

# Never change a winning chunk

Benedikt Szmrecsanyi, University of Freiburg

## Abstract

This paper deals with persistence in language production. By persistence I refer to the fact that speakers are likely to re-use structural, morphological, or phonological chunks and structures from previous discourse. Analyzing, as case study, the alternation between gerundial and infinitival complementation, I utilize a logistic regression analysis of naturalistic corpus data to demonstrate that (i) corpus data can match psycholinguistic data, (ii) that persistence is sufficiently patterned to be predicted by the analyst, and (iii) that the magnitude of the persistence effect depends on some secondary variables, such as textual distance between two chunks. The paper concludes by discussing implications of the existence of the phenomenon for linguistic theory and practice.

## Résumé

Cet article traite de la persistance à l'oral. Par persistance j'entends le fait qu'un orateur aura tendance à réutiliser des éléments phonologiques et morphologiques, ainsi que des structures, du discours antérieur. Dans le cadre d'une étude de cas, j'analyse l'alternance entre la complémentation gérondive et la complémentation infinitive à l'aide d'un modèle de régression logistique appliqué à des données linguistiques de corpus afin de démontrer que (i) les bases de données de corpus sont tout aussi utilisables que les bases de données psycholinguistiques, (ii) les schémas de persistance sont suffisamment clairs pour pouvoir être prédits, et (iii) la magnitude de la persistance dépend de variables secondaires telle que la distance entre les deux éléments. Pour conclure, cet article s'intéressera à la signification de l'existence de ce phénomène pour la théorie et la pratique linguistique.

## 1 Introduction

This paper is an extended empirical argument that speakers tend to repeat chunks (i.e. phonological or morphological material, or syntactic structure) from previous discourse whenever they can. Consider, for instance, the conversational snippet in (1):

- (1) I think it may be 10 years from now when people *start seeing* the long-term effects from it, and you *start having* problems with it. (Corpus of Spoken Professional American English, text Comm8a97)

In (1), there are two occurrences where the verb *to start* takes gerundial, as opposed to infinitival, complementation. A sizable body of literature is devoted to answering the question why the verb *to start* sometimes takes infinitival complements, and

sometimes gerundial complements. Traditionally, a linguist interested in this kind of variation would look at both *start seeing the long-term effects* and *start having problems with it* in isolation and attempt to determine, for each of the two chunks, semantic or other intralinguistic constraints that presumably caused the speaker to go for gerundial complementation in both cases. By contrast, the main argument of the present study is that an important reason why gerundial complementation is used in the second slot in (1) is because gerundial complementation has just been used immediately before.

This comparatively simple kind of explanation has received some attention in two major branches of linguistics, *psycholinguistics* and *discourse analysis*. Psycholinguists have known for some time now that the human speech apparatus is hard-wired to go for recently activated linguistic patterns whenever possible. This is because the network of memory is organized in terms of lexical, morphological, phonological/formal, or syntactic similarity. When a word, morpheme, phonological form, or syntactic structure is recognized, some site in the network is activated, and this activation subsequently spreads to nodes of related patterns or tokens (cf. Tanenhaus et al. 1980). In psycholinguistic parlance, this phenomenon is known as *production priming* (for instance, Bock 1986). Discourse analysts (in particular conversation analysts), on the other hand, have pointed out that repetitiveness plays an important role in managing discourse: repetitiveness in conversational interaction maintains involvement, connection, and interaction; repetitiveness is speaker-economical in that it provides for planning time, and hearer-economical in that it can help relax the processing load that comes with otherwise informationally dense discourse (Tannen 1989). Because of its corpus-based approach, the present study avoids referring to discourse analytic and especially psycholinguistic terminology *a priori*; this is why I will refer to the underlying phenomenon – namely, that speech production is inertial and that linguistic chunks tend to persevere – by the relatively neutral term *persistence*.<sup>1</sup>

Unlike to the discourse-analytic and psycholinguistic literature, there are few systematic corpus-linguistic, quantitative studies of persistence (some of the rare exceptions are Sankoff & Laberge 1978 and Weiner & Labov 1983), and no standard methodology exists to integrate the factor into corpus-based research designs. In attempt to begin to remedy this shortcoming, the present study's analysis has three main objectives:

1. to show that corpus data can match psycholinguistic data;
2. to suggest a methodology to integrate persistence into variationist research designs;

---

<sup>1</sup> The reason for this terminological caution is that corpus-based study may be inappropriate to explicitly investigate psycholinguistic mechanisms such as production priming. In naturalistic data, speakers' output may exhibit persistence effects for reasons of rhetoric, politeness, because speakers feel like intentionally repeating items from previous discourse, or because they have been primed in preceding discourse – but it is not easily possible to disentangle the above motivations through corpus study in a waterproof fashion.

3. to demonstrate that consideration of the phenomenon can increase the linguist's ability to account for linguistic variation, and to predict speakers' linguistic choices more accurately.

## 2 Method and data

To this end, I will conduct – as a case study – an analysis of the alternation between infinitival complementation (henceforth: *V+inf.*), as in (2), and gerundial complementation (henceforth: *V+ger.*), as in (3), after a number of head verbs (such as *begin*) after which either complementation pattern is possible:

- (2) John *began to wonder*
- (3) John *began wondering*

There appears to be more or less of a consensus that these complementation patterns themselves have semantic content. In this spirit, Quirk et al. (1985, p. 1191) state that

where both constructions [. . .] are admitted, there is usually felt to be a difference of aspect or mood which influences the choice. As a rule, the infinitive gives a mere 'potentiality' for action, as in *She hoped to learn French*, while the participle gives a sense of the actual 'performance' of the action itself, as in *She enjoyed learning French*.

Yet, it has proven to be notoriously hard to pin down these differences (cf. Quirk 1974, pp. 66-67: "There ought to be a big award for anyone who can describe exactly what makes him say 'I started to work' on one occasion and 'I started working' on another."). This is not the place to even start a review of the voluminous literature on semantic or pragmatic differences between the two complementation types. Suffice it to say that loci of variation in the sense of the present study are verbs whose complementation behavior is maximally unconditioned by semantic factors, *viz.* emotive verbs and aspectual verbs. This means that I will analyze the complementation patterns the following 11 head verbs:

- (4) *begin, cease, continue, dread, hate, intend, like, loathe, love, prefer, start*

The basic idea underlying the present study's empirical approach is that complementation strategy choice after the head verbs in (4) is conditioned by several factors, one of which is persistence. The main method that is going to be utilized is *binary logistic regression*, which is a tool that closely resembles the so-called Variable-Rules approach (cf. Sankoff & Labov 1979). The method provides a more or less theory-neutral heuristic tool of analysis which integrates probabilistic statements into the description of performance. It is applicable "wherever a choice can be perceived as having been made in the course of linguistic performance" (Sankoff 1998, p. 151). Logistic regression has the following advantages over more unsophisticated analysis methods:

- it seeks to predict a binary outcome (=a linguistic choice) given several independent (or predictor) variables;
- it quantifies the influence of each predictor;
- it specifies the direction of the effect of each predictor;
- it states how much of the empirically observable variance is explained by the predictors considered;
- it states how well the model fares in predicting actual speakers' choices.

More specifically, the following information is provided by a logistic regression model:

*The magnitude and the direction of the influence of each predictor on the outcome.* This information is provided by *odds ratios* that are associated with each individual independent. Odds ratios indicate how the presence or absence of a feature (for categorical independents) or how a one-unit increase in a scalar independent influences the odds for an outcome. Because odds ratios can take values between 0 and  $\infty$ , three cases can be distinguished: (i) if  $\exp(b) < 1$ , an increase in the independent makes a specific outcome less likely; (ii) if  $\exp(b) = 1$ , the independent has no effect whatsoever on the outcome; (iii) if  $\exp(b) > 1$ , an increase in the independent makes a specific outcome more likely.

*Predictive efficiency of the model as a whole.* The percentage of correctly predicted cases vis-à-vis the baseline prediction (% correct (baseline)) indicates how accurate the model is in predicting actual outcomes. The higher this percentage, the better the model fares in this endeavor.

*Variance explained by, or explanatory power of, the model as a whole ( $R^2$ ).* The  $R^2$  value can range between 0 and 1 and indicates the proportion of variance in the dependent variable (i.e. in the outcomes) accounted for by all the independent variables included in the model. Bigger  $R^2$  values mean that more variance is accounted for by the model and that the model is substantially more significant.<sup>2</sup>

In short, it will be the present study's job to quantify, probabilistically, how persistence impacts the choice speakers have between infinitival and gerundial complementation. The corpus that will serve as data source for the empirical investigation of complementation strategy choice is the spoken *demographically sampled spoken section of the British National Corpus* (henceforth: BNC-DS). This corpus spans ca. 4.5 million words and consists of colloquial British English - more precisely, "informal encounters recorded by a socially stratified sample of respondents, selected by age-group, sex, social class and geographic region" (Aston & Burnard 1998, p. 31).

In a first step, all occurrences of the head verbs in (4) that were either followed by an infinitive or gerund VP were identified and extracted from the BNC-DS. This yielded a database of  $N=1,876$  relevant head verbs with optional complementation. This database was then coded for the following predictor variables that have previously been claimed to influence complementation strategy choice:

*Hypothetical meaning* (henceforth: HYPOTHETICAL). Is the VP used in a hypothetical context, i.e. is the head verb preceded by *would*, *would not*, *wouldn't*, or *'d not*, as in (5)?

---

<sup>2</sup> The specific  $R^2$  measure which is going to be reported is the so-called *Nagelkerke  $R^2$* , a pseudo  $R^2$  statistic for logistic regression.

- (5) [. . .] he voiced an opinion he *would not like to be put to sleep* if at all possible. (CSPA Wh97a)

(coded 1 if the context is hypothetical, and 0 otherwise)

*Hypothesis:* According to Biber et al. (1999, pp. 757-758), *V+ger.* is unlikely in hypothetical contexts.

*Horror aequi* (henceforth: INF HORROR AEQUI and ING HORROR AEQUI). Is the head verb itself an infinitive, as in (6), or is it an *-ing* form, as in (7)?

- (6) The President has indicated that he was going *to start using* a cane Monday. (CSPA Wh97a)

- (7) The states *are just starting to test* that idea. (CSPA Wh97a)

(coded 1 if the head verb is an infinitive/*-ing* form, and 0 otherwise)

*Hypothesis:* If the head verb itself is an *-ing* form, we expect a *horror aequi* effect such that *V+ger.* is then avoided (cf. Mair 2003, p. 333; Rohdenburg 2003, p. 236). By the same token, we expect that *V+inf.* is avoided when the head verb is itself used in the infinitive.

Inclusion of the above predictors will control for what is best called ‘baseline’ variation, i.e. variation that has nothing to do with persistence. By contrast, the following predictors clearly fall into the domain of persistence:

*Which variant was employed last time there was a choice?* (henceforth: PREVIOUS). Given two successive head verbs in discourse, was the first one complemented in the same way as the second head verb (henceforth: CURRENT) or was the alternative option used? This is the most basic persistence predictor. For illustration, let us return to example (1), re-printed as (8) below for convenience:

- (8) I think it may be 10 years from now when people *start seeing* the long-term effects from it, and you *start having* problems with it. (Corpus of Spoken Professional American English, text Comm8a97)

In (8), there is a match between CURRENT (... *you start having problems* ...) and PREVIOUS (... *people start seeing* ...) with regard to the complementation strategy chosen – in both slots, *V+ger.* is employed.

*Hypothesis:* Use of a given complementation option in PREVIOUS increases the likelihood that the same option will be used in CURRENT.

*Textual distance* between CURRENT and PREVIOUS, i.e. between two choice contexts (henceforth: TEXTDIST). Psycholinguistic research has indicated that the more recently subjects were primed, the greater the priming effect (cf. Bock & Griffin 2000); corpus studies have suggested a similar effect (for instance, Sankoff & Laberge 1978). TEXTDIST will be measured in the *natural logarithm* (henceforth: *ln*) of the number of interjacent words between PREVIOUS and CURRENT<sup>3</sup> and is a proxy for recency of use of an alternating variable. For illustration, again consider (8): textual distance between the two slots in this utterance is seven words (... *the long-term effects from it, and you* ...), thus TEXTDIST would be  $\ln 7=1.95$ .

*Hypothesis:* TEXTDIST interacts with PREVIOUS such that persistence effects are stronger if TEXTDIST is small.

---

<sup>3</sup> The reason that this variable is going to be modeled logarithmically and not, say, in a linear fashion is that many psycholinguistic priming phenomena have been shown to decay this way; ‘forgetting’ functions are rarely linear (see, e.g. Cohen & Dehaene 1998 with regard to inappropriate repetitions due to brain damage; McKone 1995 with regard to decreasing exponential decay of repetition priming).

*Same verb lemma* in PREVIOUS and CURRENT (henceforth: VLEMMALD). This predictor concerns whether two successive complementation slots involve the same head verb lemma (though not necessarily the same head verb form – coded 1 if the lemma is the same, and 0 if it is not). Pickering and Branigan (1998) showed that production priming is stronger when the priming verb lemma and the target verb lemma are the same. This is the case, for instance, in (8), where the head verb *start* is used in two successive complementation slots.

*Hypothesis:* When there is a head verb lemma match between two successive complementation slots, persistence is even stronger than it would be otherwise.

*Textual distance to the last -ing form in the discourse* (henceforth: TEXTDIST-ING). This predictor measures the *ln* of the textual distance between CURRENT and the last generic *-ing* form. The idea is that a generic *-ing* form (as in *reading the book, John became tired*), although not an optional *-ing* complement, might help trigger an option *V+ger.* after a head verb nearby because it shares morphological substance with the *V+ger.* pattern.

*Hypothesis:* The odds for *V+ger.* in CURRENT increase when a gerundial trigger was used recently – i.e. when TEXTDIST-ING is small.

*Number of words starting in <t> in the discourse preceding CURRENT* (henceforth: T-ALLIT). In a context of 50 words before CURRENT, how many tokens are there that – like *to* – start in <t>? There is both discourse analytic evidence (cf. Sacks 1971 and Tannen 1989 on ‘sound coordination’) and psycholinguistic evidence (cf. Dell 1986 and Cohen & Dehaene 1998 on ‘phoneme perseveration’) that speakers prefer alliterating options; the variable thus checks on phonological persistence.

*Hypothesis:* Infinitival complementation categorically involves the token *to*; therefore speakers are more likely to use the infinitival option when T-ALLIT is high – in other words, when there are many words in CURRENT’s context that start in <t>, as does infinitival *to*.

### 3 Results

Table 1a<sup>4</sup> displays logistic regression estimates of how the ‘traditional’ predictors (meaning those hitherto discussed in the literature) influence comparison strategy choice in the BNC-DS. All of these predictors have the expected effect, given the literature. First, if a given head verb is used in a hypothetical context (HYPOTHETICAL), the odds that *V+ger.* will be used are reduced by 98%. Thus, much as claimed by Biber et al. (1999, pp. 757-758), *he would not like to be put to sleep* is clearly more typical than *he would not like being put to sleep*. Second, if a given head verb is itself an *-ing* form (ING HORROR AEQUI), the odds that its complement will be an *-ing*, too, are reduced by 98%. Third – and in a similar vein – if the head verb is an infinitive (INF HORROR AEQUI), this increases the odds that its complement will be *V+ger.* (and *not* an infinitive) by roughly 350%. Therefore, my data bear clear evidence for a *horror aequi* effect such that language users “avoid the use of formally (near-) identical and (near-) adjacent (non-coordinate) grammatical elements or structures” (Rohdenburg 2003, p. 236).

---

<sup>4</sup> In this table, the value in brackets following categorical independents indicates which category of the independent has been tested. Therefore, HYPOTHETICAL(1) tests the presence (as opposed to the absence) of a hypothetical context; this presence is associated with an odds ratio of 0.02.

Table 1. *Complementation strategy choice: logistic regression estimates*

	odds ratio
<i>a. 'traditional' predictors</i>	
HYPOTHETICAL(1)	<b>0.02</b> ***
ING HORROR AEQUI(1)	<b>0.02</b> ***
INF HORROR AEQUI(1)	<b>3.47</b> ***
<i>b. persistence-related predictors</i>	
PREVIOUS( <i>V+inf.</i> )	<b>0.15</b> ***
PREVIOUS( <i>V+inf.</i> ) * VLEMMALD	<b>0.60</b> ***
PREVIOUS( <i>V+inf.</i> ) * TEXTDIST	<b>1.25</b> ***
TEXTDIST-ING	<b>0.82</b> ***
T-ALLITERATIONS	<b>0.89</b> ***
model intercept	15.28 ***
<hr/>	
<i>N</i>	1,876
<i>model chi-square</i>	912.09 ***
<i>Nagelkerke R<sup>2</sup></i>	0.513
<i>% correct (% baseline)</i>	78.1 (50.2)
<hr/>	
* significant at $p < .05$ , ** significant at $p < .01$ , *** significant at $p < .005$ . Predicted odds are for <i>V+ger.</i>	

How important are the above, 'traditional' predictors? In other words, what is the baseline variation observable in complementation strategy choice? Collectively, HYPOTHETICAL, ING HORROR AEQUI, and INF HORROR AEQUI explain ca. 45% of the observable variance in the data and help predict 76% of speakers' actual choices accurately. These figures are a quantitative sketch of the explanatory power associated with factors pointed out in previous scholarship.

Let me now discuss how persistence complements the picture of how speakers choose a complementation pattern. When the persistence-related predictors (Table 1b) are factored in into the logistic regression model, the quality of the model is enhanced significantly (step  $\chi^2=147.2$ ,  $df=8$ ,  $p < 0.001$ ). As a result, explained variance increases to 51% (up from 45%), and predictive efficiency increases to 78% (up from 76%). These increases show what is gained analytically when persistence is considered; in a nutshell, our understanding of complementation strategy choice is improved.

How exactly does persistence function? First, consider PREVIOUS, which is associated with an odds ratio of 0.15. This means that given two successive head verbs with optional complementation, if *V+inf.* was used in the first slot, the odds that the speaker will switch to *V+ger.* in the second slot are reduced by 85%. In plain English, speakers are highly disinclined to switch between structural chunks if they

can avoid it – they prefer to stick to a given chunk once they have used it. Yet, the interaction terms with PREVIOUS (PREVIOUS(*V+inf.*) \* VLEMMAID and PREVIOUS(*V+inf.*) \* TEXTDIST) indicate that the overall persistence effect actually depends on (at least) two moderator variables.

For one thing, consider the odds ratio of 0.60 with which PREVIOUS(*V+inf.*) \* VLEMMAID is associated: the interpretation of this value is that if there is a verb lemma match between two successive head verbs with optional complementation (as in *John began thinking, and then he began worrying*), persistence is 40% stronger than otherwise. It thus seems that speakers can “resist” being repetitive even less than otherwise if they are faced with identical verb lemmas. This corpus finding is consonant with experimental, psycholinguistic research (cf. Pickering and Branigan 1998). On the other hand, note that the interaction term PREVIOUS(*V+inf.*) \* TEXTDIST is associated with an odds ratio of 1.25, meaning that for every one-unit increase in TEXTDIST (thus, in the *ln* of textual distance between two successive head verbs), the persistence effect weakens by 25%. This finding, too, is in accordance with previous psycholinguistic research: much as shown by, for instance, Bock and Griffin (2000), my corpus analysis finds that persistence is stronger if previous exposure to a structural chunk (in this case, a complementation pattern) was recent. Another way of saying this is that, exactly as hypothesized, persistence between two successive optional complementation sites in discourse weakens as textual distance between these sites increases.

Figure 1. *Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same complementation strategy is used) as function of textual distance (in words) between CURRENT and PREVIOUS. Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship*

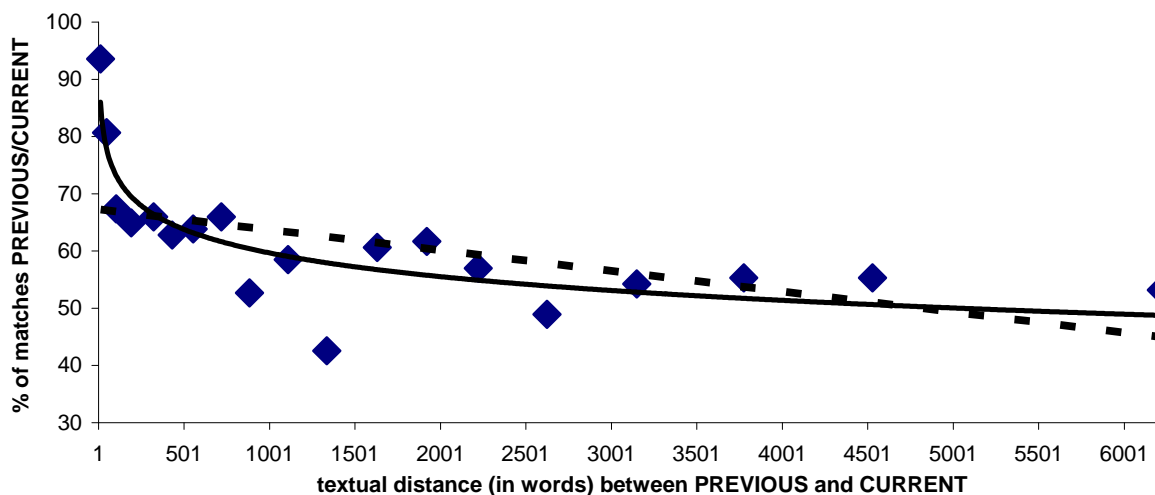
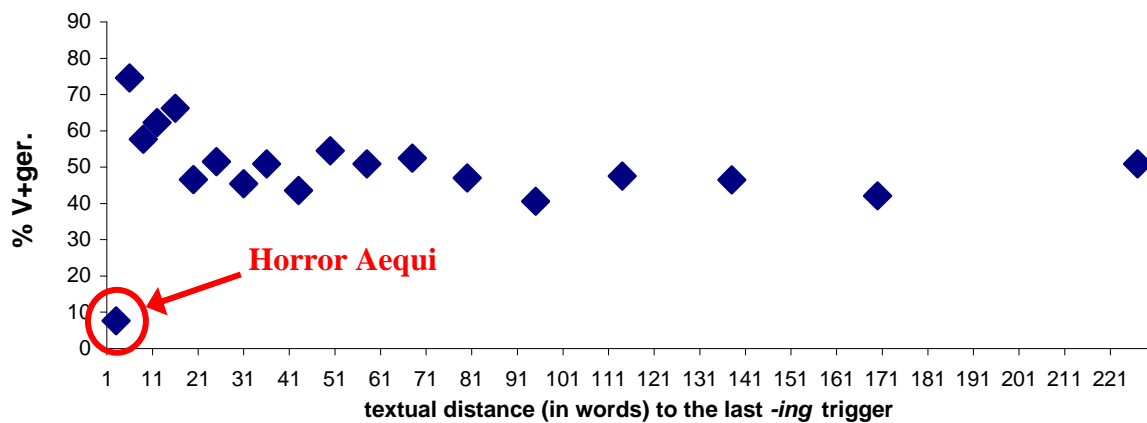


Figure 1 takes a closer look at this relationship by plotting the non-logged textual distance (in words) between PREVIOUS and CURRENT against the percentage of persistent PREVIOUS / CURRENT pairs (i.e. successive head verb pairs after which the

same complementation strategy is used),<sup>5</sup> visually confirming two things: first, the percentage of matched pairs clearly does not bob around randomly. Instead, the more recently a complementation choice has been made, the more likely speakers are to go for the same comparison strategy at the next opportunity. Recency of use clearly plays a role. Second, and also as hypothesized, the “forgetting function” that describes this relationship is logarithmic rather than linear (recall that I had *a priori* modeled TEXTDIST logarithmically in logistic regression): the heavy, logarithmic regression line fits the data much better (adjusted  $R^2=0.74$ ) than the linear estimate (dotted line; adjusted  $R^2=0.26$ ). A logarithmic forgetting function – what does this mean? Simply put, speakers forget about previous choices as time passes by, but they forget comparatively more of this information immediately after a choice is made than at later intervals.

Figure 2. *Share of gerundial complementation (on y-axis) as a function of textual distance to the last -ing trigger (on x-axis)*



Logarithmic decay, as we will see, is also evidenced in the showing of the predictor TEXTDIST-ING, which is associated with an odds ratio of 0.82 in logistic regression (cf. Table 1b). TEXTDIST-ING measures the textual distance between a head verb with variable complementation and the last generic *-ing* form (as in *after reading the book, John began feeling sick*) in discourse. Note, now, that the odds ratio associated with the predictor suggests that for every one-unit increase in the this textual distance, the odds for *V+ger.* decrease by 18%; this is tantamount to saying that the more recently a generic *-ing* form has been used, the greater the odds for *V+ger.* Thus, not only does recent usage of *optional* gerundial complementation encourage usage of *V+ger.* when there is a choice – so does recent usage of any (not necessarily optional) *-ing* form in general. Figure 2, in much the same way as Figure 1, seeks to visualize the relationship between the likelihood for *V+ger.* and textual distance to the last generic *-ing* form in the discourse by plotting the share of *V+ger.* against textual distance to the last generic *-ing* form. As can be seen, from the second

<sup>5</sup> Figure 1 – exactly like Figure 2 below – is based on 19 measuring points. These have been arrived at by dividing the observed textual distance between PREVIOUS and CURRENT into 20-tiles, i.e. into 20 equal groups (which have 19 cut-off points); the percentage of matches between PREVIOUS and CURRENT was then determined separately for each of 20-tiles.

measuring point onwards, the share of *V+ger.* decreases quite steadily as textual distance to the last *-ing* trigger increases; this is as expected. Also observe that a logarithmic decay, once again, can be surmised. Why is it, though, that *V+ger.* is so rare at the first measuring point? This is, of course, the phenomenon known as *horror aequi*: if an *-ing* form has *just* been used, the likelihood that it will be used again is much lower than otherwise (cf. Rohdenburg 2003; Mair 2003). Thus, *-ing* forms are morphologically persistent in that they have the ability to trigger *V+ger.* chunks in nearby slots unless – crucially – they are immediately adjacent to such a slot.

Finally, the predictor T-ALLITERATIONS has a statistically significant effect on complementation strategy choice as well. This effect, too, is the theoretically expected one: *V+inf.* always involves the infinitive marker *to*, which starts in <t>. Logistic regression indicates, then, that for every additional word starting in <t> in a head verb's preceding phonological context, the odds for *V+ger.* decrease by 11% (odds ratio=0.89). Another way of saying this is that when the phonological context is such that a lot of words start in <t>, the likelihood that *V+inf.* will be used in a given complementation slot is greater than otherwise. I submit that the reason is that *V+inf.* is, under such circumstances, better sound coordinated with its environment. This line of thought is consonant with previous discourse analytic (Sacks 1971; Tannen 1989) and psycholinguistic (Dell 1986; Cohen & Dehaene 1998) research.

#### 4 Summary and conclusion

By ways of a multivariate corpus analysis of the alternation between gerundial and infinitival complementation, I believe to have demonstrated that language users, indeed, have a marked tendency to re-use chunks from previous discourse when they have a choice, and that this kind of repetitiveness is sufficiently patterned to be predicted by the analyst.

More specifically, the present study would seem to have lent empirical substance to the following claims:

- If a speaker uses the *V+inf.* pattern (as in *John began to wonder*) at one point, the odds that he or she will switch to the *V+ger.* pattern (as in *John started wondering*) at the next opportunity are reduced substantially.
- This effect is even more sizable if both of the two successive head verb slots are dominated by the same head verb lemma (as in *John began to wonder, and Mary began to think*).
- The more recently a choice for either *V+inf.* or *V+ger.* was made, the more likely it is that the same complementation pattern will be used again.
- The forgetting function that describes the decay of persistence is logarithmic.
- Speakers are more likely to use *V+ger.* if they have just used a generic *-ing* form, hence *-ing* forms can trigger chunks with similar morphological properties (namely, *V+ger.*).

- In environments with lots of words starting in <t>, speakers have a tendency to choose the option, *V+inf.*, that is better sound coordinated with such environments; this is an instance of phonological persistence.

To the extent that apples can be compared to oranges (cf. fn. 1), these findings suggest that corpus data can match experimental, psycholinguistic evidence that the human speech apparatus is geared towards mechanical repetition of previously used or heard linguistic substance, a phenomenon which is known as *production priming*. Observe, along these lines, that there is no way that the forgetting functions that we have encountered could follow from discourse-functional factors. Instead, the logarithmic decay of persistence strongly suggests that the phenomenon must be – to a considerable extent – due to spreading and decaying energy levels (cf. Tanenhaus et al. 1980). Much like, for instance, radioactivity, information about previous chunks seems to decay logarithmically (that is, governed by laws of nature) in the human speech processing “hardware.”

How is persistence relevant to linguistics? The existence of the phenomenon plays havoc with a standard assumption underlying most empirical linguistic inquiry: namely, that an occurrence of a linguistic phenomenon can and should be considered the result of a new throw of the dice, and that it can be investigated in isolation and out of the wider discourse context. This is a problem for qualitative linguistic research where, often, utterances are extracted from some corpus and it is asked, ‘why did the speaker use this specific chunk here?’ My findings leave us good reason to think that the answer might often be as simple as ‘because the speaker has just used this chunk before.’

More generally, persistence is, I believe, of theoretical interest to linguists engaged in very diverse research programs. For mainstream functionalists, persistence is interesting since issues such as online processing constraints and discourse management are involved in motivating surface structure. For less mainstream, more extreme functionalists who view grammar as an emergent system of meaningful repetition and as a “vast collection of hand-me-downs that reaches back in time to the beginnings of time” (Hopper 1998, p. 159), persistence is certainly even more interesting. For Chomskians, the fact that surface persistence may actually yield dysfunctional outcomes (for instance, if functional factors would license some option A, but because of persistence it is option B that is actually used) seems, of course, to support some central tenets of their research program; also, the fact that speech generation is sometimes heavily mechanical (a claim that the present study has certainly not contradicted empirically) and thus self-contained can be interpreted to constitute evidence, albeit somewhat indirect, for the autonomy of syntax hypothesis. Surely, behaviourists – had they not disappeared from the linguistic scene long ago – would find the stimulus-response pattern of persistence, repetitiveness, and prime-target pairs intriguing. Ultimately, persistence might also have implications for historical linguistics: the multiplicative and self-enforcing effect of persistence, coupled with logarithmic forgetting functions, might very well be involved in the S-curve patterns so often observable in language change.

## 5 References

### Primary Sources

British National Corpus (BNC II). Distributed by Oxford University Computing Services,  
<http://www.natcorp.ox.ac.uk/>.

Corpus of Spoken Professional American English. Distributed by athelstan,  
<http://www.athel.com/cspa.html>.

### Secondary Sources

ASTON G. & BURNARD L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.

BIBER D., JOHANSSON S., LEECH G., CONRAD S. & FINEGAN E. (1999) *Longman grammar of spoken and written English*, Longman, Harlow.

BOCK K. (1986) "Syntactic Persistence in Language Production." *Cognitive Psychology*, **18**, 355-387.

BOCK K. & GRIFFIN Z. (2000) "The Persistence of Structural Priming: Transient Activation or Implicit Learning?" *Journal of Experimental Psychology: General*, **129**, 177-192.

COHEN L. & DEHAENE S. (1998) "Competition between past and present: Assessment and interpretation of verbal perseverations." *Brain*, **121**, 1641-1659.

DELL G. (1986) "A Spreading-Activation Theory of Retrieval in Sentence Production." *Psychological Review*, **93**, 283-321.

HOPPER P. (1998). "Emergent Grammar." In *The new psychology of language: cognitive and functional approaches to language structure*. (Ed., Tomasello, M.) , Lawrence Erlbaum Associates, Mahwah, NJ, pp. 155-176.

MAIR C. (2003) "Gerundial Complements after *Begin* and *Start*: Grammatical and Sociolinguistic Factors, and How They Work against Each Other." In *Determinants of Grammatical Variation in English* (Eds., Rohdenburg, G. & Mondorf, B.) Mouton de Gruyter, Berlin, New York, pp. 329-346.

McKONE E. (1995) "Short-term implicit memory for words and non-words." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1108-1126.

PICKERING M. and BRANIGAN H. (1998) "The representation of verbs: evidence from syntactic priming in language production." *Journal of Memory and Language* **39**: 633-651.

QUIRK R., GREENBAUM S., LEECH G. & SVARTVIK J. (1985) *A Comprehensive Grammar of the English Language*, Longman, London, New York.

ROHDENBURG G. (2003) "Cognitive Complexity and horror aequi as factors determining the use of interrogative clause linkers in English." In *Determinants of Grammatical Variation in English* (Eds., Rohdenburg, G. & Mondorf, B.) Mouton de Gruyter, Berlin, New York, pp. 205-250.

- SACKS H. (1971) *Unpublished Lecture Notes*.
- SANKOFF D. (1998) "Sociolinguistics and Syntactic Variation." In *Linguistics: the Cambridge Survey*, Vol. 4 (Ed, Newmeyer, F.) Cambridge University Press, Cambridge, pp. 140-161.
- SANKOFF D. & LABERGE S. (1978) "Statistical Dependence among Successive Occurrences of a Variable in Discourse." In *Linguistic Variation: Models and Methods* (Ed., Sankoff, D.) Academic Press, New York, pp. 119-126.
- SANKOFF D. & LABOV W. (1979) "On the use of variable rules." *Language in Society*, **8**, 189-222.
- TANENHAUS M., FLANIGAN H. P. & SEIDENBERG M. (1980). "Orthographic and phonological activation in auditory and visual word recognition." *Memory and Cognition*, **18**, 513-520.
- TANNEN D. (1989) *Talking voices: Repetition, dialogue, and imagery in conversational discourse*, Cambridge University Press, Cambridge.
- WEINER J. & LABOV W. (1983) "Constraints on the agentless passive." *Journal of Linguistics*, **19**, 29-58.