DiaHClust: an iterative hierarchical clustering approach for identifying stages in language change

Christin Schätzle
University of Konstanz
christin.schaetzle@uni-konstanz.de

Hannah Booth
Ghent University
hannah.booth@ugent.be

Abstract

Language change is often assessed against a set of pre-determined time periods in order to be able to trace its diachronic trajectory. This is problematic, since a pre-determined periodization might obscure significant developments and lead to false assumptions about the data. Moreover, these time periods can be based on factors which are either arbitrary or non-linguistic, e.g., dividing the corpus data into equidistant stages or taking into account language-external events. Addressing this problem, in this paper we present a data-driven approach to periodization: 'DiaHClust'. DiaHClust is based on iterative hierarchical clustering and offers a multi-layered perspective on change from text-level to broader time periods. We demonstrate the usefulness of DiaHClust via a case study investigating syntactic change in Icelandic, modelling the syntactic system of the language in terms of vectors of syntactic change.

1 Introduction

In historical linguistics, it is now generally acknowledged that language change proceeds gradually rather than abruptly (e.g., Kroch, 2001). Nevertheless, in order to achieve meaningful comparisons and generalizations, it is useful to be able to identify stages in a change's trajectory. In traditional approaches, the progress of a change is typically assessed against a pre-determined and somewhat arbitrary periodization scheme which segments a language's diachrony into discrete periods (e.g., 'Old', 'Middle' and '(Early) Modern'). The problematic nature of this methodology is well known, though rarely made explicit (see, e.g., Curzan, 2012). Such an approach may yield results which conceal the true trajectory of a phenomenon. For instance, relying on a discrete periodization may give misleading findings indicative of abrupt change, e.g., with a certain year as a turning point. Moreover, transitional stages, which are often of great interest, can be easily obscured. Despite such issues, for a long time this 'periodization problem' was accepted as an unfortunate but unavoidable aspect of historical linguistics.

With the boom in corpus-based and computational studies of language change over recent decades, the periodization problem has been readdressed, as new data-driven methodologies have emerged, particularly in relation to historical English (see, e.g., Gries and Hilpert, 2008, 2012; Degaetano-Ortlieb and Teich, 2018). Instead of applying a pre-determined periodization to the data at the outset, in such approaches the data is first assessed and periods then suggested based on assessment of this data. The periodization scheme can be arrived at via a range of statistical methods, e.g., hierarchical clustering and relative entropy. This yields objective data-driven periodization schemes which are faithful to the corpus data and can still be used to arrive at meaningful generalizations.

In this paper, we present DiaHClust, a new approach which can be used to identify stages in diachronic change based on quantitative corpus-derived data. As a basis we take the hierarchical clustering approach for historical data from Gries and Hilpert (2008, 2012) and develop this further, specifically for investigating syntactic change. In addition to implementing the methodology from Gries and Hilpert in the software environment R (R Core Team, 2014), we add an extra iterative approach to the hierarchical clustering which results in a multi-layered perspective on change, from text-level to broader periods, while also respecting outliers and genre effects. With DiaHClust, we show that a data-driven periodization methodology can also be applied to a language like Icelandic, where syntactic change is not as extreme as

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 126–135
Florence, Italy, August 2, 2019.

in other Germanic languages, and where the available annotated corpus data is relatively sparse.

2 Data-driven approaches to periodization

Often, the factors which go into determining a traditional top-down periodization have no direct connection to the linguistic phenomena under investigation. Moreover, the vast majority of traditional periodizations also take into account language-external factors, e.g. historical milestones or migrations. A classic example is the 'Middle English' period, which is often delimited by the onset of the Norman invasions in 1066 and the arrival of printing in the late 15th century. A further issue is that time stages within a periodization scheme are sometimes designed to be equal in length. This results in a periodization scheme whose time stages are not necessarily a best fit with the actual linguistic characteristics. Moreover, since a traditional periodization is a linear sequence of time stages, transitional periods which may overlap with certain time stages cannot readily be identified, despite the fact that understanding these transitions is vital for explaining language change.

In response to such issues, alternative approaches have emerged which are exclusively derived from the data at hand. For example, Degaetano-Ortlieb and Teich (2018) present a data-driven approach which uses relative entropy by calculating the Kullback-Leibler Divergence (KLD) between lexical and grammatical features in texts from temporally adjacent time periods to identify stages in language change. KLD is an information-theoretic measure which is used to compare probability distributions and detect differences between them. Degaetano-Ortlieb and Teich (2018) apply KLD to the detection of periods of change by selecting a starting year and a sliding window of several years to compare the probability distributions of corpus data from the preceding and subsequent years in the sliding window. The KLD models are based on the distributions of lemmas and Part-of-Speech trigrams in historical texts to track changes at the lexical and grammatical level. A change is identified by means of relative peaks or troughs in KLD.

Another methodology is the bottom-up clustering approach to periodization developed by Gries and Hilpert (2008, 2012), 'Variability-based Neighbor Clustering' (VNC) (see also Hilpert and Gries, 2009, 2016; Perek and Hilpert, 2017). In contrast to standard hierarchical clustering, VNC is sensitive to the temporal ordering of data. The basic principle is that parts of the data which exhibit similar linguistic characteristics should form part of the same period, i.e., cluster, and that breaks between periods should be inserted at points where the characteristics of the data show a quantifiable shift.

The VNC algorithm groups together temporally adjacent data points which are most similar to each other in a stepwise fashion. First, the two neighboring data points which exhibit the highest degree of similarity are identified and merged into a single data point. The similarity between data points is measured via the calculation of standard deviations or other distance measurements, e.g. Euclidean distance when the data points represent single values, or correlation measurements such as Pearson's r when the data points represent vectors of values. The data in Gries and Hilpert (2008, 2012) consists of either individual frequency values which represent the occurrence of a given structure over time, e.g., the *get*-passive in historical English, or vectors of values representing the collocations of a linguistic structure with multiple linguistic items, mostly at the lexical level.

The neighboring data points are merged into a single data point according to an amalgamation rule chosen by the researcher. The amalgamation can, for example, be achieved via averaging values or choosing the minimum/maximum of the values. Next, the two neighboring data points with the highest degree of similarity are merged. This process is repeated until all data points have been merged, grouping the data into larger time stages along the way. The result of this process is a hierarchical clustering of all data points which is generally graphically represented as a dendogram. The dendrogram shows the sequence in which the data points were merged into clusters, providing insights into how much the clusters differ from one another. The hierarchical nature of the output is a particular advantage, which – unlike traditional linear periodizations – allows transitional and overlapping stages to be identified and represented. In order to identify the most useful number of clustered time periods, Gries and Hilpert (2012) use a scree plot. Applied to VNC, the scree plot displays how much variability in the data can be

explained after each merging step, allowing the researcher to choose the most accurate periods.

Despite the data-driven focus of the VNC methodology, in order to have a suitable number of initial input clusters Gries and Hilpert (2008, 2012) aggregate individual texts into larger temporal episodes, e.g., decades or fifty-year periods. This still involves imposing an abitrary classification on the data at the outset and may bias the clustering, obscuring significant insights about transitional periods – particularly if applied to corpora where data sparsity is an issue. In this paper we present DiaHClust, a method for periodization which implements the VNC algorithm for the analysis of syntactic change, but avoids the a priori aggregation of texts by adding a second level of iteration outside the VNC. One can thus trace the clustering from text-level through several iterations until the final periodization scheme is reached, gaining detailed insights about the progress of change, possible outliers and genre effects.

3 DiaHClust: Methodology

We have developed DiaHClust for a study of syntactic change in Icelandic based on data from the Icelandic Parsed Historical Corpus ('IcePaHC', Wallenberg et al., 2011). The main objective is to provide a better understanding of the progression of previously identified changes in the language in terms of a data-driven periodization. DiaHClust extends Gries and Hilpert's (2008; 2012) vector-based approach to VNC to factor in syntactic changes. Instead of clustering with respect to the distributional features of a single phenomenon, we include multiple known syntactic changes in the vectors to create a model of the syntactic system at different stages. Moreover, we present our implementation of the VNC in R as the DiaHClust package. DiaHClust is readily usable with any kind of diachronic data suitable for hierarchical clustering. In addition to the standard VNC approach, DiaHClust provides an extra iterative approach by calculating silhouette values (Rousseeuw, 1987) to automatically identify the optimal numbers of clusters. This allows us to begin at text-level, tracing the clustering until the final larger time stages are identified, and enables the ad hoc identification of outliers and genre effects. Furthermore, this methodology avoids misleading statistics which may arise when one otherwise aggregates the data into small temporal sequences at the outset.

3.1 Vectors of syntactic change

In the vector-based approach by Gries and Hilpert (2008, 2012) and in the KLD-approach by Degaetano-Ortlieb and Teich (2018), differences in the occurrence of a linguistic feature across various contexts, i.e., its distributional properties, are assessed. These contextual differences often reflect functional, lexical and stylistic factors which are independent of grammar. In generative approaches to syntactic change, a common idea is that multiple 'surface' word order changes which show up in the data often reflect a single 'underlying' change in clause structure (e.g. Kroch, 1989). Syntactic change is thus viewed as deeply interactional, and distributional properties are less relevant in its assessment. Our syntax-specific methodology uses vectors which are packed with information about multiple interrelated syntactic developments.

In our proposal, a vector is created for each text in a given diachronic corpus. Each vector contains relative frequencies of syntactic features which change over time, see (1).

(1) Text A = {feature$_1$, feature$_2$, ..., feature$_n$}
Text B = {feature$_1$, feature$_2$, ..., feature$_n$}
⋮

In this way, existing knowledge about a language's syntactic system across time informs the data-driven periodization. Furthermore, using changing syntactic features to describe the language system at a given point of time is supported by recent work to train a classifier for the dating of early English texts (Zimmermann, 2014; Ecay and Pintzuk, 2016). We provide a more concrete example in Section 4.

3.2 Implementation of VNC

We implemented our DiaHClust methodology using R. The source code and the DiaHClust package, including a detailed documentation, are available on GitHub.[1] DiaHClust implements the VNC approach in the form of the `vnc()` function by manipulating individual steps in the workflow behind R's standard agglomerative hierarchical clustering function `hclust()`.

[1] https://github.com/christinschaetzle/diaHClust

In the vector-based approach to VNC by Gries and Hilpert (2008, 2012), a correlation statistic is calculated before clustering the data. This is generally done when applying a hierarchical clustering approach to vectorial data (see, e.g., Baayen, 2008). Thus, a correlation matrix is calculated first in the DiaHClust approach, using Pearson's *r* as correlation coefficient.[2] In DiaHClust, the correlation matrix is calculated based on a data matrix where each column represents a vector containing the changing syntactic features extracted from a text. The vectors are ordered from left to right according to the time stamp of the text. The time stamp is encoded in the vector name, i.e., the name of the corresponding column in the data matrix. For the DiaHClust package to work, the vector name should begin with a four digit year date followed by a dot and the text name, e.g., "1250.STURLUNGA", allowing one to easily identify individual texts in the clustering.[3] Following this, the correlation matrix is transformed into a distance matrix by calculating Euclidean distances between the data points, since hierarchical clustering, including VNC, requires a distance measure to determine the (dis-)similarity between two objects (see, e.g., Gries and Hilpert 2012).

Hierarchical clustering usually begins by clustering together the two most similar objects, i.e., the data points with the smallest distance to one another, merging these two data points. This process continues until all data points have been clustered. This process is illustrated in lines 6–12 in Algorithm 1.[4] Different methods for agglomeration, i.e., the merging or amalgamation of two data points, such as averaging over two data points or taking the minimum value can be applied in hierarchical clustering. When averaging is chosen as the agglomeration method, cluster similarity between two clusters is assessed based on the average of the data points in the clusters. Moreover, the two data points with the smallest distance are merged into a new data point by averaging the corresponding values after each iteration, see lines 8 and 12 in Algorithm 1. In general, all agglomeration methods available with `hclust()` are available with our implementation of the VNC. We recommend using averages – following Gries and Hilpert (2008, 2012) – since, in quantitative corpus linguistics, (co-)occurence frequencies are usually assessed by averaging frequencies over texts/time periods.

Algorithm 1 Implementation of VNC

1: **function** VNC
▷ Manipulation of distance matrix (dist):
2: **for** $i = 1$ to $numberOfRows(dist)$ **do**
3: **for** $j = 1$ to i **do**
4: **if** not $i = j$ **then**
5: $dist[i, j] = \max(dist)$
▷ Clustering process:
6: **for** $k = 1$ to $numberOfRows(dist)$ **do**
7: find m, n for which $dist[m, n] = \min(dist)$
8: $dist[, n] = \frac{(dist[,n]+dist[,n+1])}{2}$
9: **if** $dist[1, 1] = \min(dist)$ **then**
10: delete $dist[1,]$
11: **else**
12: $dist[m,] = \frac{(dist[m-1,]+dist[m,])}{2}$

In order to allow only temporally-adjacent data points (i.e., texts) to be clustered with one another in VNC, we manipulate the distance matrix before clustering the data. This is done by setting all distance values which describe distances between non-temporally adjacent data points to the value which equals the maximum value of the distance matrix, see lines 2-5 in Algorithm 1. As similarity is measured in terms of the minimum distance, it is highly unlikely that two data points which have these maximized distances to one another will be merged in the clustering process. This in turn allows us to use the standard `hclust()` function for clustering according to the ideas of VNC, instead of having to implement a separate clustering algorithm. Moreover, `vnc()` adjusts the permutations of the data points which arise during the merging process in order to guarantee the diachronic ordering of data points for plotting as in the dendrogram in Figure 1.

The most appropriate number of clusters for the data, i.e., the time stages the data points fall into, can now be identified via visual inspection of the dendrogram or by generating a scree plot as proposed by Gries and Hilpert (2008, 2012). In

[2]The correlation matrix has to be squared when negative correlation coefficients are produced. Depending on the data distribution, one has to use Spearman correlations instead of Pearson's *r* (see Baayen, 2008, 150–152 for details on correlation statistics and hierarchical clustering).

[3]This corresponds to token IDs in IcePaHC and other Penn-style treebanks. One could add more information to the vector names, e.g., genre or author, but the longer the vector names, the more difficult it is to read the dendrograms.

[4]Distance matrices in R are designed such that distances between neighboring, in our case temporally-adjacent, data points are depicted on the diagonal of the matrix. Moreover, the cells above the diagonal are empty since they mirror the cells below the diagonal. Line 9 handles the case when the first two data points, i.e., the first two columns, are merged. The first row in a distance matrix in R corresponds to the second data point from the original data matrix. Thus, the first row has to be deleted when it is merged so that the formerly second row can take its place.

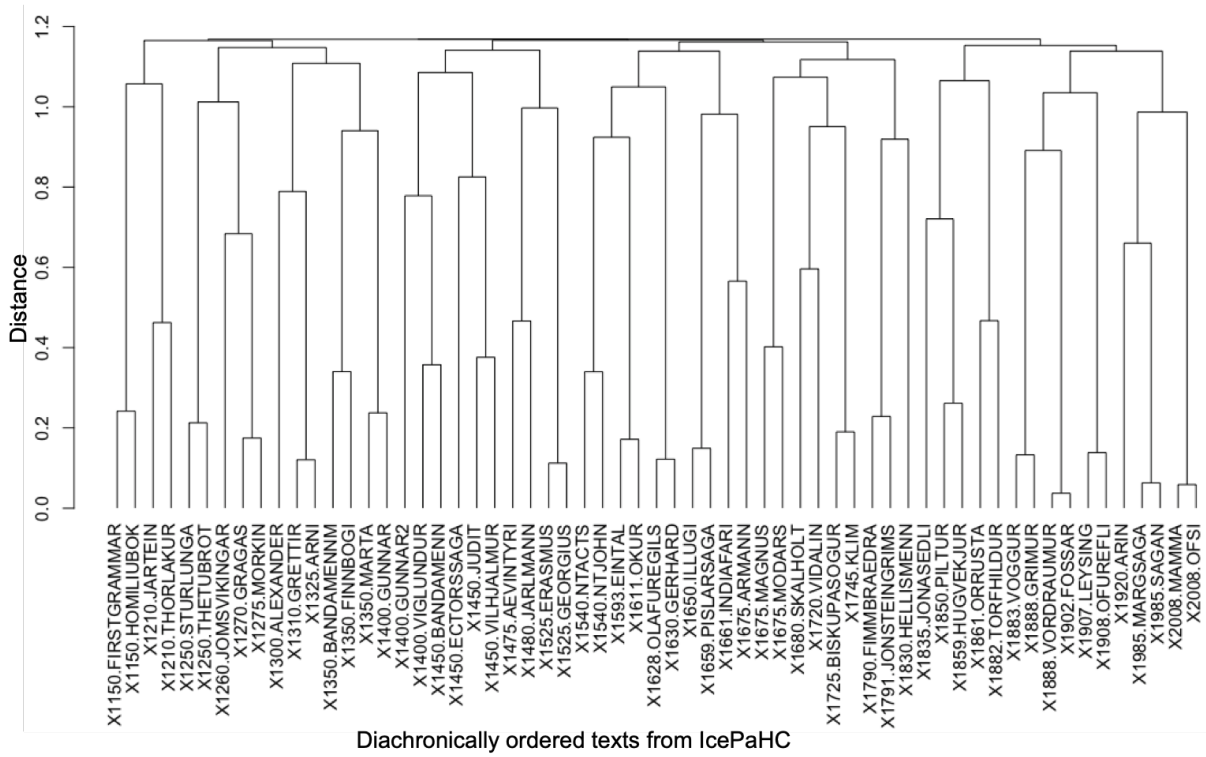

Figure 1: Dendrogram showing the results of the text-based VNC with respect to syntactic change in IcePaHC.

both cases, a decision about the horizontal cut-off point of clusters in the dendrogram has to be made. However, such a visual exploration of the data is often difficult, particularly when the input for the clustering is a large number of individual data points, which is usually the case when clustering individual texts from an entire corpus. Moreover, screeplots rely on the calculation of yet another statistical analysis, i.e., principal components or standard deviations. We therefore decided to calculate silhouettes instead, which provide a quantitative measure of the quality of clusters at different cut-off points. Calculating silhouettes is a standard method for cluster validation. Silhouettes can be used to identify the optimal number of clusters for a given data set, as we explain next.

3.3 Cluster Validation for Cluster Identification

Silhouette values provide information about the consistency of clusters by measuring the dissimilarity of an object to the cluster that it is in, compared to its dissimilarity to other clusters. The silhouette value *s(i)* of an object *i* is calculated according to the formula in Equation 1, where *a(i)* is the average dissimilarity of *i* to all other objects in the cluster *i* has been assigned to, and where *b(i)* corresponds to the average dissimilarity of *i* to its next closest cluster (cf. Rousseeuw, 1987, 56). A large silhouette value, i.e., a value close to 1, indicates that the object is clustered well as it is, and a negative *s(i)* indicates that *i* has been assigned to the wrong cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

The silhouette coefficient of a cluster is moreover defined as the average of silhouettes in a cluster. We implement the calculation of silhouette coefficents in the `optimal_clust()` function as part of DiaHClust, in order to be able to find the optimal number of clusters after VNC clustering has been applied.[5] `optimal_clust()` iterates through all clustering possibilities according to the possible number of merges throughout the clustering process, and calculates the average of silhouette coefficients of all clusters in a clustering. Eventually, the clustering with the highest average silhouette coefficient is identified as the best candidate, and returns information about the cluster memberships of data points with respect to the optimal clustering.

When clustering a large number of data points

[5] Silhouettes can be easily calculated with R using the `silhouette()` function.

– as is usually the case when the input data represents vectors for texts from an entire corpus – the silhouette coefficient may still imply a large number of optimal clusters. Although this may generate insights about the temporal grouping of the data, such a fine-grained periodization is not suitable for frequency-based investigations of syntactic change. We therefore continue the clustering process iteratively, until the optimal number of clusters is smaller than 10. This is implemented as `diahclust()`, as we now describe.

3.4 Iterative DiaHClust Approach

When the results of the `optimal_clust()` function indicate that the initial VNC clustering yields 10 or more clusters, the clustering process can be continued via the `diahclust()` function. The methodology behind the function is illustrated by the pseudocode in Algorithm 2. Before continuing the clustering process, data points which belong to a single cluster according to the previously assessed optimal clustering are aggregated by averaging the corresponding syntactic vectors in the underlying dataset. To keep track of the texts and time stages which form clusters across the iterations, the names of the new vectors consist of the sequence of the names of the aggregated vectors. The previously applied process of VNC clustering with respect to the new dataset is then repeated, including the recalculation of a correlation statistic and a new distance matrix.[6] Moreover, `diahclust()` automatically plots the clustering as a dendrogram. The labels on the dendrogram are abbreviated for better visibility, representing the range of previously aggregated vectors, with the oldest and the youngest text in the range connected via a hyphen, see Figure 2. The resulting clustering is again evaluated using the `optimal_clust()` function, which returns the cluster memberships listing the full range of texts in the clusters. The application of this process is repeated until the final evaluation arrives at an optimal number of clusters less than 10. In this iterative process, the clusters, i.e., time stages, can be inspected at each step of the iteration, allowing one to track the composition of the clusters with respect to the individual texts from the first iteration onwards. This provides insights into how similar the texts in the individual clusters are to one another. We find that this iterative approach very well facilitates the identification of outliers and time stages affected by a genre effect.

In the next section, we illustrate the functionalities of the DiaHClust package by applying the method to a case study which investigates syntactic change in the history of Icelandic.

Algorithm 2 DiaHClust methodology

```
function DIAHCLUST
    repeat
        aggregate(data)
        dist = distanceMatrix(cor(data))
        clust = vnc(dist)
        plot(clust)
        computeOptimalClustering(clust)
    until numberOfClusters < 10
```

4 Case study: syntactic change in Icelandic

Icelandic is generally acknowledged to be the most conservative of the present-day Germanic languages with respect to syntactic change. Yet, several recent corpus studies using IcePaHC have brought to light a series of syntactic changes which interact with one another along the diachrony. These changes comprise the increasing use of dative subjects (see, e.g., Schätzle, 2018), an increase in the frequency of the expletive *það* (Booth, 2018), a decrease in the occurrence of declarative V1 (verb-first) structures (Butt et al., 2014), and an increasing preference of subjects to occur in the clause-initial, prefinite position (see Booth et al., 2017). These studies employ a predetermined top-down periodization scheme akin to that suggested by Haugen (1984), which is influenced by language-external factors such as the first Icelandic translation of the New Testament (1540) and separates the corpus data into more or less equidistant time periods. These studies have in common that the frequencies of the individual phenomena seem to change rather abruptly at a similar point in the diachrony, indicating that a series of drastic changes have occurred in Icelandic clause structure during the past two centuries.

The case study presented in this section is intended to shed more light on the trajectories of these changes by applying the DiaHClust method to data extracted from IcePaHC. We create syntactic vectors on the basis of occurrence frequencies with respect to dative subjects, expletives, V1

[6] When the agglomeration method chosen for VNC clustering is not "average", a different aggregation method, e.g., the minimum with single linkage clustering, should be applied.

and subject position in IcePaHC, and also include data which we extracted for two further phenomena of change which have been previously identified in the history of Icelandic: the change from OV (object-verb) to VO (verb-object) order in the verb phrase (see, e.g., Hróarsdóttir, 2000) and a decrease in the Stylistic Fronting phenomenon (Hróarsdóttir, 1998; Rögnvaldsson, 1996).

4.1 IcePaHC

The IcePaHC corpus, from which the data for this case study is drawn, is a Penn-style treebank (Marcus et al., 1993) which is lemmatised, part-of-speech tagged and annotated for constituent structure, with additional tagging for certain grammatical functions (e.g. subject, object). The corpus contains approximately 1,000,000 words, from 61 text extracts spanning 10 centuries (1150-2008), thereby covering all attested stages of Icelandic.

Despite the significant advantages of the IcePaHC annotation scheme for syntactic research, the corpus does have some limitations. Firstly, the texts included represent only a very small sample of attested historical Icelandic. Secondly, these texts are not evenly distributed across time, so that certain centuries are affected by relative data sparsity. Thirdly, although the corpus texts span various genres, there is a strong bias towards narrative texts overall, while in certain centuries other genres (religious, biographical) dominate. These limitations make the application of a top-down periodization extremely difficult. Thus, IcePaHC represents an ideal test case for the application of our DiaHClust method.

4.2 Syntactic factors under investigation

We obtained relative frequencies for the following phenomena to create a syntactic vector for each text from IcePaHC: dative subjects, overt expletives, V1, subjects in the prefinite position, VO order, and Stylistic Fronting. The data was gathered using the CorpusSearch tool (Randall, 2000) and our own programming scripts. In general, we extracted the proportion of matrix declarative sentences in each text in which the respective phenomenon occurred, and calculated average frequencies by means of the total amount of matrix declarative clauses in the corresponding text. For the expletives, we calculated relative frequencies on the basis of the proportion of expletives occurring in presentationals and impersonals, based on the findings of a recent IcePaHC study (Booth, 2018). As an approximation of the frequency of Stylistic Fronting, we counted the matrix declarative clauses with a non-finite verb, verbal particle or negation in the clause-initial position (e.g., Maling, 1990). In order to track the rise of VO in the verb phrase at the expense of OV, we calculated the occurrence frequencies of VO and OV in matrix declaratives with a finite auxiliary and a nonfinite lexical verb, in order to abstract away from the verb-second property (see, e.g. Pintzuk, 2005). For each text, the proportion of VO versus OV was included in our syntactic vectors. The resulting data was loaded into R in the form of a data matrix, where each column represents the syntactic vector of an IcePaHC text.

4.3 Application of DiaHClust

Before applying our implementation of VNC in R via the `vnc()` function, we calculated a correlation and distance matrix for our syntactic vectors. Since we start our clustering process with 61 vectors (IcePaHC texts), the resulting number of clusters is quite large. `optimal_clust()` proposes to cluster the data into 28 clusters via the calculation of silhouette coefficients. Although the silhouettes suggest that the clusters are well structured (average silhouette coefficient > 0.5), analyzing the data quantitatively on the basis of 28 time stages is not sensible. Moreover, the visual exploration of a dendogram with such a high number of vectors is rather difficult, see Figure 1.

Therefore, we iteratively continue the VNC clustering process via the application of the `diahclust()` function until the optimal number of clusters is smaller than 10. In this way, we obtain a clustering which suggests 6 time stages: 1150–1210, 1250–1450, 1475–1630, 1830–1830, and 1835–2008. These groups can also be visually detected in the dendogram in Figure 2. Although the resulting time stages are discontinuous, we do not view this as a problem, as this reflects the distribution of texts over time and how these texts behave with respect to the syntactic phenomena. The time stage '1830–1830' consists of a single text, '1830.HELLISMENN', while the neighboring clusters are quite large. This suggests that '1830.HELLISMENN' is an outlier. This is also captured in the dendogram in Figure 2, where '1830.HELLISMENN' clusters strikingly late. The divergent behaviour of this text is likely explained by the fact that it is a 19th cen-

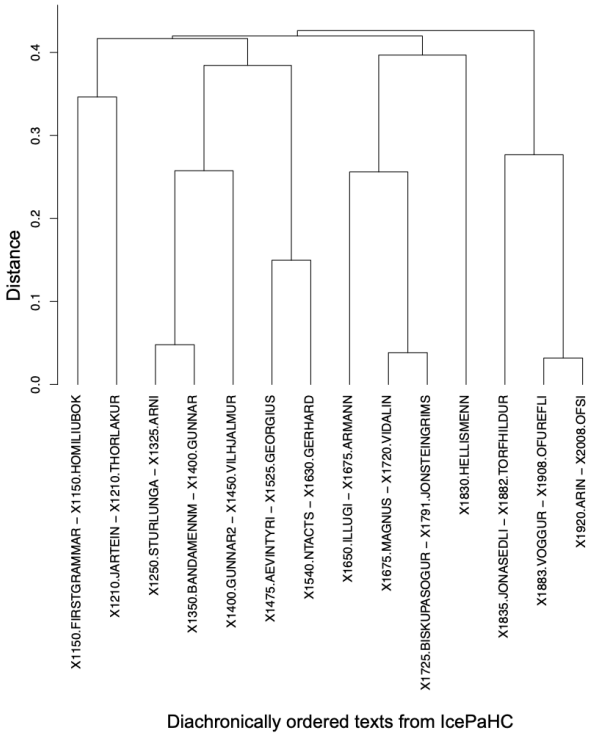

Figure 2: Dendrogram showing the results of the iterative DiaHClust approach with respect to syntactic change in IcePaHC.

tury composition which aims to imitate the older saga style. Moreover, by browsing through the dendrograms generated at each iteration, significant insights about cluster correspondences and genre effects can be obtained. For example, the third time stage (1475–1630) mainly consists of religious texts, which show a close similarity to one another already from the first iteration.

We decided to exclude '1830.HELLISMENN' and repeated the clustering process. This yielded five well-clustered time stages: 1150–1210; 1250–1450; 1475–1630; 1650–1882; 1883–2008. Whereas the first three time stages remain the same, the new clustering sheds more light on the developments occurring in the late 19th century, since '1830.HELLISMENN' no longer blocks the clustering of the surrounding texts. Moreover, the new clustering performs better in terms of average silhouette coefficients in that, with '1830.HELLISMENN' excluded, the coefficient increases from 0.4 to 0.5, indicating a more coherent clustering.

4.4 Investigating syntactic change

Once an appropriate periodization has been identified, the frequencies for the syntactic phenomena can be reassessed against this scheme which respects the corpus design and is faithful to the language-internal developments. Table 1 presents the relative frequencies for the syntactic changes under investigation averaged over the five new time stages obtained via DiaHClust. Compared to previous corpus-based investigations of these changes which made use of a top-down periodization scheme, the changes have a more gradual trajectory, cf. Table 2, which shows comparable findings from Booth et al. (2017) using a predetermined periodization. Whereas the occurrence frequencies for the syntactic changes in Table 2 remain rather stable until the last time stage, i.e., until 1900 where drastic changes can be observed, investigating the same phenomena via DiaHClust provides a more nuanced picture. Firstly, the most striking developments can be pinned down more precisely to 1650–1882 and 1883–2008. Moreover, some level of change is visible in earlier periods too. In Table 1, the frequencies in the third time stage (1475–1630) deviate from the overall trajectories. This can be attributed to a genre effect, as the DiaHClust method offers easy access to the composition of this time stage, which as mentioned consists almost exclusively of religious texts. Although this genre effect has been noted of IcePaHC by Booth et al. (2017), this effect could not be so clearly isolated using a top-down periodization, leading to a significant loss of information compared to the DiaHClust periodization method.

5 Conclusion

This paper presents a new method for the data-driven periodization of historical corpus data. Our method, DiaHClust, is implemented in R and further develops the VNC approach by Gries and Hilpert (2008, 2012). We use vectors of syntactic change as input to create knowledge-informed models of the syntactic system at different stages of the language. Furthermore, DiaHClust adds an extra iterative layer of clustering, which allows one to start the clustering at text-level, and provides significant insights about the clustering process at different levels of detail.

In order to demonstrate its value, we applied DiaHClust to a corpus-based study of syntactic change in Icelandic. Using DiaHClust reveals that syntactic change follows a more gradual trajectory in Icelandic than has been previously assumed.

Change	1150-1210	1250-1450	1475-1630	1650-1882	1883-2008
Dative subjects	3.4%	4.0%	2.6%	4.1%	5.5%
Expletives	0.1%	0.1%	0.2%	0.5%	1.5%
V1	23.7%	23.2%	6.9%	15.6%	2.3%
Prefinite subjects	44.0%	52.6%	56.2%	55.8%	72.0%
VO	48.1%	56.2%	59.9%	71.2%	83.8%
Stylistic Fronting	1.8%	1.5%	1.2%	1.2%	0.6%

Table 1: Distribution of dative subjects, expletives, V1, prefinite subjects, VO and Stylistic Fronting in IcePaHC according to the periodization scheme obtained via DiaHClust after outlier removal.

Change	1150-1349	1350-1549	1550-1749	1750-1899	1900-2008
Dative subjects	3.9%	3.2%	3.7%	3.8%	5.8%
V1	20.6%	19.9%	14.8%	18.4%	2.7%
Prefinite subjects	51.4%	55.0%	54.2%	57.6%	73.0%

Table 2: Distribution of dative subjects, V1, and prefinite subjects in IcePaHC as per Booth et al. (2017).

Moreover, DiaHClust carves out the effect which genre has on the syntactic phenomena in question and allows the researcher to track changes along the diachrony more easily, without obscuring transitional periods. Finally, we have shown that DiaHClust offers valuable insights into a language like Icelandic, where the available corpus data is relatively sparse and where syntactic change is relatively subtle. As such, applying DiaHClust to a language like English – for which there are several diachronic corpora and where syntactic change is more 'extreme' – should be relatively unproblematic. Testing this, we leave for future work.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 251654672 – TRR 161 (Project D02) for their financial support. We would also like to thank Aikaterini-Lida Kalouli and Miriam Butt for the valuable feedback that they have provided for this paper.

References

R. Harald Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Hannah Booth. 2018. *Clause Structure and Expletives: Syntactic Change in Icelandic*. Ph.D thesis, University of Manchester.

Hannah Booth, Christin Schätzle, Kersti Börjars, and Miriam Butt. 2017. Dative subjects and the rise of positional licensing in Icelandic. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG'17 Conference, University of Konstanz*, pages 104–124. CSLI Publications, Stanford, CA.

Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe, and Daniel Keim. 2014. V1 in Icelandic: A multifactorical visualization of historical data. In *Proceedings of the LREC 2014 Workshop "VisLR: Visualization as added value in the development, use and evaluation of Language Resources"*, Reykjavik, Iceland.

Anne Curzan. 2012. Periodization in the history of the English language. In Alexander Bergs and Laurel J. Brinton, editors, *The History of English: Vol.1 Historical Outlines from Sound to Text*, pages 8–35. de Gruyter, Berlin.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico, USA.

Aaron Ecay and Susan Pintzuk. 2016. The syntax of Old English poetry and the dating of Beowulf. In Leonard Neidorf, Rafael J. Pascual, and Tom Shippey, editors, *Old English Philology: Studies in Honour of R.D. Fulk*, pages 219–258. D.S Brewer, Cambridge.

Stefan Th. Gries and Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1):59–81.

Stefan Th. Gries and Martin Hilpert. 2012. Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In Nevalainen

Terttu and Elizabeth Closs Traugott, editors, *The Oxford Handbook of the History of English*, pages 134–144. Oxford University Press, Oxford.

Einar Haugen. 1984. *Die skandinavischen Sprachen: Eine Einführung in ihre Geschichte*. Hamburg: Buske.

Martin Hilpert and Stefan Th. Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4).

Martin Hilpert and Stefan Th. Gries. 2016. Quantitative approaches to diachronic corpus linguistics. In Merja Kytö and Päivi Pahta, editors, *The Cambridge Handbook of English Historical Linguistics*, Cambridge Handbooks in Language and Linguistics, page 36–53. Cambridge University Press, Cambridge.

Thorbjörg Hróarsdóttir. 1998. *Setningafræðilegar Breytingar á 19. Öld: þróun þriggja málbreytinga*. Málvísindastofnun Háskóla Íslands, Reykjavík.

Thorbjörg Hróarsdóttir. 2000. *Word Order Change in Icelandic. From OV to VO*. John Benjamins, Amsterdam.

Anthony Kroch. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.

Anthony Kroch. 2001. Syntactic change. In Mark Baltin and Chris Collins, editors, *The Handbook of Contemporary Syntactic Theory*, pages 699–729. Blackwell, Oxford.

Joan Maling. 1990. Inversion in embedded clauses in Modern Icelandic. In Joan Maling and Annie Zaenen, editors, *Syntax and Semantics: Modern Icelandic Syntax*, pages 71–91. Academic Press, San Diego, CA.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Florent Perek and Martin Hilpert. 2017. A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics*, 22:490–520.

Susan Pintzuk. 2005. Arguments against a universal base: evidence from Old English. *English Language & Linguistics*, 9(1):115–138.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Beth Randall. 2000. CorpusSearch: a Java program for searching syntactically annotated corpora. Dept. of Linguistics, University of Pennsylvania, Philadelphia.

Eiríkur Rögnvaldsson. 1996. Word order variation in the VP in Old Icelandic. *Working Papers in Scandinavian Syntax*, 58:55–86.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, (20):53–65.

Christin Schätzle. 2018. *Dative Subjects: Historical Change Visualized*. Ph.D. thesis, University of Konstanz.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC), version 0.9. http://linguist.is/icelandic_treebank.

Richard Zimmermann. 2014. Dating hitherto undated Old English texts based on text-internal criteria. *Ms.*, University of Geneva.