Natalia Levshina

Anybody (at) home? Communicative efficiency knocking on the Construction Grammar door

https://doi.org/10.1515/gcla-2018-0004

To Doris, wishing her a sense of literal and figurative arrival, when she can finally be (at) home more.

Abstract: The present study focuses on the locative adverbials *home* and *at home*, which are interchangeable in some contexts, e.g. *She decided to stay (at) home today*. Using data from the spoken component of COCA and different multivariate statistical techniques, such as conditional inference trees and dichotomous logistic regression, I investigate the differences between *home* and *at home* with regard to several contextual variables, such as the syntactic function of *(at) home*, the presence of particular adverbs, e.g. *back (at) home*, figurativeness of semantics and the presence of presupposed arrival in the context. Moreover, special attention is paid to the variables that represent predictability of Verb + *(at) home* given a verbal predicate and the other way round, as well as linguistic distance between the predicate and the locative adjunct. The effects of these variables are interpreted as a manifestation of the universal tendency to maximize communicative efficiency and minimize cognitive complexity. I also argue that these effects represent an important social aspect of language use that should be taken into account by contemporary Cognitive Linguistics and Construction Grammar.

Keywords: home, information theory, predictability, multifactorial grammar, cognitive complexity

1 Aims of the study

The present study investigates the use of locative adverbials *home* and *at home* in American English. When the meaning is directional, e.g. *go/return/bring*

Natalia Levshina, Universität Leipzig, natalevs@gmail.com

(someone) home or *a long way home*, no preposition is used. The forms *home* and *at home* can only be interchangeable when the meaning is locative, as in (1):

(1) a. Dads who stay *at home* [COCA, Magazines]
b. Stories abound of men *staying home* to look after newborns [COCA, Magazines]

For brevity, this variation will be called the domative alternation. To the best of my knowledge, it has not been studied systematically by linguists. One of the few mentions of this variation can be found in Huddleston and Pullum (2002: 683). They claim that *home* marks location only as a subject-oriented complement, as in *Are you home? We stayed home*, but not in other contexts, e.g. **I kept my computer home* or **Home, the children were playing cricket*.

At the same time, the use of these expressions attracts language learners' attention, judging from numerous discussions on Internet fora.[1] One of the aims of the present study is to fill this gap and investigate the linguistic factors that influence the use of *(at) home*. I will focus on American English, where this variation seems to be more common, as one can conclude from language users' intuitions and experts' comments.[2] In this study, I will test some of the factors that are mentioned in these discussions, such as figurative vs. literal meaning and the semantics of arrival (see Section 2.2). The data, which will be described in Section 2.1, come from the spoken component of the *Corpus of Contemporary American English* (COCA) (Davies 2008–). I use conditional inference trees (Hothorn et al. 2006) to test the impact of the above-mentioned factors.[3] As Section 3 will demonstrate, these factors turn out to be highly relevant for the domative alternation.

The second aim is more theoretical and has to do with universal principles of language behaviour. Zooming in on the uses of *(at) home* after intransitive predicates in the contexts with the greatest flexibility regarding the choice between the domative variants, I test if the classification can be further improved with the help of variables that reflect the predictability and ease of identifiability of the adjuncts as such. The first hypothesis is that the shorter form *home* is preferred when either the verb or the adjunct is more predictable. Predictability is understood and measured as the conditional probability of a verbal predicate given the adjunct with or without *at*, and as the probability of the adjunct given the verb. These measures are similar

1 E.g. https://english.stackexchange.com/questions/21286/im-home-or-im-at-home, https://www.quora.com/What-is-the-difference-between-I-am-home-and-I-am-at-home, https://www.usingenglish.com/forum/threads/68883-Correct-Usage-home-or-at-home and numerous others.

2 E.g. https://forum.wordreference.com/threads/home.883256/, http://www.bbc.co.uk/worldservice/learningenglish/grammar/learnit/learnitv240.shtml

3 All analyses are done with the help of R (R Core Team 2017).

to Schmid's (2000) notions of Attraction and Reliance. The second hypothesis follows from Rohdenburg's (1996) principle of avoidance of cognitive complexity. In more cognitively complex environments, speakers tend to provide additional formal coding in order to facilitate comprehension. One can expect the longer form *at home* to be used when the distance between the head predicate and the adjunct is larger, which means that the latter is more difficult to identify as such. These theoretical considerations, which will be presented in Section 4, are supported by a dichotomous logistic regression model and less formal analyses reported in Section 5.

Although it is in principle possible to consider several different explanations of predictability effects, they are most likely to be based on the universal bias towards communicative efficiency and economy, as will be argued in Section 6.1. The fact that constructional variation can be influenced by such factors has important consequences for Construction Grammar, which has mostly focused on the semantic and conceptual relationships between constructions and their collexemes (e.g. Gries et al. 2005). In line with another recent trend (cf. Divjak et al. 2016), this study can be regarded as a contribution to the social turn in Cognitive Linguistics (see Section 6.2).

2 Data extraction and variables

2.1 Data extraction procedure

First, I extracted all instances of the wordform *home* from the spoken component of COCA. I also extracted the contexts, which included 25 words on the left and 25 words on the right from the target form, as well as the information about the broadcasting channel and TV or radio program where each observation occurs. Next, the instances were inspected manually, and all spurious hits were discarded. In addition to contexts with verbs of self- and caused motion (e.g. *go home*, *drive someone home*), I removed the instances with such verbs as *expect*, *get, want, allow, call, invite* and *welcome* (*someone home*), where the directional semantics was strong. I also excluded idiomatic expressions, such as *drive/hammer a point/message home, home and dry* and *home free*. In addition, names of films, books, songs and programmes (e.g. *Home Alone* and *Home on the Range*) were removed, as well as the lexicalized uses of *stay at home*, as in *a stay at home mom/dad*.

After that, I still had over 10,000 instances of locative *(at) home*. From this dataset, I took a random sample of 1,000 instances and coded them for several variables, which are described in Section 2.2. The longer variant *at home* in this random sample is almost twice as frequent as the variant with zero marking *home* (more exactly, 652 occurrences of *at home* and 348 occurrences of *home*).

2.2 Contextual variables

This subsection describes the contextual variables. See an overview in Table 1.

Table 1: Overview of the contextual variables.

Variable	Label	Values
1. Locative adverbial or particle before *(at) home*	*PlaceAdv*	*No, back*, *here*, *Other*
2. Literal or figurative meaning	*Figurative*	*Literal*, *Metaph* (metaphorical), Gener (generic)
3. Semantics of arrival	*Arrival*	*Yes*, *No*
4. Syntactic function	*SyntFun*	*Pred_Intr* (intransitive predicate), *Pred_Tr* (transitive predicate), *Sent* (sentence adjunct), *Exist* (existential construction), *Attr* (NP attribute), *Ellipsis* (elliptical structure)
5. Channel	*Channel*	*ABC*, *CNN*, *CBS*, etc.

1. *Locative adverbial or particle before (at) home*. This variable shows whether the domative adjunct was preceded by another locative adverbial or particle. The values were "No", "back", "here" and "Other". The adverbs *back* and *here* were both frequent (95 and 51 occurrences in the sample, respectively), and therefore were taken into account individually. Although they displayed strong preferences for one or the other variant (i.e. *back home* and *here at home*), there were several exceptions. Compare, for example, (2a) and (2b):

   (2) a. I'd rather be poor back *home* than here ... [NPR Tell more]
   b. But back *at home*, the party was over. [CBS 48 Hours]

2. *Literal or figurative meaning of (at) home*. The category "Literal" means that *(at) home* indicates being in a place where someone actually lives. An example is provided in (3).

   (3) I also have geraniums *at home*. [ABC Primetime]

   The second sense is metaphorical. It is used when *(at) home* expresses one's feeling of being comfortable and at ease in a particular situation:

   (4) a. And he's probably more comfortable and *at home* with his stage makeup every day. [Ind Geraldo]
   b. ... so that if they know if anything goes wrong, they're going to be able to survive, and it's like being *home*. [CNN Talkback]

Although the *at*-variant is usually preferred in these situations, as in (4a), there are a few cases when the bare variant is used with metaphorical meaning, as in (4b). The third type is a semantic generalization, when *(at) home* is used to refer to the city or country where one lives. An example of this type is (5a), which discusses a politician, who faces problems in his own country:

(5) a. But for all his achievements on the international scene, the problems he faces *at home* seem insurmountable. [ABC Nightline]
b. ... Republicans I talked to, lawmakers, several of them who are *home* with their constituents, home with potential voters ... [CNN Zahn]

In the generalized contexts like this, the *at*-variant is used more frequently, as in (5a) than the bare variant, as in (5b).

3. *Semantics of arrival*. This variable stands for the presence or absence of contextual clues that suggest that the person or object staying at home has recently arrived there. Compare (6a), where the speaker signals his or her arrival home, with (6b), where this information is not available or relevant:

(6) a. Darling, I'm *home*!
b. Is anybody *home*?

To code this variable, I relied on different contextual clues, including certain temporal expressions (e.g. *soon*, *by now*, *for Christmas*, *finally*, *after a journey to X*) and the previous location (from place X). An example is provided in (7).

(7) I was *home* from college for the summer, and I said I'd do it. [NPR Weekend]

One can expect that arrival should be more often implied in the contexts with *home* than in those with *at home*. The reason is the closeness of such uses to directional semantics, which is expressed by bare *home*.

4. *Syntactic function of (at) home*. This variable is inspired by Huddleston and Pullum's (2002) observation that the unmarked locative *home* can be used only with subject-oriented predicates (see Section 1). Coding the orientation of the predicate as subject- or object-oriented turned out to be very difficult in practice. As a proxy, I decided to use transitivity of the predicate because intransitive predicates are usually subject-oriented. There were also many other functions. The full list is as follows:
   - adjunct of an intransitive predicate, i.e. one without a direct object: *I'm home*;
   - adjunct of a transitive predicate, i.e. one with a direct object: *I build furniture at home*;
   - sentence adjunct: *At home, I drink only tea*;

- attribute that post-modifies a nominal phrase: *Their stores at home are even emptier than here*;
- adverbial modifier in the existential construction *there + BE*: *There is too much stress at home*;
- part of an elliptic structure: *Finally, at home!*

Following Huddleston and Pullum's (2002) claim, we can expect *home* to be used predominantly as an adjunct of intransitive predicates.

5. *Channel*, which stands for the broadcasting channel that the observation comes from, e.g. ABC, CNN and Fox Broadcasting Company. This was done in order to take into account possible variation across the media.

3 Conditional inference tree model of the entire dataset

This section tests the variables which were introduced in the previous section. I use a non-parametric method of conditional inference trees. This is a classification and regression method which has been used in sociolinguistics (e.g. Tagliamonte and Baayen 2012) and variational probabilistic grammar (e.g. Szmrecsanyi et al. 2016). One of the main advantages of this method is that it helps to model complex interactions between predictors in a very intuitive and easily interpretable way (see also Levshina in press). Conditional inference trees are grown based on *binary recursive partitioning*. The algorithm starts with the entire dataset and tries to find the predictor that is the most strongly associated with the response. Then the algorithm makes a binary split in that variable, such that the strength of association or correlation between the predictor and the response is maximized. After that, the procedure is repeated again as long as certain criteria are met. Most importantly, a split can be made when a certain level of statistical significance is achieved, which serves as the minimum criterion for splitting.

To fit a conditional inference tree model, I used the package *party* (Hothorn et al. 2006).[4] The default settings are used (i.e. the minimum criterion for splitting is 0.95, which corresponds to the maximal *p*-value 0.05, the minimum number of

4 I also tried a more recent package *partykit*, which is claimed to have more up-to-date algorithms. The results are identical.

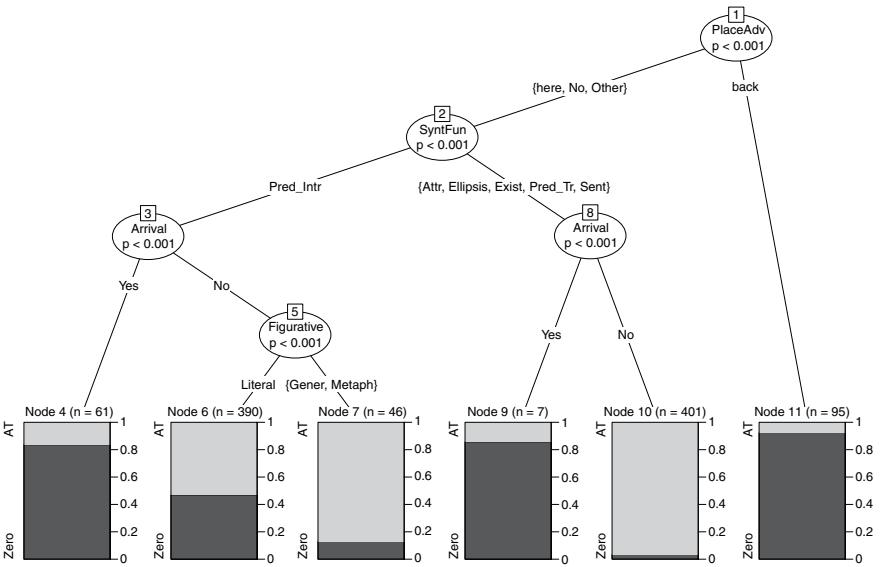

Figure 1: Conditional inference tree of *(at) home*.

observations in a node is seven, the minimum number of observations for a node to be considered for further splitting is 20). The resulting tree is shown in Figure 1.

The first split at the very top (see Node 1) is made in the variable *PlaceAdv*, which stands for the presence of a location adverb modifying *(at) home*. If we have *back* before the domative adjunct, no other splits are made. The corresponding final Node 11 on the extreme right contains all observations with the adverb *back* + *(at) home* in the dataset. The proportions of the bare and *at*-variants are shown in the barplot. As one can see, the variant with zero marking is predominant. This means that the chances of the bare variant after *back* are very high. Consider an example:

(8) See, I tend bar back *home* in Indiana. [NPR Fresh Air]

In the absence of *back*, other variables play a role. Let us examine them. First, consider Node 2, where the split is made in the syntactic function of *(at) home*. It separates adjuncts of intransitive predicates (*Pred_Intr*), e.g. *be (at) home*, which form the left branch, from the other functions, which form the right branch and are then split in the variable related to arrival (Node 8). If the semantics of arrival is prominent, the chances of the bare variant are very high (see Node 9). An example is given in (9).

(9) We can't wait to have your kids *home* safe as well, and for that wedding, just incredible. [ABC GMA]

If it is not prominent, *at home* is used almost exclusively (Node 10), for example:

(10) And you *at home*, go have fun with eggs, and happy Easter, everybody. [ABC GMA]

Let us now go back to the contexts with intransitive predicates. Here, again, a split is made in *Arrival* (Node 3). As in the previous case, the semantics of arrival is associated with a very high proportion of the bare variant (Node 4). Consider an example:

(11) But he promised he would be *home* in a year and he never came home. [NPR Morning]

In all other remaining cases, the distinction between figurative and literal meanings plays a role (Node 5). If the semantics is figurative (generalized or metaphoric), the *at*-variant is almost exclusively used (Node 7), as in the following example:

(12) Well, you are a party animal. ... You were right *at home*. [NBC Today]

If *(at) home* is used in the literal sense, the proportions of the bare and *at*-variants are almost equal (Node 6). Examples are given in (13).

(13) a. I'm lucky because I'm a writer and I work *at home*. [NPR Talk of the Nation]
b. ... if you were sitting *home* on a sunny day while a lot of other boys were playing baseball ... [NPR Fresh Air]

There are 390 such observations, which will be explored further in Section 5.

The classification accuracy of this tree is 77.9 %. This number stands for the proportion of observations where the predictions of the model and the actual variants used in the contexts coincide.[5] If an observation is in a node with predominantly *home*, e.g. Node 11, then the model will predict the bare variant for this observation, as well. In an observation is in a final node with predominantly *at home*, then all observations in that final node will obtain the *at*-variant. The accuracy is higher than the baseline of 65.2 %, which represents the accuracy that can be achieved if one always predicts the more frequent variant *at home*.

Another statistic, which is based on predicted probabilities, is the concordance index *C*. Predicted probabilities for an individual observation are computed on the basis of the proportions of each variant in a given final node. In our case, $C = 86.3$ (with 0.5 for a useless model and 1 as perfect discrimination). This number shows that the model discriminates well between the variants.

5 Based on out-of-bag prediction.

4 Communicative efficiency, information theory and grammar

There is ample evidence that predictability plays an important role at all linguistic levels: phonology, morphology, lexicon and syntax. In particular, high predictability of a word or another phonological unit from the left or right context triggers phonological and morphological reduction, while low predictability increases the likelihood of full variants (e.g. Jurafsky et al. 2001; Aylett and Turk 2004; Mahowald et al. 2013). Similar effects are observed in syntactic variation with optional use of grammatical markers. For example, the object marker in Japanese is omitted in typical agent-patient configurations (Kurumada and Jaeger 2015). Also, Wasow et al. (2011) show that the relativizer *that* or *which* in non-subject relative clauses is more likely to be omitted after definite head nominal phrases (NPs) and with superlative adjectives, as in (14a). This omission is more likely because such NPs are frequently followed by relative clauses. In contrast, the relativizers are less frequently omitted after indefinite NPs, as in (14b), because such NPs are less commonly followed by a relative clause.

(14) a. The most difficult course I ever had was on Kartvelian morphology.
b. I'm having lunch with a colleague *that* I met at the conference.

Thus, speakers tend to provide more formal coding to express less predictable meanings, and less coding to express more predictable meanings. A similar tendency has been observed by typologists. In particular, there is a correlation between the relative frequency of grammatical categories and the degree and optionality of their formal marking (e.g. Greenberg 1966; Haspelmath 2008). The more frequent and therefore more predictable categories, such as singular number, positive degree of comparison and cardinal numerals, usually have shorter formal markers, including zero, than their less frequent counterparts (e.g. plural number, comparative degree and ordinal numerals, respectively). This correlation is explained by a universal bias towards efficient, or economical communication, which is reflected in Zipf's (1949) Principle of Least Effort or Du Bois' (1985) dictum "Grammars code best what speakers do most".

In more recent information-theoretic studies of contextual predictability, a related idea has been expressed as the smooth signal redundancy hypothesis (Aylett and Turk 2004) and the hypothesis of Uniform Information Density (Levy and Jaeger 2007). According to these hypotheses, speakers manage the quantity of information per linguistic unit, providing more formal coding for more informative (i.e. less predictable) units and less coding for less informative ones. As a result, information content is spread (more) evenly across the signal, which results in more efficient communication.

In this paper, I want to focus on a specific type of information content, which reflects the conditional probability of a construction given its collexemes and the other way round. Since language learners and users are very sensitive to such co-occurrence information (e.g. Gries et al. 2005; Ellis and Ferreira-Junior 2009), it would only be natural to expect that it plays a role in constructional alternations. A previous study (Levshina 2018) has shown informativeness effects in the alternation *help* + (*to*) Infinitive. More exactly, when *help* as the control predicate is less expected given the total uses of a verb in the Infinitive slot, the *to*-infinitive is more frequently preferred. Such verbs are highly frequent verbs, e.g. *be*, *have*, *go*, *say* and *think*. A saying attributed to St. Augustine serves as an example:

(15) O Lord, help me to be pure, but not yet.

In addition, in some varieties of English and in some subschemata (e.g. *helping* + (*to*) Infinitive) the information content of a verb given the form of *help* also plays a role. In this paper, I want to test whether the same effects can be observed in the construction $\text{Verb}_{\text{INTR}}$ + (Z) + (*at*) *home*.

Another aspect of communicative efficiency is inspired by Rohdenburg's (1996) principle of cognitive complexity. According to this principle, speakers help hearers to process cognitively complex contexts by providing more explicit formal clues. Complexity is defined structurally, in terms of length and ordering of the constituents, e.g. heavy nominal phrases are more complex than light ones, passives are more complex than active forms, and long syntactic dependencies are more complex than short dependencies. According to one of the manifestations of this principle, speakers use more formal coding when the distance between two parts of a construction is large. Such contexts are more complex because it is more difficult to identify the second part of the construction. Similar ideas are expressed by Mondorf (2009), who argues that the longer and more analytic comparative forms of adjectives are used to mitigate processing demands (the so called *more*-support). Therefore, one can expect the longer form *at home* to be preferred when the locative adjunct appears at some distance from the head verb.

5 Effects of predictability and identifiability

5.1 Generalized linear models of *(at) home* after intransitive predicates

This section reports the results of quantitative analyses, which are based on 390 observations from Node 6 in the conditional inference tree presented in Section 3.

Recall that these observations have no adverb *back*, contain only intransitive predicates, exhibit no semantics of arrival and occur only in the literal meaning. In addition, the variants *home* and *at home* are almost equally distributed in these contexts. I will test, based on the considerations presented in Section 4, whether one can further improve the classification by adding such parameters as constructional information content and Rohdenburg's cognitive complexity. These parameters are represented by the three variables: information content of locative *(at) home* given the verbal predicate, information content of the verb given *(at) home*, and linguistic distance in words between the predicate and the adjunct. These variables are described below.

The first variable is information content of a verb given the domative. This measure is defined as a negative log-transformed conditional probability of the verb given *(at) home*: *InfoVerb* = *–ln P(Verb|Domative)*.[6] The conditional probability *P(Verb|Domative)* was computed as the frequency of the verb with the locative *(at) home* in the sample divided by the total frequency of *(at) home* in the same sample, which equals 1,000. This operationalization represents in a simplified way *P(Verb|Domative) = P(Verb, Domative)/P(Domative)*. It is easy to see that this measure is a negative log-transformed version of Schmid's (2000) Attraction. In other words, greater information content means lower Attraction, and smaller information content means greater Attraction. The somewhat more sophisticated information-theoretic measure is preferred here to simple probabilities because it provides a conceptual link with numerous studies, where information theory is used to explain linguistic phenomena (see examples in Section 4). The lowest scores belong to verbs that occur very frequently with the domative, such as *be*, *stay, have* and *sit*. Consider the verb *stay*. It occurs with (*at*) *home* 100 times, which represents 0.1, or 10 % of the entire sample. Therefore, its information content is $-ln(0.1) \approx 2.3$. An example of a verb with high information content is *cook*, which occurs only once in the sample, which corresponds to the conditional probability of 0.001. Its information content is then $-ln(0.001) \approx 6.9$. This is the maximal value in this sample.

The variable that represents the opposite direction of association is information content of the domative given a verb. This measure represents a negative log-transformed conditional probability of locative *(at) home* given a specific verb: *InfoDomative* = *–lnP(Domative|Verb)*. The conditional probability

6 The present study used the natural logarithm. As a result, information content is measured in nats (so-called natural units of analysis). Information theory often uses the logarithm with the base 2, which means that information content is measured in bits. This difference is not important for the statistical analyses.

P(Domative|Verb) = P(Verb, Domative)/P(Verb) was computed as the frequency of the verb with *(at) home* in the sample divided by the total frequency of the verb in the entire corpus.[7] This measure is the opposite of Schmid's (2000) Reliance and Gries et al.'s (2005) Faith. That is, the greater information content, the smaller the Reliance and Faith score. The lowest scores belong to the verbs that occur infrequently in the corpus, such as *hunker*, *brew* and *dwell*. For instance, *hunker* occurs once in the sample, and 83 times in the spoken subcorpus in total. Therefore, its conditional probability is 1/83 ≈ 0.012, or 1.2 %, and its information content is $-ln(0.012) \approx 4.42$. Among the relatively low-scoring verbs are also such verbs as *stay* and *sit*. For example, *stay* occurs 100 times in the domative sample and 28,161 times in the entire spoken subcorpus, which represents the conditional probability of approximately 0.0035, or 0.35 %. This corresponds to information content of $-ln(0.0036) \approx 5.64$. In contrast, the highest scores belong to such high-frequency 'promiscious' verbs as *do* and *think*, which are used in a multitude of different contexts. For example, *think* occurs only once in the domative sample, and 426,878 in the entire spoken subcorpus, which makes its information content equal to $-ln(0.00000234) \approx 12.96$.

Note that the second measure is not absolute information content, but rather a relative one because the co-occurrence frequencies of a verb and *at (home)* are based on a sample of 1,000 instances of the domative alternation, whereas the verb frequencies are obtained from the entire spoken corpus. This fact matters for interpreting the effect size of this variable, but it does not play a significant role in testing the null hypothesis of no effect in the statistical analyses that follow, however.

Finally, I also measured the distance from the verb in words, which are defined as character strings separated by spaces. This variable represents Rohdenburg's cognitive complexity. An example in (16) has two words between the verb *hit* and the locative part *at home*.

(16) It is what is hitting them most *at home*. [PBS News Hour]

The further *(at) home* is from the head, the more difficult it would be for the hearer to identify the former as an adjunct of the verb. Therefore, one would expect the more explicit form *at home* to be preferred.

All these variables were centred. I also removed one observation with the linguistic distance of 11 words. It was an outlier which skewed the results due to its high leverage.

7 It was practically difficult to obtain automatically the frequencies of phrasal verbs, such as *go on*. Instead, I used the first part (e.g. *go*) to measure the total frequencies.

After trying out several models of different complexity,[8] I fit a normal logistic model with fixed effects (package *rms*, Harrell 2017). The coefficients of the logistic model are shown in Table 2. All three predictors have a significant effect on the choice between the variants. All pairwise interactions between the predictors were tested, but not found to be statistically significant. With every unit, information content of a verb given *(at) home* (*InfoVerb*) increase the chances of the bare variant *home* by the factor $exp(-1.58) \approx 0.21$, which, perhaps more intuitively, means that it decreases the chances of the bare variant by the factor of $exp(1.58) \approx 4.85$. Similarly, information content of *(at) home* given a verb (*InfoDom*) decreases the chances of the bare variant by the factor of $exp(0.54) \approx 1.7$. Finally, every additional word between the predicate and the locative part of the construction decreases the chances of *home* by the factor 1.7.

Table 2: Coefficients of the logistic model with fixed effects.

Terms	*Coefficient*	*S.E.*	*Wald Z*	*P-value*
Intercept	−1.16	0.28	−4.22	< 0.0001
InfoVerb	−1.58	0.26	−5.98	< 0.0001
InfoDomative	−0.54	0.1	−5.26	< 0.0001
LingDist	−0.55	0.29	−2.99	0.0028

The discriminatory power of the model is quite good. The concordance index *C* is 0.812, and the pseudo R^2 is 0.471. The prediction accuracy is 73.4 %, which is higher than the base level of 52.8 %, which is achieved when we only predict the more frequent response in all contexts. Thus, the constructional information about the slot fillers and the distance between the verb and the locative part enable us to improve the classification.

8 First, I fit several generalized additive mixed models (Wood 2006) with the help of the package *mgcv* by the same author because, based on my previous work (Levshina 2018), I assumed non-linear relationships between the logit and the information-theoretic variables. The channels and specific verbs were treated as individual intercepts. However, the models showed no evidence of non-linearity. Further, logistic mixed-effect models (package *lme4*; Bates et al. 2015) with all predictors showed that the random intercepts made no contribution to the explanatory power of the model (based on several likelihood ratio tests).

5.2 Zooming in on particular verbs

An examination of the individual verbs in the dataset representing Node 6 reveals that the bare variant *home* is mostly observed with three verbs: *be*, *stay* and *sit*. Table 3 displays the data for the intransitive predicates that occur with the domative more than five times.

Table 3: Intransitive predicates (frequency with domative > 5): frequency information.

Verb	Frequency *home*	Frequency *at home*	Total frequency with (*at*) *home*	Total frequency of verb in the corpus	InfoVerb	InfoDom
be	179 (66.5 %)	90 (33.5 %)	269 (100 %)	4,928,728	1.3	9.8
stay	76 (76 %)	24 (24 %)	100 (100 %)	28,161	2.3	5.6
sit	8 (25.8 %)	23 (74.2 %)	31 (100 %)	26,202	3.5	6.7
live	0 (0 %)	14 (100 %)	14 (100 %)	45,785	4.3	8.1
feel	0 (0 %)	15 (100 %)	15 (100 %)	67,881	4.2	8.4
work	0 (0 %)	10 (100 %)	10 (100 %)	82,258	4.6	9
watch	0 (0 %)	6 (100 %)	6 (100 %)	32,247	5.1	8.6

The only verbs with bare forms shown in the table are those that are highly frequent in the domative construction: *be*, *stay* and *sit*. The verb *stay* also has the lowest information content of *(at) home*. It also has the highest overall proportion of the bare variant in the sample.

Note that these are not the only possible intransitive verbs to occur with the bare infinitive. An additional analysis of the full data from the spoken corpus reveals several examples with other verbs, e.g. *remain* and *wait*, followed by the bare variant. However, these verbs are much less frequent than *be*, *sit* and *stay*.

(17) a. Listen, you got that dog waiting *home* for you. [CBS 48 Hours]
b. Tali is also arrested but makes bail and is allowed to remain *home* wearing an ankle bracelet tracking device. [ABC 20/20]

For the sake of completeness, I also investigated the transitive predicates. The results are shown in Table 4.

The highest proportions of bare forms are observed with *leave* and *keep*. These have mid-range values of *InfoVerb*, but also the lowest values of *InfoDomative* in the table. The frequencies, unfortunaetely, are too low for a proper statistical analysis. This is left for future research.

This informal analysis supports the conclusions made in the previous section that both types of predictability should be taken into account when predicting the domative variant, both for intransitive and transitive predicates.

Table 4: Transitive predicates (frequency with domative > 5): frequency information.

Verb	Frequency *home*	Frequency *at home*	Total frequency with *(at) home*	Total frequency of verb in corpus	InfoVerb	InfoDom
have	6 (14.6 %)	35 (85.4 %)	41 (100 %)	1,376,480	3.2	10.4
leave	5 (38.5 %)	8 (61.5 %)	13 (100 %)	47,231	4.3	8.2
keep	4 (33.3 %)	8 (66.7 %)	12 (100 %)	51,596	4.4	8.4
do	2 (7.7 %)	24 (92.3 %)	26 (100 %)	1,039,376	3.6	10.6
make	1 (6.7 %)	14 (93.3 %)	15 (100 %)	206,440	4.2	9.5
get	0 (0 %)	8 (66.7 %)	8 (100 %)	395,544	4.8	10.8
call	0 (0 %)	8 (66.7 %)	8 (100 %)	83,599	4.8	9.3
spend	0 (0 %)	6 (66.7 %)	6 (100 %)	28,539	5.1	8.5
try	0 (0 %)	6 (66.7 %)	6 (100 %)	98,063	5.1	9.7

6 Summary and discussion

6.1 Summary of the results

To summarize, the quantitative analyses presented in Sections 3 to 5 reveal the following. First, the variant *home* is strongly preferred in the combination with *back*. In other contexts, it is frequently used with the semantics of arrival is involved, as in *Darling, I'm home!* The variant *at home* is the one frequently used in the figurative meaning, e.g. *feel at home*, and in all syntactic functions with the exception of adjuncts of intransitive predicates, under the condition that the semantics of arrival is not prominent, e.g. *At home, I drink only tea*.

However, there are quite a few contexts where both *home* and *at home* are used with almost equal frequencies. These contexts have the following features: intransitive predicates, non-figurative meaning and the absence of the adverb *back*. In such contexts, information content plays an important role. In particular, the variant *home* is frequently used after the verbs *be*, *stay* and *sit*. These predicates are characterized by their low information content (or high conditional probability) given the domative adjunct. As for the verb *be*, it is actually the most frequent predicate which occurs with *(at) home* in the whole dataset. The verb *stay* is not only frequently followed by *(at) home* compared to the other verbs with the domative adjunct, but is also the one with the highest proportion of being used with *(at) home* relative to its total frequency in the corpus. This means that the information content of the domative given *stay* is low. Therefore, the predictions based on information theory are borne out. Speakers use the shorter and less explicit variant in low-information cases, and the longer and more explicit

variant when the relationship between the construction and the collexemes is more informative, or less predictable.

Additional analyses also suggest that information content may play a role in transitive predicates, especially *keep* and *leave*, the ones after which *(at) home* is the most likely to occur, relative to their total frequencies in the corpus. However, due to the low frequencies of *home* with transitive predicates, this hypothesis needs to be tested on a larger sample.

Finally, the choice between the variants also depends on the distance between the predicate and the locative part. More exactly, the closer these two parts, the higher the chances of the bare variant. This meets the expectations based on Rohdenburg's principle of cognitive complexity. The more distant the second part of a construction, the more difficult it is to recognize it as such. Extra formal marking (here, the preposition *at*) is used in order to facilitate the processing of the construction by the hearer.

6.2 Discussion: Cognition and social interaction

The predictability effect described in this paper can be explained by communicative efficiency and audience design. In the case of *home* and *at home*, the shorter variant is preferred when the hearer (according to the speaker's estimation) has sufficient cues to recognize *home* as a locative expression, and there is no need to use the full variant with *at*. This happens when the adjunct is closely associated with the predicate, and when both constructional components are located closely. Note that the shorter variant is used very infrequently in the other contexts. For instance, *back home* disregarded, all 30 sentence adjuncts in the data represent the full variant *at home*, as in (18).

(18) *At home*, he can be a terror, cutting up rugs and breaking a window. [ABC 20/20]

In such contexts, the sentence adjunct is weakly integrated in the structure of the sentence. It also normally occurs in the beginning of the sentence, so one does not have the left context which could help one to identify the role of the expression.

One might think of alternative explanations, however. The main cause of formal reduction discussed in usage-based linguistics has been neuromotor routinization (Bybee 2010: Chapter 3; Diessel and Hilpert 2016). There is substantial evidence that frequently repeated sequences of strings gradually become perceived, stored and produced as one unit and undergo phonetic reduction, e.g. *going to, want to* and *have to* become *gonna*, *wanna* and *hafta* (Krug 2000).

This may help to explain the preference for the bare variant when the adjunct immediately follows the verb, e.g. *be home*. However, one can also find examples of Verb + *home* with elements in between. An example is provided in (19), where the discussion is about slumber parties with alcohol:

(19) And worse yet, sometimes parents aren't even *home*. [NBC Dateline]

I'm not aware of examples of chunking when users skip elements in-between. Therefore, this explanation does not sound very plausible.

Another possible explanation is facilitation of production. Speaker may use additional markers or extend duration in a different way in order to get extra time for planning and preparing the next segments of the utterance when their lexical material is less available (e.g. Ferreira and Dell 2000; see also an overview in Jaeger and Buz 2018). However, it is difficult to see how this principle can be applied in the case of "backward" probabilities, when the probability of *be/stay/sit* given (*at*) *home* is measured. Generally speaking, such backward probabilities are often more important in explaining reduction effects than forward probabilities (cf. Bell et al. 2009; Seyfarth 2014).

Although obviously more research is needed in order to explain the results, it seems that the communicative efficiency explanation can be useful for this purpose. This explanation is also known as audience design: formal variation is used by the speaker to help the addressee process the utterance (cf. Jaeger and Buz 2018). Unlike neuromotor routinization and facilitation of production, which happen mostly in the language user's brain, communicatively efficient behaviour involves an interaction between the cognitive and social aspects of communication.

Recently, Cognitive Linguistics has witnessed a growing interest in the social aspects of language use (e.g. Croft 2009; Divjak et al. 2016; Schmid 2015; Geeraerts 2016, to name just a few). Communicative efficiency is relevant to socially oriented Cognitive Linguistics at two levels. The first level is that of the interaction between the interlocutors and the behaviour of the speaker, who tries to achieve his or her communicative goals by investing as little effort as possible. This behaviour is unconscious, but in principle rational. The second level of that of language as a system, which emerges in the correlated processes of entrenchment (the cognitive dimension) and conventionalization (the social dimension) (Schmid 2015). In Keller's words (1994: 57), it is an unintended result of intended actions. As language users try to maximize the communicative efficiency of their linguistic behaviour, the use of *home* after certain verbs and with small linguistic distance may become both entrenched and conventionalized. Individual communicatively efficient behaviour thus becomes a part of the common language system.

Cognitively oriented Construction Grammarians have been mostly interested in the language users' knowledge of constructions, constructional networks, their semantic features, compatibility between verbs and more schematic constructions, etc. All these aspects are vital for a cognitively plausible description of language. The present study demonstrates, in addition, that constructional predictability and ease of identifiability also play a role in the speaker's choices between different constructional variants and therefore should be incorporated in the descriptions of linguistic constructions.

References

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. http://doi.org/10.18637/jss.v067.i01

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Croft, William A. 2009. Toward a social cognitive linguistics. In Vyvyan Evans & Stéphanie Pourcel (eds.), *New directions in cognitive linguistics*, 395–420. Amsterdam: John Benjamins.

Davies, Mark. 2008 –. *The Corpus of Contemporary American English (COCA): 560 million words, 1990–present*. Available online at https://corpus.byu.edu/coca/.

Diessel, Holger & Martin Hilpert. 2016. Frequency effects in grammar. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*. New York: Oxford University Press.

Divjak, Dagmar, Natalia Levshina & Jane Klavan. 2016. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27(4). 447–464.

Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax*, 343–65. Amsterdam: John Benjamins.

Ellis, Nick C. & F. Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 188–221.

Ferreira, Victor S. & Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40. 296–340.

Geeraerts, Dirk. 2016. The sociosemiotic commitment. *Cognitive Linguistics* 27(4). 527– 542.

Greenberg, Joseph. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

Harrell, Frank E. Jr. 2017. rms: Regression modeling strategies. R package version 5.1-1. https://CRAN.R-project.org/package=rms (last access 20.07.2018)
Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.
Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.
Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Jaeger, T. Florian & Esteban Buz. 2018. Signal reduction and linguistic encoding. In Eva M. Fernández & Helen Smith Cairns (eds.), *Handbook of psycholinguistics*, 38–81. Hoboken, NJ: John Wiley.
Jurafsky, Daniel, Alan Bell, Michelle L. Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: John Benjamins.
Keller, Rudi. 1994. *On language change: The invisible hand in language.* London: Routledge.
Krug, Manfred. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin & New York: Mouton de Gruyter.
Kurumada, Chigusa & T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83. 152–178. http://doi.org/10.1016/j.jml.2015.03.003
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). 55.1–22. http://doi.org/10.5334/gjgl.294
Levshina, Natalia. in press. Conditional inference trees and random forests. In Magali Paquot & Stefan Th. Gries (eds.), *Practical handbook of corpus linguistics*. Berlin & New York: Springer.
Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt & Thomas Hoffman (eds.), *Advances in neural information processing systems (NIPS)*, Volume 19, 849–856. Cambridge, MA: MIT Press.
Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi & Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126. 313–318.
Mondorf, Britta. 2009. *More support for more-support.* Amsterdam: John Benjamins.
R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (last access 20.07.2018)
Rohdenburg, Günther. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182.
Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin & New York: Mouton de Gruyter.
Schmid, Hans-Jörg. 2015. A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association* 3. 1–27.
Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1). 140–155.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2). 109–137.

Tagliamonte, Sally & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was*/*were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Wasow, Thomas, T. Florian Jaeger & David M. Orr. 2011. Lexical variation in relativizer frequency. In Horst J. Simon & Heike Wiese (eds.), *Expecting the unexpected: Exceptions in grammar*, 175–195. Berlin: De Gruyter Mouton.

Wood, Simon N. 2006. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Zipf, George K. 1949. *Human behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.