Тенденции выбора падежной формы в конструкциях типа Фермер продал три овцы // трех овец.

Ангелина Присяжная

Введение

Вариативность (наличие нескольких вариантов чего-либо) – очень распространенное явление в русском языке. Она имеет место в различных областях языка, в том числе и в морфологии. Интересно изучить то, что именно влияет на выбор того или иного варианта из нескольких возможных, а также выявить тенденции этого выбора.

Для более подробного изучения взяты конструкции типа Фермер продал три овцы // трех овец. Целью данной работы является выявление тенденций выбора падежной формы в этих конструкциях и определение факторов, влияющих на этот выбор.

Для изучения были взяты числительные оба, два, три, четыре. Подбор материала осуществлялся с помощью НКРЯ и корпуса SketchEngine. Я использовала лексико-грамматический поиск и вводили следующий запрос: оба/два/три/четыре (в именительном или винительном падеже) + существительное (женского рода, во множественном числе, одушевленное, семантический класс - животное), расстояние между словоформами от 1 до 3. При таком запросе также находятся и диминутивы существительных.

В данных конструкциях может быть использована одушевленная или неодушевленная форма винительного падежа. Формы обе/две/три/четыре будем называть “неодушевленными”, а формы обеих/двух/трех/четырех – “одушевленными”:

*Завтракал Орлов (деж.). Долго гулял, убил **три вороны**. Занимался и писал, а после обеда читал Аликс вслух.* [Николай II. Дневники 1904-1907 (1904-1907)]

*Гейден. Погулял еще и убил **трех ворон**. Занимался с успехом.* [Николай II. Дневники 1904-1907 (1904-1907)]

Из найденных примеров не рассматривались те, в которых использовались существительные пиявка, креветка (для данных существительных в Грамматическом словаре русского языка А.А.Зализняка указаны колебания по одушевленности):

*А теперь припустил себе к носу **две пиявки** да воображает, что у него усы!* [М. Н. Загоскин. Москва и москвичи (1842-1850)]

*Сергей выбрал еще **три креветки**, стараясь найти среди них самые крепкие и привлекательные, наживил каждую из них, тщательно продев крючок сквозь всетуловище, и осторожно опустил за борт свою снасть.* [Фазиль Искандер. Морской скорпион (1977)]

Необходимые библиотеки.

```
library('languageR')
library('Hmisc')
library(party)
library(lattice)
library(rms)
library(ggplot2)
```

Файлы с данными.

```
nk = read.csv('/Users/angelinaprisyazhnaya/Desktop/ovtsy_ruscorpora.csv', sep=';')
se = read.csv('/Users/angelinaprisyazhnaya/Desktop/ovtsy_sketch.csv', sep=';')
ovtsy_all = read.csv('/Users/angelinaprisyazhnaya/Desktop/ovtsy_all.csv', sep=';')
```

Данные из НКРЯ

В НКРЯ было найдено 729 подходящих примеров. На графике показано количество вхождений для различных форм и числительных.

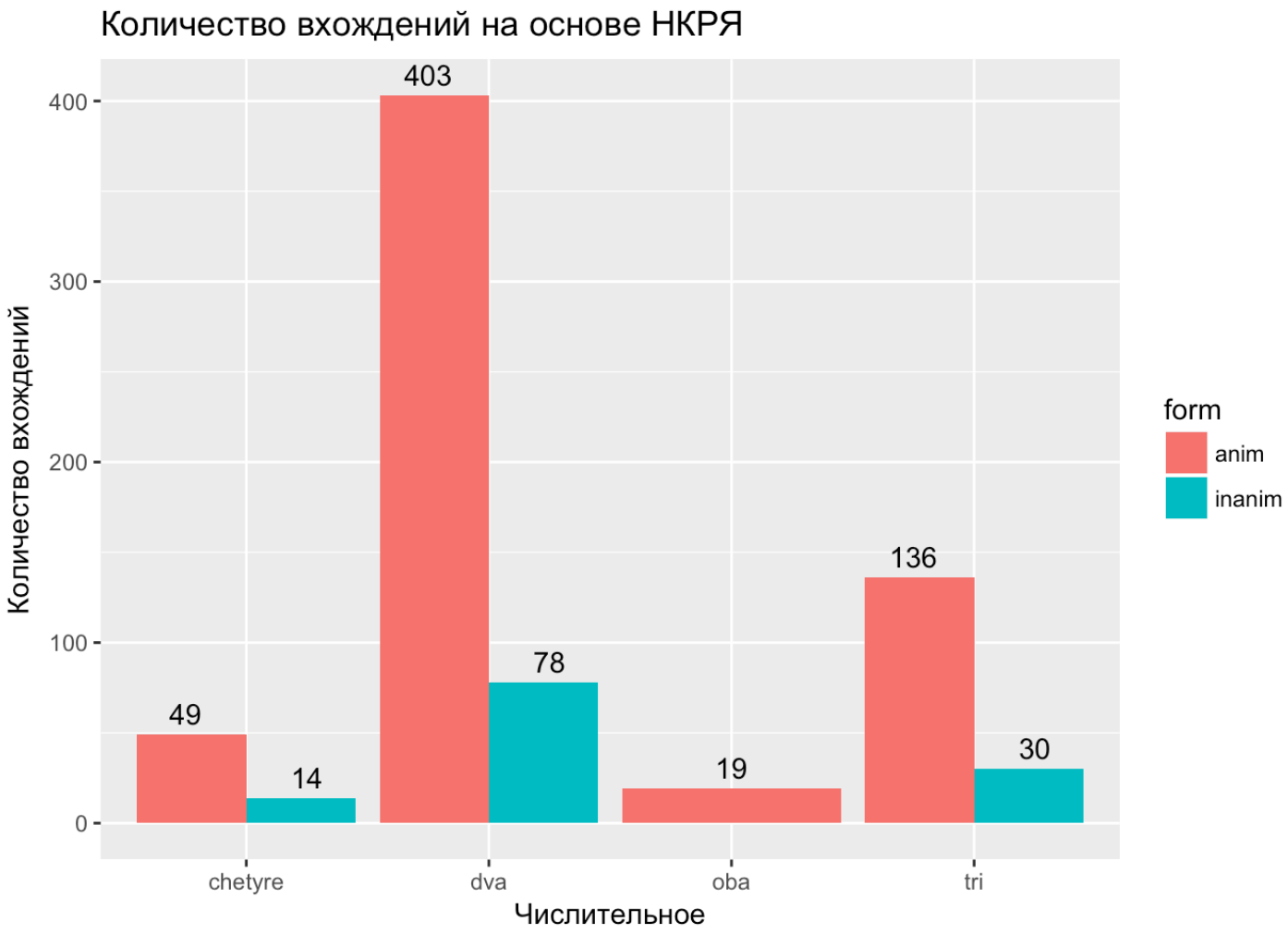
```
summary(nk)
```

```
##     numeral        form          year       century     homogenious_parts
##  chetyre: 63   anim  :607   Min.   :1709   XIX  :197   no         :575
##  dva    :481   inanim:122   1st Qu.:1892   XVIII: 13   yes_anim   :117
##  oba    : 19                Median :1932   XX   :429   yes_inanim: 37
##  tri    :166                Mean   :1929   XXI  : 90
##                             3rd Qu.:1977
##                             Max.   :2011
##  pair_numerals definiteness adjectives
##  no :711       high  :586   no :577
##  yes: 18       low   : 15   yes:152
##                medium:128
##
##
##
```

```
head(nk)
```

```
##   numeral   form year century homogenious_parts pair_numerals definiteness
## 1     dva inanim 2011     XXI                no            no          high
## 2     dva inanim 2003     XXI                no           yes           low
## 3     dva inanim 2003     XXI                no            no        medium
## 4     dva inanim 2001     XXI                no            no          high
## 5     dva inanim 2001     XXI          yes_anim            no        medium
## 6     dva inanim 2001     XXI                no           yes           low
##   adjectives
## 1         no
## 2        yes
## 3         no
## 4         no
## 5        yes
## 6         no
```

```
ggplot(nk, aes(numeral)) + geom_bar(aes(fill = form), position="dodge") + geom_tex
t(stat='count',aes(label=..count.., hjust=0.5, vjust=-0.5, group=form), position=p
osition_dodge(width = 1)) + xlab("Числительное") + ylab("Количество вхождений") +
ggtitle("Количество вхождений на основе НКРЯ")
```



Данные из SketchEngine

В SketchEngine было найдено 784 подходящих примера. На графике показано количество вхождений для различных форм и числительных.

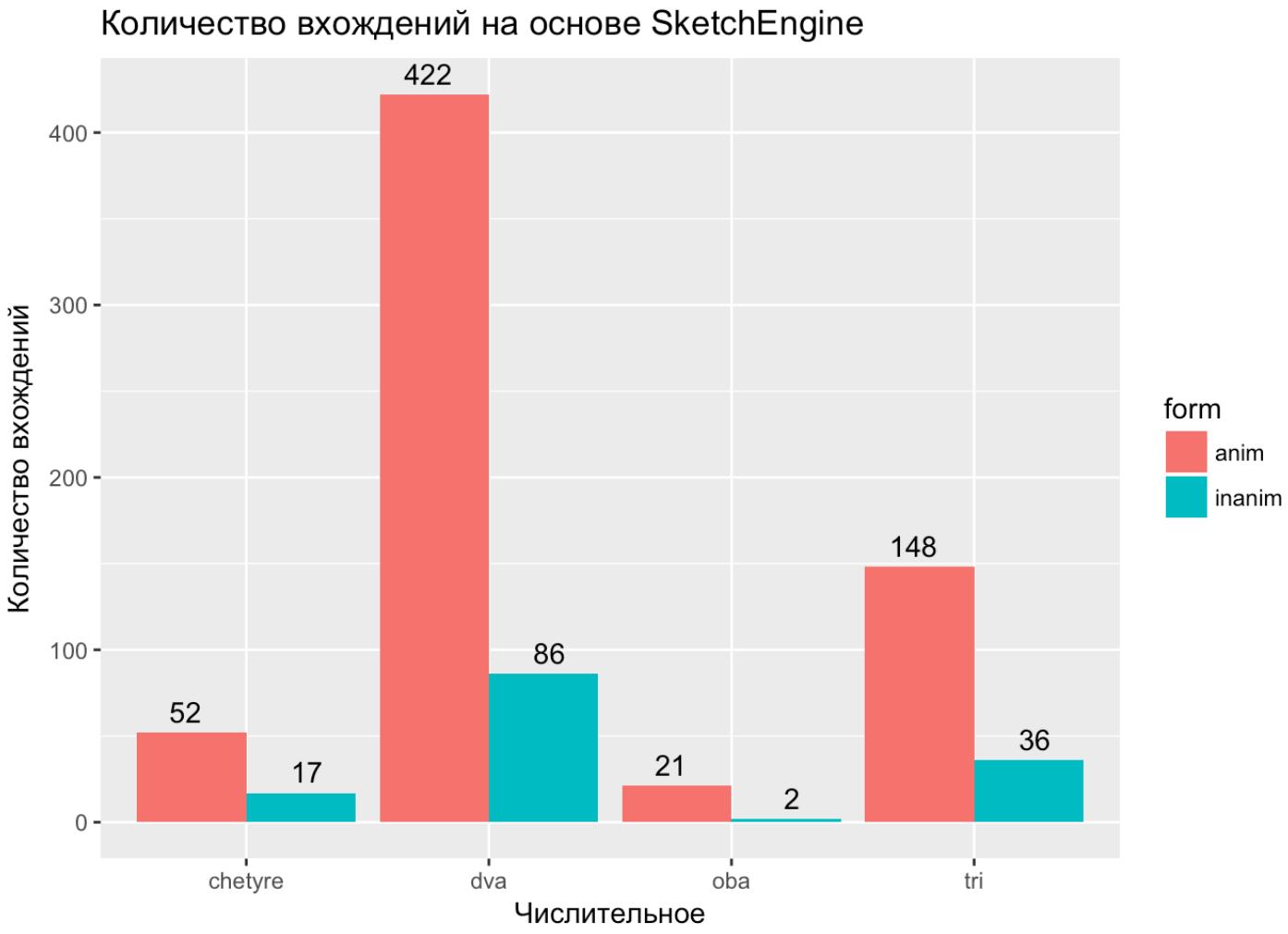
```
summary(se)
```

```
##     numeral        form      homogenious_parts pair_numerals definiteness
##  chetyre: 69   anim  :643   no        :669     no :720       high  :528
##  dva    :508   inanim:141   yes_anim  : 86     yes: 64       low   : 64
##  oba    : 23                yes_inanim: 29                   medium:192
##  tri    :184
##  adjectives
##  no :633
##  yes:151
##
##
```

```
head(se)
```

```
##   numeral   form homogenious_parts pair_numerals definiteness adjectives
## 1     dva inanim               no            no         high         no
## 2     dva inanim               no            no         high         no
## 3     dva inanim               no            no         high        yes
## 4     dva inanim               no            no         high         no
## 5     dva inanim               no            no         high        yes
## 6     dva inanim               no            no         high         no
```

```
ggplot(se, aes(numeral)) + geom_bar(aes(fill = form), position="dodge") + geom_tex
t(stat='count',aes(label=..count.., hjust=0.5, vjust=-0.5, group=form), position=p
osition_dodge(width = 1)) + xlab("Числительное") + ylab("Количество вхождений") +
ggtitle("Количество вхождений на основе SketchEngine")
```



Факторы

Я предположила, что выбор формы может зависеть от следующих факторов:

- наличие однородных членов;
- двойные числительные;
- определенность;
- наличие определений;
- дата.

Наличие однородных членов

Возможны три варианта для данного фактора:

- есть одушевленный однородный член (yes_anim)

  Сии последние стреляли по двум полкам, оставшимся верными королю, и убили до смерти шестерых всадников и двух лошадей. [Журнал событиям, совершившимся в Париже в 11-го по 17-е июля 1789 года (1789)]
- есть неодушевленный однородный член (yes_inanim)

  Один из грабителей убит, двое скрылись, оставив двух лошадей и ценные вещи. [неизвестный. Вести (1911.02.13) // «Новое время», 1911]
- нет однородных членов (no)

Гипотеза: если в предложении содержатся однородные члены в “неодушевленной” форме, то название животного подвергается их влиянию и употребляется в “неодушевленной” форме (и наоборот).

Двойные числительные

Возможны два варианта для данного фактора:

- двойное числительное (yes)
  Лишь некоторые зажиточные вогуличи держали при юртах одну-две коровы. [Финно-угорские народы. Манси (2001) // «Жизнь национальностей», 2001.12.28]
- обычное одиночное числительное (no)

Гипотеза: двойные числительные чаще употребляются в “неодушевленной” форме, чем в “одушевленной”.

Определенность

Возможны три варианта для данного фактора:

- Речь о конкретных объектах, их количество точно определено (high)
  Привязав обеих лошадей к прутьям ограды, монах вошел в палисадник. [Роберт Штильмарк. Наследник из Калькутты (1950-1951)]
- Речь не о конкретных объектах, но их количество точно определено (medium)
  На эти деньги мы должны приобрести четырех собак. [Марта Баранова, Евгений Велтистов. Тяпа, Борька и ракета (1962)]
- Речь не о конкретных объектах, их количество точно не определено (low)
  И не зря: как ни пойдешь, всегда тут две-три гадюки увидишь. [В. В. Бианки. Лесные были и небылицы (1923-1958)]

Гипотеза: с повышением определенности чаще употребляются “одушевленные” формы.

Наличие определений

Возможны два варианта для данного фактора:

- есть определение (yes)
  Нюра еще издали окинула взглядом путик и поняла, что рыжая лесовая собака не успела пробежаться по нему и напроказить, может, погрызла двух-трех тундровых птиц, никак не больше. [Владимир Личутин. Вдова Нюра (1973)]
- нет определения (no)

> Гипотеза: при наличии определений чаще употребляются “одушевленные” формы.

Дата

Возможные варианты для данного фактора: XVIII, XIX, XX, XXI века.

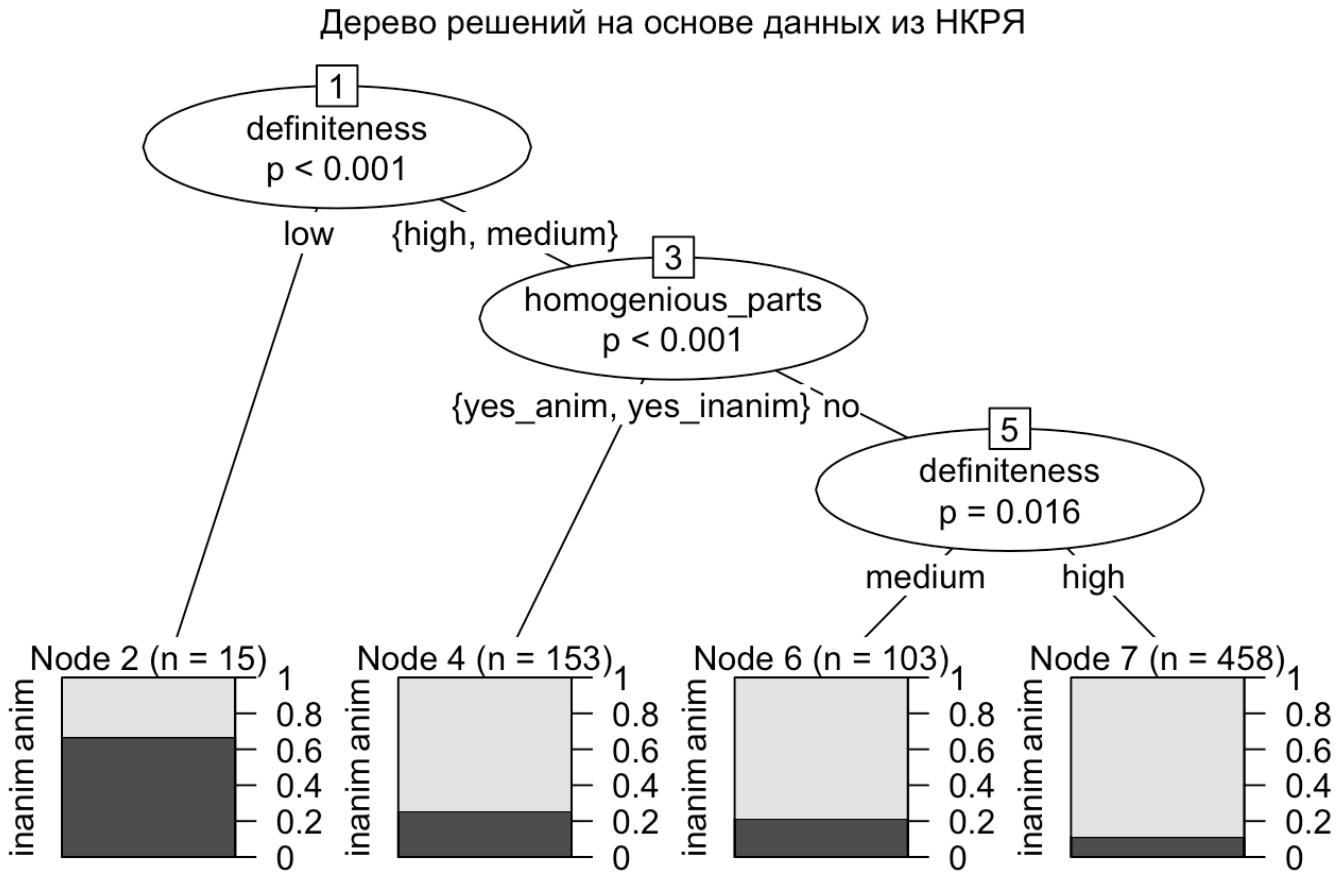
> Цель - проверить, есть ли временная тенденция.

Дерево решений

Для того, чтобы выяснить, зависит ли выбор падежной формы от каких-либо рассматриваемых факторов, я решила использовать дерево решений.

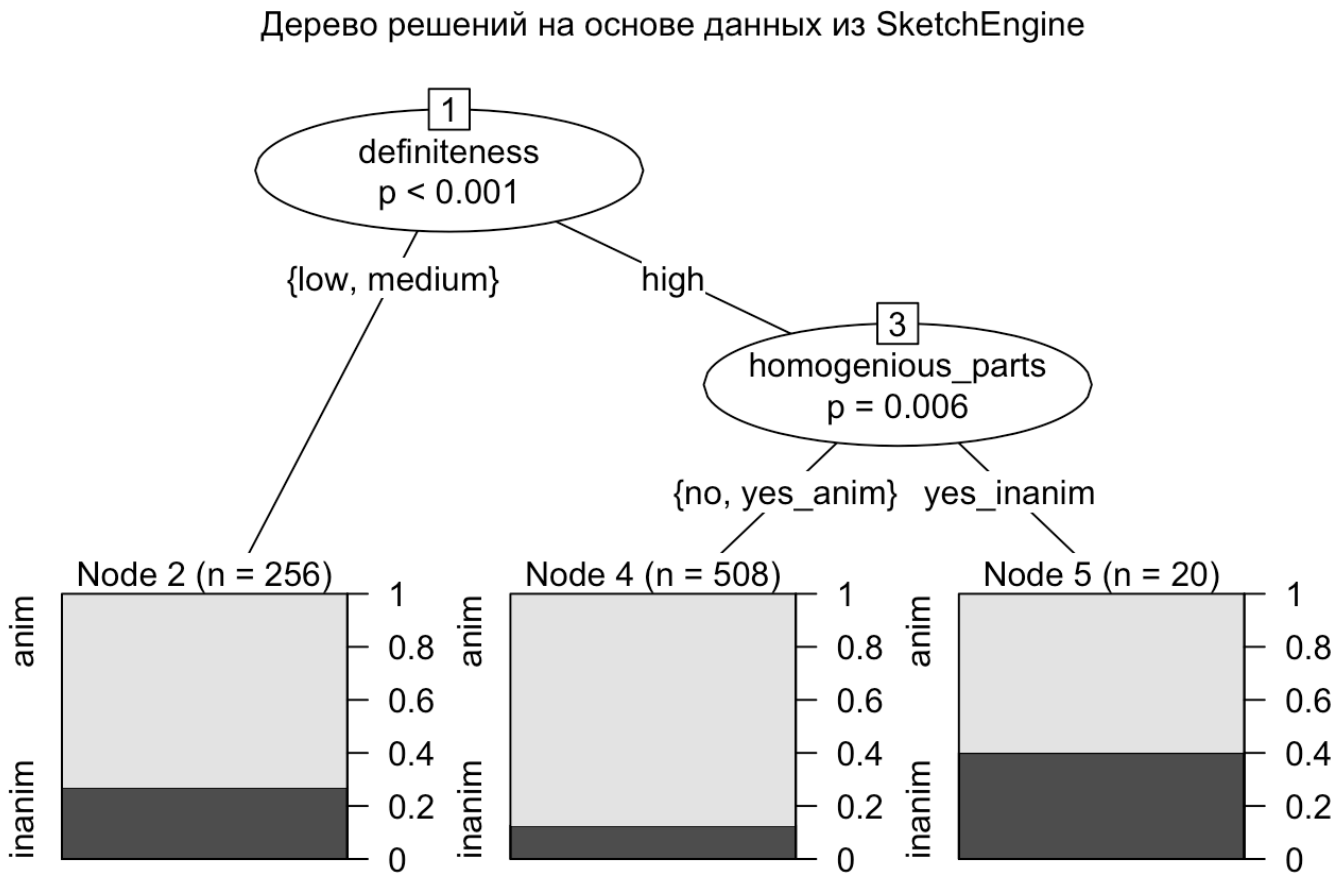
Дерево решений на основе данных из НКРЯ

```
nk.ctree=ctree(form ~ definiteness + homogenious_parts + adjectives, nk)
plot(nk.ctree, main="Дерево решений на основе данных из НКРЯ")
```



Дерево решений на основе данных из SketchEngine

```
se.ctree=ctree(form ~ definiteness + homogenious_parts + adjectives, se)
plot(se.ctree, main="Дерево решений на основе данных из SketchEngine")
```



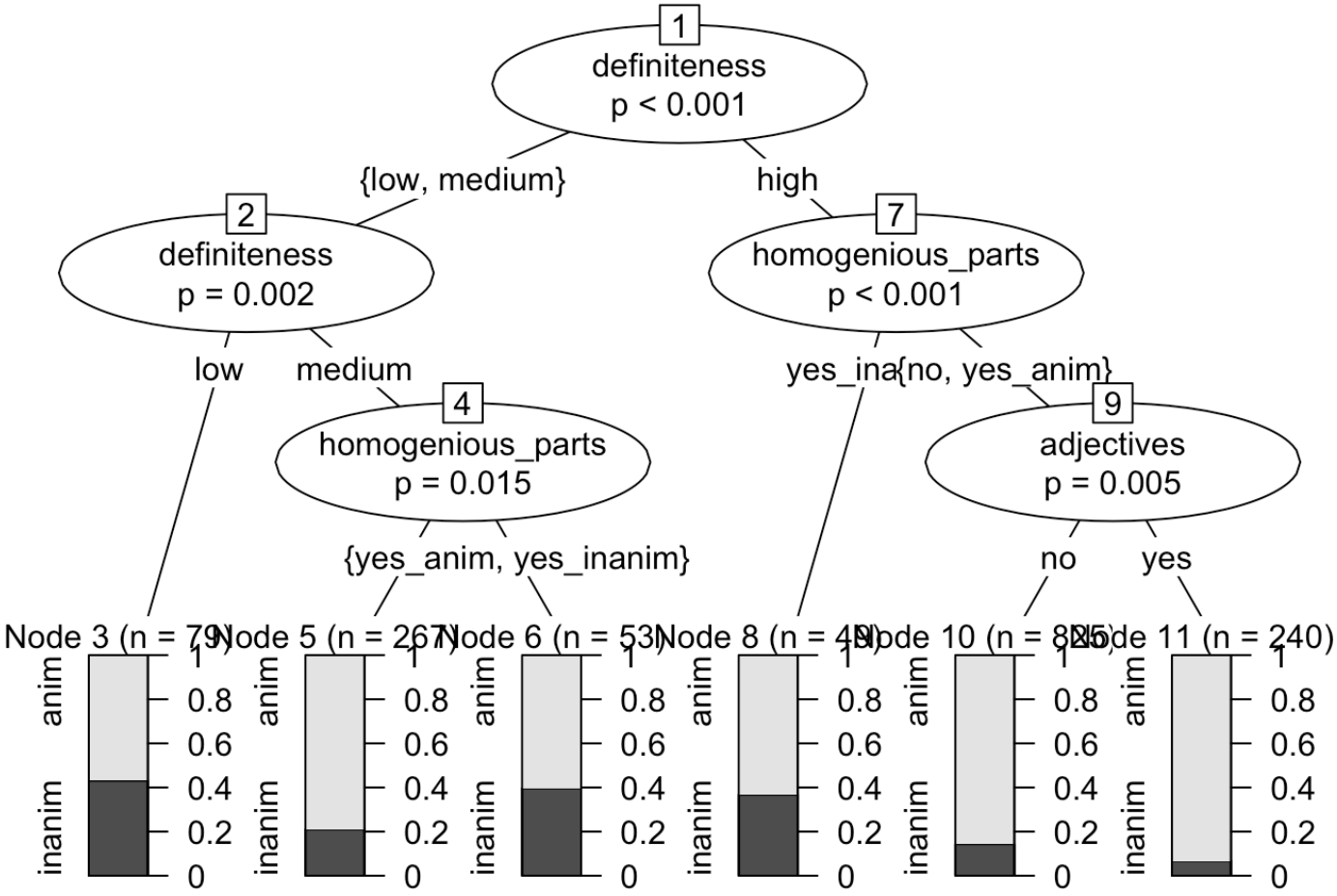
Результаты показывают, что на выбор падежной формы действительно влияют следующие факторы:

- определенность (при низкой определенности чаще употребляются “неодушевленные” формы);
- наличие однородных членов (при наличии однородных членов в “неодушевленной” форме чаще употребляются “неодушевленные” формы).

Дерево решений на основе всех данных

Также я построила дерево для агрегированных данных - и из НКРЯ, и из SketchEngine. Но в этом случае анализ неточный, поскольку некоторые примеры могут встречаться в обеих выборках (дублироваться) - и в НКРЯ, и в SketchEngine. При этом, если это пересечение есть, то оно неравномерно, так как SketchEngine содержит только современные тексты.

```
ovtsy_all.ctree=ctree(form ~ definiteness + homogenious_parts + adjectives, ovtsy_
all)
plot(ovtsy_all.ctree)
```



Логистическая регрессия на основе данных из НКРЯ

```
nk.lrm=lrm(form ~ definiteness + homogenious_parts + adjectives, data=nk, x=T, y=T
, linear.predictors=T)
nk.lrm
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ definiteness + homogenious_parts + adjectives,
##      data = nk, x = T, y = T, linear.predictors = T)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
##  Obs            729    LR chi2      49.63    R2        0.111    C         0.675
##   anim          607    d.f.             5    g         0.657    Dxy       0.351
##   inanim        122    Pr(> chi2) <0.0001    gr        1.929    gamma     0.456
##  max |deriv| 8e-09                          gp        0.095    tau-a     0.098
##                                             Brier     0.128
##
##                               Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept                    -1.9518 0.1511 -12.91 <0.0001
##  definiteness=low              2.7519 0.5721   4.81 <0.0001
##  definiteness=medium           0.7344 0.2424   3.03 0.0025
##  homogenious_parts=yes_anim    0.6643 0.2580   2.57 0.0100
##  homogenious_parts=yes_inanim  1.3023 0.3731   3.49 0.0005
##  adjectives=yes               -0.6693 0.3051  -2.19 0.0283
##
```

Определенность и наличие однородных членов оказывают значимое влияние на выбор падежной формы.

Логистическая регрессия на основе данных из SketchEngine

```
se.lrm=lrm(form ~ definiteness + homogenious_parts + adjectives, data=se, x=T, y=T
, linear.predictors=T)
se.lrm
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ definiteness + homogenious_parts + adjectives,
##      data = se, x = T, y = T, linear.predictors = T)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
##  Obs            784    LR chi2      41.02    R2        0.084    C        0.651
##   anim          643    d.f.             5    g         0.543    Dxy      0.302
##   inanim        141    Pr(> chi2) <0.0001    gr        1.721    gamma    0.380
##  max |deriv| 1e-09                          gp        0.087    tau-a    0.089
##                                             Brier     0.139
##
##                               Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept                    -1.9205 0.1490 -12.89 <0.0001
##  definiteness=low              1.3780 0.2921   4.72 <0.0001
##  definiteness=medium           0.6691 0.2165   3.09 0.0020
##  homogenious_parts=yes_anim    0.1673 0.3029   0.55 0.5807
##  homogenious_parts=yes_inanim  1.6137 0.3962   4.07 <0.0001
##  adjectives=yes               -0.2155 0.2655  -0.81 0.4169
##
```

Результаты очень близки к результатам на основе НКРЯ.

Логистическая регрессия на основе всех данных

Регрессию я также применила и к агрегированным данным (но здесь та же проблема - данные могут дублироваться).

```
ovtsy_all.lrm=lrm(form ~ definiteness + homogenious_parts + adjectives, data=ovtsy
_all, x=T, y=T, linear.predictors=T)
ovtsy_all.lrm
```

```
## Logistic Regression Model
##
##  lrm(formula = form ~ definiteness + homogenious_parts + adjectives,
##      data = ovtsy_all, x = T, y = T, linear.predictors = T)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
##  Obs           1513    LR chi2      82.58    R2        0.088    C        0.659
##   anim         1250    d.f.             5    g         0.585    Dxy      0.319
##   inanim        263    Pr(> chi2) <0.0001    gr        1.795    gamma    0.405
##  max |deriv| 8e-09                          gp        0.090    tau-a    0.092
##                                             Brier     0.135
##
##                                Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept                     -1.9298 0.1056 -18.27 <0.0001
##  definiteness=low               1.6473 0.2475   6.66 <0.0001
##  definiteness=medium            0.6915 0.1603   4.31 <0.0001
##  homogenious_parts=yes_anim     0.4403 0.1940   2.27 0.0233
##  homogenious_parts=yes_inanim   1.4107 0.2685   5.25 <0.0001
##  adjectives=yes                -0.4187 0.1986  -2.11 0.0350
##
```

Временная тенденция

Отдельно от всех рассмотренных факторов рассмотрим дату, чтобы выяснить, существует ли какая-либо зависимость употребления падежной формы от времени. В данном случае будут рассматриваться только данные НКРЯ, так как в датасете из SketchEngine нет информации о годе написания текста.

Временная тенденция (на основе данных из НКРЯ)

```
ggplot(nk, aes(form, year)) + geom_violin(scale = "count", draw_quantiles = c(0.25
, 0.5, 0.75)) + xlab("Форма") + ylab("Год написания текста") + ggtitle("Временная
тендеция выбора падежной формы")
```

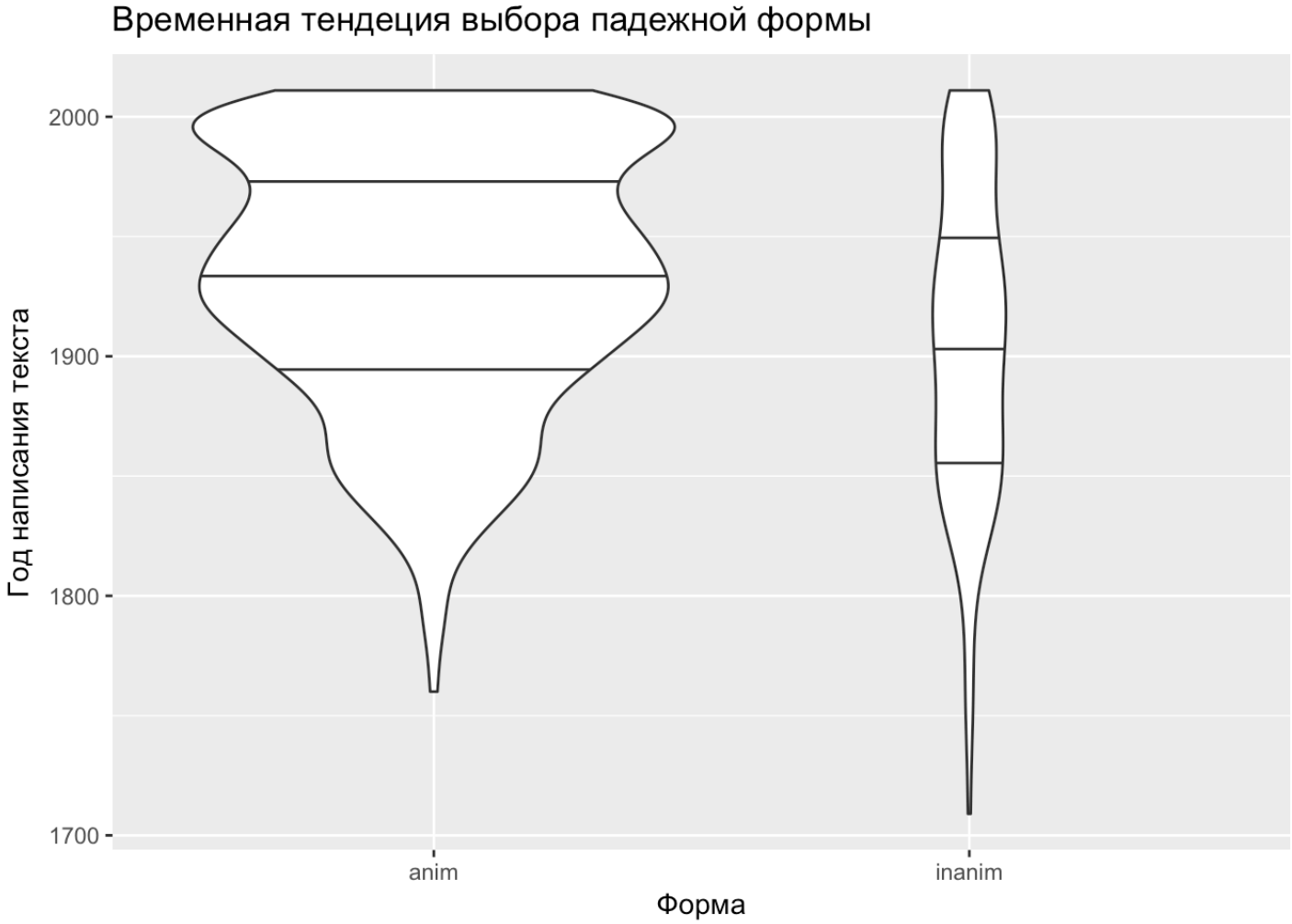
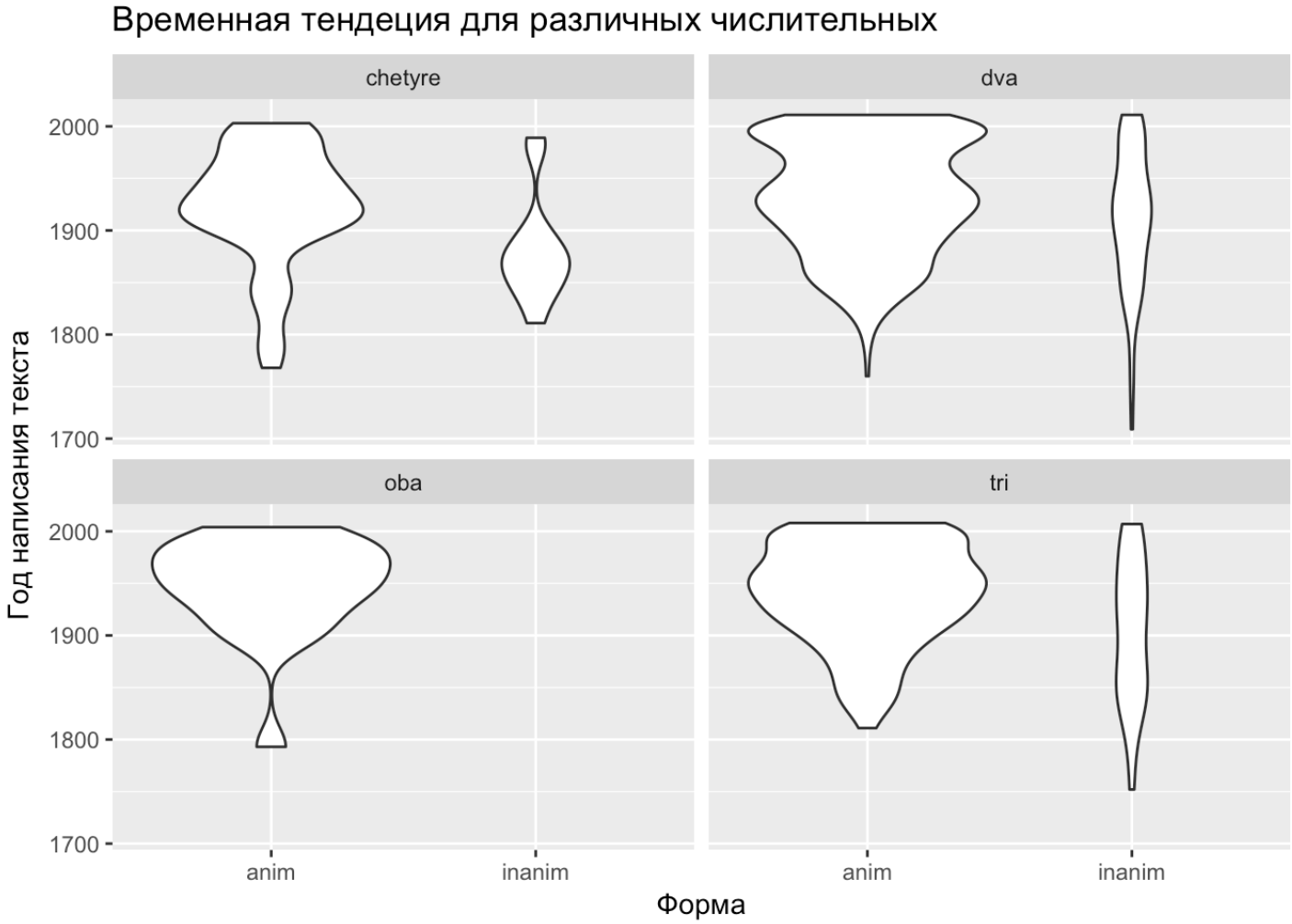


График показывает, что “неодушевленные” формы употребляются на всем временном промежутке. “Одушевленные” формы начинают употребляться значительно позже.

Временная тенденция для разных числительных (на основе данных из НКРЯ)

Рассмотрим также зависимость от времени для различных числительных.

```
ggplot(nk, aes(form, year)) + geom_violin(scale = "count") + xlab("Форма") + ylab(
"Год написания текста") + ggtitle("Временная тендеция для различных числительных")
+ facet_wrap(~numeral)
```



Для числительного оба в НКРЯ вообще не нашлось неодушевленных форм. Для всех остальных числительных видна четкая временная тенденция - чаще начинают употребляться одушевленные формы.

Заключение

Удалось выяснить, что:

- с течением времени количество употреблений “одушевленных” форм возрастает, а “неодушевленных” – убывает;
- если однородные члены в “неодушевленной” форме, то исследуемая конструкция подвергается их влиянию (чаще используется в “неодушевленной” форме);
- с возрастанием определенности чаще употребляются “одушевленные” формы.