

# Developing methods for reproducible research in linguistics: A first step

Bradley McDonnell and Patrick Hall

University of Hawai'i at Mānoa and University of California, Santa Barbara

## Introduction

- ▶ Practical methods for *Reproducible Research* have been developed for other fields:
  - ▷ The R package *knitr*<sup>[1]</sup> and  $\LaTeX$  provides a good model for many in linguistics.
  - ▷ Dynamic Documents<sup>[2]</sup>: R source code (resulting in numeric and graphical output) written alongside literal writings (in  $\LaTeX$ , HTML, or markdown).
    - ▶ *knitr* allows R to compile at the same time as  $\LaTeX$ .
    - ▶ allows data, source code, and analysis to be linked or 'live' in the same place.
    - ▶ already common in fields that rely on statistical analysis (including linguistics).
- ▶ For RR, Dynamic Documents crucially allow computation to be *portable*.
  - ▷ Data and dynamic document are in a single directory.
  - ▷ Workflow is streamlined with less room for error.
    - ▶ Necessary changes to the data are done in a one-step process.

## The Problem

- ▶ For linguistics, there is a lack of practical methodologies for Reproducible Research:
  - ▶ typically linguists manually input and format data (e.g., interlinearized glossed examples) into a document.
    - ▷ Manually inputting/formatting linguistic examples is *tedious* and *error-prone*.
    - ▷ Citing examples adds more manual tasks that are again tedious and error-prone.
    - ▷ Leaves no link between corpus and example.
    - ▷ Inevitable changes to the data require a two-step process.
- ▶ Glossbox<sup>[3]</sup> sets out to provide methods for RR for linguists using interlinearized glossed examples.
  - ▷ Glossbox is a Python script that allows linguists to insert examples from a Toolbox corpus into a  $\LaTeX$  document.

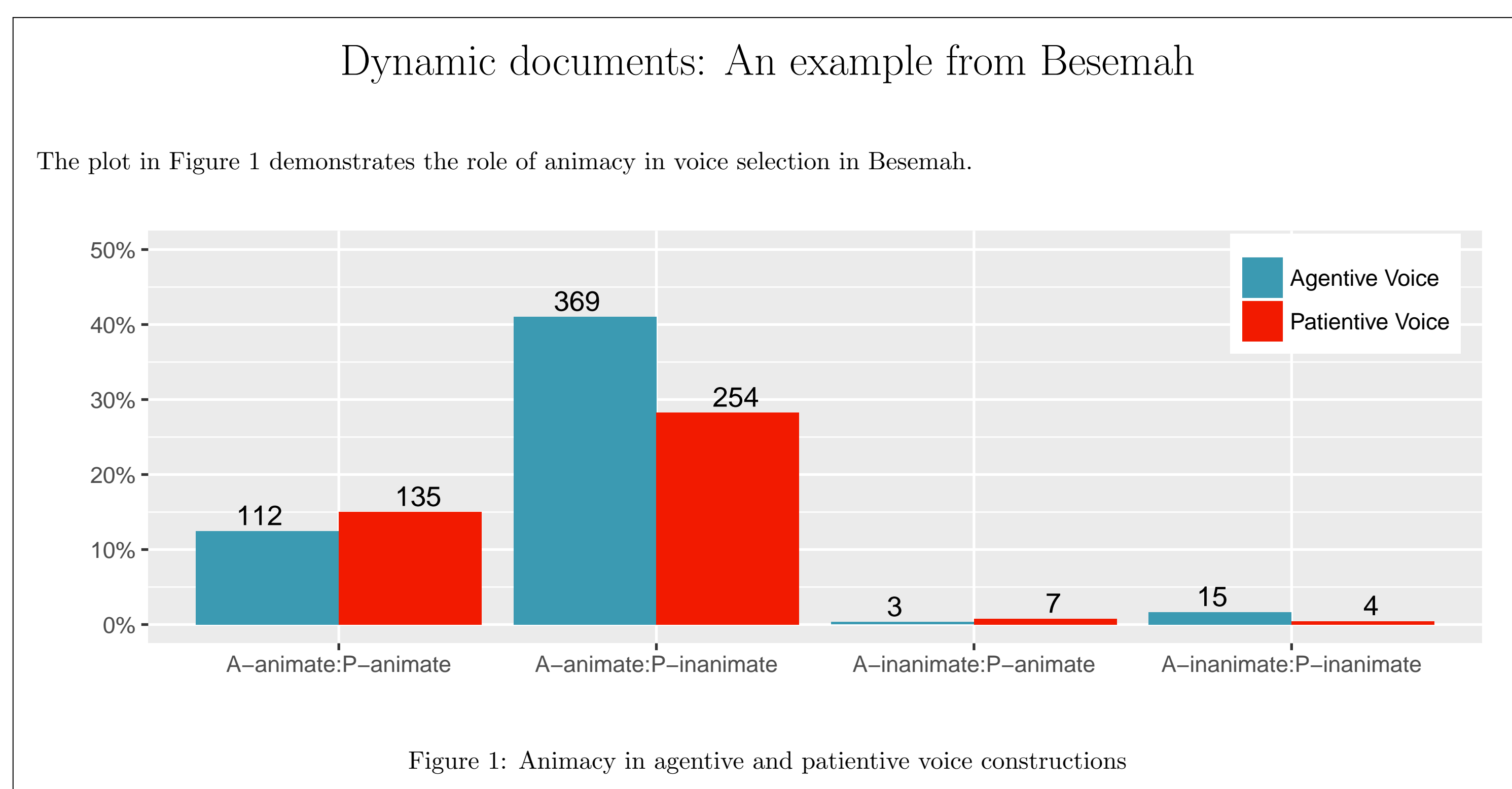
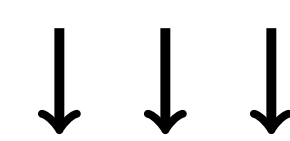
## Dynamic Document with knitr: A good model

```
Dynamic Document Source Code

\documentclass{article}
\title{Dynamic documents: An example from Besemah}
\begin{document}
\maketitle

<<setup, include=FALSE>>=
library(knitr); library(ggplot2); library(wesanderson)
@

\noindent The plot in Figure \ref{fig:pse-animacy-plot} demonstrates the role of animacy in voice selection
in Besemah.
\begin{figure}
<<model, fig.width=8, fig.height=3.5, fig.align='center', echo=FALSE>>=
animacy_table <- read.delim("animacy_table.csv")
animacy_plot <- ggplot(animacy_table,
  aes(x=Pairs, y=Frequency, fill=Voice)) +
  geom_bar(stat="identity", position=position_dodge()) +
  scale_y_continuous(labels=scales::percent, limits=c(0,0.5)) +
  xlab("") +
  ylab("") +
  labs(title = "") +
  theme(legend.title=element_blank(), legend.justification=c(1,0), legend.position=c(1,0.65)) +
  geom_text(data=animacy_table,
    aes(x=Pairs, y=Frequency, label=c(112,3,369,15,135,7,254,4)),
    vjust = -0.3, position = position_dodge(width=1)) +
  scale_fill_manual(values=wes_palette(name="Darjeeling"),
    labels=c("Agentive Voice", "Patientive Voice"))
animacy_plot
@
\caption{Animacy in agentive and patientive voice constructions}\label{fig:pse-animacy-plot}
\end{figure}
\end{document}
```



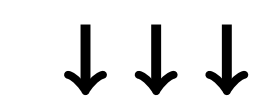
## Glossbox example

```
Dynamic document source code before gloss inclusion

\documentclass{article}
\usepackage{expex}
\title{Glossbox: An example from Besemah}
\begin{document}
\maketitle

\noindent The example in (\ref{ex:pse-pv-unrealized-a}) is a case where A is unrealized in the patientive
voice.\x

\ex~\underline{Patientive voice with unrealized A argument}\x
% GLOSSBOX_BJM01-011_00:00:42.000_00:00:44.000
\label{ex:pse-pv-unrealized-a}
\end{document}
```



```
Dynamic document source code after gloss inclusion

\documentclass{article}
\usepackage{expex}
\title{Glossbox: An example from Besemah}
\begin{document}
\maketitle

\noindent The example in (\ref{ex:pse-pv-unrealized-a}) is a case where A is unrealized in the patientive
voice.\x

\ex~\underline{Patientive voice with unrealized A argument}\x
\begin{gloss}
\gla e'e lah di-pulibik-i.//
\glb uhuh \textsc{pfv} \textsc{pv}-plastic.bag-\textsc{loc.appl}//
\glft uhuh, (the seedlings) have already been potted.\\
\trailingcitation{(oai:paradisec.org.au:BJM01-011, 00:00:44.008--00:00:44.008, Rili)}//
\end{gloss}
% GLOSSBOX_BJM01-011_00:00:42.000_00:00:44.000
\label{ex:pse-pv-unrealized-a}
\end{document}
```

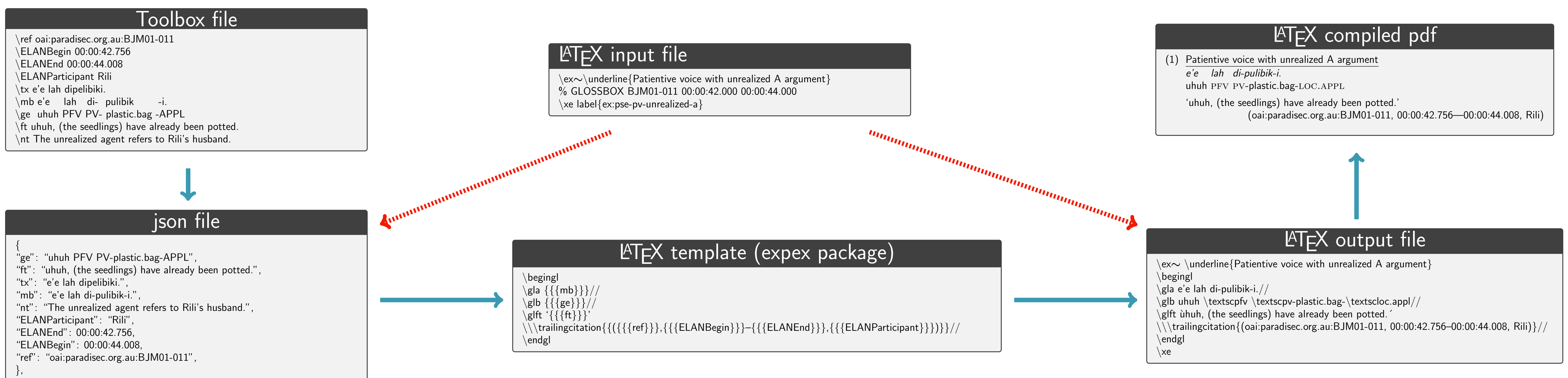


Glossbox: An example from Besemah

The example in (1) is a case where A is unrealized in the patientive voice.

(1) Patientive voice with unrealized A argument  
e'e lah di-pulibik-i.  
uhuh PFV PV-plastic.bag-LOC.APPL  
'uhuh, (the seedlings) have already been potted.'  
(oai:paradisec.org.au:BJM01-011, 00:00:44.008--00:00:44.008, Rili)

## Glossbox workflow



## Future directions

- ▶ R package *glossr* is already in planning stages by McDonnell and graduate students at the University of Hawaii.
  - ▷ *glossr* integrates better with *knitr*.
  - ▷ Ideally, *glossr* works with ELAN.
  - ▷ Can format IGEs as needed in document

## References

- Xie, Yihui. *knitr: A general-purpose package for dynamic report generation in R*. R package version 1.11. 2015.
- Xie, Yihui. *Dynamic Documents with R and knitr*. 2nd ed. Chapman and Hall/CRC, 2015.
- Hall, Patrick and Bradley McDonnell. *Glossbox*. 2016. URL: <https://github.com/amundo/glossbox>.