title: “The variability of number in Latin borrowings in the English language” author: “Arsentyev Denis” output: html_document: theme: journal

Intro

The main aim of the work is to show which forms of Latin borrowings in the English language are the dominant ones based on the modern corpora (COCA, BNC).

Data

The data comprises the following:

Data + two predicates,

Agenda + one predicate,

cacti/cactuses variability,

indexes/indices variability,

formulae/formulas variability.

Methods

In this particular work I used decision trees and ggplot graphs to determine the corellation between the following factors: register, year and dialect of English.

```
library(party)
```

```
## Warning: package 'party' was built under R version 3.3.3
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Warning: package 'strucchange' was built under R version 3.3.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.3.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Warning: package 'sandwich' was built under R version 3.3.3
```

```
library(lattice)
library(languageR)
```

```
## Warning: package 'languageR' was built under R version 3.3.3
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.3.3
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.3.3
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

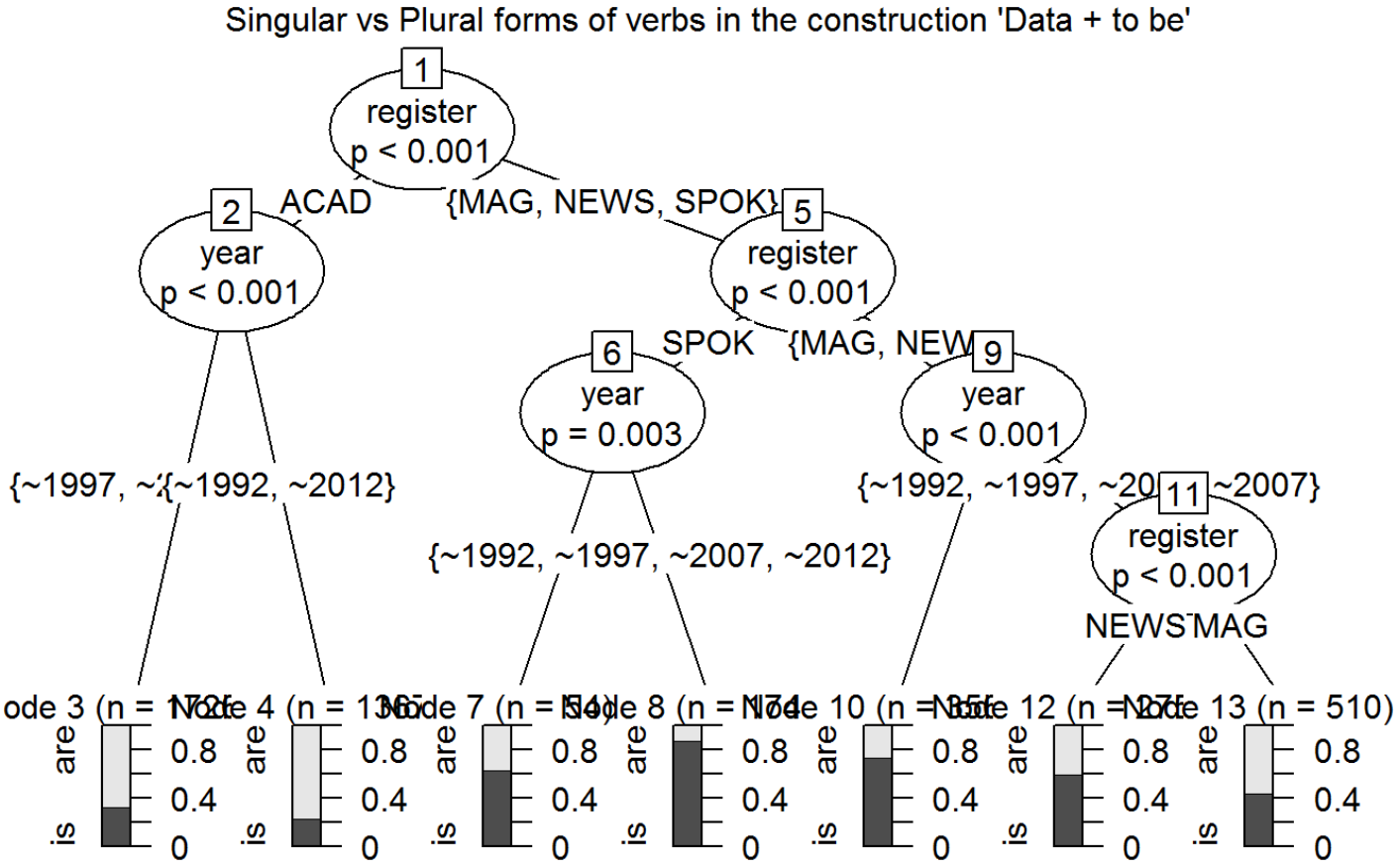
```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

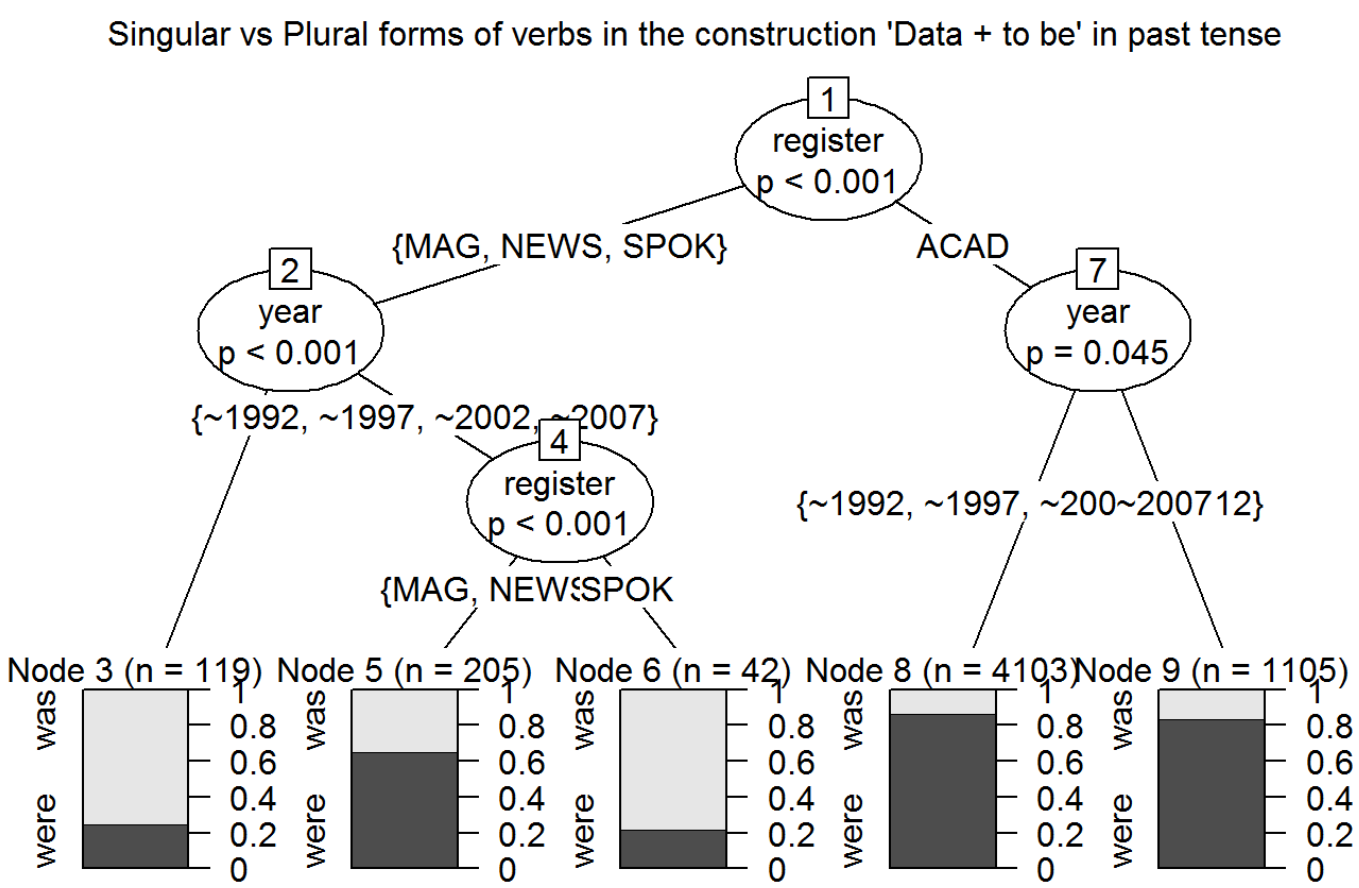
```
library(ggplot2)
```

The noun ‘data’ with two different predicates (Desicion trees)

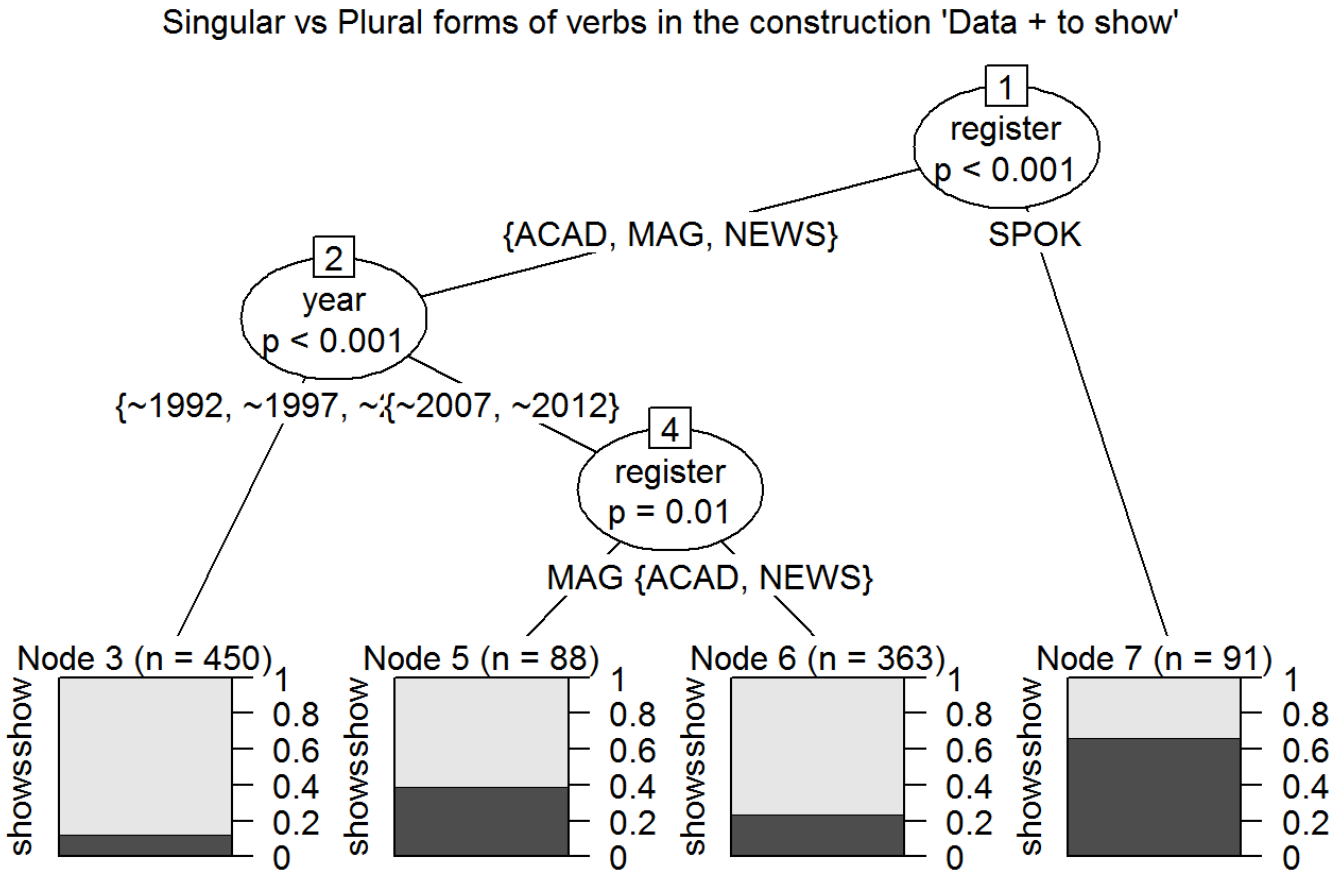
```
data_to_be <- read.csv(file='Z:\\HSE\\R\\Project\\latin2.csv', sep="\t", hea
der = TRUE)
data_to_be.ctree = ctree(verb ~ register + year, data_to_be)
plot(data_to_be.ctree, main = "Singular vs Plural forms of verbs in the cons
truction 'Data + to be'")
```




```
data_to_be_past <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\data_to be_past
.txt', sep="\t", header = TRUE)
data_to_be_past.ctree = ctree(verb ~ register + year, data_to_be_past)
plot(data_to_be_past.ctree, main = "Singular vs Plural forms of verbs in the
construction 'Data + to be' in past tense")
```




```
data_to_show <- read.csv(file='Z:\\HSE\\R\\Project\\latin1.csv', sep="\t", h
eader = TRUE)
data_to_show.ctree = ctree(verb ~ register + year, data_to_show)
plot(data_to_show.ctree, main = "Singular vs Plural forms of verbs in the co
nstruction 'Data + to show'")
```

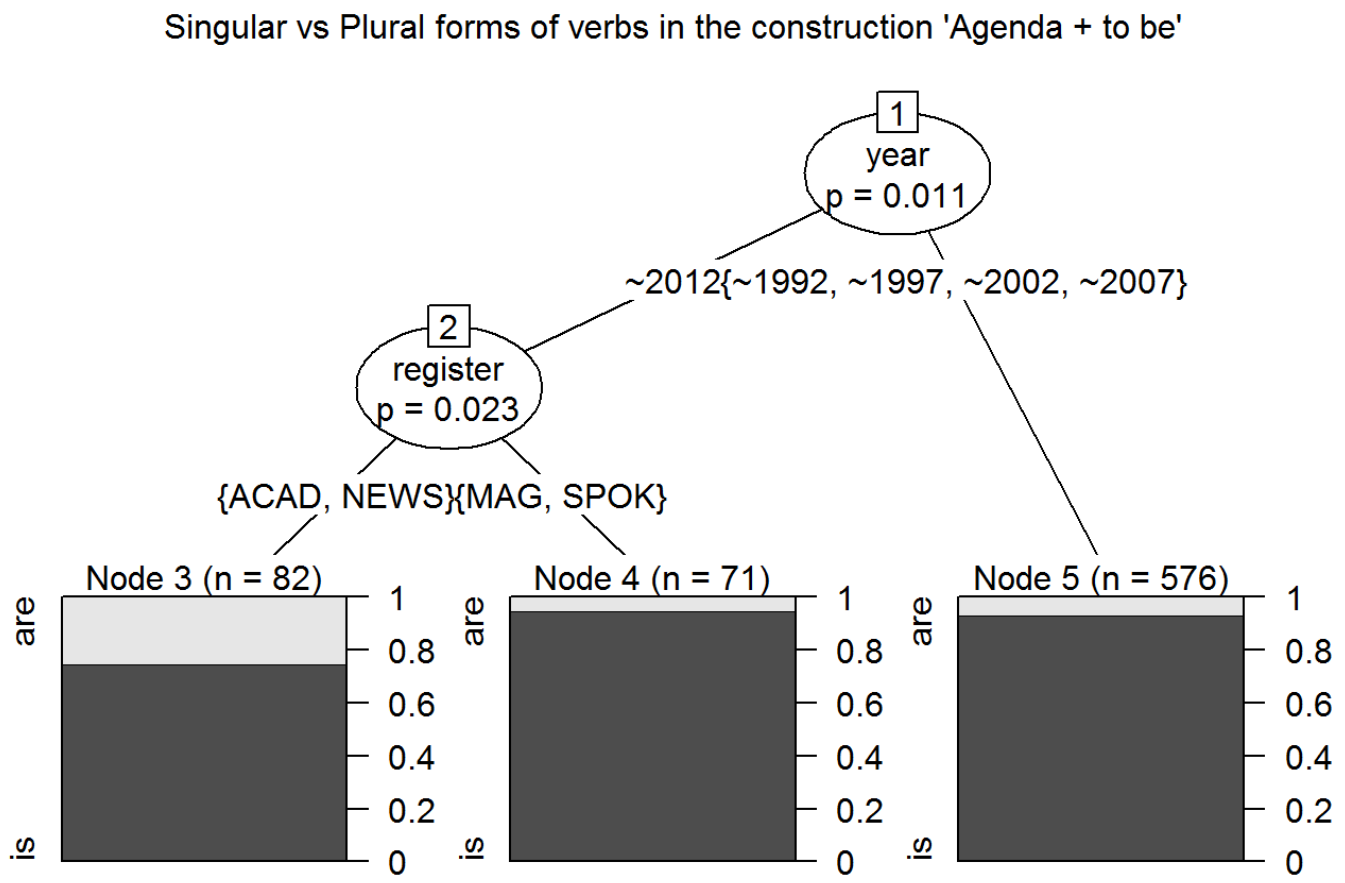


Linguistic evaluation of ‘data’

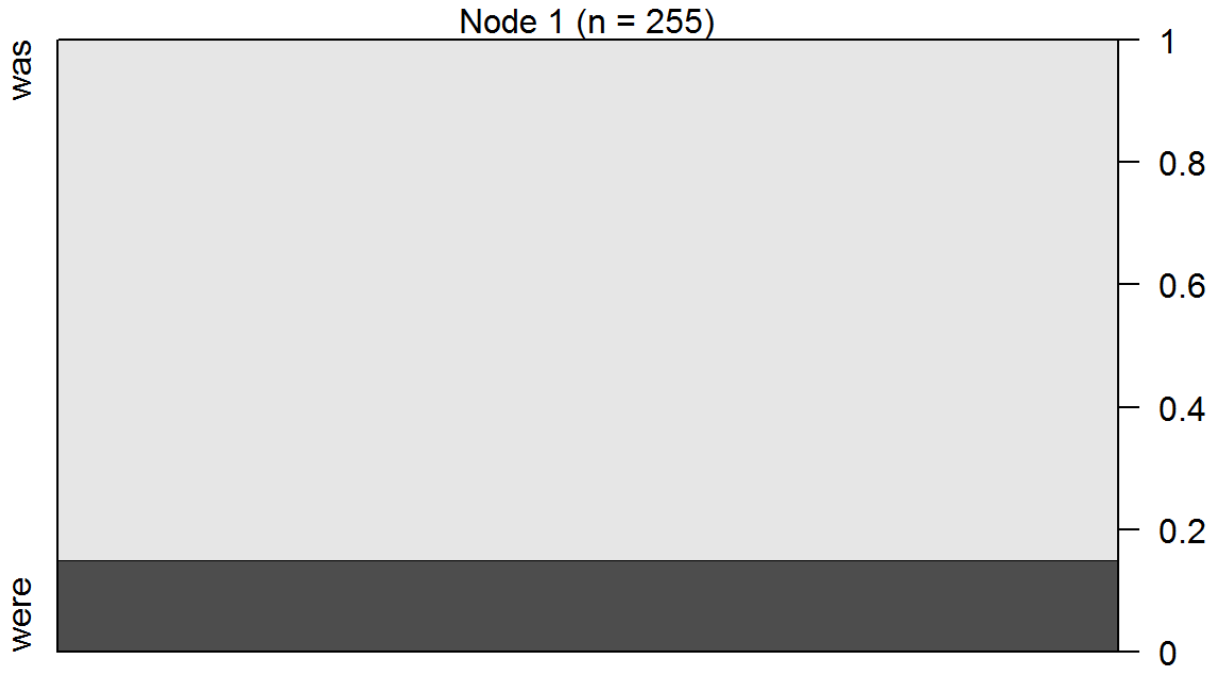
As we can see, in the academic papers ‘data’ is considered to be a plural noun, and the verb in the plural form is prefered. In the spoken register, however, ‘data’ is mostly used as a singular noun, which tells us a lot about the perception of the word by most people. In magazines and newspapers ‘data’ is percieved more or less equally as both a singular and a plural form.

The noun ‘agenda’ with one predicate in different tenses (Desicion trees)

```
agenda_to_be <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\agendabe.txt', sep
="\t", header = TRUE)
agenda_to_be.ctree = ctree(verb ~ register + year, agenda_to_be)
plot(agenda_to_be.ctree, main = "Singular vs Plural forms of verbs in the co
nstruction 'Agenda + to be'")
```




```
agenda_to_be_past <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\agenda_to be_
past.txt', sep="\t", header = TRUE)
agenda_to_be_past.ctree = ctree(verb ~ register + year, agenda_to_be_past)
plot(agenda_to_be_past.ctree, main = "Singular vs Plural forms of verbs in t
he construction 'Agenda + to be' in past tense")
```



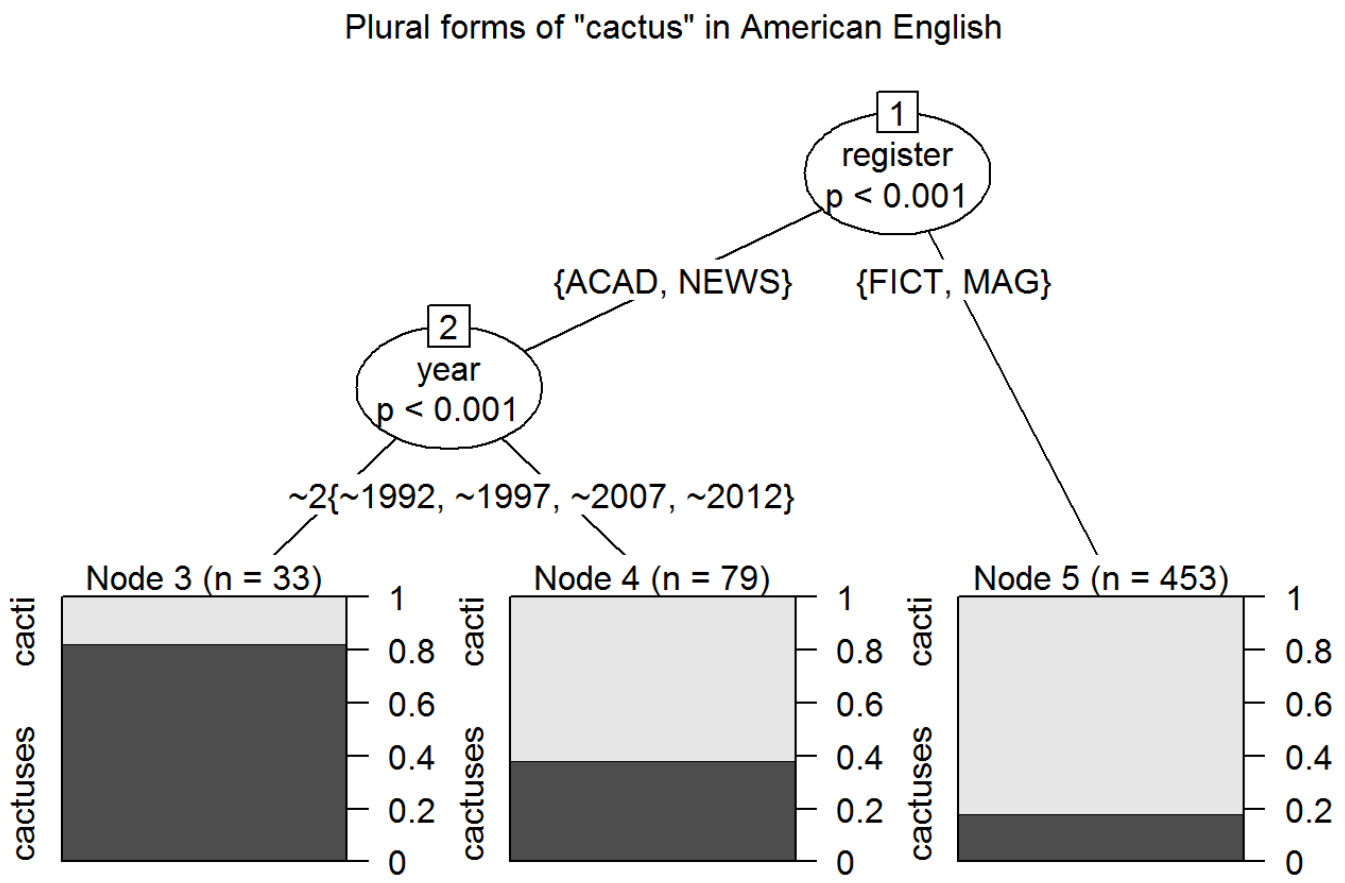
Linguistic evaluation of ‘agenda’

‘Agenda’ even though it is also a plural form (as ‘data’) is almost never percieved so. From this we can conclude that ‘agenda’ is not really a frequent siginificant word of academic papers, so the perception of ‘agenda’ is overall closer to the spoken register. And because, according to the English language, the word ‘agenda’ definitely looks more like a singular form than a plural form, it is understandable that in casual speech people percieve it as a singular noun.

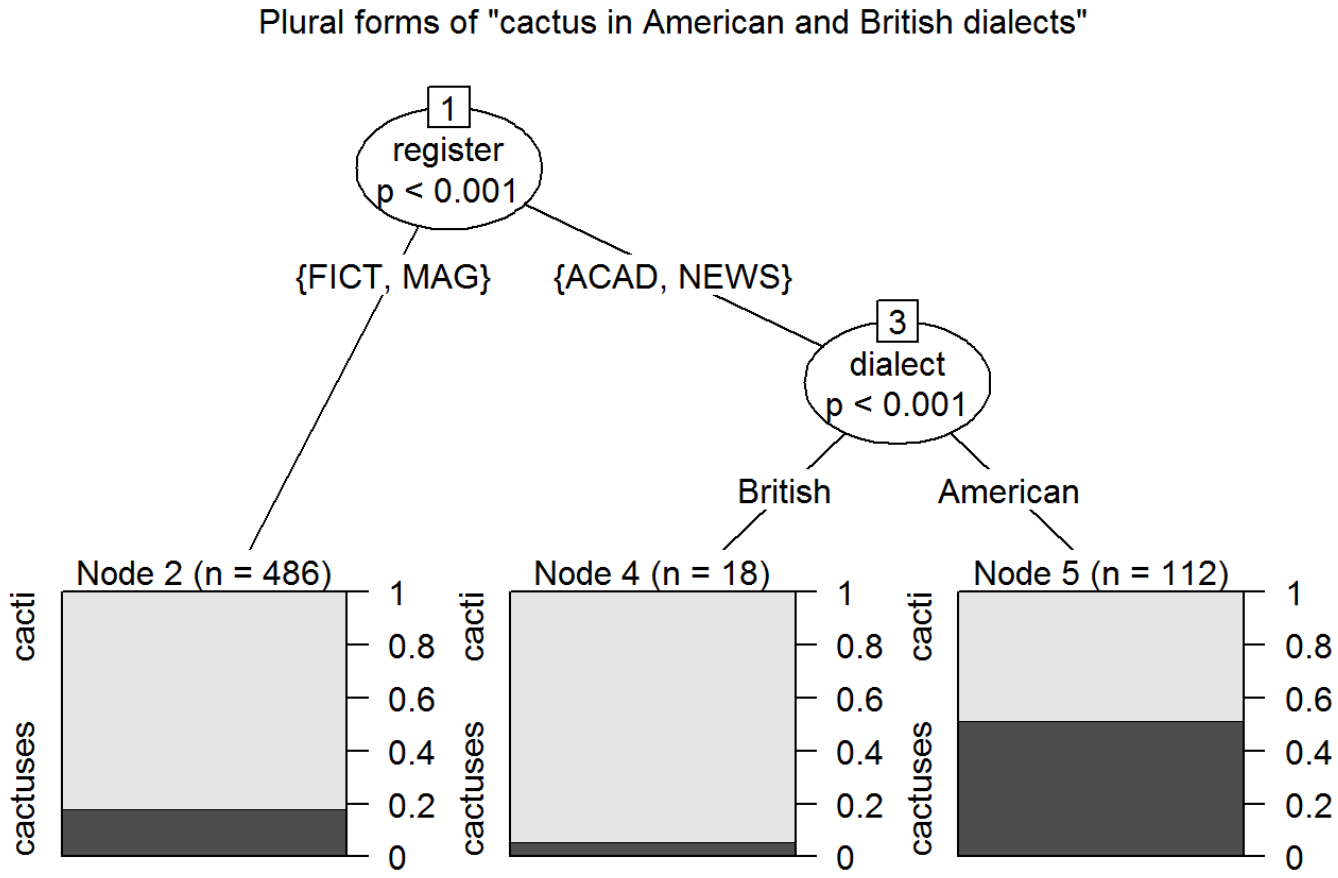
Variability of ‘cacti/cactuses’ in American English and comparing to British English (Desicion trees, ggplot)

For the variability of noun forms in the English language, I also decided to use ggplot graphs as I found them more illustrative when it comes to the correlation between the dialects and the nouns.

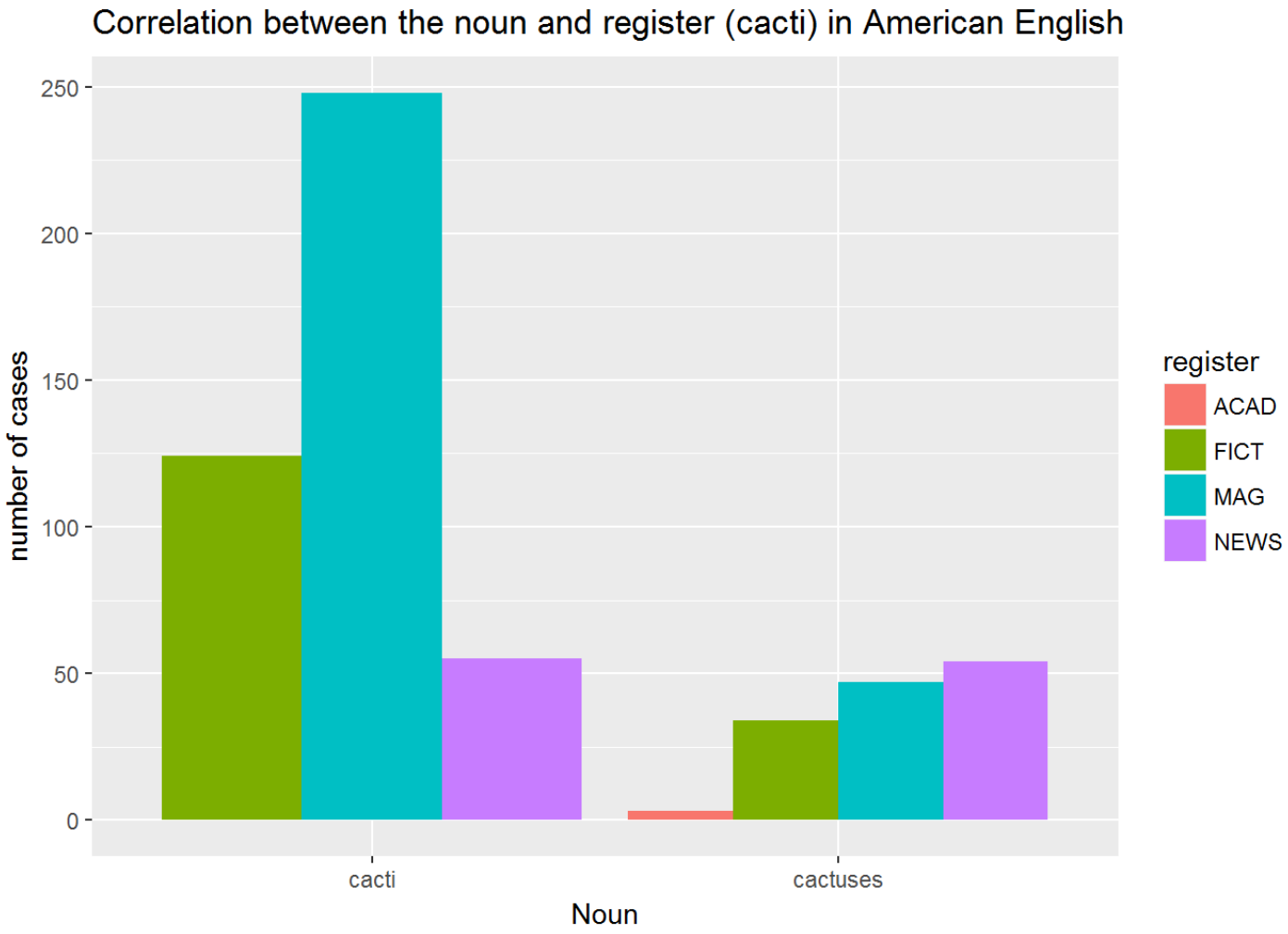
```
cacti_AM <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\Cactuses_cacti_AM.txt'
, sep="\t", header = TRUE)
cacti_AM.ctree = ctree(noun ~ register + year, cacti_AM)
plot(cacti_AM.ctree, main = 'Plural forms of "cactus" in American English')
```



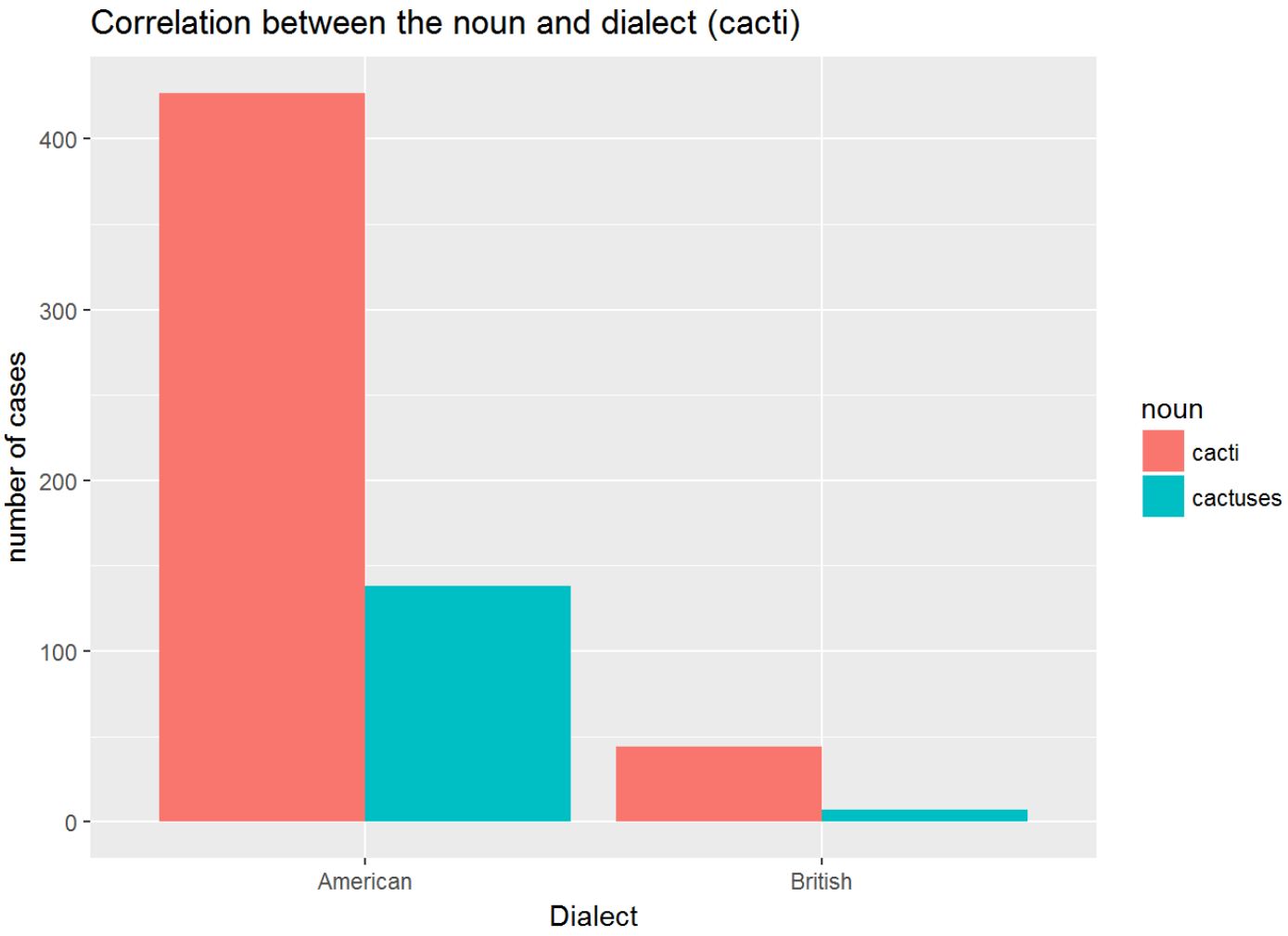

```
cacti_int <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\Cactuses_cacti_BR.txt
', sep="\t", header = TRUE)
cacti_int.ctree = ctree(noun ~ register + dialect, cacti_int)
plot(cacti_int.ctree, main = 'Plural forms of "cactus in American and Britis
h dialects"')
```




```
ggplot(data=cacti_AM, aes(x=noun, fill=register)) +
geom_bar(stat="count", position=position_dodge()) +
xlab("Noun") + ylab("number of cases") +
ggtitle("Correlation between the noun and register (cacti) in American Engli
sh")
```




```
ggplot(data=cacti_int, aes(x=dialect, fill=noun)) +
  geom_bar(stat="count", position=position_dodge()) +
  xlab("Dialect") + ylab("number of cases") +
  ggtitle("Correlation between the noun and dialect (cacti)")
```

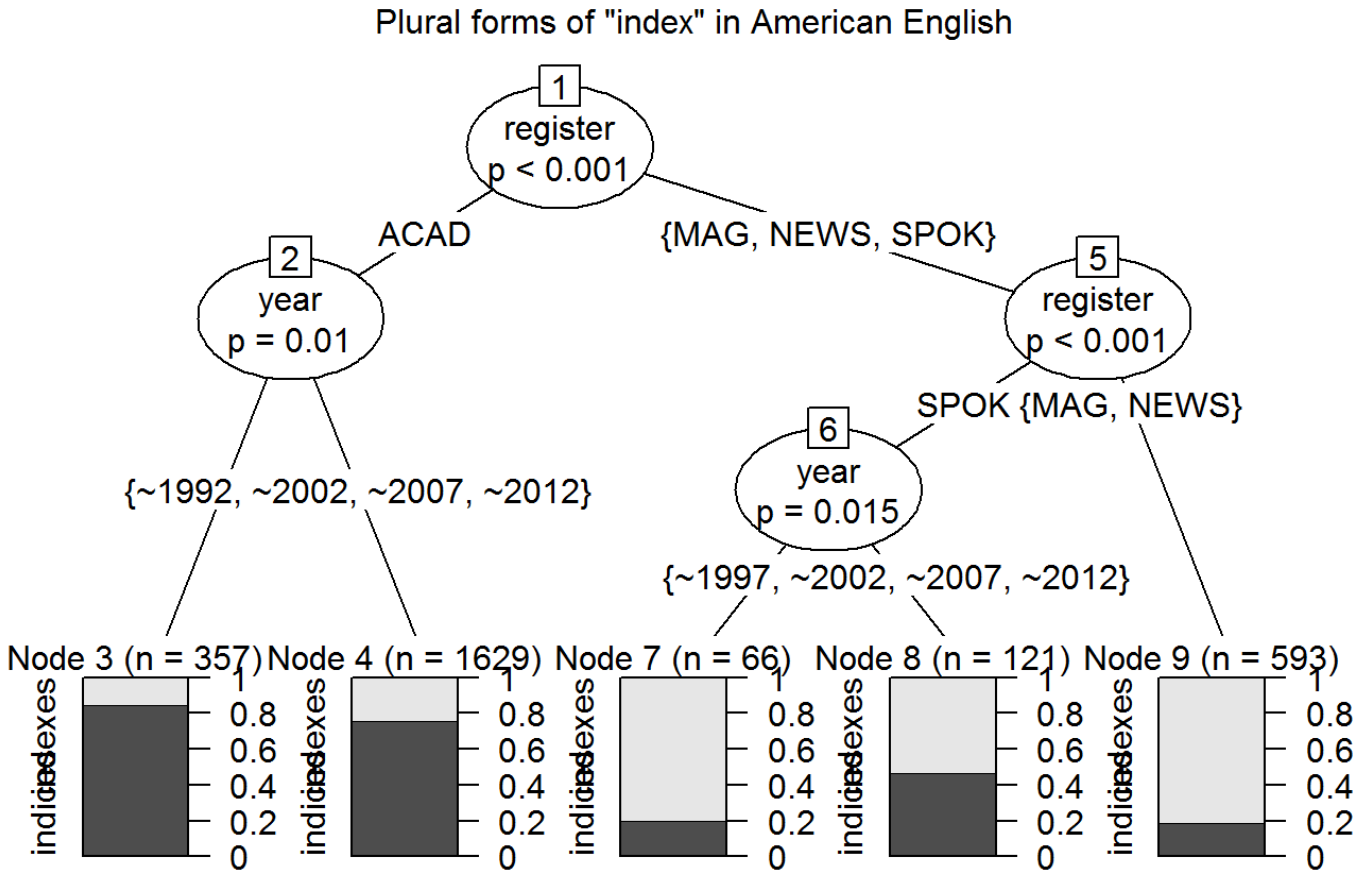


Linguistic evaluation of ‘cacti’

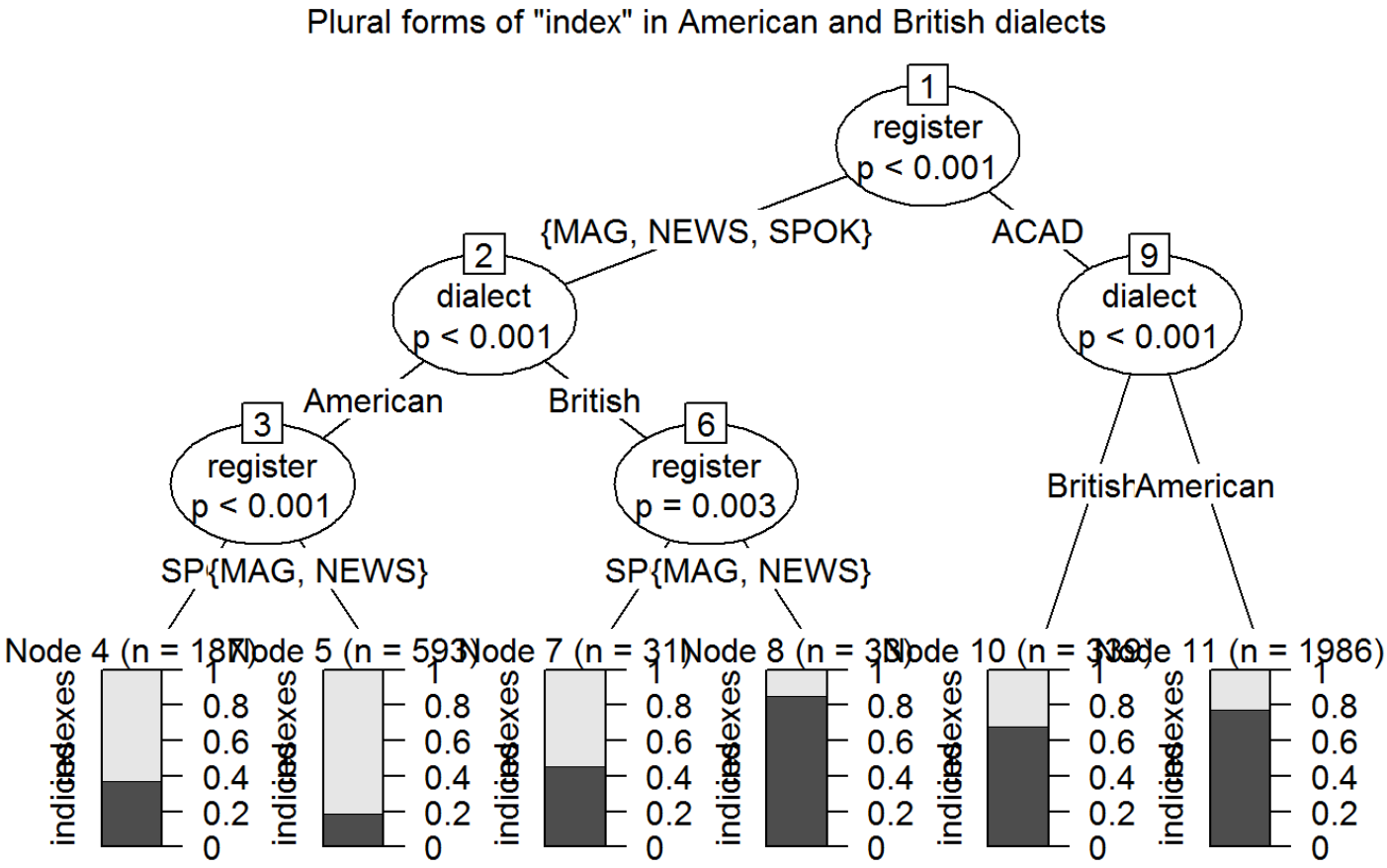
The most important thing here is that in the British dialect the form ‘cacti’ strongly prevails over the form ‘cactuses’, while in the American dialect or rather in the American corpora the distribution is not so radical.

Variability of ‘indices/indexes’ in American English and comparing to British English (Desicion trees, ggplot)

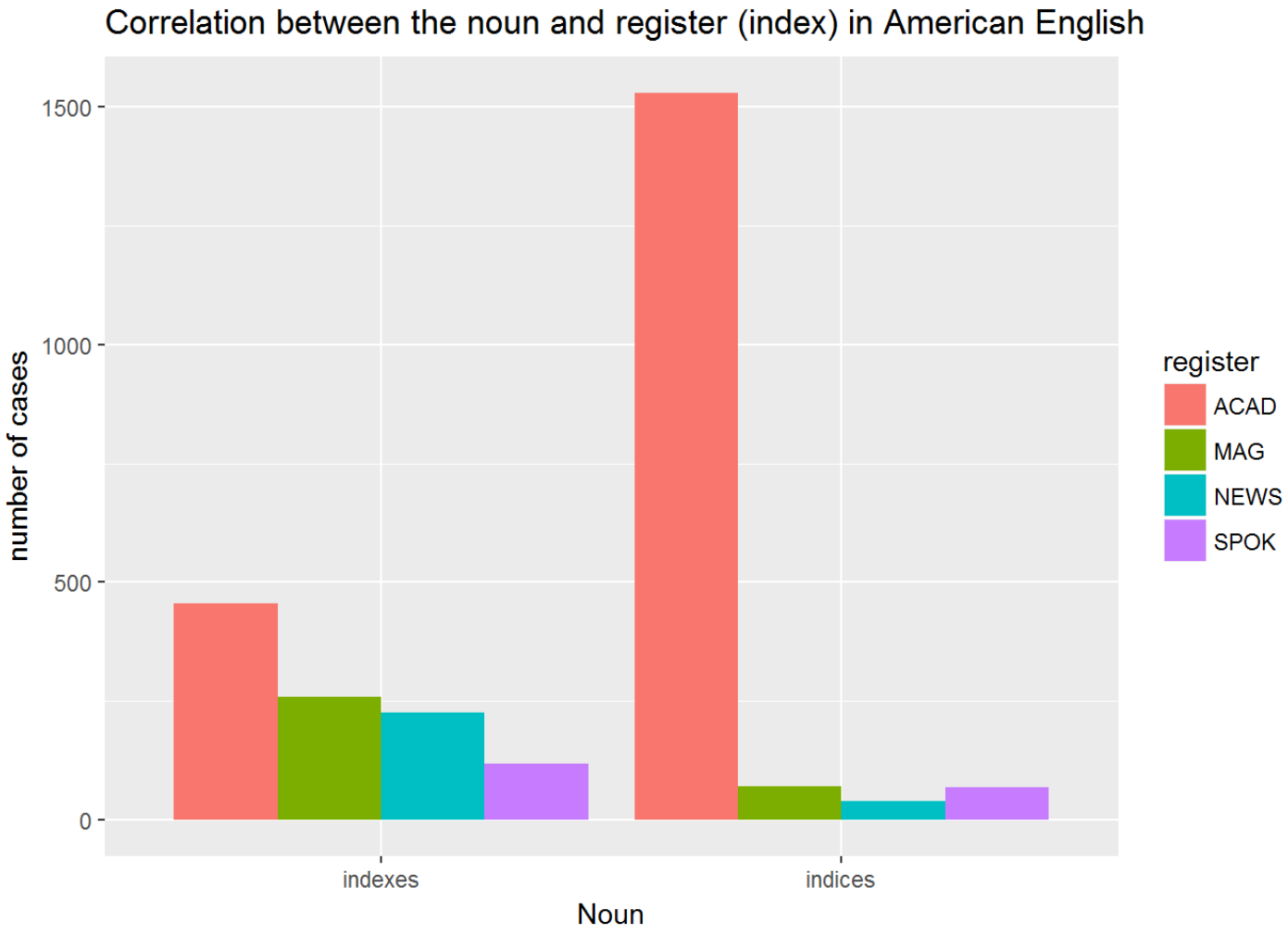
```
indices_AM <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\indices_AM.txt', sep
="\t", header = TRUE)
indices_AM.ctree = ctree(noun ~ register + year, indices_AM)
plot(indices_AM.ctree, main = 'Plural forms of "index" in American English')
```



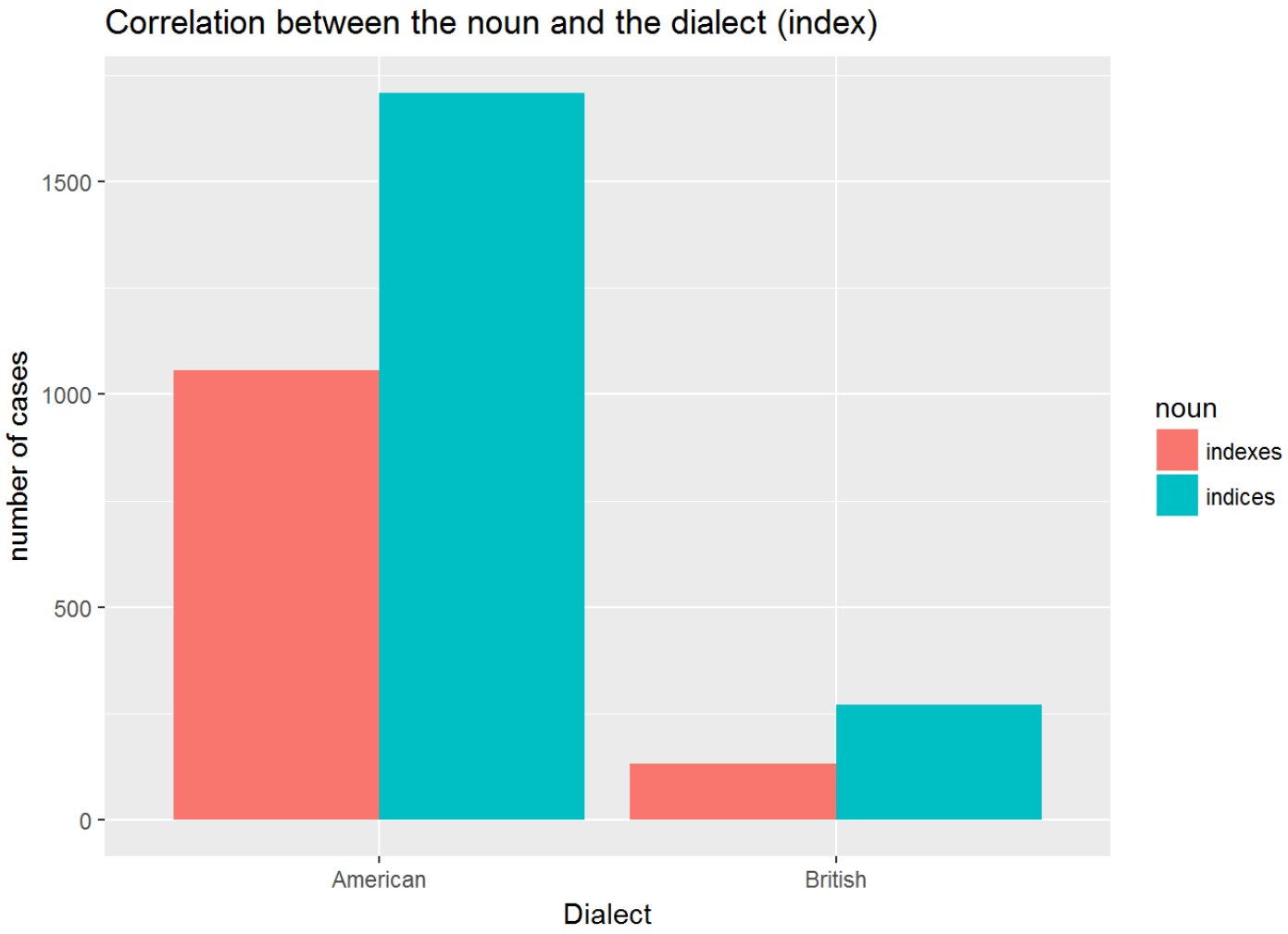

```
indices_int <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\indices_int.txt', s
ep="\t", header = TRUE)
indices_int.ctree = ctree(noun ~ register + dialect, indices_int)
plot(indices_int.ctree, main = 'Plural forms of "index" in American and Brit
ish dialects')
```




```
ggplot(data=indices_AM, aes(x=noun, fill=register)) +
  geom_bar(stat="count", position=position_dodge()) +
  xlab("Noun") + ylab("number of cases") +
  ggtitle("Correlation between the noun and register (index) in American Eng
lish")
```




```
ggplot(data=indices_int, aes(x=dialect, fill=noun)) +
  geom_bar(stat="count", position=position_dodge()) +
  xlab("Dialect") + ylab("number of cases") +
  ggtitle("Correlation between the noun and the dialect (index)")
```

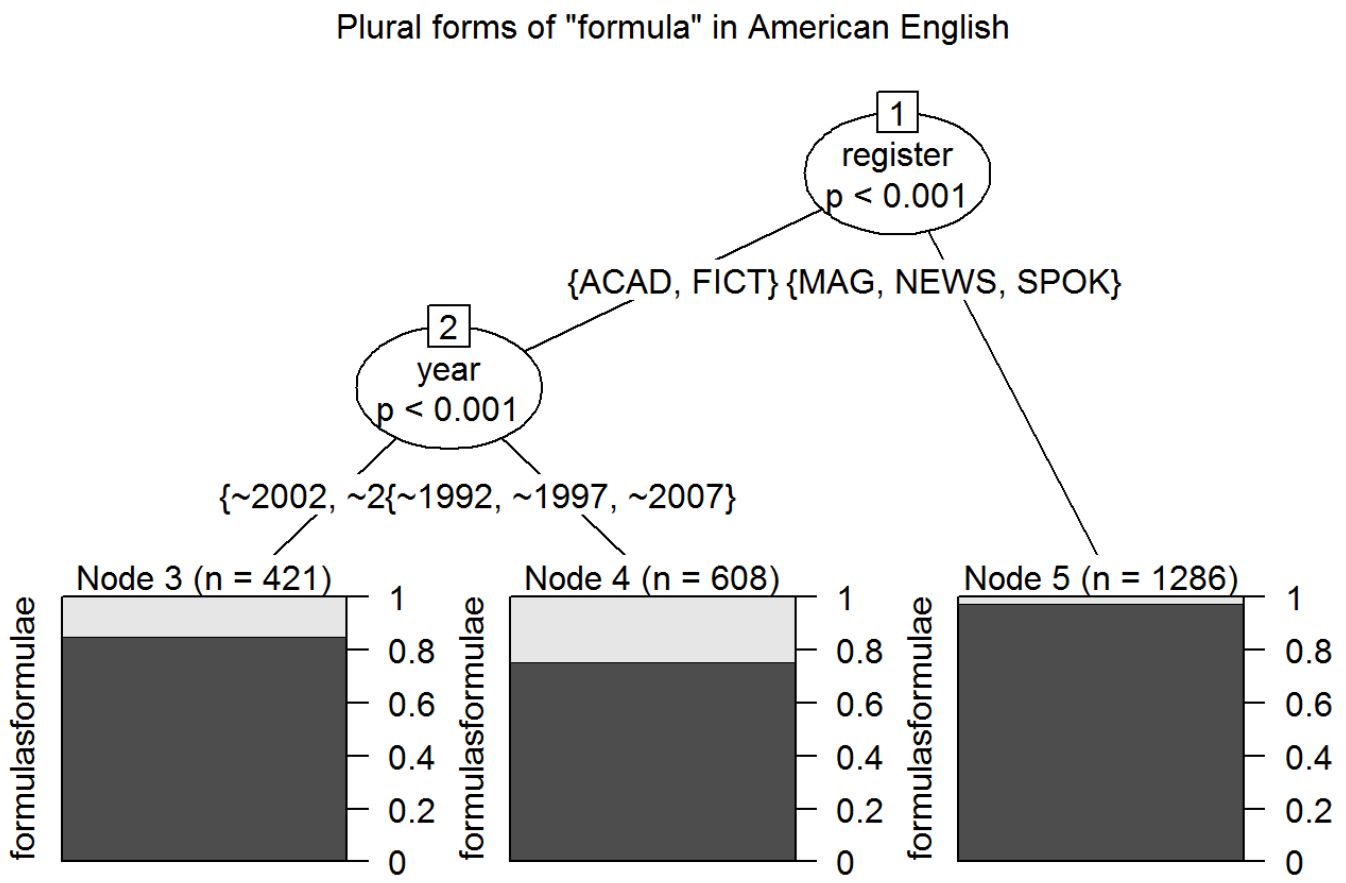


Linguistic evaluation of 'indices'

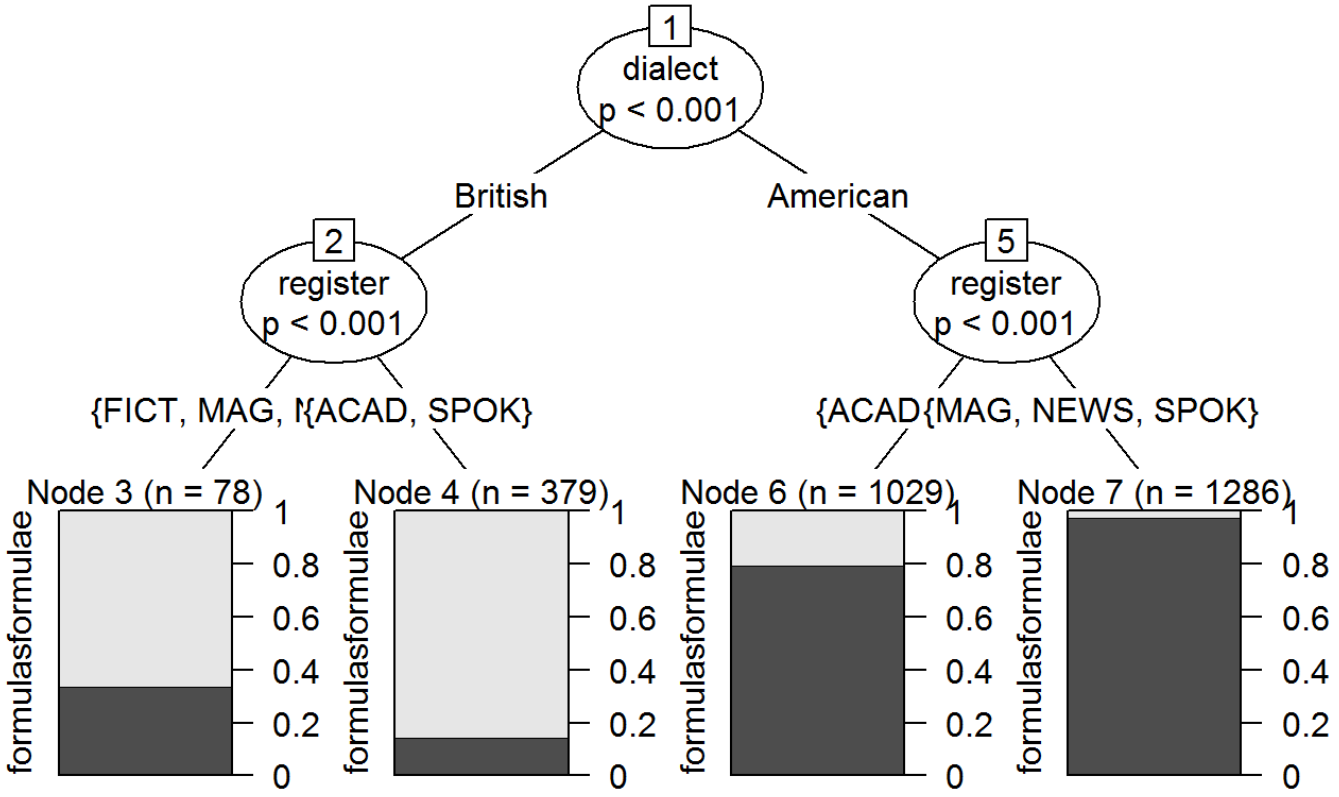
The case of 'indices' strongly resemble that of 'data' as it is a very canonic academic word, and therefore in the academic papers the form 'indices' strongly prevails, while in other registers the form 'indexes' is more popular. The second important observation is that in the British dialect the difference is more noticable - 'indices' is more prefered than 'indexes'.

Variability of 'formulae/formulas' in American English and comparing to British English (Desicion trees, ggplot)

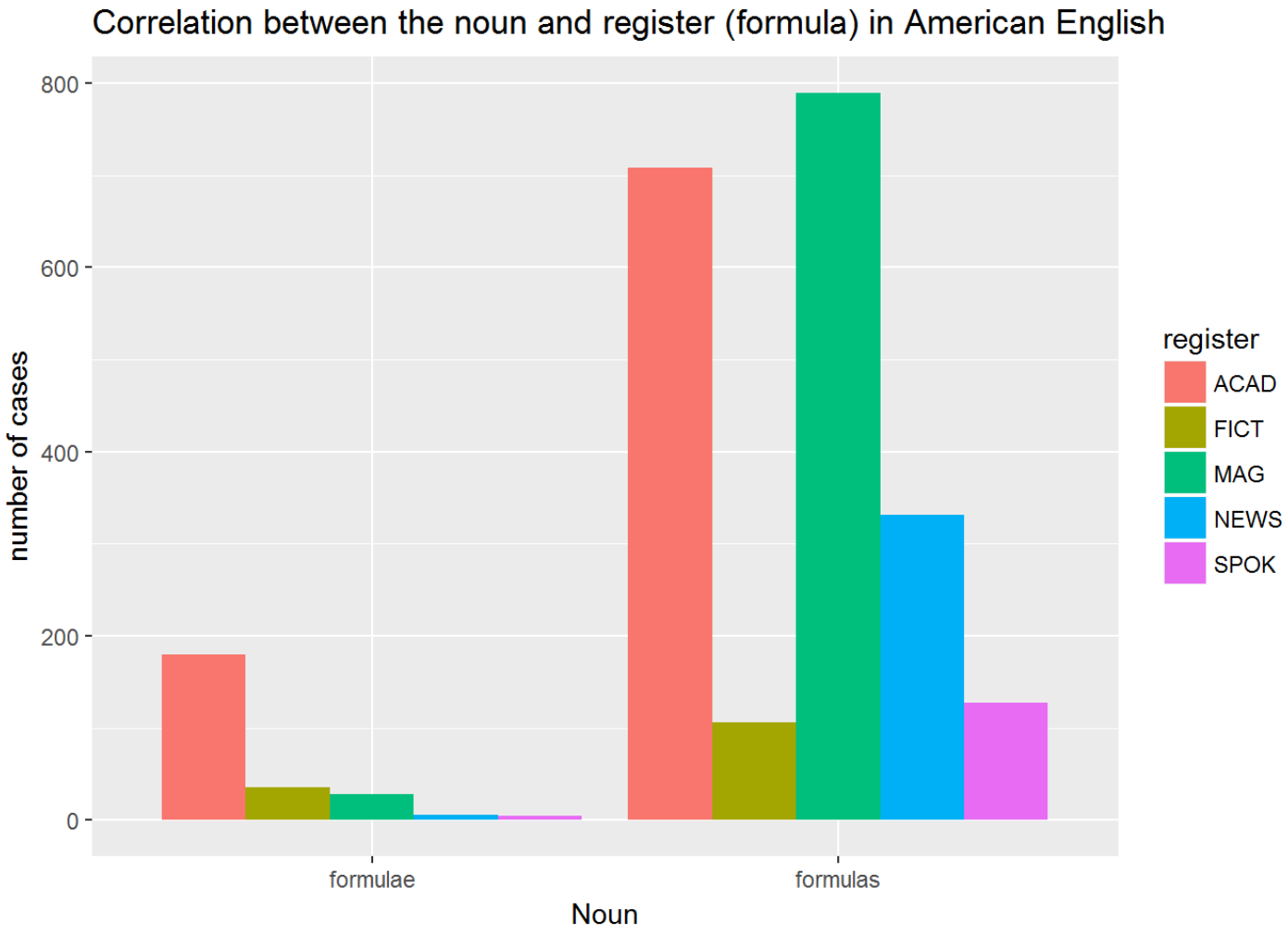
```
formulae_AM <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\formulae_AM.txt', s
ep="\t", header = TRUE)
formulae_AM.ctree = ctree(noun ~ register + year, formulae_AM)
plot(formulae_AM.ctree, main = 'Plural forms of "formula" in American Englis
h')
```



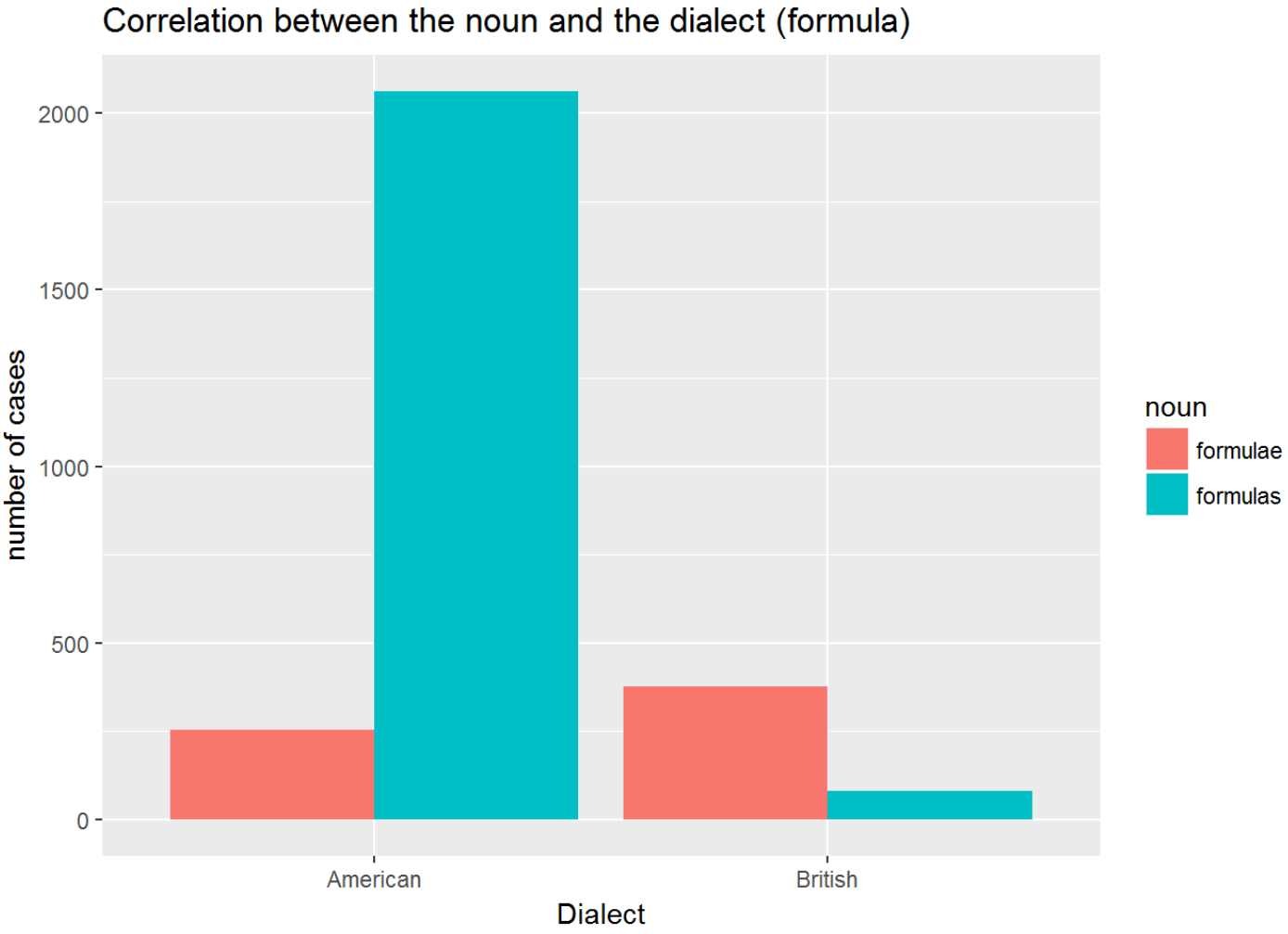

```
formulae_int <- read.csv(file='Z:\\HSE\\R\\Project\\Data\\formulae_int.txt',
sep="\t", header = TRUE)
formulae_int.ctree = ctree(noun ~ register + dialect, formulae_int)
plot(formulae_int.ctree, main = 'Plural forms of "formula" in American and B
ritish dialects')
```




```
ggplot(data=formulae_AM, aes(x=noun, fill=register)) +
  geom_bar(stat="count", position=position_dodge()) +
  xlab("Noun") + ylab("number of cases") +
  ggtitle("Correlation between the noun and register (formula) in American E
nglish")
```




```
ggplot(data=formulae_int, aes(x=dialect, fill=noun)) +
  geom_bar(stat="count", position=position_dodge()) +
  xlab("Dialect") + ylab("number of cases") +
  ggtitle("Correlation between the noun and the dialect (formula)")
```



Linguistic evaluation of 'formulae'

The case of 'formulae' is very important for this work as it is a perfect example of how native speakers percieve words that are very different from most words. So, this type of a plural form with -e at the end is very unusual for the English language, and that is why, even though 'formula' is popular concept for academic papers, even scientists prefer the 'formulas'.

However, this theory is somewhat debatable as in the British dialect the situation is almost contrary. 'Formulae' strongly dominates there over 'formulas'. This might be connected to the overall tendency to choose the traditional form in the British dialect and a smaller corpus.

Conclusion

I believe that the main results of this work are the following:

The register is the main factor of variability of the forms,

The variability changes over time and sometimes drastically,

Dialect is important mainly because in the American dialect people avoid very weird (according to the English language) forms such as ‘cacti’ and ‘formulae’, while in the British dialect these forms strongly prevail over the new-constructed forms such as ‘cactuses’ and ‘formulas’.