Predicting the propagation of language change using regular time series

Sigríður Sæunn Sigurðardóttir
University of Iceland, sigridursaeunn@hi.is

Language forecasting, i.e., predicting the future state of a language, has long been regarded with a fair amount of skepticism (Keller 1994:72; Bauer 1994:25; Labov 1994:10; Croft 2000:3). In recent years, however, more positive views have emerged with scholars focusing on the predictive power of the S-curve (Sanchez-Stockhammer 2015; Van de Velde 2017).

In this paper the potentials of language forecasting for studying language change are discussed. An overview is provided of each step of the forecasting task, including identification of the forecasting problem, forecast length, type of data required, which methods can be used and how forecasts are eventually generated. Furthermore, it is shown how regular time series and tried and tested forecasting methods (e.g., Box, Jenkins, Reinsel & Ljung 2016; Hyndman and Athanasopoulos 2021) can be used to predict the propagation of selected linguistic variants.

The process of forecasting is demonstrated with an example from Icelandic. While nominative is the unmarked subject case in Icelandic, oblique subjects (accusative, dative or genitive) also occur (Andrews 1976 and subsequent work). Throughout the history of the language, there has been a tendency to replace oblique subjects with nominative (e.g., Svavarsdóttir 1982). This change is termed Nominative Substitution and it has become more widespread in recent years. The change centered on in this case study is the opposite of Nominative Substitution, i.e., Oblique-Case Substitution, involving generalization of oblique case (accusative or dative) with a limited number of experiencer predicates originally taking nominative subject (e.g., Svavarsdóttir 1982; Jónsson & Eythórsson 2005). One such predicate is *hlakka til* 'look forward to' which was originally used with a nominative (1), but is now also attested with accusative (from 1892 onwards) and dative (from 1942 onwards) as illustrated in (2).

(1) ***Ég*** *hlakka til sumarsins.*
I.NOM look.forward to the.summer
'I look forward to the summer.'

(2) ***Mig***/***Mér*** *hlakkar til sumarsins.*
me.ACC/DAT looks.forward to the.summer
'I look forward to the summer.'

Focusing on first-person subjects with *hlakka til*, a total of 22.198 examples from the period 2003–2021 were extracted from selected sources in the Icelandic Gigaword corpus (IGC, cf. Steingrímsson et al. 2018) and a total of 9.629 examples from the period 2012–2022 from X (formerly Twitter). All examples were annotated according to whether they had a nominative or an oblique subject. The data was then projected into a regular time series with yearly (IGC) and quarterly (X) observations, showing the proportion of oblique subjects at each period. The IGC and the X time series were used as input for forecasting models that are designed to pick up on trend-signals in historical data. Simple models such as a Naïve and Drift model were

used as a baseline to contrast more complex models with. The more complex models were types of ETS and ARIMA models (e.g., Hyndman & Athanasopoulos 2021) which use lagged values of historical data to predict future observations. Before predictions were made, the IGC and X time series were split into training and test sets. Models were fitted to the training data which amounted to 89% (17 observations) of the ICG series and 91% (40 observations) of the X series. The test set was used to evaluate how well each model performed. Based on the best performing model for each series, a forecast was made until Q4 2029 (X) and 2041 (IGC) using the entire historical data. A forecast for the IGC data is shown in Figure 1 where the proportion of oblique 1st person subjects is predicted to become minimal in 2041.

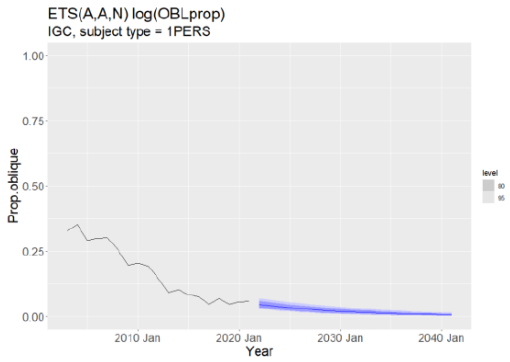

Figure 1. A forecast for the proportion of 1st person oblique subjects with *hlakka til* until 2041. An ETS(A, A,N) model was used.

Although the accuracy of predictions can only be properly evaluated once the specified time has come to pass, the process of forecasting offers a novel and important insight into language change. Systematic forecasting methods impose requirements on the type of data that can be used for making predictions. In the case of time series analysis, data must be gathered at regular intervals over an extended period of time, and this generally leads to a more thorough documentation of the linguistic variant of interest. Ultimately, this leads to a better understanding of how to extract information on language change from data that shows variation. The Icelandic example discussed here shows that observing variation at regular intervals over 10–18 years can give a consistent signal as to the propagation of a linguistic variant. This suggests that time series analysis and forecasting is a promising step in language forecasting.

References

Bauer, Laurie. 1994. *Watching English change: An introduction to the study of linguistic change in Standard Englishes in the twentieth century*. London: Longman.

Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel & Greta M. Ljung. 2016. *Time series analysis: Forecasting and control* (Fifth edition). Hoboken, New Jersey: Wiley.

Croft, William. 2000. *Explaining language change: An evolutionary approach*. London: Longman.

Hyndman, Rob J. & George Athanasopoulos. 2021. *Forecasting: Principles and practice* (3rd edition). Melbourne, Australia: Otext. https://otexts.com/fpp3/.

Jónsson, Jóhannes Gísli & Thórhallur Eythórsson. 2005. "Variation and change in subject case marking in Insular Scandinavian." *Nordic Journal of Linguistics*, 28(2):223–245.

Keller, Rudi. 1994. *On language change: The invisible hand in language*. London: Routledge.

Labov, William. 1994. *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell.

Sanchez-Stockhammer, Christina. 2015. "Can we predict linguistic change? An introduction." *Studies in Variation, Contacts and Change in English*, 16:15.

Svavarsdóttir, Ásta. 1982. "Þágufallssýki." *Íslenskt mál og almenn málfræði*, 4:19–62.

Van de Velde, Freek. 2017. Retro-predicting language change with binomial regression analysis. [Conference Presentation]. ICHL, Austin, Texas, United States.