Grammatical Change Begins within the Word:
Causal Modeling of the Co-evolution of Icelandic Morphology and Syntax

Fermín Moscoso del Prado Martín (fmoscoso@linguistics.ucsb.edu)
Department of Linguistics, UC – Santa Barbara
Santa Barbara, CA 93106 USA

Abstract

I introduce a combination of information-theoretical and causal modeling to study the cascading of changes between the morphology and the syntax of a language on a diachronic scale. Through the analysis of a historical treebank of Icelandic language ranging from the XII to the XXI century, I show that it is changes in the inflectional morphology of the language that triggered changes in its syntax. This offers a novel and powerful approach to draw conclusions in historical linguistics from a macroscopic perspective. In addition, these findings have implications for the dynamical properties of the linguistic system.

Keywords: Granger-causality; Historical Linguistics; Icelandic; Information Theory; Morphological paradigms; Syntactic complexity.

Introduction

A common observation in the field of Historical Linguistics is that grammatical changes in language tend to be cascaded (e.g., Biberauer & Roberts, 2008; Lightfoot, 2002). Changes at a given level of language (e.g., phonology, morphology, syntax, …) disturb the unstable equilibrium at which human languages reside, triggering a cascade of further linguistic changes as a result. These cascaded changes continue until the system reaches a new meta-stable state (e.g, Croft, 1995; Smith, 1996). For instance, it is widely documented that the loss of the grammatical case markers (morphology) in Late Middle English led to a more rigid word order (syntax) in Early Modern English (cf., Fisiak, 1984).

Traditionally, historical linguists have tracked these cascaded changes by looking for the earliest document at which a particular grammatical innovation can be found or –conversely– at the latest time in which a later extinct construction was documented. Historical linguistics often relies on hard dichotomies on the presence or absence of individual words, affixes, or constructions. This approach is clearly useful for documenting the approximate time at which particular constructions appeared or disappeared. The approach is however limited in its power to detect more subtle forms of grammatical change. For instance, as I will argue in this study, Icelandic morphological paradigms are remarkably resilient, having survived with little change for a such a long period, that most of its current system can be directly traced –virtually unchanged– all the way up to the Old West Norse of the XII century. However, despite the striking conservativeness of the Icelandic paradigms, their patterns of usage have changed along this period.

One can obtain a higher degree of sensitivity by studying the frequencies (and implicitly the probabilities) of usage of different constructions in diachronic scale. These often reveal gradual changes in the grammar of a language. For instance, Ellegård (1953) documented how the usage of the English periphrastic *do* construction (e.g., *I do not speak* vs. *I speak not,* or *Do you speak?* vs. *Speak you?*) arose gradually –rather than abruptly– during a period of two hundred years, from the late XV century to early XVIII century. With the wide availability of diachronic corpora in electronic form, these frequency-based methods have gained much prominence in recent years (see Hilpert & Gries, in press, for a recent survey of quantitative methods in historical linguistics). Admittedly, frequency methods offer an improved sensitivity to the gradualness of grammatical changes along time. Most often, these methods are applied to relate the evolution of the usage of a particular construction in different contexts. The natural tools to achieve this type of inferences are different types of regression analysis. These analyses study the correlations between different factors that might affect the emergence or demise of a construction. For instance, Hilpert (2013) uses such tools to analyze the evolution of (among many others) the patterns of usage of the English future constructions *will* vs. *be going to*. Using a logistic regression analysis, he identifies several factors that significantly co-occur with the uses of either construction.

One must –however– keep in mind the old adagio: "correlation does not imply causation". Even in the cases when one finds significant correlation between the frequencies of use and co-occurrence of different linguistic patterns and constructions, it still remains problematic to argue that one pattern *causes* another. Back to example above, the causal connection between the loss of grammatical case and the fixation of word order is in fact not so trivial on the basis of the historical data alone. Evidently, I cannot infer that I have just had dinner because it stopped raining, even if I observed both events in a sequence. To make such arguments, I would require some form of statistical evidence on how reliably do I start eating whenever it stops raining. Similarly, just the temporal sequentiality between the loss of case marking and the emergence of rigid word order does not –by itself alone– necessarily warrant causality. For instance, Kiparsky (1996) discusses how fixed word order arose also in Icelandic while the case markers were preserved.

Fortunately, the inherent temporality of diachronic data lends itself naturally to the use of more explicitly causal methods. In the presence of temporal sequences, one can use *Granger-causality* (Granger, 1969) analyses to assess whether one time series is significantly affecting the evolution of another. Importantly, in contrast with plain correlations, Granger-causality is asymmetrical, and takes time as a crucial factor in its computation. This reflects the old philosophical constraint that causes must always temporally precede their effects.

Co-evolution of complexities

In the discussion above I have assumed that the object of study should be individual constructions, as has been so productively exploited in the philological tradition. However, a parallel and also important line of investigation concerns not so much the individual constructions themselves, but the more general macroscopic balance in the system. This approach, although more detached from descriptive linguistics, is of great importance for drawing inferences on the types of constraints and mental processes that drive the very nature of human languages. As important as is the study of the individual constructions, it is to be able to generalize between those specific instances. Jointly considering strict causality measures with a detached macroscopic look at the morphological and syntactic systems, is the main objective of this study. With this goal, I advance an hypothesis of how language change would be reflected in measures of language complexity.

Consider that we have two coupled measures, one measuring the overall complexity of a language's inflectional morphology system, and another measure of the syntactic complexity of the language (i.e., irrespective of individual constructions). As is illustrated by Fig. 1, plotting the joint temporal evolution of these two measures along the language's history would reveal the periods in time at which change is occurring in the language. Notice that grammatical change implies that, during the temporal interval at which change is taking place, the old grammatical forms on their way out of the language, would co-exists with the newer forms. Therefore, these critical periods of change would be revealed by peaks on the complexity measures, whereas the meta-stable periods in which language is not changing so much would be represented by flat valleys in the complexity curves.

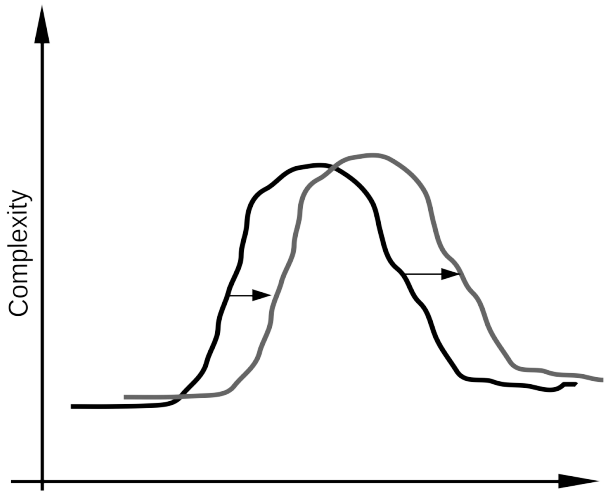

Figure 1. Expected pattern of a causal relation in language change. Change on the hypothetical grammatical component plotted by the black line triggers change in the component plotted by the gray line.

When one simultaneously considers two (or more) of these curves, there are multiple patterns that could become apparent:

1. Change could be happening at just one of the curves. This would be indicative of change at one level of the grammar which is not cascaded into other levels.
2. Change happens simultaneously in both curves, as would be indicated by roughly overlapping peaks. This would signal co-temporal change at both levels in the language. In turn this could be due to both of the changes being driven by some other event. Whether this event would be endogenous (e.g., changes in the phonology of the language) or exogenous (e.g., increased language contact during this period due to geo-political events) is beyond the scope of this study. Another possibility in this case could be that of two perfectly coupled systems, in which changes propagate across levels almost instantly.
3. Changes at one level are temporally followed by changes at the other, which is the pattern illustrated in Fig. 1. In this case it would be possible to talk about cascading of the changes across the levels. However, prior to making such argument one would need explicit statistical evidence that the second level is indeed "following" the first one. Granger-causality is very well suited to test such an hypothesis.

In this study, I investigate the presence of such patterns between the morphology and the syntax of Icelandic. In what follows, I begin by introducing the measures of diversity used to compute the curves. I then compute the curves for a large diachronic corpus of Icelandic with full morphosyntactic annotation (the *IcePaHC* corpus; Wallenberg, Ingason, Sigurðsson & Rögnvaldsson, 2011), and finally I assess whether or not a causal relation can be statistically argued on the basis of the curves.

Grammatical Diversities

With the goal studying the co-evolution of the richness of the inflectional paradigms (inflectional diversity), and the diversity of the syntactic structures (syntactic diversity) in a

language, I introduce here suitable macroscopic measures of these two diversities. Importantly, both of these measures are also of demonstrated psychological relevance for language processing, a factor that will become important in the discussion.

Inflectional Diversity

In recent years, it has become evident that the richness of a base word's morphological paradigms (i.e., the set of forms that can be created from it using consistent inflectional process, such as prefixing and suffixing) is an important factor in the mental representation of the word. Moscoso del Prado, Kostić, and Baayen (2004) showed that this richness was best demonstrated by the entropy (Shannon, 1948) of an inflectional paradigm. If p_i denote the probabilities of occurrence of the of the inflected forms of a word *w* having *n* distinct inflectional variants, they defined the *inflectional entropy* as,

$$H(w) = -\sum_{i=1}^{n} \frac{p_i}{\sum_{i=1}^{n} p_i} \log \frac{p_i}{\sum_{i=1}^{n} p_i}$$

This measure has been demonstrated to be an major co-determiner of human lexical recognition latencies, resiliency and regularity of inflectional processes (Baayen & Moscoso del Prado, 2005), and crucially, predicts the relative ease with which the paradigms are acquired (Stoll et al., 2012).

Inflectional entropy is therefore a good measure of the diversity of inflected forms in which a word can appear. Therefore, the average of the inflectional entropy across all the words in a language will provide an accurate measure of the richness of its inflectional paradigms. This average entropy is in fact just the conditional entropy of the inflected forms given the lemmas. If ℓ denotes a lemma in a language, and *w* denotes an inflected form, we can define the entropy of the lemmas as

$$H(L) = -\sum_{\ell} p(\ell) \log p(\ell)$$

where $p(\ell)$ denotes the probability of encountering the lemma ℓ. Further, we can define the joint entropy between the lemmas and the word forms

$$H(W, L) = -\sum_{w,\ell} p(w, \ell) \log p(w, \ell)$$

where $p(w, \ell)$ denotes the probability with which the inflected word form *w* belonging to lemma ℓ is found in the language. On the basis of these two measures, the average inflectional entropy is exactly given by the conditional entropy

$$H(W|L) = H(W, L) - H(L).$$

which measures the richness of the inflectional paradigms in the language.

Syntactic Diversity

I refer to syntactic diversity as the diversity of the syntactic structures that can be used in the language. According to some theories of grammar, this diversity is potentially infinite for human languages. For instance, if the grammar is defined by a simple recursive phrase-structure grammar, there is an unbounded number of possible grammatical structures that could be generated by the grammar. This however is not the case when one considers that the grammars themselves must be probabilistic. Even if – formally speaking– all structures generated a probabilistic grammar with infinite generative capacity are possible, reality dictates that many such structures will be abysmally improbable, and can for all practical reasons can be overlooked.

If we describe the grammar of a language at a particular point in time using a probabilistic context-free grammar (PCFG; Booth & Thompson, 1973; Charniak, 1997), it is also possible to compute the entropy of the parse trees generated by this grammar, which I will denote by *H(G)*. Notice here that, as discussed above, the number of syntactic trees that could *potentially* be generated by a PCFG is possibly infinite. However, the number of trees that will be generated with a non-negligible (cf., Shannon, 1948) probability is actually finite, and given by the exponent of the entropy. Making use of a theorem by Grenander (1976), Chi (1999) showed that the entropy of a PCFG can be computed exactly from its rules and the associated probabilities. Chi further demonstrated that this entropy is finite for any PCFG that can be inferred by maximum-likelihood from a corpus.

Similar to what was found for inflectional entropies, Hale (2006) showed that the *syntactic entropy H(G)* is a predictor of the relative ease or difficulty with which a sentence is understood by people.

Corpus Analysis

Materials & methods

Materials. I obtained the Icelandic Parsed Historical Corpus (*IcePaHC v0.9;* Wallenberg et al., 2011). This corpus contains morpho-syntactically annotated (i.e., parsed and lemmatized) samples of Icelandic language ranging from AD 1,150 to AD 2,008. This period covers basically the whole history of the Icelandic language, from its origins as an Old West Norse dialect, to the current official language of Iceland. The corpus contains roughly 1,000,000 word tokens of parsed Icelandic, divided into 61 files of similar sizes (around 18,000 words per file). Each of the files correspond to documents from one particular year. The

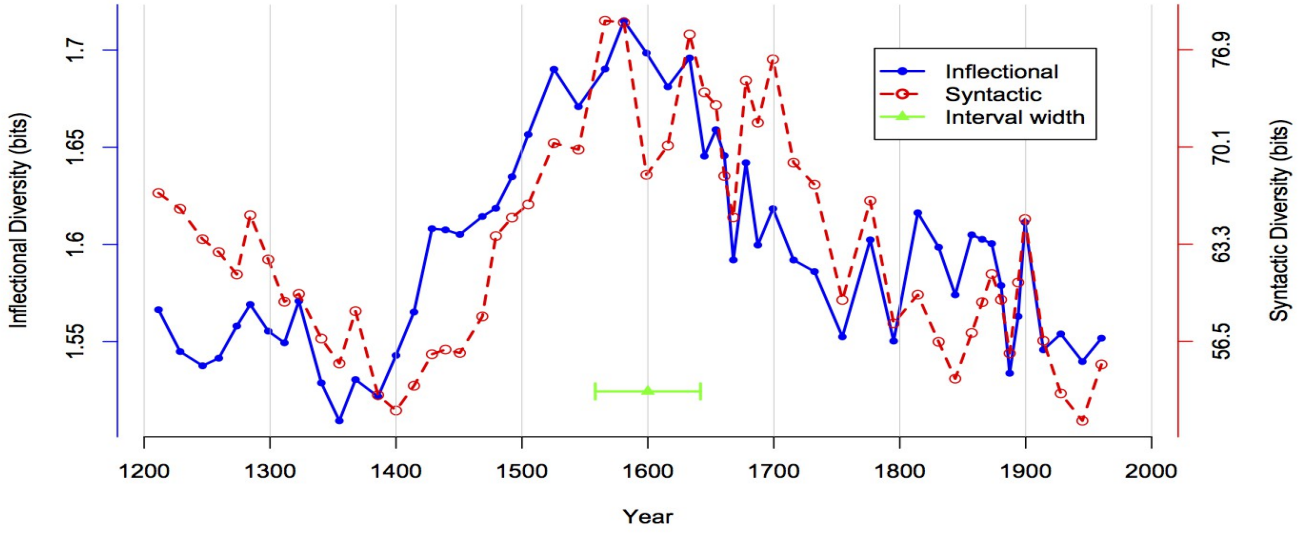

Figure 2. Diachronic co-evolution of morphological and syntactic diversity in Icelandic, according to the *IcePaHC* corpus. The solid blue line plots the evolution of the richness of the inflectional paradigms. The dashed red line plots the evolution of syntactic diversity in the same period. The central segment plots the average width of the time interval over which the temporal smoothing was performed. Notice that the vertical scales are different for inflectional diversity (left vertical axis), and syntactic diversity (right vertical axis).

documents were chosen to cover the period in an approximately uniform manner, sampling from different genres at each period (inasmuch as possible).

Preprocessing. In order to increase the available sample size for each particular year, and also to account for the gradual nature of language change (e.g., Ellegård, 1953), I constructed new samples based on a sliding window technique. Taking the files in chronological order, I slid a window of seven files across the list, at each step constructing a sub-corpus with all the documents in each file in the window. In this way, I obtained 55 sub-corpora, each containing an average of 101,123 ± 1,974 word tokens. Notice that, in this way, each sub-corpus overlaps strongly with the previous one. Using a window of seven files amounts to considering samples spanning an average of 84 ± 4 years, which is indicated by the width of the green segment on Fig. 2.

Computation of the diversities. For each of the files, I estimated the inflectional diversity ($H(W|L)$) and syntactic diversity ($H(G)$) measures described above.

In order to estimate $H(W|L)$, one must consider that the probability values obtained from the corpus are maximum-likelihood estimators, and are therefore certain to result in substantial underestimations on the entropy. For this reason, I made use of the Coverage-Adjusted Entropy Estimator of Chao and Shen (2003). This techniques relies on classical Good-Turing frequency smoothing (Good, 1953) of the word frequencies combined with an additional correction of the entropy equations to account for unseen terms.

Similarly, estimation of $H(G)$ required first performing a simple maximum-likelihood grammar induction on each of the 55 treebanks (i.e, by simply counting the number of instances of each grammatical rule, discounting the lexical productions). The entropy was then estimated using the technique of Chi (1999). In addition, to also account for possibly unobserved context-free rules, the Chao-Shen technique was also applied to the rule probabilities.

Results and Discussion

Fig. 2 plots the chronological evolution of the estimated inflectional (blue solid line, solid dots) and syntactic (red dashed lines, open dots) diversities for the 55 sub-corpora. The year for each of the sub-corpora was computed as the average of the years included in the seven sampled files. Notice the different scales of the diversities, plotted on the left and right vertical axes.

The pattern is remarkably reminiscent of what was predicted in Fig. 1. The inflectional diversity begins a rapid increase in the mid XIV century, reaching a peak in the late XVI century, from which it follows a downwards stabilizing pattern down to contemporary Icelandic. This indicates that a change in the usage of the paradigms was occurring during this period. We will return later to what these changes might be reflecting, here I will just remark that the peak of the

change coincides with the historical point on which the Icelandic bible was printed, a major milestone in the standardization of modern Icelandic.

The syntactic diversity measure on the other hand, follows a pattern very similar to that observed for the inflectional diversity, only apparently lagged by some 20-30 years. If this lagging was found to be reliable, it would mean that the changes in Icelandic syntax were indeed a consequence of the changes in its morphology, with a lag of approximately one generation between the start of the change in the inflectional morphology, and the beginning of the syntactic change.

Causality analyses

In order to assess whether there was indeed a causal relation between the inflectional and syntactic entropy time series, I turn now to a Granger-causality analysis.

Granger-causality. If ***x*** and ***y*** are *stationary* time series, in order to test whether ***x*** *Granger-causes* ***y***, one starts by fitting autoregressive models that predict the values of ***y*** from its own *m* values lagged into the past,

$$y[t]=a_0 + a_1y[t-1] + a_2y[t-2] + ... +a_my[t-m] + \varepsilon[t].$$

One then augments the autoregression by including *v* lagged values of ***x***:

$$y[t]=a_0 + a_1y[t-1] + a_2y[t-2] + ... +a_my[t-m] + \\ + b_1x[t-1] + b_2x[t-2] + ... +b_vx[t-v] + \varepsilon[t],$$

where the $\varepsilon[t]$ terms denote uncorrelated (white) gaussian noises. If the second regression is a significant improvement over the first, then it can be said that ***x*** Granger-causes ***y***, indicating that past values of ***x*** significantly predict future values of ***y*** over and above any predictive power of ***y***'s own past values. This is tested using an *F*-test, with the null hypothesis being that the second model does not improve on the first one.

Analysis. Despite the Chao-Shen corrections, both entropy measures retained significant correlations with the size of the sub-corpora and with the richness of its vocabulary. Furthermore, the width of the chronological interval covered by each particular sub-corpus was also linearly related to the entropy measures. In order to avoid these factors acting as confounds linking the two entropy measures, I residualized them away by performing linear regressions with each of the entropies as the dependent variable, and the variables mentioned above as co-predictors. I retained for the causality analyses the residuals from these regressions. Also, in order to improve the comparability of the results, and for better fitting the autoregressive models, both residual time series (which were roughly normally distributed) were standardized to zero mean and unit variance.

As can be appreciated in Fig. 2, the peaking trend present in both time series makes them non-stationary, a fact that was confirmed by a Kwiatkowski-Phillips-Schmidt-Shin test for trend stationarity (Inflectional Entropy: KPSS Level = *0.55, p=.03*; Syntactic Entropy: KPSS Level = *.40, p=.08*). As the Granger-test requires that both series are stationary, I differentiated the sequences (i.e., took the differences between consecutive values). These differentiated sequences were not significantly deviating from stationarity. A further level of differentiation eliminated the unit roots, while retaining the stationarity (Inflectional entropy: KPSS level=*.27, p=.1*; Syntactic entropy: KPSS level=*.14, p=.1*). These doubly differentiated sequences were deemed suitable for Granger-causality analyses.

I selected the optimal autoregressive orders (values *m* and *v*) in the autoregression equations by fitting autoregressive models with different number of lags, and choosing the model with the lowest AIC value (in fact all other model selection criteria coincided in this choice). This indicated that the optimal models were those with a single lagged value in time. In order to be safe, I included two terms (*m=2, v=2*) in both autoregressions (the results did not change when these lags were anywhere between 1 and 6).

I fitted four autoregressive models (i.e, two models to test whether inflectional entropy Granger-causes syntactic entropy, and two models to test the reverse hypothesis, whether syntactic entropy predicts inflectional entropy). Model comparisons revealed that inflectional entropy significantly Granger-causes syntactic entropy ($F(2,94)=4.59$, $p=.01$), whereas –crucially– syntactic entropy does not Granger-cause inflectional entropy ($F<1$). These results confirm the intuition from observing Fig. 2, the changes in the syntax are the consequence of the changes in the inflectional morphology.

Conclusion

I have shown that there is substantial statistical evidence to assert that the changes in Icelandic grammar were triggered by changes in its inflectional morphology. This confirms previous intuitions that grammatical change spreads outwards from the words into the rest of the grammar (e.g., Lightfoot, 2002). In doing so, I have introduced I new method for studying grammatical change from a macroscopic perspective.

The periods where change is more marked coincide with major changes in the history of Icelandic. The most dramatic changes correspond to the end of the Middle Icelandic period (AD 1350 – AD 1550), reflecting the formation of modern Icelandic (AD 1550 onwards). Interestingly, it would superficially seem as if the morphology of Icelandic had changed very little. Even today, Icelandic inflection is, except for very slight details,

virtually identical to the inflection of Old West Norse. The change in the complexity of the inflectional paradigms must have taken place within the relative frequencies of use of each of the forms (e.g., nominal cases, verbal forms). This is indicative of those inflectional markers changing their functions, which is known to affect the mental representation of words (Kostić, Marković & Baucal, 2003).

Another possibility is that, either both the inflectional, or both the inflectional and the syntactic changes were themselves triggered by changes in the phonology of Icelandic. Indeed, phonology is the component of Icelandic grammar that has undergone the strongest change in historical times. Therefore it is very likely that phonology also played an important role in triggering these changes.

Importantly, changes in the entropy of an inflectional paradigm are known to have important consequences for the cost of processing (Moscoso del Prado et al., 2004) and acquiring words (Stoll et al., 2012). This would be in line with theories positing that problems and changes in language acquisition are crucial drivers of grammatical change (e.g., Lightfoot, 2002). A problem in the acquisition of morphological paradigms could in turn trigger problems in the acquisition of syntax. This is reinforced by the lag of approximately one generation between the onset of inflectional and grammatical change, which as be seen in Fig. 2. One can therefore speculate that it is the changes in the frequencies of the paradigm members in the parent generation leading to more inconsistent evidence for the children. This would in turn trigger syntactic changes in this generations, causing a sort of chain reaction that results in syntactic changes (word order, etc.) across the grammar.

In sum, I have shown how information theory and causal modelling are useful for the study of grammatical change at a macroscopic scale.

References

Baayen, R. H. and Moscoso del Prado Martín, F. (2005) Semantic density and past-tense formation in three Germanic languages. *Language, 81*, 666-698

Biberauer, T., & Roberts, I. (2008). Cascading parameter changes: internally driven change in Middle and Early Modern English. In Th. Eythórsson (Ed.) *Grammatical Change and Linguistic Theory: The Rosendal Papers* (pp. 79-114). Philadelphia, PA: John Benjamins.

Booth, T. L. & Thompson, R. A. (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers*, *C-22,* 442–450.

Chao, A., & Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in the sample. *Environmental and Ecological Statistics,10*, 429-443.

Chi, Z. (1999) Statistical properties of probabilistic context-free grammars. *Computational Linguistics, 25*, 131–160

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press/MIT Press

Ellegård, A. (1953). *The Auxiliary Do: The Establishment and Regulation of its Use in English*. Stockholm: Almqvist & Wiksell.

Fisiak, A. (Ed.) (1984) *Historical Syntax.* Berlin: de Gruyter.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40,* 237–264.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37, 424–438.

Grenander, U. (1976) *Lectures in Pattern Theory, Volume 1, Pattern Synthesis*. New York, NY: Springer-Verlag.

Hale, J. T. (2006). Uncertainty about the rest of the sentence. *Cognitive Science, 30* , 643–672.

Hilpert, M. (2013). *Constructional change in English. Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press

Hilpert, M. & Gries, S. Th. (in press). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.), *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press.

Kiparsky, P. (1996). The shift to head-initial VP in Germanic. In S. Epstein H., Thrainsson, & S. Peters (Eds.), *Studies in Comparative Germanic Syntax*, *2,* Dordrecht: Kluwer.

Kostić, A., Marković, T., & Baucal, A. (2003) Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In R. H. Baayen & R. Schreuder (Eds.) *Morphological Structure in Language Processing.* Berlin: Mouton de Gruyter.

Lightfoot, D. (Ed.) (2002). *Syntactic Effects of Morphological Change*. Oxford, UK: Oxford University Press.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition, 94*, 1–18.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27,* 379–423

Smith, J. (1996). *A Historical Study of English*. New York, NY: Routledge.

Stoll, S., Bickel, B., Lieven, E., Paudyal, N. P., Banjade, G., Bhatta, T. N., Gaenszle, M., Pettigrew, J., Rai, I. P., Rai, M. and Rai, N. K. (2012). Nouns and verbs in Chintang: Children's usage and surrounding adult speech. *Journal of Child Language, 39*, 284–321.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9*. Available from `http://www.linguist.is/icelandic_treebank`