Case and Cause in Icelandic: Reconstructing Causal Networks of Cascaded Language Changes

Fermín Moscoso del Prado Martín and **Christian Brendel**
University of California, Santa Barbara
Department of Linguistics, South Hall 3521
Santa Barbara, CA 93106, USA
fmoscoso@linguistics.ucsb.edu and cdbrendel@gmail.com

Abstract

Linguistic drift is a process that produces slow irreversible changes in the grammar and function of a language's constructions. Importantly, changes in a part of a language can have trickle down effects, triggering changes elsewhere in that language. Although such causally triggered chains of changes have long been hypothesized by historical linguists, no explicit demonstration of the actual causality has been provided. In this study, we use co-occurrence statistics and machine learning to demonstrate that the functions of morphological cases experience a slow, irreversible drift along history, even in a language as conservative as is Icelandic. Crucially, we then move on to demonstrate –using the notion of Granger-causality– that there are explicit *causal* connections between the changes in the functions of the different cases, which are consistent with documented processes in the history of Icelandic. Our technique provides a means for the quantitative reconstruction of connected networks of subtle linguistic changes.

1 Introduction

Sapir (1921/2014, p. 123) noticed that "Language moves down on a current of its own making. It has a *drift*" (emphasis added). In Sapir's view, the formation of different dialects requires that the small changes constantly being introduced by the speakers are not just plain white noise, but rather random walks in which minute changes accumulate over time. The very high dimensionality on which languages operate makes cumulative linguistic changes irreversible. Once a change has been effected there is very little chance that the language will ever return to its original state before the change, in the same way that a diffusion process in a very high dimensional space is never going to return to the exact same point in the space. Drift in language is in this respect reminiscent of random genetic drift from evolutionary biology (Wright, 1955). However, Sapir's idea of drift goes further in that he viewed it as a *directional* process, more similar to Wright's (1929) concept of a directional drift related to selectional pressures. In Sapir's view, "language has a 'slope"'; the small changes that accumulate in linguistic drift are not fully random, but rather they reflect the speakers' unconscious cognitive tendency to increase the consistency within their languages. This idea is currently challenged by some researchers (Croft, 2000; Lupyan and Dale, 2015), who are of the opinion that purely random drift –of the same type as that found in genetics–, when coupled with adequate selection mechanisms, is sufficient to account for the diachronic changes observed in the world's languages. Sapir motivated the need for directional change in what he saw as apparent causal chains in language change, which he illustrated with the progressive loss and functional shift of English oblique case markers, into an absolutive case-free system encoding animacy and position relative to the head noun.

'Chain reactions' along the history of a language are particularly well-studied in phonology. Chain shifts (Martinet, 1952) are processes by which the position in perceptual/articulatory space of a phoneme changes in response to the change in position of another phoneme (either moving away from the second phoneme, in a 'push' chain, or moving to occupy the space left void by the other, in a 'pull' chain). A famous example of a chain shift is the Great English Vowel Shift. In a similar fashion, one could think of functional chain shifts

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2421–2430,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

in morphology, by which a certain morphological category takes over some of the functions of another, triggering a chain of 'push' and/or 'pull' movements in other categories. Such cascaded changes have often been reported in diachronic linguistics (Biberauer and Roberts, 2008; Fisiak, 1984; Lightfoot, 2002; Wittmann, 1983).

Icelandic is a famously conservative language. Compared to most other languages, its grammar has experienced remarkably little change since the high middle ages. For instance, Barðdal and Eythórsson (2003) argue that the changes it has experienced from its old phase (Old West Norse; mid XI century to mid XIV century) to its current phase are comparable to the slight changes occurring from early Modern English (late XV century into early XVIII century) into Modern English (from early XVIII century). In terms of inflectional morphology, change in Icelandic has been minimal. For instance, one finds that the nominal paradigms of Old West Norse, are mostly the same as those of Modern Icelandic. Notwithstanding the apparent formal stability of Icelandic cases, there is evidence that they are experiencing subtle changes in their *functions* (Barðdal, 2011; Barðdal and Eythórsson, 2003; Eythórsson, 2000). In particular, Barðdal argues that an accumulation of small syntatico-semantic shifts has finally resulted in a shift in the Icelandic dative's functions (i.e., 'dative sickness'), possibly triggered by earlier changes in nominatives and accusatives (e.g., 'nominative sickness').

In this study, we investigate whether one can reliably detect a drift in the functions of Icelandic case and –crucially– whether there is evidence for causal chain shifts in these functions. In Section 2, we describe the processing of a diachronic corpus of Icelandic to obtain co-occurrence representations of the functions of case types and tokens. Section 3 uses machine learning on the co-occurrence vectors of tokens to demonstrate that the usage of Icelandic cases has been subject to a constant drift along history, a drift that is distinguishable from the overall changes experienced by the language in this period. We then go on –in Section 4– to demonstrate using Granger-causality (Granger, 1969) that there are causal relations between the changes in the different cases, and it is possible to reconstruct a directed network of chain shifts, which is consistent with the directions of causality hypothesized by Barðdal (2011). Finally, in Section 5, we discuss the theoretical implications of our results for theories of language change, as well as the possibilities offered by the technical innovations presented here.

2 Corpus Processing

2.1 Corpus

We used the Icelandic Parsed Historical Corpus (Wallenberg et al., 2011), a sample of around one million word tokens of Icelandic texts that have been orthographically standardized, manually lemmatized, part-of-speech tagged, and parsed into context-free derivation trees. An example of the lemmatization and part-of-speech tagging for a sentence is shown in Fig. 1. The dating of the text samples ranges from 1,150 CE to 2,008 CE, covering most of the history of Icelandic (from its origins in Old West Norse, to the current official language of Iceland). The corpus is divided into 61 files of similar sizes (around 18,000 words per file), each file corresponding to a single document. The documents were chosen to cover the period in a roughly uniform manner, sampling from similar genres across the periods.

2.2 Preprocessing

We collapsed into a single file all documents that were dated on the same year. This left us with 44 files containing texts from distinct years. From each of the files, we discarded all tokens that contained anything but valid Icelandic alphabetic characters or the dollar sign (used for marking enclitic breaks within a word, such as the clitic determiner in *krossins* from the example in Fig. 1). All remaining word tokens were lower-cased, and the '$' character was removed from the stem elements of broken stem plus clitic pairs (e.g., *kross$* was changed into *kross*).

2.3 First Order Co-occurrence Vectors

Ideally, for constructing co-occurrence vectors, it is best to choose as features those words with highest overall informativity, which in fact tend to be those words with the highest occurrence frequencies (Bullinaria and Levy, 2007; Bullinaria and Levy, 2012; Lowe and McDonald, 2000). In our case, however, using plain token frequencies runs the risk of creating a representational space that is strongly uninformative about particular periods in the history of the language. Instead, we used document frequencies, as these still provide a measure

En	*armar*	*kross-*	*-ins*	*merkja*	*ást*	*við*	*Guð*	*og*	*menn*
en	*armur*	*krosur*	*hinn*	*merkja*	*ást*	*við*	*guð*	*og*	*maður*
CONJ	NS-N	N-G	D-G	VBP	N-A	P	NPR-A	CONJ	NS-A

Figure 1: Tagging and lemmatization of the Old West Norse sentence (number 819) from the *Íslensk Hómilíubók* ("Icelandic Book of Homilies"; late XII century): *En armar krossins merkja ást við Guð og menn.* ("But the arms of the cross mark the love of God and man.")

of word frequency (and therefore informativity), while at the same time ensuring that those words chosen as features are most representative across the history of the language. We selected as features all word types that occurred in at least 75% of the 44 by-year files, that is, all 529 distinct (unlemmatized) word forms that had a document frequency in the corpus of at least 33 documents.

For each (unlemmatized) word type (w) occurring at least three times in the whole corpus (17,741 distinct word types), we computed its co-occurrence frequency with each of the feature words (t). In this way, we obtained a matrix of 17,741 $\times$ 529 word by feature co-occurrence frequencies (f$[w, t]$) within a symmetrical window including the two preceding and following words.[1] The plain co-occurrence matrix was converted into a matrix of word-feature pointwise mutual informations $M = (m_{i,j})$, such that,

$$m_{i,j} = \log \frac{N \cdot \mathrm{f}[w_i, t_j]}{(W_1 - 1) \cdot \mathrm{f}[w_i] \cdot \mathrm{f}[t_j]},$$

where N = 899,763 tokens is the total number of tokens in the corpus, $W_1 = 5$ is the total sliding window size considered, and f$[w_i]$ and f$[t_j]$, are the overall corpus frequencies of words w_i and t_j, respectively. In this manner, the row $M_{i,\cdot} = (m_{i,1}, \ldots, m_{i,529})$ represents the contexts in which the word type w_i is found across the whole corpus.

2.4 Second Order Co-occurrence Vectors

The co-occurrence vectors computed above provide representations for the average contexts in which a given word type is found. In order to represent the specific context of each word token, we used second order co-occurrence vectors (Schütze and Pedersen, 1997). These provide important information about the aspects of a context that are relevant for inflectional morphology (Moscoso del Prado Martín, 2007). The second order vectors were computed by passing a symmetrical sliding window including, for each token, the immediately preceding and following word. The vector for each token was computed as the average between the first order vectors (of Subsection 2.3) of the preceding and following words. If no first order vector was available for either the preceding or the following word, the second order vector directly corresponded to the plain first order vector of the word for which there was a first order vector. We excluded those tokens for which we had first order vectors for neither the previous nor the following word type. We computed such second order vectors for all tokens in the corpus that had been tagged for grammatical case (a total of 419,910 vectors, on average 9,453 vectors per year, of which 38.14% were nominatives, 10.91% were genitives, 26.38% were accusatives, and 24.56% were datives).

2.5 Representation of the Case Prototypes

In order to represent the prototypical usages each grammatical case (i.e., nominative, genitive, accusative, and dative) in a given year, we used the first order co-occurrence technique. For each of the 44 distinct years –using the same features identified in Subsection 2.3– we computed first order co-occurrence vectors collapsing all word tokens in each grammatical case, and using a reduced window size including just the preceding and following words (i.e., $W_2 = 3$).[2] For each year (y) we obtained a 4×529 element matrix of co-occurrence frequencies (f$_y[c, t]$), indicating the number of times that each case (c)

[1] To avoid $\log 0$ values, all frequency counts in this paper were increased by one.

[2] The optimal window sizes for the first order co-occurrence vectors for words and for case prototypes were different because they were chosen to optimize different tasks. The window size for first order vectors for words were chosen to optimize the machine learning algorithm for identifying case identities of second order vectors (Section 3), whereas the first order vectors for case prototypes were optimized for clustering cases across the years (Section 4).

was found to co-occur (within the specified window) with feature (t). These matrices were transformed into case to feature pointwise mutual informations, resulting in a series of 44 matrices ($M[y] = (m[y]_{i,j})$ such that,

$$m[y]_{i,j} = \log \frac{N \cdot \mathrm{f}_y[c_i, t_j]}{(W_2 - 1) \cdot \mathrm{f}[c_i] \cdot \mathrm{f}[t_j]},$$

where N is the total number of tokens in the corpus, $\mathrm{f}[c_i]$ is the number of instances of case c_i, and $\mathrm{f}[t_j]$ denotes the number of instances of word t_j in the corpus. In this way, the rows of the $M[y]$ matrices provided a representation of the contexts in which each grammatical case was used in each year.

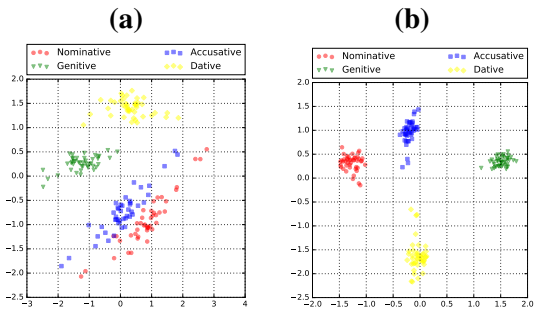

Figure 2: **(a)** Representation of the raw case vectors in SVD-reduced space (i.e., SVD dimension 1 vs. dimension 2). **(b)** Representation of the case vectors after discounting the average vector for each year (i.e., SVD dimension 1 vs. dimension 2).

Fig. 2a plots the spatial organization of the resulting vectors (after reducing to a bidimensional projection using Singular Value Decomposition; SVD). Notice that the prototypes for each case very naturally cluster together across the years. The scatter is however asymmetric, hinting at a process of change along the years common for all four cases. If we compute a yearly overall prototype vector as the average vector for the cases in each year, and we substract it from the corresponding case prototypes, we find that the case identities become clearly differentiated in space (see Fig. 2b), demonstrating that the case prototype vectors do indeed capture the contextual properties of all four cases, which are highly distinctive.

3 Functional Drift in Icelandic Cases

As was discussed in the Introduction, the inflectional paradigms marking case and number have barely –if at all– changed along the history of Icelandic. On the basis of this fact alone, one could conclude that the grammatical case system is not actually experiencing any linguistic drift, but has rather remained basically static throughout the last millennium. There is, however, another possibility. Linguistic drift could have been affecting the *functions* of grammatical cases in Icelandic. If this were the case, one would expect to observe a slow –constant rate– diachronic change in the contexts in which each of the four cases is used.

To investigate this latter possibility, for each of the 44 years documented in the corpus, we trained a basic logistic classifier in the task of assigning grammatical case to the second order co-occurrence vectors developed in Subsection 2.4. Once each of the classifiers had been trained, we tested the classifiers' performances on the vectors obtained from each of other 43 years on which they were not trained. On the one hand, if the functions of the cases have indeed remained constant along the history of the language, one would expect that the performance of a classifier tested on the data from a given year, should remain approximately constant when tested on vectors from all other years. If, on the other hand, the functions of Icelandic grammatical cases have been subject to linguistic drift, the irreversible and cumulative nature of the drift (Sapir, 1921/2014) implies that the classifier error should grow –if only so slightly– with each year passed. The reason for this is that the contexts in which one would use each case should be slightly different from year to year. One should then predict that the error of the classifier should depend on the temporal distance between the year of the testing vectors and that of the training ones. Furthermore, the change in error should be of a linear nature, with a very slight slope.

When tested on the same years in which they had been trained, the classifiers performed rather well in inferring the case to which each of the context vectors belonged (the distribution of errors across the 44 years was well approximated by a normal distribution with a mean error of 26.67%, a standard deviation of 1.99%, and best and worst classification errors of 22.17% and 30.39%, respectively).[3]

[3]Although we chose the best model among different learning algorithms, including multiple versions of Support Vector Machines, Classification Trees, and a Softmax Classifier, we have no doubt that the learning performance can be improved upon. For our purposes, however, it was sufficient

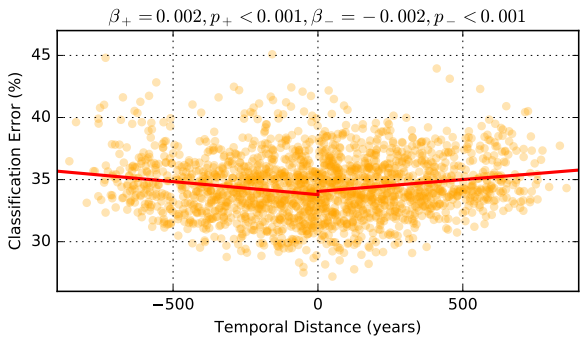

Figure 3: Correlation between the classifier error and the temporal distance from the year from which the training vectors were obtained to the year when the testing vectors were obtained. The solid lines plots a linear regressions.

We then tested the classifiers on the vectors obtained from different years. The results are plotted in Fig. 3. The scatter plots the difference in years (i.e., the difference values are positive when the classifier was tested on vectors obtained after those used for training, and negative when testing with vectors obtained before the training ones). When testing on data different from the training sets, there is a logical loss in performance (of about 8%) from the baseline of testing on the same training set. We fitted two linear regressions, one to the positive differences and another to the negative differences (plotted by the solid lines in Fig. 3). The first thing that stands out is that the performance of the classifier is remarkably good when tested on vectors obtained at considerable temporal distances from the time when the training vectors were obtained. While the error of the classifier is of about 34% when tested on vectors from the year after or before the training vectors, the error remains at 35% for vectors originating from texts that are five centuries apart. Once again, this speaks to the remarkable conservativeness of the Icelandic language. However, these small differences are in fact reliable: There are significant slopes in both regressions (positive differences: $R_+ = .161, p_+ < .001$; negative differences: $R_- = -.164, p_- < .001$). A second remarkable thing is that both regressions are substantially symmetrical, in fact their slopes are basically identical ($|\beta_+| = |\beta_-| = .002$). This indicates that the degree to which the usages of the cases at different

to have a classifier with a decent performance, as our goal was showing that the error is time-dependent.

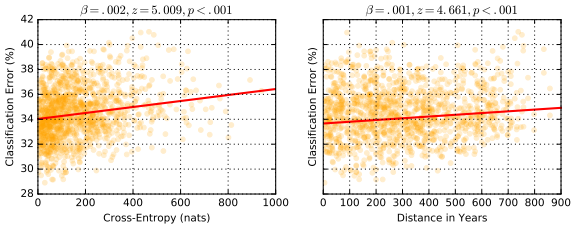

Figure 4: Independent effects of cross-entropy (left panel) and distance in years between the training and testing sets (right panel) as estimated by the linear mixed-effects regression model.

time points have diverged depends on the amount of time that has intervened, irrespective of whether it was the training or the testing set that was collected before.

One could argue that the slow drift observed may not be really due to changes in the functions of the grammatical cases themselves, but just to overall changes either in the overall language, or in the very topics that are addressed (e.g., one might guess that talk of swords, slaves, and longships was more frequent in XII century Norse than it is in Modern Icelandic). To investigate this possibility we used an information-theoretical measure of the prototypicality of a set of second order vectors for a particular year (based on that used in Moscoso del Prado Martín, 2007). From the vectors of each year, irrespective of their case, we fitted a 529-dimensional Gaussian distribution (by estimating the mean vector for that year, $\boldsymbol{\mu}_y$, and the covariance matrix, Σ_y). The average inadequacy of a given set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ obtained in year z to the distribution fitted to the vectors obtained in year y is measured by the *cross-entropy*, a Monte Carlo estimator of which is given by,[4]

$$H(z, y) \approx K + \frac{1}{2}\log|\Sigma_y| + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{v}_i - \boldsymbol{\mu}_y)^T\Sigma_y^{-1}(\mathbf{v}_i - \boldsymbol{\mu}_y),$$

where K is a constant.[5] In addition, one should also take into account the fact that, for some years, the classifier might generalize better or worse than for others (due to irrelevant idiosyncrasies of one specific text used for training), which could lead to a distortion of the results.

To investigate whether, after discounting for the inadequacy of the vectors to the overall distribution of those in which the classifier was trained,

[4] Assuming Σ_y is definite positive so that its inverse exists.
[5] $K = \frac{529}{2}\log 2\pi$.

there was still evidence for drift in the functions of the cases, while also accounting for the different generalization powers of the classifiers, we fitted a linear mixed-effects model to the classifier errors, including fixed-effect predictors of the cross-entropy described above, and the absolute value of the difference in years between the training set and testing set dates (as indicated above, the effects were equivalent for positive and negative values in years), and the dating of the testing set as a random effect. As expected, we found that the cross-entropies had a significant positive effect ($\beta = .002, z = 5.009, p < .001$; left panel if Fig. 4), indicating that the performances of the models were indeed worse for less adequate sets of testing vectors, irrespective of any aspect of grammatical case. However –crucially– even after considering the effect of cross-entropy, there remained a significant positive effect of the temporal distance ($\beta = .001, z = 4.661, p < .001$; right panel if Fig. 4).[6] This result therefore supports the hypothesis that the function of grammatical case has been subject to a slight constant change during the history of the language: a functional drift.

4 Functional Chain Shifts in Case

In the previous section we have demonstrated that the functions of Icelandic cases have been subject to slow linguistic drift. The question now arises of whether this drift is purely random, or rather it has some degree of directionality arising from endogenous linguistic factors. It is possible that changes in the functions of some cases caused changes in the functions of others. We investigate this possibility using the notion of Granger-causality

4.1 Granger-causality

Granger-causality (Granger, 1969) is a powerful technique for assessing whether one time series can be said to be the cause of another one. The basic idea is that one time series $\mathbf{x}$ is said to *Granger-cause* another series $\mathbf{y}$ if the past of series $\mathbf{x}$ predicts the future of series $\mathbf{y}$ over and above any predictive power that can be found on $\mathbf{y}$'s own past. This idea has proven of great value to investigate the causal connections between economic variables, sequences of behavioral responses, neural spikes, or electroencephalographic potentials. Often, the technique is used to reconstruct directional networks of variables and processes that have causal connections.

If $\mathbf{x}$ and $\mathbf{y}$ are stationary time sequences on time (τ), in order to test whether $\mathbf{x}$ Granger-causes $\mathbf{y}$, one begins by fitting *autoregressive models* (AR) that predict the values of $\mathbf{y}$ from its own n values lagged into the past. This consists on finding values $a_1, a_2, \ldots, a_n$ that minimize the error ε in the equation,

$$y[\tau] = a_0 + \overbrace{a_1 y[\tau-1] + a_2 y[\tau-2] + \ldots + a_n y[\tau-n]}^{\text{past of } \mathbf{y}} + \varepsilon[\tau].$$

One then augments the autoregression by including m lagged values of $\mathbf{x}$, with additional parameters $b_1, \ldots, b_m$ to be fitted,

$$\begin{aligned} y[\tau] = & a_0 + \overbrace{a_1 y[\tau-1] + a_2 y[\tau-2] + \ldots + a_n y[\tau-n]}^{\text{past of } \mathbf{y}} + \\ & + \underbrace{b_1 x[\tau-1] + b_2 x[\tau-2] + \ldots + b_m x[\tau-m]}_{\text{past of } \mathbf{x}} + \varepsilon[\tau]. \end{aligned}$$

where the ε sequences are uncorrelated (white) gaussian noises, reflecting the fully random or chaotic part of the system, which cannot be predicted from its past (i.e., the error, that is termed by some the 'creativity' of the model). If the second regression is a significant improvement over the first, then it can be said that $\mathbf{x}$ *Granger-causes* $\mathbf{y}$, indicating that past values of $\mathbf{x}$ significantly predict future values of $\mathbf{y}$ over and above any predictive power of $\mathbf{y}$'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. The selection of the autoregressive order parameters n and m is achieved by model comparisons using information criteria.

When one is interested in reconstructing a network of causal relations between multiple variables, one can use a mutivariate generalization of the AR model, the *vector autoregressive model* (VAR). The VAR model consists of mutiple AR equations (one for each variable in the model). If we consider an autoregressive order of one (i.e., $m = n = 1$), when we are simultaneously considering p variables $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_p\}$, the VAR[1] model to be fitted can be expressed in matrix no-

[6] The estimated covariance matrices were not definite positive for two of the years, which were excluded from the analyses. In addition, in 552 out of the remaining 1,849 estimates, the cross-entropy took unusually large values, orders of magnitude larger than the rest (likely reflecting inadequacy of the multidimensional Gaussian approximation for these cases), which distorted the effect estimates. The analyses reported exclude these 552 points. However, keeping these outlying values in the regression, both key effects remained significant, but the slope estimates were less trustworthy.

tation as,

$$\begin{pmatrix} y_1[\tau] \\ \vdots \\ y_p[\tau] \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} A_{1,1} & \dots & A_{1,p} \\ \vdots & \ddots & \vdots \\ A_{p,1} & \dots & A_{p,p} \end{pmatrix} \begin{pmatrix} y_1[\tau-1] \\ \vdots \\ y_p[\tau-1] \end{pmatrix} + \begin{pmatrix} \varepsilon_1[\tau] \\ \vdots \\ \varepsilon_p[\tau] \end{pmatrix}.$$

This model enables testing for Granger-causality between any pair of variables $\mathbf{y}_i \in \mathcal{Y}$ and $\mathbf{y}_j \in \mathcal{Y}$, after partialling out the possible confounding effects of $\{\mathbf{y}_t,\, t \neq i, t \neq j, 1 \leq t \leq p\}$. $\mathbf{y}_j$ is said to Granger-cause $\mathbf{y}_i$ if the model coefficient $A_{i,j}$ is significantly different than zero, and the reverse holds if $A_{j,i}$ significantly different than zero (i.e., $\mathbf{y}_i$ Granger-causes $\mathbf{y}_j$).

4.2 Granger-causality in Case Drift

To investigate whether the pattern of change in one case triggers (i.e., Granger-cause) the pattern of change in another, we made use of the prototype vectors for the cases in each of the years developed in Subsection 2.5. As a measure of the amount of contextual change for a given case in a given year, we computed the city-block distances between the case prototypes from each year to the next available time point, which are plotted in Fig. 5a. Notice that there is an overall pattern of change equally affecting all cases, and the changes are therefore strongly correlated. This reflects the overall pattern of historical changes affecting Icelandic as a whole, as well as changes in the topics that would be discussed in the different time periods, as was documented in Subsection 2.5 and Section 3. Considering the changes in each case as a component in a four-dimensional vector, the modulus of this vector (plotted by the dashed orange line in Fig. 5a) gives the overall magnitude of the changes that are unspecific to the cases themselves. To remove this component from the changes, we fitted a linear regression to the sequence of changes in each case, using the overall pattern of change as a predictor. Fig. 5b plots the resulting residuals, indicating the amount of change that was specific to each case, over and above the overall pattern.[7]

A precondition for testing for Granger-causality is that the time series under consideration are stationary. In our case, the series depicted in Fig. 5b are significantly non-stationary; they exhibit, for instance, significant temporal trends. In order to remove the non-stationarities, the series were differentiated (i.e., we considered the difference between each two consecutive points). The result of this differentiation, plotted in Fig. 5c, removed the non-stationary trends from the original series.

Table 1: Results of the Granger-causality analyses. Causality directions that remained significant after FDR correction are highlighted in bold.

Direction	$F[1,144]$	p	p (FDR)	Direction	$F[1,144]$	p	p (FDR)
Nom. → Gen.	2.614	.108	.184	**Gen. → Nom.**	5.618	.019	**.046**
Nom. → Dat.	8.295	.005	**.018**	Dat. → Nom.	3.834	.052	.104
Gen. → Acc.	.566	.453	.454	Acc. → Gen.	2.408	.123	.184
Acc. → Nom.	6.802	.010	**.030**	Nom. → Acc.	.644	.424	.454
Acc. → Dat.	10.249	.002	**.018**	Dat. → Acc.	.563	.454	.454
Dat. → Gen.	1.354	.246	.329	**Gen. → Dat.**	9.034	.003	**.018**

We fitted a VAR[n] model to the four differentiated time-series. The autoregressive order found to maximize Akaike's Information Criterion (Akaike, 1974) was $n = 1$.[8] The F statistics and significance values for the coefficients in the resulting VAR[1] model are given in Tab. 1. In order to reconstruct the causality network, we also need to consider that we started out with only very vague predictions on the possible directions of causality. As the model involved twelve separate p-value tests, the p-value estimates need to be corrected for multiple comparisons. This correction was done using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995), resulting in the corrected p-value estimates listed in the last column of Tab. 1.

The Granger-causality analysis leads us to reconstruct the causality network depicted in Fig. 6. It appears that the drift in the functions of Icelandic case is not plainly random. Instead, we find evidence that changes in the functions of the accusatives and genitives have had a domino effect, triggering further changes in the functions of nominatives. Finally, changes in all other three cases result in changes in the functions of the dative. In summary, the changes observed are consistent with the idea discussed in the Introduction of a *functional chain shift* affecting the morphological case system of Icelandic.

5 Discussion

We have presented evidence for a steady drift –of the precise kind advocated by Sapir (1921/2014)– even in a language as remarkably conservative as is Icelandic. This supports the claim that human languages are in a state of 'perpetual motion' (Beckner et al., 2009; Dediu et al., 2013; Hawkins and Gell-Mann, 1992; Hopper, 1987;

[7] Negative values in this figure indicate changing less than the average, rather than 'negative change'.

[8] In fact $n = 1$ was also found to maximize both Akaike's Final Prediction Error and Hannan-Quinn Criteria.

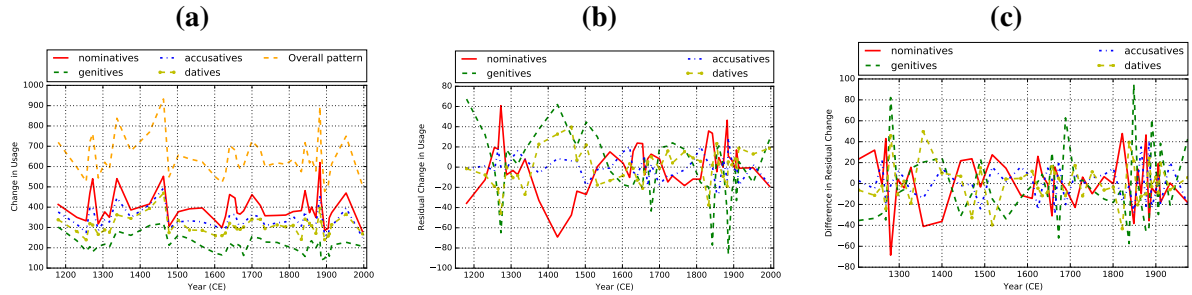

Figure 5: **(a)** Overall value of the city-block distances between the prototypical case vectors for consecutive years. **(b)** Residual value of the distances specific to each case after residualizing the overall pattern of change. **(c)** Differentiated values of the residualized distances, removing non-stationarities.

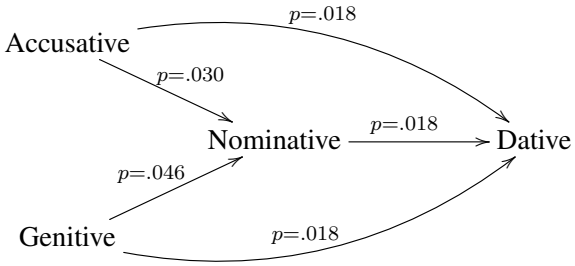

Figure 6: Reconstructed network of Granger-causal connections between diachronic changes in case functions. The p-values indicated on the causal arrows are FDR-corrected.

Larsen-Freeman and Cameron, 2007; Niyogi and Berwick, 1997). Although we have found that functional change in Icelandic case has proceeded at a constant rate, we do not think, as argued by Nettle (1999a, 1999b), that this rate of change needs to be constant across languages. There are strong arguments suggesting that in other languages such rates might be different (Wichmann, 2008; Wichmann and Holman, 2009).

The crucial innovation presented in this paper is the reconstruction of the causality network linking the changes in the four cases. Previous applications of the notion of Granger-causality to diachronic language change (Moscoso del Prado Martín, 2014) have focused on the macroscopic relation between sudden changes in syntax and morphology. Here, we have demonstrated that Granger-causality can also be used to reconstruct detailed networks of slow changes within the morphological system, at a more microscopic scale. The techniques developed offer a mechanism for investigating subtle changes in the functions of linguistic constructions, and the causal relations between them. Traditionally, historical linguists have focused on 'narrative' accounts of the the chains of change within a language. Although such type of accounts are extremely useful, the often very subtle changes in usage that can occur from one time-point to another cannot always be described with such clearcut patterns. Nevertheless, we have shown that those very small changes do accumulate in meaningful ways.

An important question addressed by this study is the presence of endogenous causal chains in language change. Lupyan and Dale (2015) argue that languages are constrained by their 'ecological niches', the communities in which they are spoken, and the extralinguistic properties of those niches can trigger exogenous change in the morphology of the languages. Following on Lupyan and Dale's ecosystem analogy, one should see that, as well as being part of ecosystems, languages are also ecosystems in themselves, in a nesting similar to that found in natural ecosystems (i.e., an animal is part of a particular ecosystem, and its body is an ecosystem in itself). Sounds, words and constructions have their own ecological niches within the language, and disturbances in the system can trigger cascaded changes, leading to readaptation (evolution) of the constructions. This contrasts with the view of changes in the function of Icelandic cases expressed by Eythórsson (2000). He showed that verbs whose arguments exhibit 'nominative sickness' and 'accusative sickness' tend to be clustered along certain syntactic and semantic lines. That it is in these particular niches that accusatives and datives ended up settling is not, however, the *cause* of the language changes. As we have shown, the case system was subject to a string of cascaded pressures. That the cases ended

up settling in new syntactico-semantic niches was the *result*, rather than the cause of the changes.

References

Hirotogu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Jóhanna Barðdal and Thórhallur Eythórsson. 2003. The change that never happened: the story of oblique subjects. *Journal of Linguistics*, 39:439–472.

Jóhanna Barðdal. 2011. The rise of dative substitution in the history of Icelandic. *Lingua*, 121:60–79.

Clay Beckner, Nick C. Ellis, Richard Blythe, John Holland, Joan Bybee, Jinyun Ke, Morten H. Christensen, Diane Larsen-Freeman, William Croft, and Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language Learning*, 59:1–26.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Theresa Biberauer and Ian Roberts. 2008. Cascading parameter changes: internally driven change in Middle and Early Modern English. In Thórhallur Eythórsson, editor, *Grammatical Change and Linguistic Theory: The Rosendal Papers*, pages 79–114. John Benjamins, Philadelphia, PA.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44:890–907.

William Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Longman, London, England.

Dan Dediu, Michael Cysouw, Stephen C. Levinson, Andrea Baronchelli, Morten H. Christensen, William Croft, Nicholas Evans, Simon Garrod, Rusell D. Gray, Anne Kandler, and Elena Lieven. 2013. Cultural evolution of language. In Peter J. Richerson and Morten H. Christensen, editors, *Cultural Evolution: Society, Technology, Language, and Religion*, pages 303–331. MIT Press, Cambridge, MA.

Thórhallur Eythórsson. 2000. Dative vs. nominative: changes in quirky subjects in Icelandic. *Leeds Working Papers in Linguistics*, 8:27–44.

Adam Fisiak, editor. 1984. *Historical Syntax*. de Gruyter, Berlin.

Clive W. J. Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438.

John Hawkins and Murray Gell-Mann, editors. 1992. *The Evolution of Human Languages*. Santa Fe Institute Studies in the Sciences of Complexity. Addison Wesley, Reading, MA.

Paul J. Hopper. 1987. Emergent grammar. *Proceedings of the Berkeley Linguistic Society*, 13:139–157.

Diane Larsen-Freeman and Lynne Cameron. 2007. *Complex Systems and Applied Linguistics*. Oxford University Press, Oxford, UK.

David Lightfoot, editor. 2002. *Syntactic Effects of Morphological Change.* Oxford University Press, Oxford, UK.

Will Lowe and Scott McDonald. 2000. The direct route: Mediated priming in semantic space. In Lila Gleitman and Aravind K. Joshi, editors, *Proceedings of the XXII Annual Conference of the Cognitive Science Society*, pages 806–811, Austin, TX. Cognitive Science Society.

Gary Lupyan and Rick Dale. 2015. The role of adaptation in understanding linguistic diversity. In Rik De Busser and Randy J. LaPolla, editors, *Language structure and environment: Social, cultural, and natural factors*, pages 289–316. John Benjamins Publishing Company, Philadelphia, PA.

André Martinet. 1952. Function, structure, and sound change. *Word*, 8:1–32.

Fermín Moscoso del Prado Martín. 2007. Co-occurrence and the effect of inflectional paradigms. *Lingue e Linguaggio*, 6:247–263.

Fermín Moscoso del Prado Martín. 2014. Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. In Paul Bello, Marcello Guarini, Marjorie McShane, and Brian Scassellati, editors, *Proceedings of the XXXVII Annual Conference of the Cognitive Science Society*, pages 2657–2662, Austin, TX. Cognitive Science Society.

Daniel Nettle. 1999a. Using social impact theory to simulate language change. *Lingua*, 108:95–117.

Daniel Nettle. 1999b. Is the rate of linguistic change constant? *Lingua*, 108:119–136.

Partha Niyogi and Robert C. Berwick. 1997. A dynamical systems model for language change. *Complex Systems*, 11:161–204.

Edward Sapir. 2014. *Language: An Introduction to the Study of Speech.* Dover Publications, Mineola, NY. (Original work published 1921).

Hinrich Schütze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33:307–318.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. Icelandic parsed historical corpus (IcePaHC – v. 0.9).

Søren Wichmann and Eric W. Holman. 2009. Population size and rates of language change. *Human Biology*, 81:259–274.

Søren Wichmann. 2008. The emerging field of language dynamics. *Language & Linguistics Compass*, 2:1294–1297.

Henri Wittmann. 1983. Les réactions en chaîne en morphologie diachronique ("Chain reactions in diachronic morphology"). In *Actes du colloque de la Societé Internationale de Linguistique Fonctionnelle*, volume 10, pages 285–292, Québec, Canada. Université Laval.

Sewall Wright. 1929. The evolution of dominance. *The American Naturalist*, 63:556–561.

Sewall Wright. 1955. Classification of the factors of evolution. *Cold Spring Harbor Symposia on Quantitative Biology*, 20:16–24.