# Explorative error analysis of the manual evaluation of UDPipe-tagger

Silvie Cinkova March 17, 2019

#### Introduction

This document analyzes the manual evaluation of tagging provided by UDPipe version 2.0. To the date of March 18, the following languages have been fully evaluated (i.e. at least a 5000-token sample annotated):

- English;
- French;
- German;
- Portuguese;
- Slovene.

We also have a small sample of Nynorsk (600 tokens) and a 5000-token Hungarian sample from a different tagger. The Nynorsk sample has been tentatively included.

We are waiting for Czech and Hungarian-UDPipe. The fate of any variety of Norwegian is unknown at the moment.

Here I present a bird's eye view on the languages in comparison, as well as more detailed views on individual languages. The detailed language-specific word cloud plots are meant to facilitate a manual error analysis of the individual languages.

#### Datasets and libraries

```
library(tidyverse)
library(ggwordcloud)

errs <- read_tsv("all_errors.tsv") # one Portuguese word has no ID, that's ok
err_freqs <- read_tsv("frequencies_errors_all_langs.tsv") # all unique tokens
#with frequencies in the given language
err_freqs$combs <- factor(err_freqs$combs)
levels(err_freqs$combs)[2] <- "features"
err_lines_01 <- read_tsv("summary_table_all_langs.tsv") # summary for each lang
altok_df <- read_tsv("all_tokens_together.tsv") %>% na.omit()#all annotated tokens
```

### Raw frequencies of different error types

This table compares raw frequencies of different error types. Explanation of column names:

- rowsN: number of rows; i.e. tokens in the given language sample;
- errowsN: number of tokens where at least one error has occurred;
- lem\_errN: number of tokens with a lemmatization error;
- tok\_errN: number of tokens with a tokenization error;
- upos\_errN: number of tokens with a POS-tagging error.

The samples (except Nynorsk) are of comparable size. The French sample has been shrunken by a random selection of 5,100 rows from the annotated data of about 11,100 rows.

The table does not contain the raw frequencies of errors in features, since these are hardly comparable across languages.

```
err_lines_01 %>% select(language, ends_with("N"))
## # A tibble: 6 x 6
##
     language rowsN errowsN lem errN tok errN upos errN
##
     <chr>>
                                  <int>
                                            <int>
               <int>
                        <int>
                                                       <int>
## 1 EN
                5010
                          194
                                     60
                                               22
                                                          135
## 2 FR
                5100
                          545
                                    330
                                               54
                                                         270
## 3 GER
                5043
                          986
                                    363
                                              144
                                                         375
                                                          67
## 4 NYN
                 599
                           81
                                     12
                                                1
## 5 POR
                5083
                          575
                                    261
                                               31
                                                         346
## 6 SLV
                         1036
                                                         430
                5081
                                    485
                                                9
```

### Relative frequencies and geometric mean of feature errors

This table shows the relative frequencies of different error types in the individual languages, in percent. The columns are the following:

- prc\_errows: percent of tokens with at least one error of all tokens;
- prc\_lem\_err: percent of tokens with a wrong lemma, of all tokens;
- prc\_tok\_err: percent of tokens with a tokenization error, of all tokens;
- prc\_upos\_err: percent of tokens with a wrong POS, of all tokens;
- geomean\_feat\_err: geometric mean of percentages of feature errors per token.

Note that the geometric mean comes in its specific units that have nothing to do with the number of tokens, number of features, or number of errors. We cannot interpret them but only sort the languages according to them. We use the geometric mean because the amount of features varies both between languages and in each individual language, where the numbers vary between POS as well as other specific word groups (e.g. quantifiers): 100% features wrong of one feature is quite a difference from 100% features of 10 features. Thus it does not make sense to compute the arithmetic mean.

Two more remarks

## 6 SLV

- 1) These errors are exclusively precision errors, since the annotators were not asked to indicate missing features.
- 2) We merge two types of errors: a wrong feature (e.g. tense with an adverb) and a wrong feature value (e.g. Gender="Masculine" in a feminine noun, such as the German Frau).

```
err_lines_01 %>% select(language, starts_with("prc"), starts_with("geom"))
## # A tibble: 6 x 6
##
     language prc_errows prc_lem_err prc_tok_err prc_upos_err
##
     <chr>
                    dbl>
                                 <dbl>
                                              <dbl>
                                                            <dbl>
                     3.90
                                                            2.70
## 1 EN
                                  1.20
                                             0.400
## 2 FR
                    10.7
                                  6.50
                                             1.10
                                                            5.30
## 3 GER
                                  7.20
                                                            7.40
                    19.6
                                             2.90
## 4 NYN
                    13.5
                                  2.00
                                             0.200
                                                            11.2
## 5 POR
                    11.3
                                  5.10
                                             0.600
                                                            6.80
```

0.200

8.50

## # ... with 1 more variable: geomean\_feat\_err <dbl>

9.50

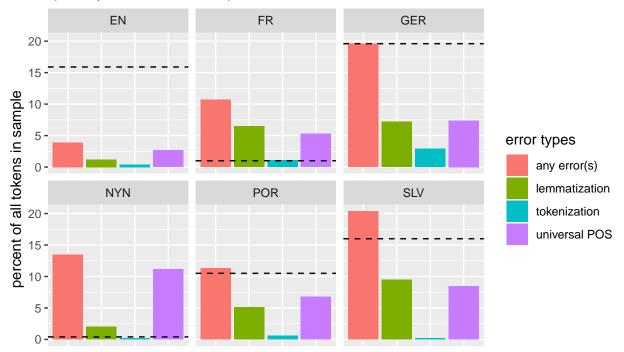
20.4

A small restructuring of the data for ggplot2 (tidy)

```
err_lines <- dplyr::select(err_lines_01, language, starts_with("prc"),
                           starts_with("geo")) %>%
  tidyr::gather(key = "key", value = "percent", starts_with("prc"))
ggplot(err_lines, aes(x = key, y = percent, fill = key)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_abline(mapping = aes(intercept = geomean_feat_err, slope = 0),
              linetype = 2) +
  facet_wrap(~ language) + theme(axis.text.x = element_blank(),
                                 axis.ticks.x = element blank(),
                                 axis.title.x = element_blank()) +
  scale_y_continuous(name = "percent of all tokens in sample" ) +
  scale_fill_discrete(name = "error types",
                      labels = c("any error(s)", "lemmatization",
                                 "tokenization",
                                 "universal POS",
                                 "geometric mean of feature errors")) +
  ggtitle(label = "UD tagging errors for individual languages",
          subtitle = "Line represents the geometric mean of feature
          errors per token in generic units \n (not in percent of all tokens)")
```

### UD tagging errors for individual languages

Line represents the geometric mean of feature errors per token in generic units (not in percent of all tokens)

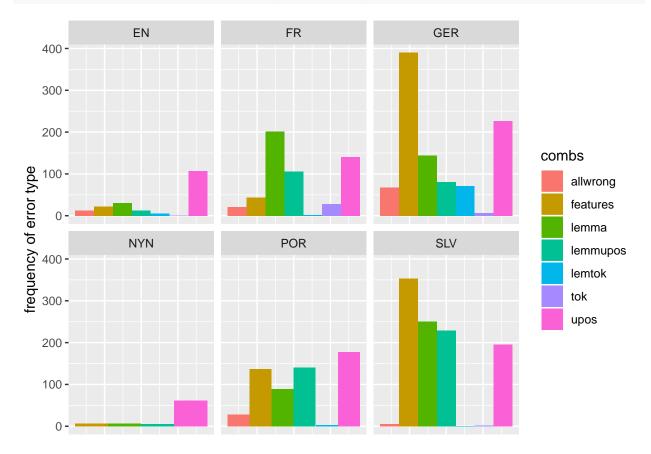


## Comparison of languages according to different combinations of error types

First summarize the occurrences

```
sumfreqs <- err_freqs %>% group_by(language, combs) %>%
   summarize(freqy = sum(freq)) %>% na.omit()
sumfreqs$combs <- factor(sumfreqs$combs)
levels(sumfreqs$combs)[2] <- "features"

ggplot(sumfreqs, aes(x = 1, y = freqy, fill = combs)) +
   geom_bar(stat = "identity", position = "dodge") + facet_wrap(~ language) +
   theme(axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        axis.title.x = element_blank()) +
   scale_y_continuous(name = "frequency of error type")</pre>
```



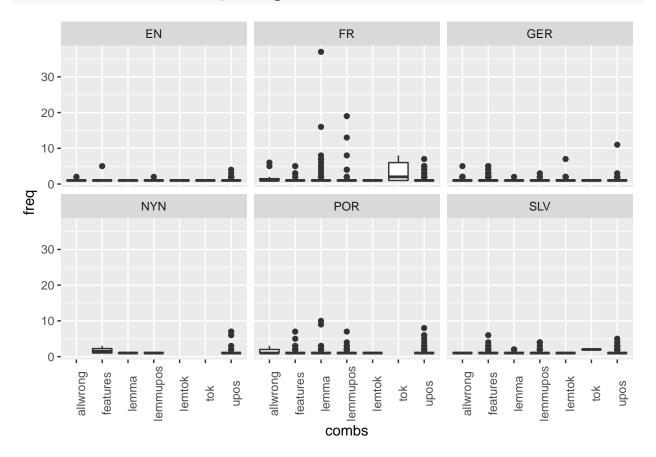
### Distribution of errors across lemmas

The faceted boxplot below shows that errors of a certain type are sometimes associated with a few recurrent word forms, typically function words, as we can see in the French sample. The most extreme case in the entire multilingual collection is one incorrectly determined French lemma that has occurred over 40 times! Among the wrongly tokenized French words, one quarter of the observations are words that have occurred 2-5 times. On the other hand, the errors in English are rather spread across different words. The case of French suggests a grammatical difference between the 19th-century French (or just a different spelling of

function words), while the errors in English may be caused by many different content words that have been unknown to a tagger trained on late-20th-century newspapers.

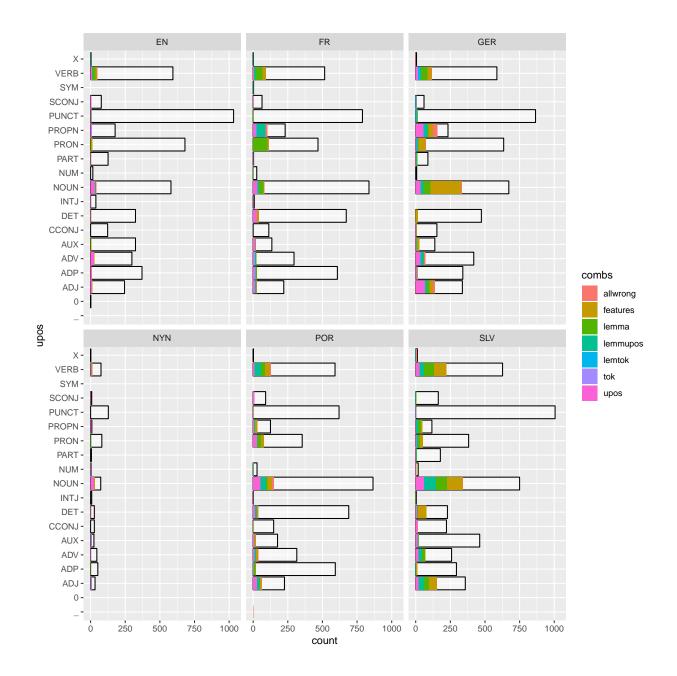
```
freqs <- select(err_freqs, language, combs, freq) %>% na.omit()
freqs$combs <- factor(freqs$combs)
levels(freqs$combs)[2] <- "features"

ggplot(freqs, aes(x = combs, y = freq)) + geom_boxplot() +
  facet_wrap(~language) +
  theme(axis.text.x = element_text(angle = 90))</pre>
```



## Error types in each POS for individual languages

This flipped and faceted barplot shows raw frequencies of each upos (white bar). Colorful stacked bars inside indicate raw frequencies of the individual error types.



## Word clouds for individual languages

```
vars <- err_freqs$language %>% unique()
titles <- c("English", "French", "German", "Nynorsk", "Portuguese", "Slovene")</pre>
```

The plot code snippets are very repetitive, but different languages require different combinations of errors to be plotted together or separately. The code snippets will be presented in the English section. Then we will extract functions from these snippetes and use these functions throughout to keep the code shorter. Originally we wanted to run the script as a loop, but different plots require different image size, so we produce each plot separately.

#### EN - Across parts of speech

```
i <- 1
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower form, color = combs, size = freq)) +
  geom text wordcloud(show.legend = TRUE, shape = "square") +
  scale size area(max size = 12) +
 ggtitle(label = titles[i], subtitle = "All error types")
```

lemma

lemtok

lemmupos

#### **English** All error types

```
worrying
                    together when
                                                                  waxed
                                                         throes
    singing swathing
                                                                             combs
                                          therefore
                                 swelled
                                                                   tender
                                                 seaward rushed
                                  scorned
                                                                              a allwrong
  o'reilly yer provided
                                                                   remains
                           paradise
                                          place stir priceless
                                                                                features
                                                               more
                                   mind
                                          just charming
  interferences
                     invigorating
                                                           hesitated
                                  hooted
                   faint frail faint done
that kneeling farine
                                          entreated
                                                         fierce
                                                                    parting
                                                     do farther mal
                                                                              a
                   burned o
     more folle
                             bitter anglo-
                                         did blazed
                deer alarming
    lammie
                                           administer dared
                                                            hers overheard
                                                                             a upos
 round left flings we
                               11th -given
   mute flouted crept alight
                                           afraid browsing grub must that
                                'ere
                              'd 's 's 'i
                                         're all bade encourager
 sha listen for
                                                                             freq
                  bound
                         all 're
                                         afloat committed n't mine to
                 closed an be
                                 absent
                                                                               a 1
                                         better bed o escaped
   ye look in fake bast no
                                before
                                                                   not
                                                                               a 2
talks mine
          enough craddock
                                                      do hate-
                                  drawled
                                             egremo
                                                     flashed kindred shot
                       further
                                         nt chafed
                              housewifely
                   loose one
                                                        laughs
                                          to meantime
                                                                 morrow
                    picturesque ornament
                                 psammead Propos
                                                                  rovina
      remembered
                    savage
                                             sound
                             soi-disant
                                                     slanting
                                                                   terrified
   stamp
                                         wharves
                                                       where where
                              tongued
whom
                                            woven
```

```
df <- dplyr::filter(err_freqs, language == vars[i], combs != "upos") %>%
  na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
  geom_text_wordcloud(show.legend = TRUE, shape = "square") +
  scale_size_area(max_size = 12) +
  ggtitle(label = titles[i],
          subtitle = "Selected error types except POS errors") +
  facet_wrap(~ combs)
```

English
Selected error types except POS errors



```
POS tagging errors
```

```
tender swathing soi-disant
where together therefore
                    seaward
 round provided
yer to lammie stir kindred chafed
                                                            freq
                                  mal housewifely whom
 restless farther
                                                             а
                 encourager done craddock
                                                             a
   o'reilly fierce
                  before bed be bast faint
           folle deer absent 11th all bitter
  sha mind fake afraid o 'ere afloat do
                                         flings o remains
 vet papa frail better -given, re
                                 alight faint must savage
sound mute in administer all
                              closed did no laughs that
  we nine browsing
                                                            combs
                      farine
                             for
                                  just enough
          grub
                                                            a lemmupos
               charming
                           more
        never
                                                            a upos
               paradise
                          place tear NOt
  priceless
               psammead
     though
```

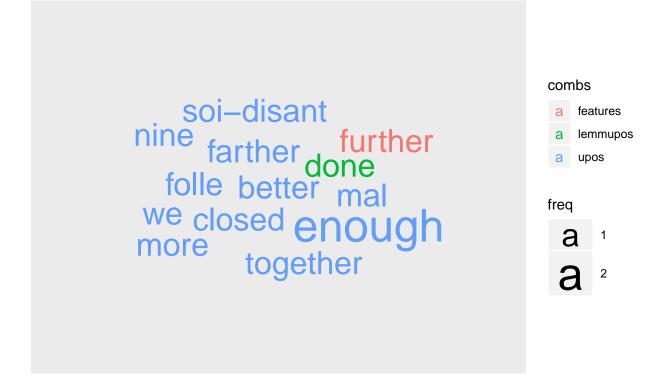
#### Function to extract different combinations of errors across all POS.

#### EN - Individual parts of speech

Word clouds of errors within one guessed part of speech.

```
ups <- err_freqs$upos %>% unique() %>% sort()
#The English sample does not contain any erroneous tokens with this POS.
i <- 1
y <- 1
df <- dplyr::filter(err freqs, language == "EN", upos == ups[y]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
  geom_text_wordcloud(show.legend = TRUE, shape = "square") +
  scale_size_area(max_size = 12) +
 ggtitle(label = titles[i], subtitle = paste0("Errors in POS \"", ups[y], "\""))
i <- 1
y <- 2
df <- dplyr::filter(err_freqs, language == "EN", upos == ups[y]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
  geom_text_wordcloud(show.legend = TRUE, shape = "square") +
  scale_size_area(max_size = 12, breaks = 1:10) +
  ggtitle(label = titles[i], subtitle = paste0("Errors in POS \"", ups[y], "\""))
```

## English Errors in POS "ADJ"



Again, we extract a function to shorten the code.

Errors in POS "ADP"

```
to to mind before for 0 round

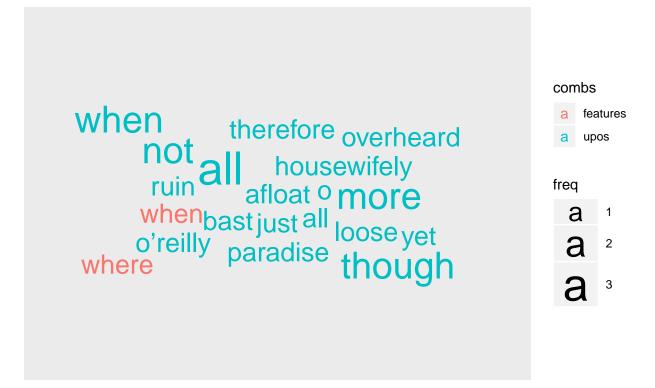
combs

a allwrong
a upos
```

```
# df <- dplyr::filter(err_freqs, language == langua, upos == ups[y]) %>% na.omit()
# set.seed(123)
# ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
# geom_text_wordcloud(show.legend = TRUE, shape = "square") +
# scale_size_area(max_size = 12) +
# ggtitle(label = titles[i], subtitle = paste0("Errors in POS \"", upos[y], "\""))
```

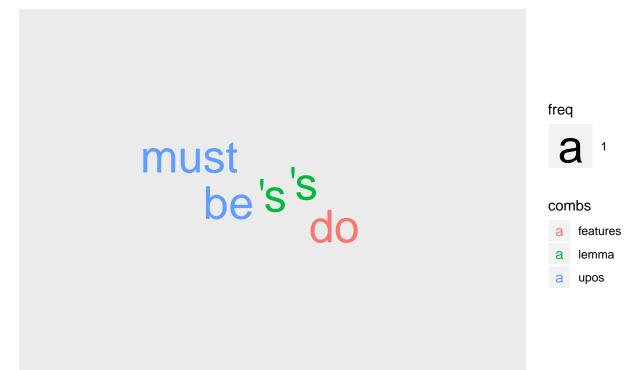
```
i <- 1
y <- 4
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

Errors in POS "ADV"



```
# df <- dplyr::filter(err_freqs, language == "EN", upos == ups[y]) %>% na.omit()
# set.seed(123)
# ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
# geom_text_wordcloud(show.legend = TRUE, shape = "square") +
# scale_size_area(max_size = 12) +
# ggtitle(label = titles[i], subtitle = pasteO("Errors in POS \"", upos[y], "\""))
i <- 1
y <- 5
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### Errors in POS "AUX"



```
# upos <- err_freqs$upos %>% unique() %>% sort()
# df <- dplyr::filter(err_freqs, language == "EN", upos == ups[y]) %>% na.omit()
# set.seed(123)
# ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
# geom_text_wordcloud(show.legend = TRUE, shape = "square") +
# scale_size_area(max_size = 12, breaks = c(1,2)) +
# ggtitle(label = titles[i], subtitle = pasteO("Errors in POS \"", upos[y], "\""))
#No errors in English CCONJ
# i <- 1
# y <- 6
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])

i <- 1
y <- 7
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### Errors in POS "DET"



```
i <- 1
y <- 8
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

Errors in POS "INTJ"

## never papa

```
combs
a upos
freq
1
```

```
i <- 1
y <- 9
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

## English Errors in POS "NOUN"

```
stamp savage psammead parting combs
stir flings n't fierce faint roving charming alight 11th morrow
laughs bitter administer queer sound deer interferences priceless seaward talks propos

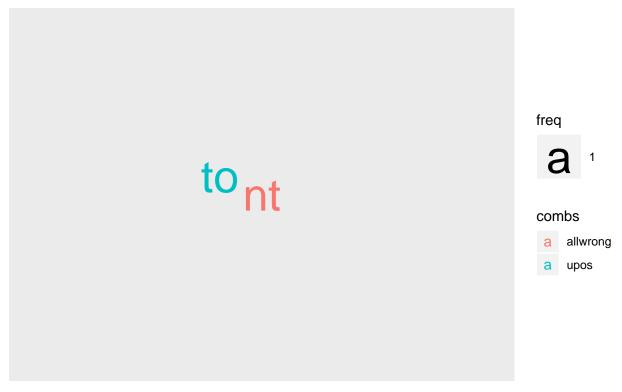
wharves

tender psammead parting combs
a allwrong a lemma a lemtok
a upos
freq
a 1
a 1
a 2
```

```
#No errors in English NUM
# i <- 1
# y <- 10
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])

i <- 1
y <- 11
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### Errors in POS "PART"



```
i <- 1
y <- 12
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

English
Errors in POS "PRON"



```
i <- 1
y <- 13
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### Errors in POS "PROPN"

# listen encourager tear afraid sha egremo meantime

freq



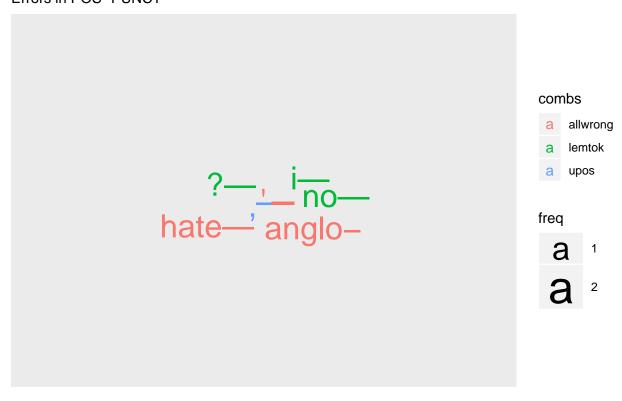
combs

a tok

a upos

```
i <- 1
y <- 14
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

English
Errors in POS "PUNCT"



```
i <- 1
y <- 15
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### Errors in POS "SCONJ"

```
of in that

combs
a upos

freq
a 1
a 2
a 3
a 4
```

```
# No errors in English SYM
# i <- 1
# y <- 16
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
i <- 1
y <- 17
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

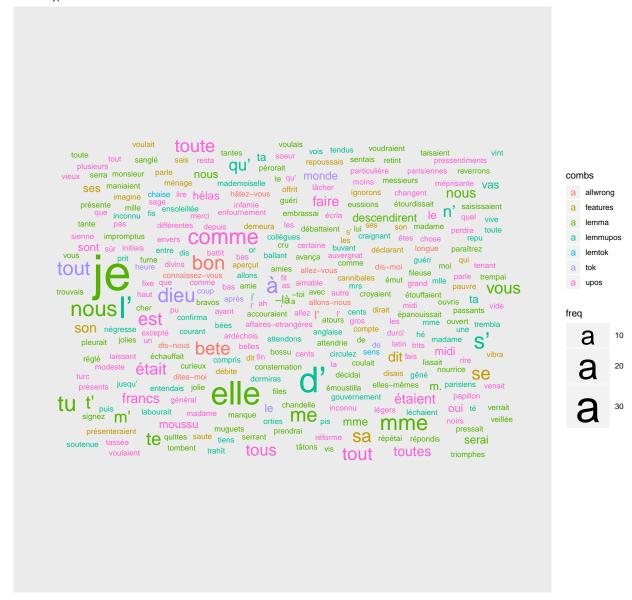
## English Errors in POS "VERB"

```
worrying
 swelled slanting terrified
 hesitated flouted
                     flashed
                                           combs
 hooteddo
                                           a allwrong
           blazed alarmi
 escaped
                                           a features
                                           a lemma
        burned're'dbade
                                           a upos
    fake bound
                                           freq
            crept
  invigorating
scorned
```

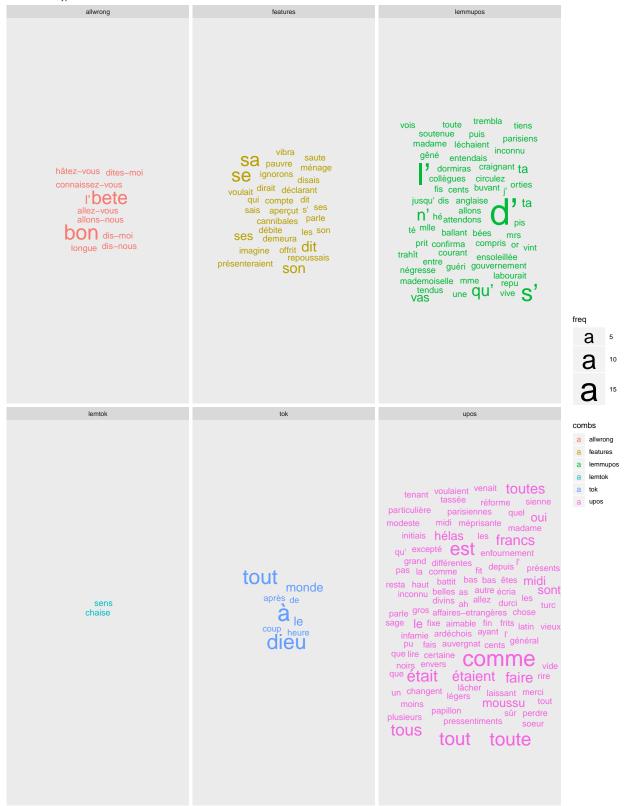
```
# No errors in English X
# i <- 1
# y <- 18
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French - errors across parts of speech

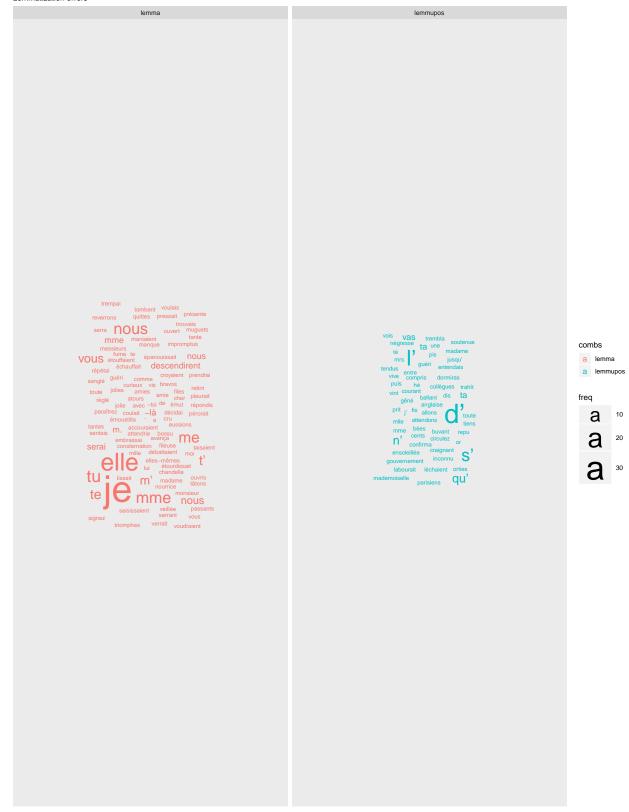
```
i <- 2
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
    geom_text_wordcloud(show.legend = TRUE, shape = "square") +
    scale_size_area(max_size = 20) +
    ggtitle(label = titles[i], subtitle = "All error types")
```



French Selected error types



French Lemmatization errors



#### French - errors in individual parts of speech

```
# No errors in French _
# i <- 2
# y <- 1
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
i <- 2
y <- 2
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French

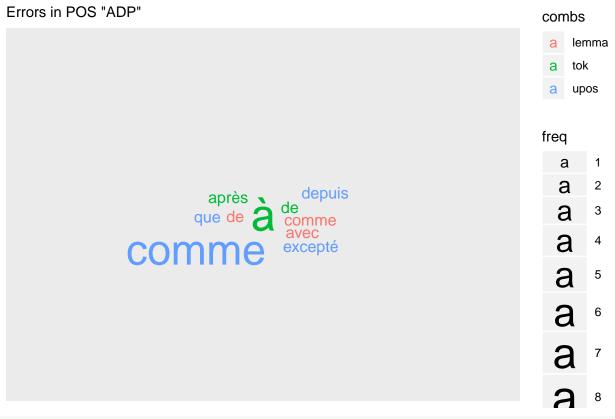
Errors in POS "ADJ"

```
présente longue présents
francs
haut sûr
cher autre fritsjolies
bas généraljolie
inconnu infamie vieux
pressentiments

freq
a 1
a 2
a 3
combs
a allwrong
a lemma
a upos
```

```
i <- 2
y <- 3
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French



```
i <- 2
y <- 4
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

## French Errors in POS "ADV"

```
toute tout comme tenant moins—làfin pas enfournement

tout

tout

toute tout tenant a 1 a 2 a 3 a 4 a 5 a 6
```

```
i <- 2
y <- 5
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

## French Errors in POS "AUX"



```
# No errors in French CCONJ
# i <- 2
# y <- 6
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
i <- 2
y <- 7
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

## French Errors in POS "DET"

```
toutes
son ses le plusieurs
tous les la le son
les certaine quel toute
différentes le les
tout

combs
a features
a tok
a upos
```

```
i <- 2
y <- 8
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French

Errors in POS "INTJ"

```
combs
a upos
freq
a 1
a 2
```

```
i <- 2
y <- 9
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French combs Errors in POS "NOUN" allwrong a features a lemma a lemtok tante sens particulière resta messieurs madame changent muguets a tok a upos sagehâtez-vous consternation connaissez-vous tâtons lâcher cents belles battit atours chandelle veillée freq monsieur cannibales amie aimable bas êtes modeste légers coup affaires-etrangères divins M. nourrice a 1 rire lire dis-moi amies ardéchois curieux midi a 2 initiais auvergnat bossu bravos noirs hélas chaise files fileuse gros heure perdre impromptus méprisante mme monde passants pauvre tantes triomphes

```
i <- 2
y <- 10
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French

### Errors in POS "NUM"

```
freq
a
1
combs
a lemma
```

```
i <- 2
y <- 11
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

French
Errors in POS "PART"

```
combs
a upos
freq
a
1
```

```
i <- 2
y <- 12
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French

### Errors in POS "PRON"

```
combs

a allwrong
a features
a lemma
a upos

freq
a 1
a 2
a 3
a 4
a 5
a 6
a 7
a 8
a 9
a 10
```

```
i <- 2
y <- 13
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

# French Errors in POS "PROPN"

```
freq
                                                  a 1
                                                  a
                                                     2
                                                     3
           réforme
sienne
                     parle latin
          chose
  un l'
                         e papillon
   midi
 turc madame ah envers grand soeur allons-nous
    dites-moi écria CIEU
moussu parisiennes
                                                 combs
                                                 a allwrong
                                                 a tok
                                                 a upos
```

```
i <- 2
y <- 14
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

# French

# Errors in POS "PUNCT"

```
freq
a
1
combs
a lemma
```

```
i <- 2
y <- 15
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

# French

# Errors in POS "SCONJ"

```
combs
a upos
freq
```

```
# No errors in French SYM
# i <- 2
# y <- 16
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
i <- 2
y <- 17
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

# French Errors in POS "VERB"

```
voulais voudraient
 tombent taisaient signez trempai
   repoussais saute saisissaient sais
                 présenteraient paraîtrez
 prendrai
   laissant imagine ignorons étourdissait
  étouffaient échauffait descendirent
vide fume débattaient
                       croyaient fixe réglé
pleuraitémut coulait
  parle demeura aperçut dirait fit ouvris disais accouraient fais répétaire attendrie cru dit énage débite
          débite
                  décidai déclarant quittes
   embrassai
                 émoustilla épanouissait
     quéri
            lissait maniaient manque
 pérorait
            répondis pressait
                                        retint
                        serai sanglé
reverrons
              sentais
                                        serra
   serrant
                                 verrait
    venait trouvais
                      vibra
                                voulait
```



```
i <- 2
y <- 18
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])</pre>
```

### French

Errors in POS "X"

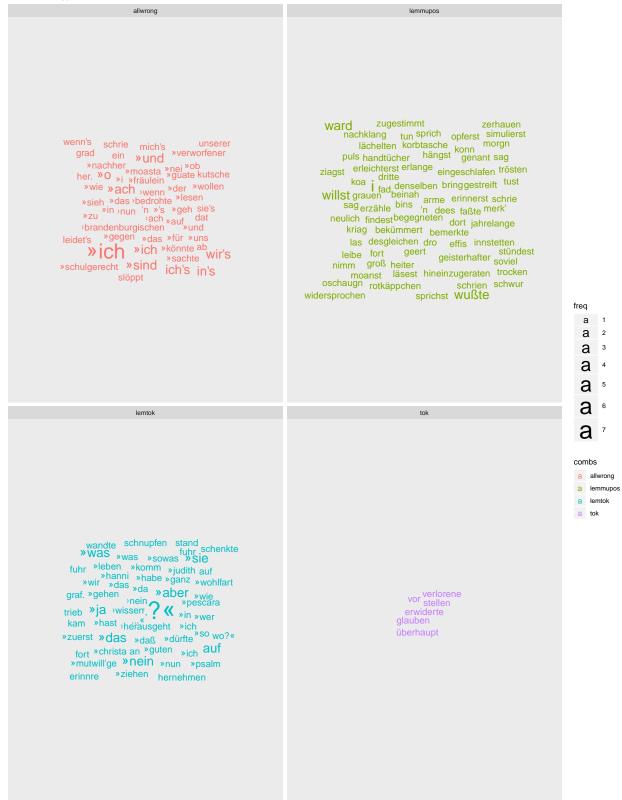


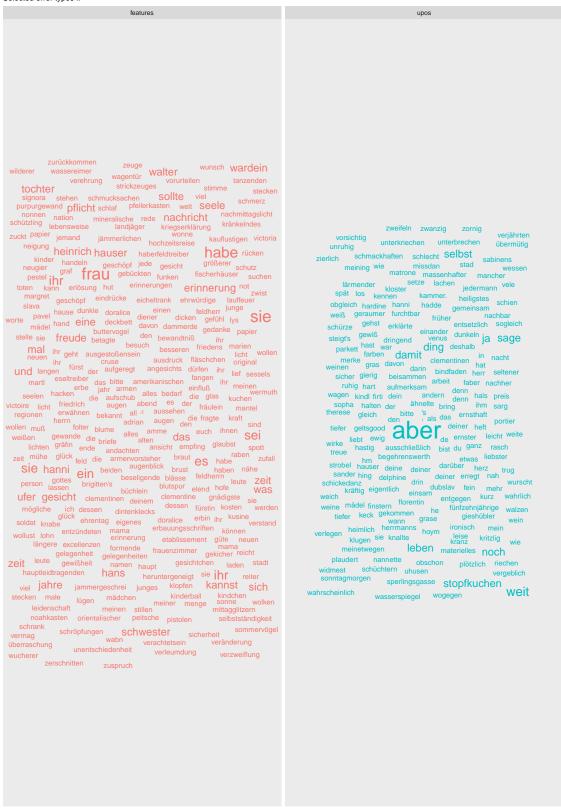
## German

```
i <- 3
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
    geom_text_wordcloud(show.legend = TRUE, shape = "square") +
    scale_size_area(max_size = 12) +
    ggtitle(label = titles[i], subtitle = "All error types")
```



#### German Selected error types I





combs

a features
a upos

freq
a 2.5
a 5.0

a 7.5

zweiten wüsste wahrhaftes vereinzelter viersitzige verdarben unterhaltung übermochte stärkeres vergißt schweiße staunen schoß plauderte Pavels operngläser munde mochtenmitleides lehrhaft leiser kinde klangen letzte lämmer krieges kinde iedes kasten ienem zugestimmt zerhauen widersprochen krieges kinde jedes kasten jenem that gewährte getroffen gestanden sprichst simulierst rotkäppchen tun oschaugn lächelten korbtasche läsest ziagst koa handtücher jahrelange that genügegenossen nehme erscholl empfandfünfzehnten komm blicken bums danach küßte genant tust geisterhafter erlange erinnerst lieber barbyschen auszudrücken laß war gedacht augenbrauendritter glieder bett angekränkelt sie sie puls fort desgleichenmerk' las dro bemerkte kriag ur bekümmert sag erzählebegegnetenkonn hirten denn abgefunden gewußt gestreift bins 'n dees innstetten laut garn aeronauten bork ihnen tiger neulich fad arme dort i opferst hirtinnenbesen anwesen diele kaum grauen beinah bring heiter lachte gebe arme armen lachte gebearme armen habnaufrichten augn fast jeder eife ausgeschüttet beicht <sup>güte</sup> reisen schrie denselben nimm effis dritte groß leibe eingeschlafen italien beobachte besonderes muaß erleichterst findest geert
hängst sag trösten
moanst hineinzugeraten soviel
morgn nachklang schritt dienstboten eingewickelt jedem ihr genährt höchst gesandten jede gesenkt wegen hervorrief hinausglitt hinauszufahren klang konnte schrien sprich stündest trocken knabenjahren kummervoll mußte orlamünde mitgetheilt nicht ruhmes onamunde s sporen schlummerschweren schürze stimm stunden taugenichtse weder vergossenvaterliebe weiser ungekanntes versetzte versteinerte verzehrt willst

freq

a <sup>1</sup>

a 3

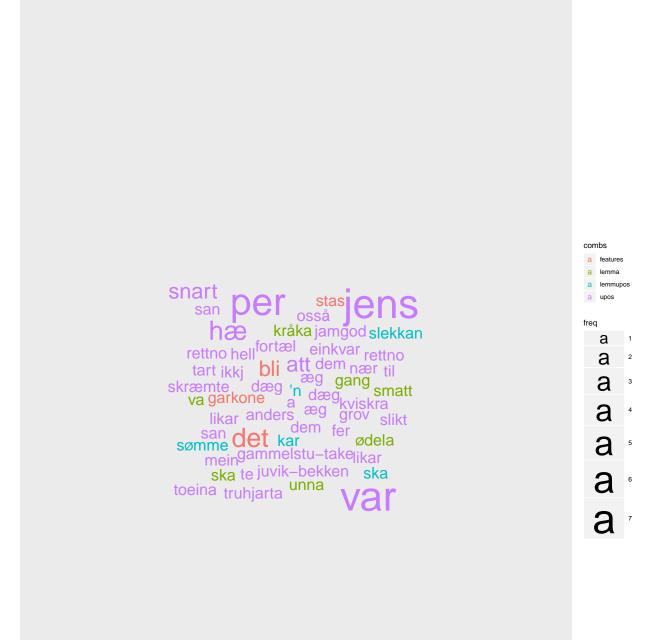
combs

a lemma

a lemmupos

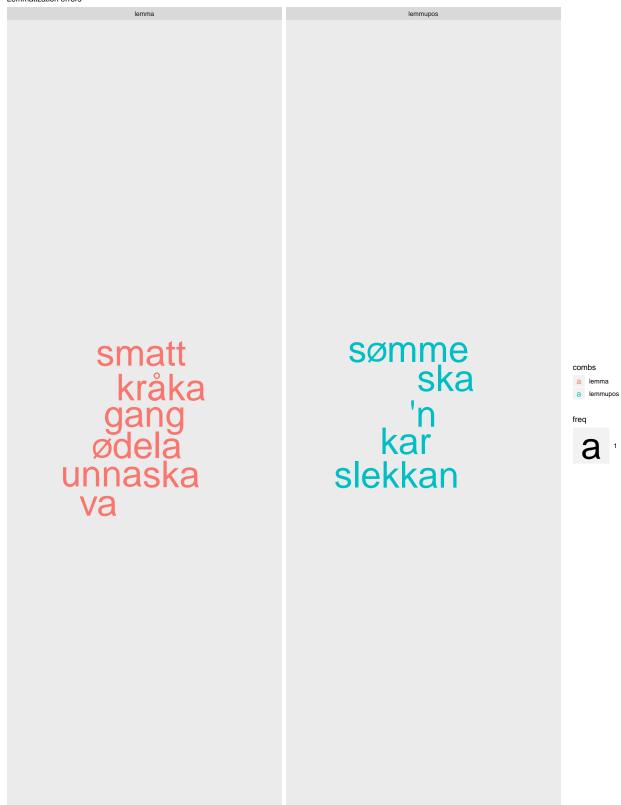
# Nynorsk

```
i <- 4
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
    geom_text_wordcloud(show.legend = TRUE, shape = "square") +
    scale_size_area(max_size = 20) +
    ggtitle(label = titles[i], subtitle = "All error types")
```



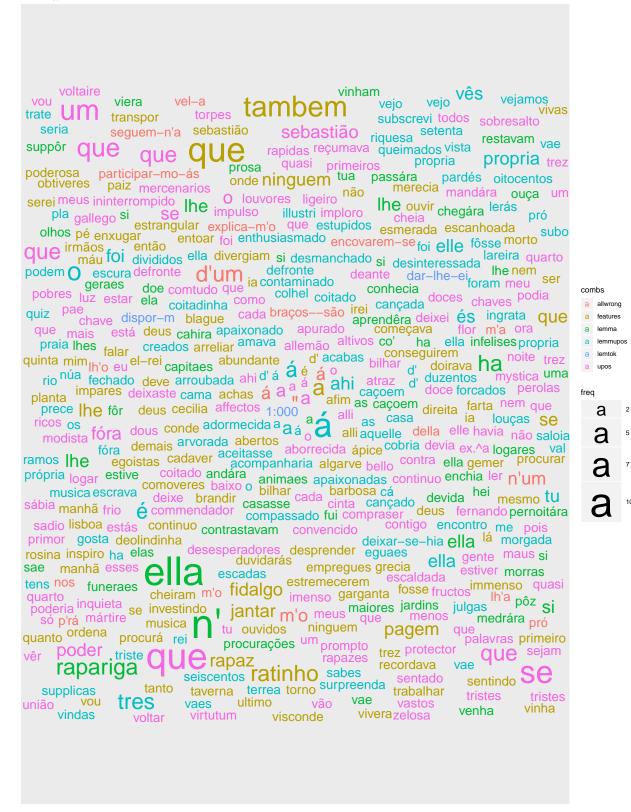
#### Nynorsk Selected error types

features	lemmupos	upos	
stas garkone bli det	sømme ska 'n kar slekkan	truhjarta snart skræmte rettno osså juvik-bekken san gammelstu-take rettno hell fortæl dæg ikkj til dem att tart likar æg likar dem a fer te mein æg kviskra anders san dæg grov hæ einkvar jamgod Jens per slikt per	combs a features a lemmupos a upos  freq a 1 a 2 a 3 a 4 a 5 a 6 a 7



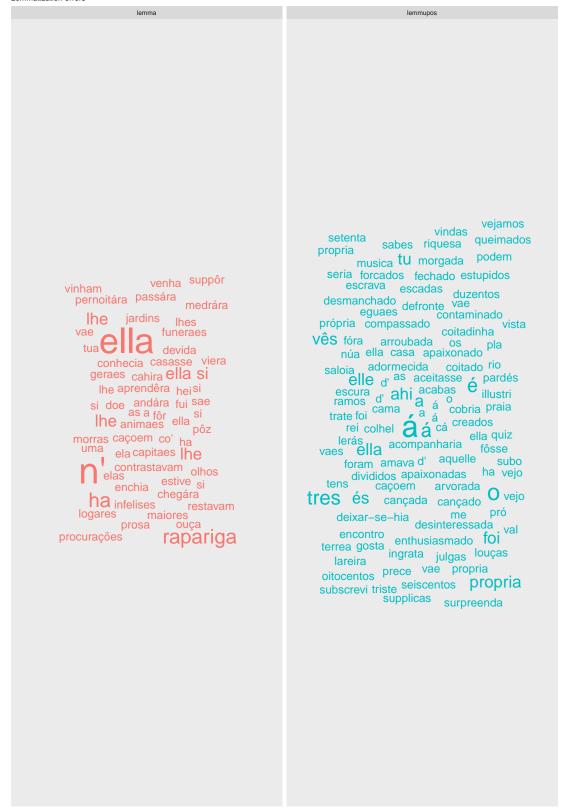
## Portuguese

```
i <- 5
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
   geom_text_wordcloud(show.legend = TRUE, shape = "square") +
   scale_size_area(max_size = 20, breaks = c(2,5,7,10)) +
   ggtitle(label = titles[i], subtitle = "All error types")
```



#### Portuguese Selected error types





combs a lemma

freq

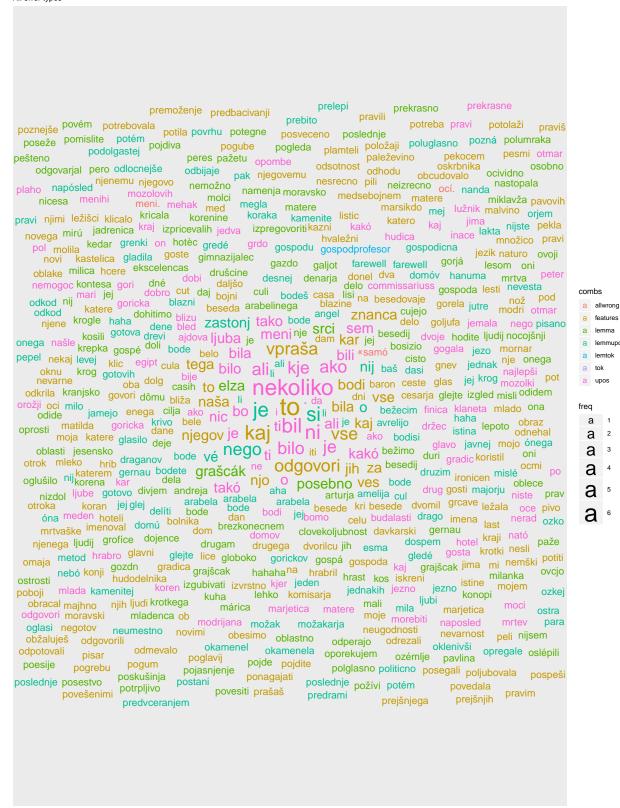
a lemmupos

a 2.5

### Slovene

There were too many erroneous words in Slovene for the word cloud to remain legible. We have therefore arranged the tokens in descending order according to frequency and truncated approx. 400 words (which have all occurred just once).

```
i <- 6
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit() %>%
    dplyr::arrange(desc(freq)) %>% dplyr::slice(1:600)
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
    geom_text_wordcloud(show.legend = TRUE, shape = "square") +
    scale_size_area(max_size = 12) +
    ggtitle(label = titles[i], subtitle = "All error types")
```



Slovene Selected error types I

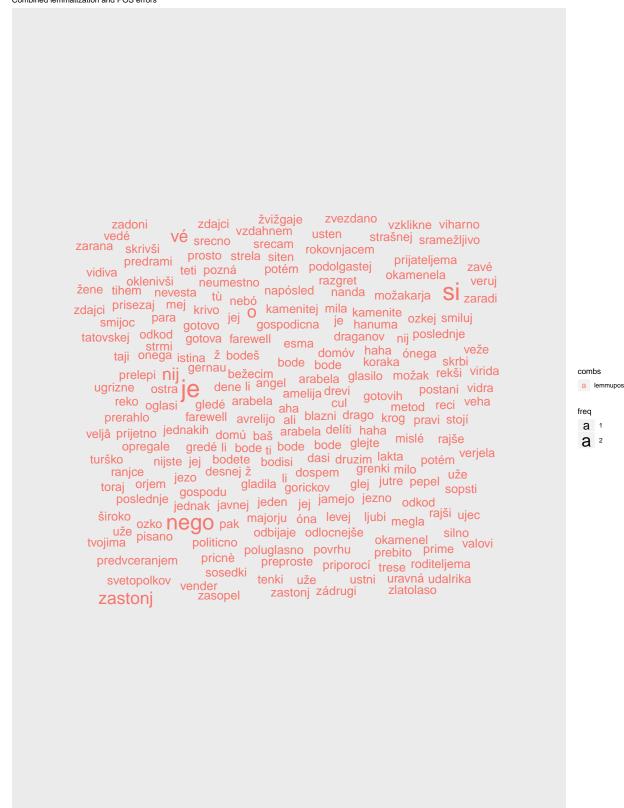


zmagálceva zbledi zapahi zganen zatajil zasoplim zaupljiva zemski zadosti vtepel vleglo vpraša vesela vega zánj verjetno vzlasti ubijmo toskane svakinja treba trobenta trenotek tebi prihodnjost resnicnega vól rudecelicni razpletel smodke skritemu prestrašila vari prenašala uroša prepiustite povém pojasnjenje rovtah pomislite odgovarjal prav poesije ozémlje vrag silno odperajo ocividno oblastno trdih stopal moravsko milica miklavža márica krotkega oni rekoc kedar kontesa polglasno mali korena izpricevalih lisi hahaha kuha izgled odide izgubivati požívi matere gospé drušcine tri globoko '2 nizdol skovir seže oni kos divjem denarja glejte dômu jej potegne jadrenica brezkonecnem daj delo gozdn mirú prevesela dné besedij je bosizio galjot jej ljudij poseže šteie nemožno hodite sveti onega krog gernau bežimo arturja dela grofice nje polumraka poslednje lesom gorjá casih andreja cujejo gospá namenja sosed poslednje lesom gorja casin dinara, odjoja kricala idji kricala sepec paže nij hotèc doli arabelinega cilja duri hrast ovcjo toge želja oprosti ljudij drugam besedij bode ona cisto gazdo misli prekrasno pojde konopi clovekoljubnost culi deje elza milanka priporod misli prekrasno gospoda si ekscelencas gimnazijalec svetel pero lepoto grajšcak grajšcak grašcak oblasti pavlina izpregovoriti korenine kosili jesensko ovoji sveti mladenca molci lehko kosili krogle tur kranjsko smijal oglušilo nastopala neizrecno nijsem nocojšnji sanjah oporekujem pažetu ostrosti peres pešteno ume pogleda prepricam posebno poskušinja posestvo povesiti primerno soseda prepricate rudecih spanjec sporocijo spozabila spremljevala tresla stegniti trenotek uzrok ubitega udati upotrebljujem veselo veslu vzbujenega zateleban vzrastel zanicljivo zarudel zaljubljen zvédela zvunaj zgenila znamenje





combs a lemma



Other ideas - Distribution of wrongly guessed upos - Compare distributions of features in contemporary texts

vs. our annot. samples as a proxy for recall