–90.
- Boot, Arnout B., Erik Tjong Kim Sang, Katinka Dijkstra & Rolf A. Zwaan. 2019. How character limit affects language usage in tweets. *Palgrave Communications* 5(76). <https://doi.org/10.1057/s41599-019-0280-3>.
- Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, 474–478 volume Accepted. Istanbul: ELRA. Available at: <https://aclanthology.org/L12-1098/>.
- Bresnan, Joan, Cueni Anna, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: KNAW.
- Brodersen, Kay Henning, ChengOng Soon, Klaas Enno Stephan & Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition*, 3121–3124.
- Bylin, Maria. 2013. *Aspektuella hjälpverb i svenskan* (Stockholm studies in Scandinavian philology. New Series 58) [Aspectual auxiliary verbs in Swedish]. Stockholm: University of Stockholm.
- Coussé, Evie. Forthcoming. De verbale constituent [The verbal constituent]. *Algemene Nederlandse Spraakkunst*, Submitted for publication.
- Chicco, Davide & Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(6). 1–13.
- Christensen, Lisa. 1997. *Framtidsuttrycken in svenskans temporal system [The future expressions in the Swedish temporal system]*. Lund: Lund University Press.
- Croft, William. 2000. *Explaining language change: An evolutionary approach*. London: Pearson Education.
- Delsing, Lars-Olof. 1993. *Kommer utan att [Kommer without att]*. In *Flyktförsök: Kalasbok till Christer Platzack på femtioårsdagen*. Lund: University of Lund.

- Falk, Cecilia. 2002. Hjälpverbet *komma* [The auxiliary verb *komma*]. In Hanna Lehti-Eklund (ed.), *Studier i svensk språkhistoria 6 (Folkmålsstudier 41)*, 89–98. Helsinki: Föreningen för nordisk filologi.
- Fischhoff, Baruch & Ruth Beyth. 1975. I knew it would happen: Remembered probabilities of once-future things. *Organizational Behavior & Human Performance* 13(1). 1–16.
- Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23(5). 389–407.
- Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg. 2022. *Sparv 5 user manual*. Research Reports from the Department of Swedish, Multilingualism, Language Technology. University of Gothenburg. <http://hdl.handle.net/2077/73604>.
- Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.
- Hilpert, Martin. 2008. *Germanic future constructions. A usage-based Approach to language change*. Amsterdam: John Benjamins.
- Hosmer, David W., Stanley Lemeshow & Rodney X. Sturdivant. 2013. *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons.
- Hyndman, Rob J. & George Athanasopoulos. 2018. *Forecasting: Principles and practice*, 2nd edn. OTexts. OTexts.com/fpp2 (accessed 9 June 2022).
- Hyndman, Rob J. & Yeasmin Khandakar. 2008. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 27(3). 1–22.
- Kjellmer, Göran. 1985. Help to/help ø revisited. *English Studies* 66. 156–161.
- Kjellmer, Göran. 2000. Auxiliary marginalities: The case of *Try*. In John M. Kirk (ed.), *Corpora Galore. Analyses and techniques in describing English*, 115–124. Amsterdam: Rodopi.
- Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.
- Koplenig, Alexander, Sascha Wolfer & Peter Meyer. 2022. Human languages trade off complexity against efficiency. *Preprint*, In preparation.
- Labov, William. 1994. *Principles of linguistic change. Vol. 1: Internal factors*. Oxford: Wiley-Blackwell.
- Labov, William. 2001. *Principles of linguistic change. Vol. 2: Social factors*. Oxford: Wiley-Blackwell.
- Labov, William. 2011. *Principles of linguistic change. Vol. 3: Cognitive and cultural factors*. Chichester: John Wiley & Sons.
- Lagervall, Marika. 1999. Jakten på det försvunna infinitivmärket. Om definitionen av modala hjälpverb och infinitiv utan att [The hunt for the lost infinitive marker. On the definition of modal auxiliaries and infinitives without att]. In *Från dataskärm och forskarpärm (Meddelanden från Institutionen för Svenska Språket 25)*, 126–134. Gothenburg: University of Gothenburg.
- Lagervall, Marika. 2015. Modala hjälpverb i språkhistorisk belysning [Modal auxiliary verbs in a language historical light]. In *Göteborgsstudier i nordisk språkvetenskap*, vol. 23. Gothenburg: University of Gothenburg.
- Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help + (to) infinitive in varieties of web-based English. *Glossa: A Journal of General Linguistics* 3(1). 55.
- Lind, Åge. 1983. The variant forms *help to/help ø*. *English Studies* 64. 263–273.
- Lohmann, Arne. 2011. *Help vs help to*: A multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15. 499–521.
- Mair, Christian. 2002. Three changing patterns of verb complementation in late modern English: A real-time study based on matching text corpora. *English Language and Linguistics* 6. 105–131.

- Malmgren, Sven-Göran. 2017. Hur upplever du *hen*? Nio lexikala resor från 1965 till 2015 [What do you feel about *hen*? Nine lexical journeys from 1965 to 2015]. In Emma Sköldberg et al. (eds.), *Svenskans beskrivning 35* (Göteborgsstudier i nordisk språkvetenskap 29), 19–35. Gothenburg: University of Gothenburg.
- McEnery, Anthony & Zhonghua Xiao. 2005. *HELP* or *HELP to*: What do corpora have to say? *English Studies* 86. 161–187.
- McMahon, April. 1994. *Understanding language change*. Cambridge University Press.
- Mjöberg, Josua. 1950. Infinitivmärke på glid [Infinitive marker on the skids]. *Moderstmåslärarnas förenings årsskrift* 1950. 71–80.
- Olofsson, Arne. 2007. An endangered marker. On the loss of Swedish *att* after *kommer* and some parallels in English. *Nordic Journal of English Studies* 6(1). 1–10.
- Olofsson, Arne. 2008. Framtid i förändring. Hur länge kommer *att* dröja sig kvar? [A future in change. How long will *att* linger?] *Språk och Stil* 18. 143–155.
- Persson, Jens. 2005. Kommer utan *att* [*Kommer* without *att*]. *Scripta Minora* 45. 28–43.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Pearson Longman.
- Rabe-Hesketh, S. & Anders Skrondal. 2012. *Multilevel and longitudinal modeling using Stata*, 3rd edn. College Station, Tex: Stata Press Publication.
- Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7. 149–182.
- 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*, 205–249. Berlin: De Gruyter.
- Rohdenburg, Günter. 2009. Grammatical divergence between British and American English in the nineteenth and early twentieth centuries. In Ingrid Tieken-Boon van Ostade & Wim van der Wurff (eds.), *Current issues in late modern English*, 301–330. Bern: Peter Lang.
- Royston, Patrick & Douglas G. Altman. 1994. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics* 43(3). 429.
- Royston, Patrick & Willi Sauerbrei. 2008. *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. (Wiley series in Probability and statistics). Chichester, England; Hoboken, NJ: John Wiley.
- Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells. From corpus to cognition*. Berlin: Mouton de Gruyter.
- Sundman, Marketta. 1983. Svenska modalverb — ett continuum från hjälverb till huvudverb? [Swedish modal verbs – a continuum from auxiliaries to main verbs?] In Erik Andersson, Mirja Saari & Peter Slotte (eds.), *Struktur och variation. Festschrift till Bengt Loman (Meddelanden från Stiftelsen för Åbo akademi forskningsinstitut 85)*, 321–334. Åbo/Turku: Åbo Akademi.
- Smith, Taylor G., 2017. *pmdarima: ARIMA estimators for Python*. Available at: <http://www.alkaline-ml.com/pmdarima>.
- StataCorp. 2022. *fp – Fractional polynomial regression*. StataCorp. <https://www.stata.com/manuals/rfp.pdf> (accessed 13 September 2022).
- Svenska Språknämnden. 2005. *Språkriktighetsboken* [The book about language correctness]. Stockholm: Nordstedts.
- Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik. 3. Fraser* [The Swedish Academy grammar, 3. Phrases]. Stockholm: Svenska Akademien .
- The Forecasting Collaborative. 2023. Insights into the accuracy of social scientists' forecasts of societal change. *Nature Human Behaviour*. Available at: <https://doi.org/10.1038/s41562-022-01517-1>.

- Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen & Hans van Halteren. 2013. Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2). 227–262.
- Tofallis, Chris. 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society* 66(8). 1352–1362.
- Van de Velde, Freek. 2015. Schijnbare syntactische feniksen [Apparent syntactic phoenixes]. *Nederlandse Taalkunde* 20. 69–107.
- Van de Velde, Freek. 2017. Limits to language change. *Nederlandse Taalkunde* 22. 79–83.
- Wasserstein, Ronald, Allen Schirm & Nicole Lazar. 2019. Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician* 73(1 Suppl). 1–19.
- Würschinger, Quirin. 2021. Social networks of lexical innovation. Investigating the social dynamics of diffusion of neologisms on Twitter. *Frontiers in Artificial Intelligence* 4. 1–20.