Анализ белорусского диалекта в русском языке

Цель: проанализировать признаки встречаемых белорусских диалектизмов в русском языке.

Для данного проекта был выбран набор данных, содержащий транскрипты бесед с жителями пгт Хиславичи, Смоленская область. Данные разделены на токены, присутствует POS-разметка conllu, но нет помет относительно диалекта. В связи с этим нами была проведена полуручная разметка диалектизмов в датасете. Также было добавлено поле `time_dif`, которое представляет собой разницу между `time_end` и `time_start`, т.е. длительность фразы.

Обзор данных и дополнительная предобработка

```
library("tidyverse")
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
df <- read_csv("khislavichi_marked.csv")
```

```
## Rows: 294018 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (11): corpus, tier_name, source, sentence, token, lemma, upos, xpos, fea...
## dbl  (8): time_start, time_end, id, term_id, token_id, head_token_id, is_dia...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df
```

```
## # A tibble: 294,018 × 19
##    corpus   tier_name time_start time_end source    id sentence term_id token_id
##    <chr>    <chr>          <dbl>    <dbl> <chr>  <dbl> <chr>      <dbl>    <dbl>
##  1 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       1        1
##  2 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       2        2
##  3 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       3        3
##  4 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       4        4
##  5 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       5        5
##  6 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       6        6
##  7 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       7        7
##  8 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       8        8
##  9 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…       9        9
## 10 data_di… zvp1934        0.073     4.30 2018_…     1 Это вон…      10       10
## # ℹ 294,008 more rows
## # ℹ 10 more variables: token <chr>, lemma <chr>, upos <chr>, xpos <chr>,
## #   feats <chr>, head_token_id <dbl>, dep_rel <chr>, misc <chr>,
## #   is_dialect <dbl>, time_dif <dbl>
```

```
str(df)
```

```
## spc_tbl_ [294,018 × 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ corpus       : chr [1:294018] "data_dialect_khislavichi" "data_dialect_khislavichi" "da
ta_dialect_khislavichi" "data_dialect_khislavichi" ...
##  $ tier_name    : chr [1:294018] "zvp1934" "zvp1934" "zvp1934" "zvp1934" ...
##  $ time_start   : num [1:294018] 0.073 0.073 0.073 0.073 0.073 0.073 0.073 0.073 0.073 0.0
73 ...
##  $ time_end     : num [1:294018] 4.3 4.3 4.3 4.3 4.3 ...
##  $ source       : chr [1:294018] "2018_Malye_Xutora_zvp1934_1_1.eaf" "2018_Malye_Xutora_zv
p1934_1_1.eaf" "2018_Malye_Xutora_zvp1934_1_1.eaf" "2018_Malye_Xutora_zvp1934_1_1.eaf" ...
##  $ id           : num [1:294018] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sentence     : chr [1:294018] "Это вон, скамеечку берите, тут (тута) я воды наносила."
"Это вон, скамеечку берите, тут (тута) я воды наносила." "Это вон, скамеечку берите, тут (тут
а) я воды наносила." "Это вон, скамеечку берите, тут (тута) я воды наносила." ...
##  $ term_id      : num [1:294018] 1 2 3 4 5 6 7 8 9 10 ...
##  $ token_id     : num [1:294018] 1 2 3 4 5 6 7 8 9 10 ...
##  $ token        : chr [1:294018] "это" "вон" "," "скамеечку" ...
##  $ lemma        : chr [1:294018] "этот" "вон" "," "скамеечка" ...
##  $ upos         : chr [1:294018] "PRON" "NOUN" "PUNCT" "NOUN" ...
##  $ xpos         : chr [1:294018] "DT" "NN" "," "NN" ...
##  $ feats        : chr [1:294018] "Animacy=Inan|Case=Nom|Gender=Neut|Number=Sing" "Animacy=
Anim|Case=Nom|Gender=Masc|Number=Sing" NA "Animacy=Inan|Case=Dat|Gender=Masc|Number=Sing" ...
##  $ head_token_id: num [1:294018] 2 0 4 2 4 7 2 9 7 9 ...
##  $ dep_rel      : chr [1:294018] "nsubj" "root" "punct" "conj" ...
##  $ misc         : chr [1:294018] NA "SpaceAfter=No" NA NA ...
##  $ is_dialect   : num [1:294018] 0 0 0 0 0 0 1 0 1 0 ...
##  $ time_dif     : num [1:294018] 4.23 4.23 4.23 4.23 4.23 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   corpus = col_character(),
##   ..   tier_name = col_character(),
##   ..   time_start = col_double(),
##   ..   time_end = col_double(),
##   ..   source = col_character(),
##   ..   id = col_double(),
##   ..   sentence = col_character(),
##   ..   term_id = col_double(),
##   ..   token_id = col_double(),
##   ..   token = col_character(),
##   ..   lemma = col_character(),
##   ..   upos = col_character(),
##   ..   xpos = col_character(),
##   ..   feats = col_character(),
##   ..   head_token_id = col_double(),
##   ..   dep_rel = col_character(),
##   ..   misc = col_character(),
##   ..   is_dialect = col_double(),
##   ..   time_dif = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(df)
```

```
##     corpus           tier_name           time_start        time_end
##  Length:294018      Length:294018      Min.   :   0.0   Min.   :   0.77
##  Class :character   Class :character   1st Qu.: 281.2   1st Qu.: 284.47
##  Mode  :character   Mode  :character   Median : 567.5   Median : 571.04
##                                        Mean   : 615.6   Mean   : 619.10
##                                        3rd Qu.: 902.5   3rd Qu.: 906.28
##                                        Max.   :2202.1   Max.   :2203.16
##     source                id          sentence           term_id
##  Length:294018      Min.   :  1.0   Length:294018      Min.   : 1.000
##  Class :character   1st Qu.: 64.0   Class :character   1st Qu.: 3.000
##  Mode  :character   Median :136.0   Mode  :character   Median : 6.000
##                     Mean   :154.4                      Mean   : 7.512
##                     3rd Qu.:228.0                      3rd Qu.:10.000
##                     Max.   :715.0                      Max.   :60.000
##     token_id         token              lemma              upos
##  Min.   : 1.000   Length:294018      Length:294018      Length:294018
##  1st Qu.: 3.000   Class :character   Class :character   Class :character
##  Median : 5.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 6.723
##  3rd Qu.: 9.000
##  Max.   :60.000
##      xpos               feats           head_token_id      dep_rel
##  Length:294018      Length:294018      Min.   : 0.000   Length:294018
##  Class :character   Class :character   1st Qu.: 2.000   Class :character
##  Mode  :character   Mode  :character   Median : 5.000   Mode  :character
##                                        Mean   : 6.079
##                                        3rd Qu.: 9.000
##                                        Max.   :58.000
##      misc             is_dialect        time_dif
##  Length:294018      Min.   :0.0000   Min.   : 0.006
##  Class :character   1st Qu.:0.0000   1st Qu.: 2.295
##  Mode  :character   Median :0.0000   Median : 3.312
##                     Mean   :0.1445   Mean   : 3.502
##                     3rd Qu.:0.0000   3rd Qu.: 4.481
##                     Max.   :1.0000   Max.   :22.931
```

Поле `is_dialect` содержит пометы о том, является ли данная лемма диалектизмом: 0 - нет, не является; 1 - да, является.

```
df %>%
  distinct(is_dialect)
```

```
## # A tibble: 2 × 1
##   is_dialect
##        <dbl>
## 1          0
## 2          1
```

Посмотрим на частеречную разметку:

```
df %>%
  distinct(upos)
```

```
## # A tibble: 16 × 1
##    upos
##    <chr>
##  1 PRON
##  2 NOUN
##  3 PUNCT
##  4 VERB
##  5 ADP
##  6 PROPN
##  7 ADJ
##  8 AUX
##  9 PART
## 10 ADV
## 11 SCONJ
## 12 CCONJ
## 13 DET
## 14 SYM
## 15 NUM
## 16 X
```

```
df %>%
  filter(upos == "X") %>%
  select(token, is_dialect)
```

```
## # A tibble: 181 × 2
##    token    is_dialect
##    <chr>         <dbl>
##  1 вуругнул          0
##  2 а                 0
##  3 всю               0
##  4 дай               0
##  5 тэй               1
##  6 давай             0
##  7 тэй               1
##  8 тэй               1
##  9 нехай             1
## 10 да.               0
## # ℹ 171 more rows
```

```
df %>%
  filter(upos == "SYM") %>%
  select(token, is_dialect)
```

```
## # A tibble: 2,669 × 2
##    token   is_dialect
##    <chr>        <dbl>
##  1 эт=              0
##  2 восемь=          0
##  3 выра=            0
##  4 ве=              0
##  5 пя=              0
##  6 не=              0
##  7 пе=              0
##  8 сме=             0
##  9 ста=             0
## 10 стар=            0
## # ℹ 2,659 more rows
```

```
df %>%
  filter(upos == "SYM" & is_dialect != 0) %>%
  select(token, is_dialect)
```

```
## # A tibble: 0 × 2
## # ℹ 2 variables: token <chr>, is_dialect <dbl>
```

Избавляемся от знаков препинания в датасете, а также от токенов с пометой `SYM` , поскольку они представляют собой “обрывочные” токены. Помимо этого, оставляем только те колонки, которые могут быть нам полезны в дальнейшем.

```
df = df %>%
  filter(upos != "PUNCT" & upos != "SYM") %>%
  select(sentence, token, lemma, upos, xpos, feats, head_token_id, dep_rel, misc, is_dialect,
time_dif)
df
```

```
## # A tibble: 202,226 × 11
##    sentence token lemma upos  xpos  feats head_token_id dep_rel misc  is_dialect
##    <chr>    <chr> <chr> <chr> <chr> <chr>         <dbl> <chr>   <chr>      <dbl>
##  1 Это вон… это   этот  PRON  DT    Anim…             2 nsubj   <NA>           0
##  2 Это вон… вон   вон   NOUN  NN    Anim…             0 root    Spac…          0
##  3 Это вон… скам… скам… NOUN  NN    Anim…             2 conj    <NA>           0
##  4 Это вон… бери… берит NOUN  NN    Anim…             4 nmod    Spac…          0
##  5 Это вон… тут   тут   NOUN  NN    Anim…             2 conj    <NA>           1
##  6 Это вон… тута  тут   NOUN  NN    Anim…             7 appos   Spac…          1
##  7 Это вон… я     я     PRON  PRP   Case…            13 nsubj   <NA>           0
##  8 Это вон… воды  воды  NOUN  NN    Anim…            13 nsubj   <NA>           0
##  9 Это вон… нано… нано… VERB  VBC   Aspe…             2 parata… Spac…          0
## 10 С капел… с     с     ADP   IN    <NA>              2 case    <NA>           1
## # ℹ 202,216 more rows
## # ℹ 1 more variable: time_dif <dbl>
```

```
df <- mutate(df, across(where(is.character), ~factor(.)))
str(df)
```

```
## tibble [202,226 × 11] (S3: tbl_df/tbl/data.frame)
##  $ sentence     : Factor w/ 30873 levels "- говорю, я уже не могу убираться, детки, дак вы
уже, говорю, на крыльце постойте, не пустила их сюда.",..: 29257 29257 29257 29257 29257 2925
7 29257 29257 29257 23964 ...
##  $ token        : Factor w/ 20417 levels "-а","-ай","-ай-ай",..: 20210 1834 16220 600 1843
9 18440 20268 1706 9156 15512 ...
##  $ lemma        : Factor w/ 15733 levels "-","-ай","-ай-ай",..: 15590 1470 12459 493 14207
14207 15627 1359 6994 11926 ...
##  $ upos         : Factor w/ 14 levels "ADJ","ADP","ADV",..: 10 7 7 7 7 7 10 7 13 2 ...
##  $ xpos         : Factor w/ 34 levels "!","AFX","AWP",..: 6 15 15 15 15 15 19 15 27 8 ...
##  $ feats        : Factor w/ 332 levels "Abbr=Yes","Animacy=Anim|Aspect=Imp|Case=Ins|Gender
=Fem|Number=Sing|Tense=Pres|VerbForm=Part|Voice=Pass",..: 173 60 121 117 169 165 320 59 208 N
A ...
##  $ head_token_id: num [1:202226] 2 0 2 4 2 7 13 13 2 2 ...
##  $ dep_rel      : Factor w/ 39 levels "acl","acl:relcl",..: 27 38 13 26 13 6 27 27 36 9
...
##  $ misc         : Factor w/ 3 levels "SpaceAfter=No",..: NA 1 NA 1 NA 1 NA NA 1 NA ...
##  $ is_dialect   : num [1:202226] 0 0 0 0 1 1 0 0 0 1 ...
##  $ time_dif     : num [1:202226] 4.23 4.23 4.23 4.23 4.23 ...
```

Посмотрим на количество диалектизмов по фразам и их длительности:

```
time_df <- df %>%
  group_by(sentence, time_dif) %>%
  summarise(cnt = sum(is_dialect))
```

```
## `summarise()` has grouped output by 'sentence'. You can override using the
## `.groups` argument.
```

```
write_csv(time_df, "khislav_time.csv")

time_df <- read_csv("khislav_time.csv")
```

```
## Rows: 33875 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): sentence
## dbl (2): time_dif, cnt
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

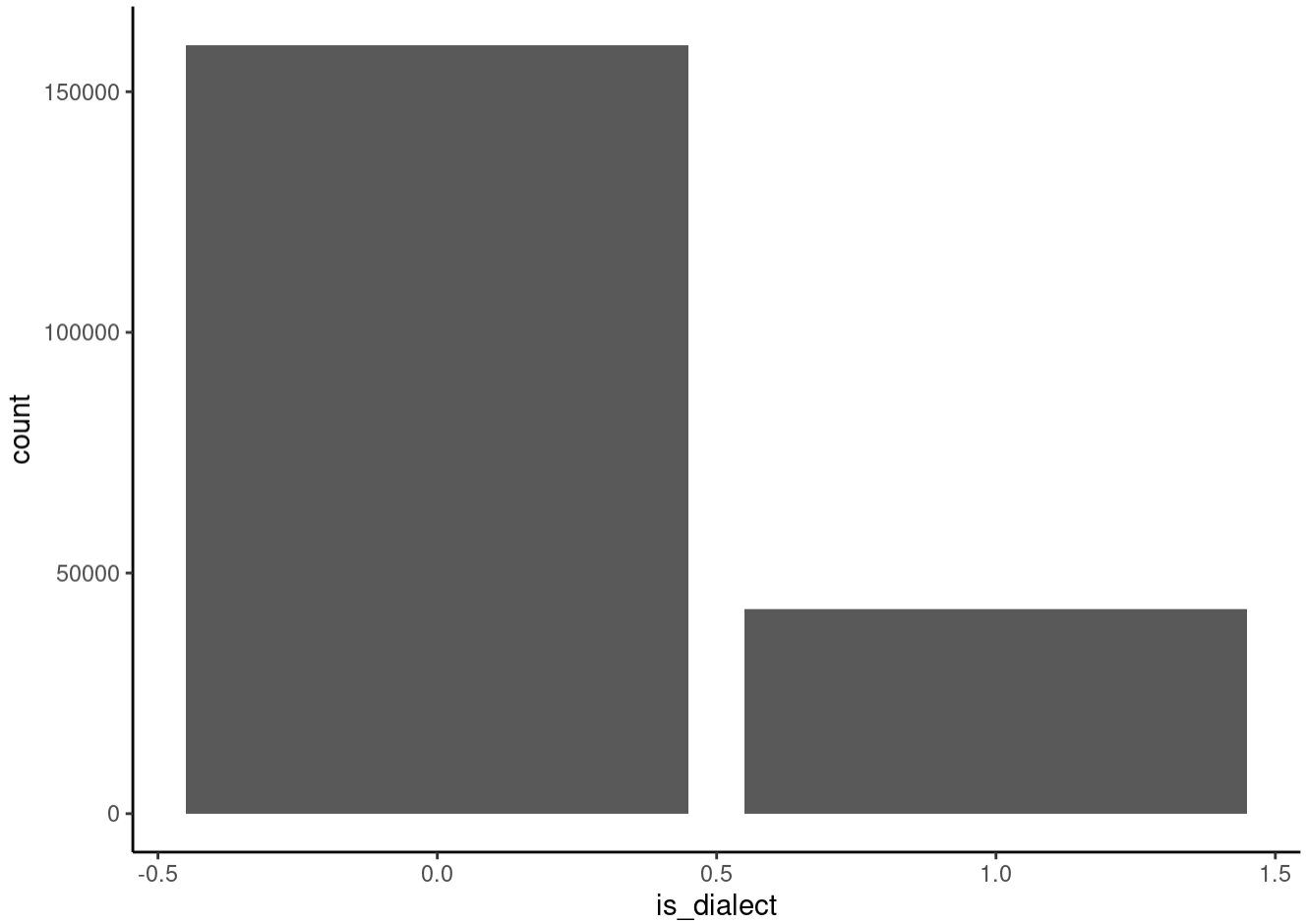
```
time_df
```

```
## # A tibble: 33,875 × 3
##    sentence                                                    time_dif   cnt
##    <chr>                                                          <dbl> <dbl>
##  1 - говорю, я уже не могу убираться, детки, дак вы уже, говорю,…     6.33     4
##  2 - Да на что? - Надо ей.                                            4.73     1
##  3 - Дай коробку, баба Клава сказала ты каб коробку.                  4.73     1
##  4 'Иди мне принеси' значит 'я иди замуж', вот.                       4.16     1
##  5 'коли', 'было', 'ёсь', 'нема'.                                     2.49     3
##  6 'яйка, млеко', пожрать надо было.                                  2.81     0
##  7 [[нрзб] всё это попорезали, ягоднику нема]                         7.06     1
##  8 [А вот если дом загорелся, что делали раньше?]                     3.57     1
##  9 [а гостим], ну это сейчас, а тогда ж.                              2.78     1
## 10 [А как (як) попадётся эти...]                                      1.54     3
## # ℹ 33,865 more rows
```

Визуализация данных

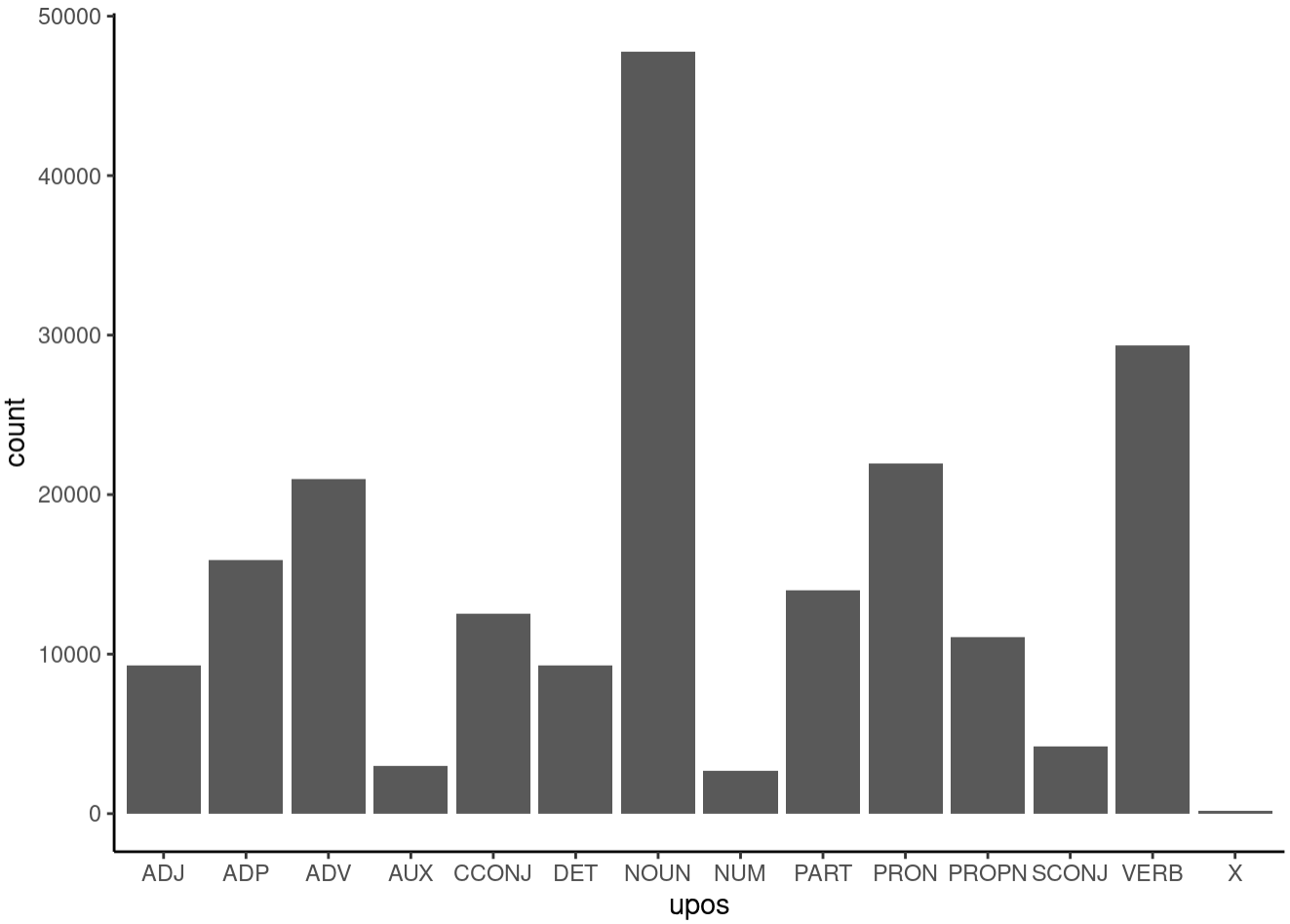
Как видно на гистограмме, в датасете преобладают “не-диалектные” леммы: их более 150-ти тыс. единиц, в то время как диалектными считаются менее 50-ти тыс. единиц.

```
ggplot(df, aes(x = is_dialect)) +
  geom_bar() +
  theme_classic()
```

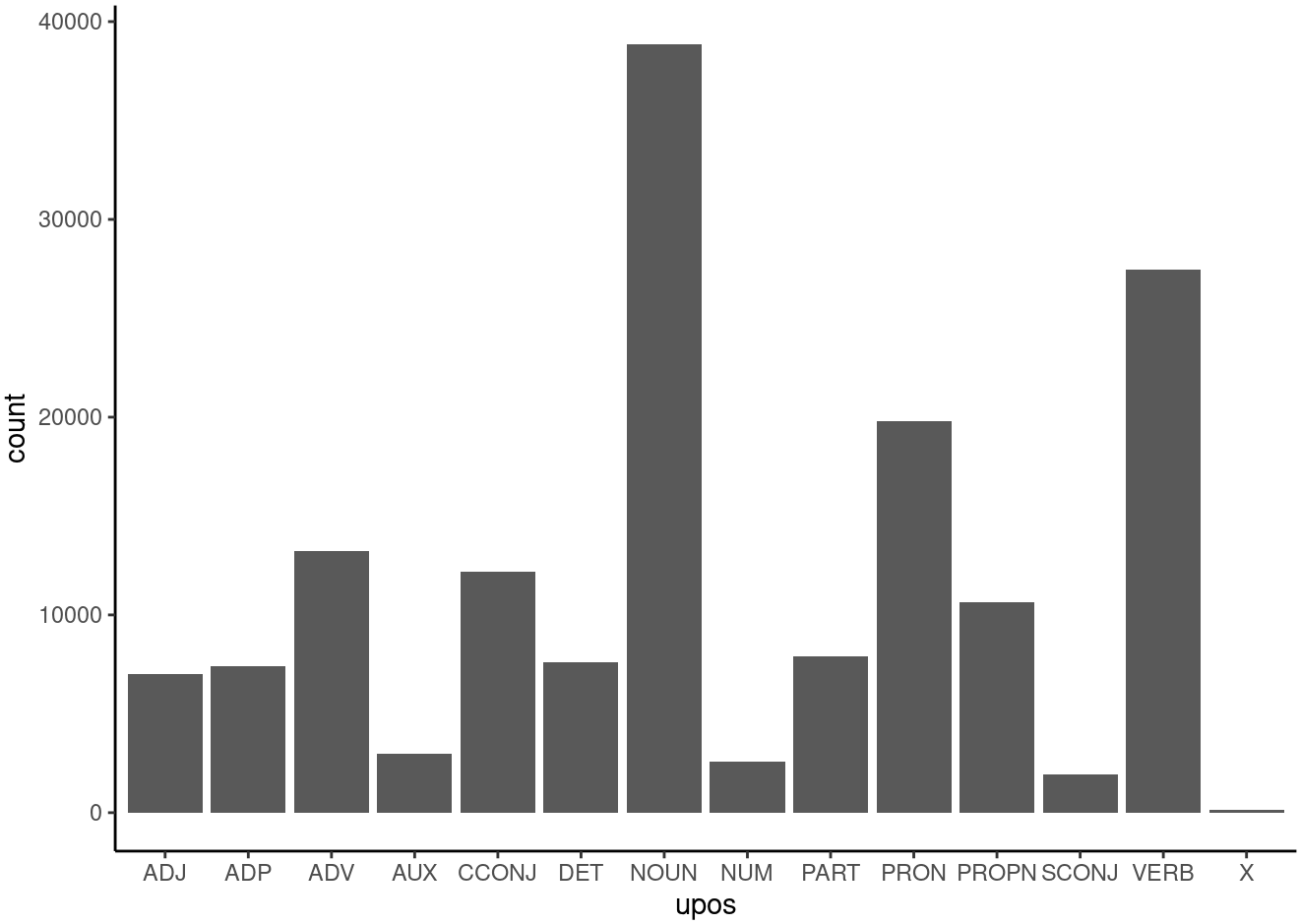


Во всем датасете преобладают существительные, на втором месте глаголы.

```
ggplot(df, aes(x = upos)) +
  geom_bar() +
  theme_classic()
```

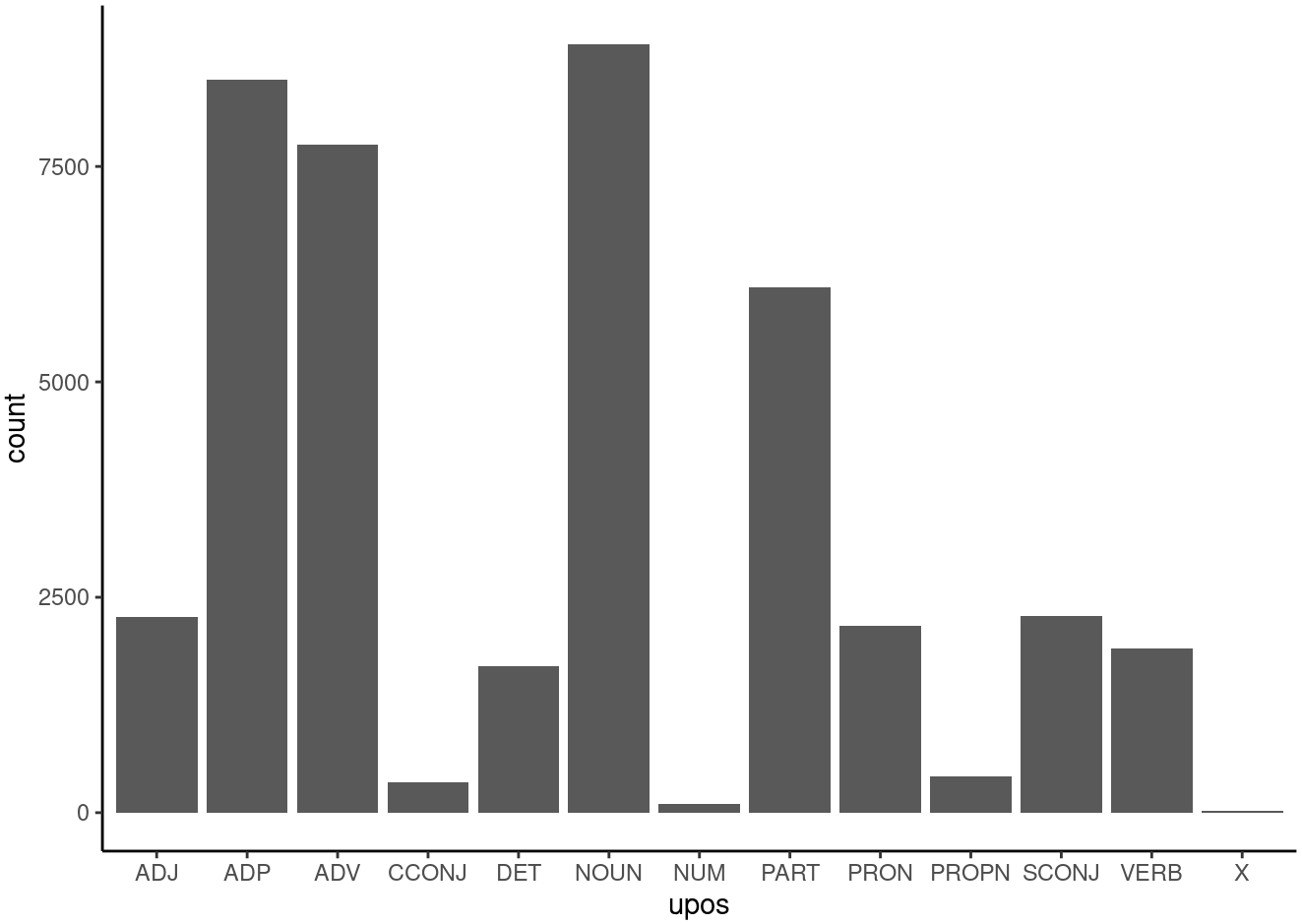



```
non_dial_df <- df %>%
  filter(is_dialect == 0)

ggplot(non_dial_df, aes(x = upos)) +
  geom_bar() +
  theme_classic()
```



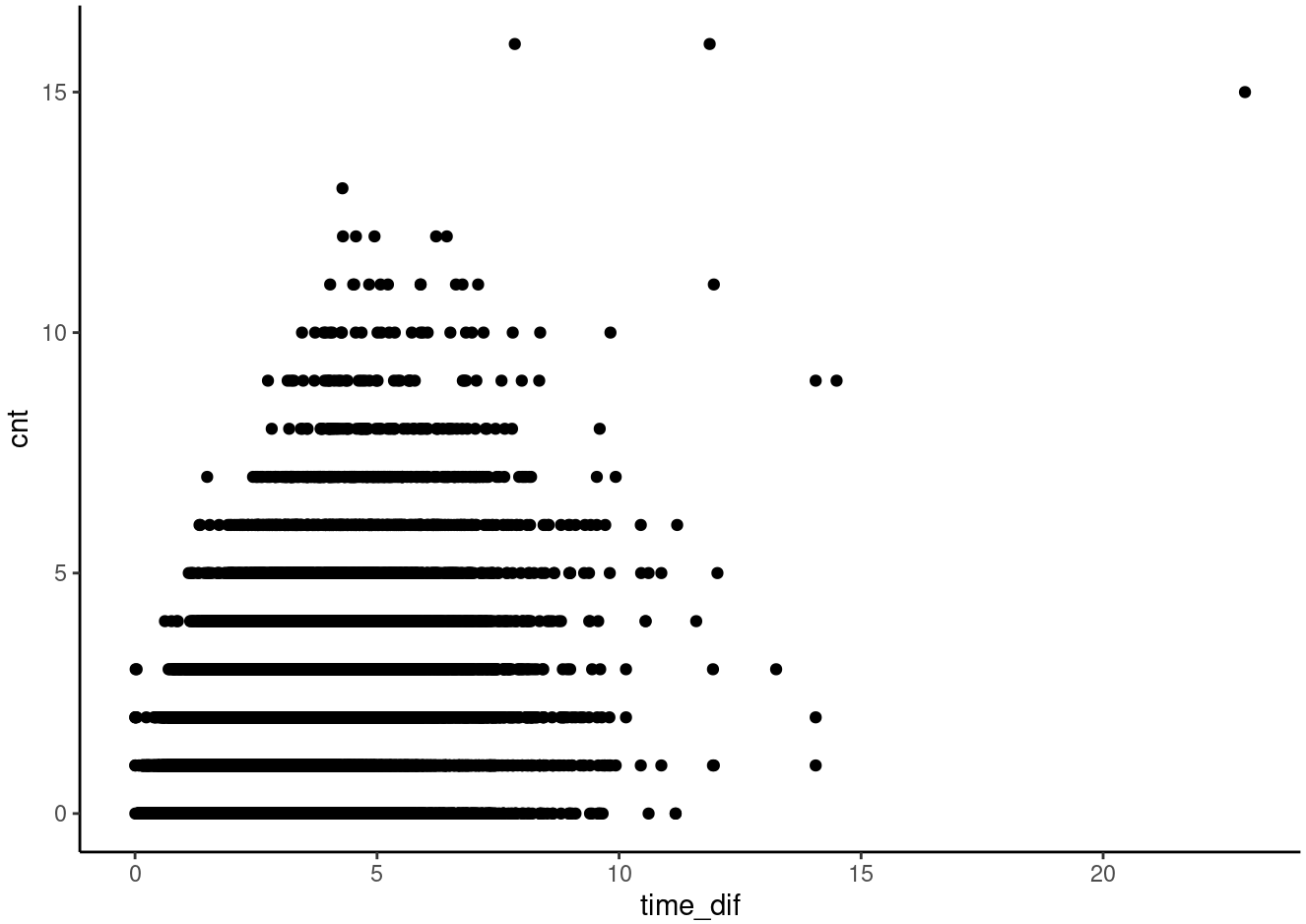
Однако среди диалектизмов такое распределение не сохраняется; существительные все еще на первом месте, однако к ним приближаются предлоги, наречия и частицы. Процентное соотношение глаголов гораздо меньше, по сравнению со всем набором данных целиком.

```
dial_df <- df %>%
  filter(is_dialect == 1)

ggplot(dial_df, aes(x = upos)) +
  geom_bar() +
  theme_classic()
```



Посмотрим на график распределения количества диалектизмов во фразе относительно ее длины:

```
ggplot(time_df, aes(x = time_dif, y = cnt)) +
  geom_point() +
  theme_classic()
```



Попробуем построить хитмэп корреляций между длительностью фразы и количеством в ней диалектизмов:

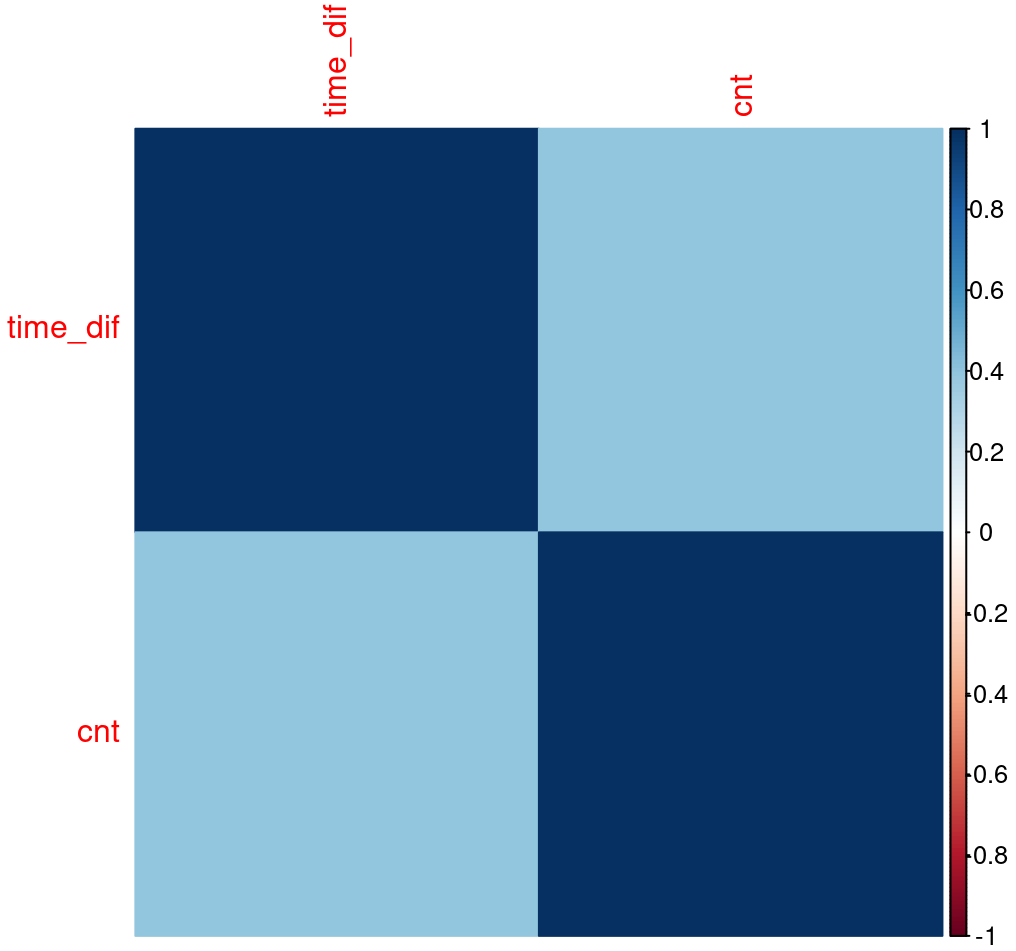
```
cor_df <- time_df %>%
  select(time_dif, cnt)

cor(cor_df)
```

```
##           time_dif       cnt
## time_dif 1.0000000 0.3906934
## cnt      0.3906934 1.0000000
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(cor_df), method = "color", order = "hclust")
```



Гипотезы

$H_0 : \tau = 0$ (между длительностью фразы и количеством диалектизмов в ней отсутствует корреляция)

$H_1 : \tau \neq 0$ (между длительностью фразу и количеством диалектизмов в ней есть корреляция)

Тестирование

```
str(time_df)
```

```
## spc_tbl_ [33,875 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ sentence: chr [1:33875] "- говорю, я уже не могу убираться, детки, дак вы уже, говорю,
на крыльце постойте, не пустила их сюда." "- Да на что? - Надо ей." "- Дай коробку, баба Клав
а сказала ты каб коробку." "'Иди мне принеси' значит 'я иди замуж', вот." ...
##  $ time_dif: num [1:33875] 6.33 4.73 4.73 4.16 2.49 ...
##  $ cnt     : num [1:33875] 4 1 1 1 3 0 1 1 1 3 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   sentence = col_character(),
##   ..   time_dif = col_double(),
##   ..   cnt = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
cor.test(time_df$time_dif, time_df$cnt, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  time_df$time_dif and time_df$cnt
## z = 73.618, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2978322
```

Глядя на параметры, можно сказать, что нулевая гипотеза опровергнута, и можно принять альтернативную гипотезу (P-value < 0.05). Имеет место корреляция между полями `time_dif` и `cnt` в представленном датасете.

Результаты

Таким образом, поставленная изначально гипотеза подтвердилась, т.е. между длительностью фразы и количеством в ней диалектизмов есть корреляцие: чем дольше фраза, тем больше в ней диалектизмов. Однако коэффициент корреляции равен 0.298, что показывает незначительность связи между двумя параметрами.