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
## 4 NYN        599      81       12        1        67
## 5 POR       5083     575      261       31       346
## 6 SLV       5081    1036      485        9       430
```

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

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*).

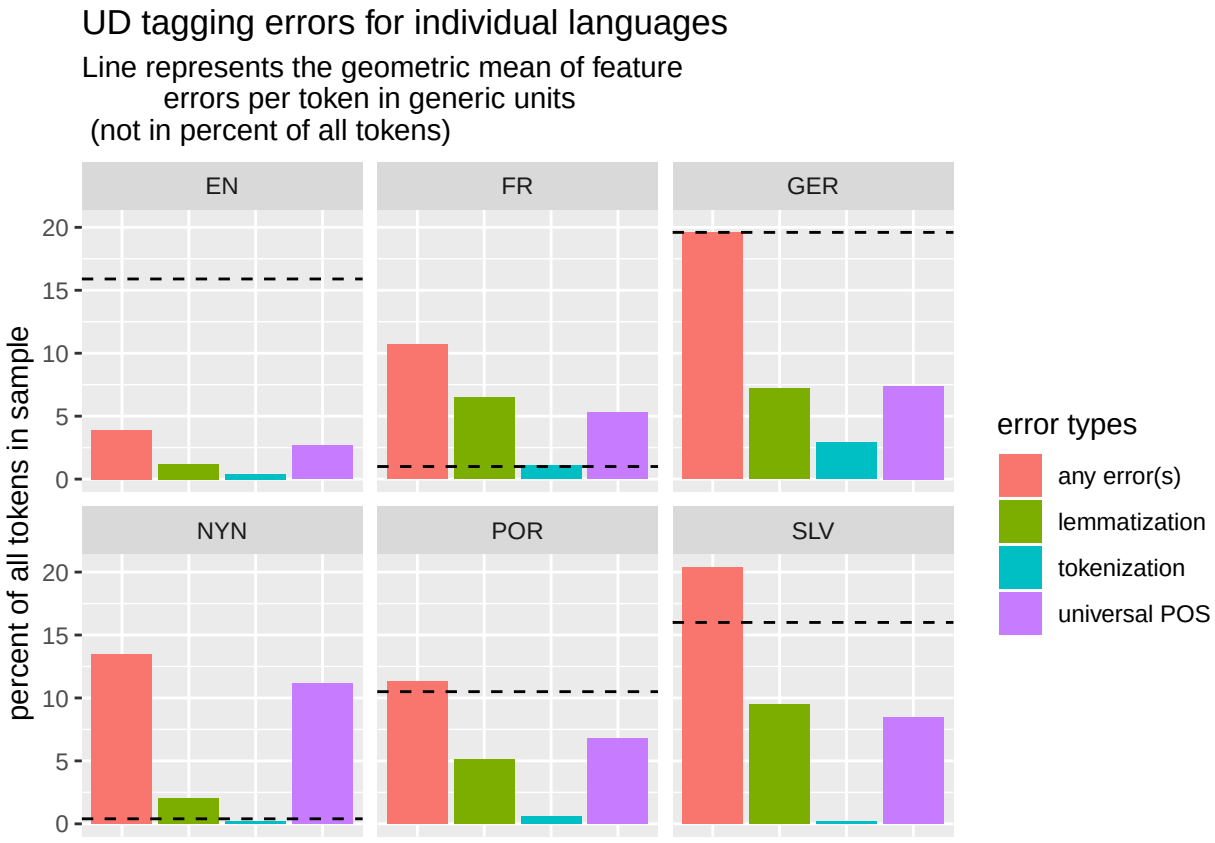
```
err_lines_01 %>% select(language, starts_with("prc"), starts_with("geom"))
```

```
## # A tibble: 6 x 6
##   language prc_errows prc_lem_err prc_tok_err prc_upos_err
##   <chr>         <dbl>       <dbl>       <dbl>        <dbl>
## 1 EN             3.90        1.20       0.400         2.70
## 2 FR            10.7         6.50       1.10          5.30
## 3 GER           19.6         7.20       2.90          7.40
## 4 NYN           13.5         2.00       0.200        11.2
## 5 POR           11.3         5.10       0.600         6.80
## 6 SLV           20.4         9.50       0.200         8.50
## # ... with 1 more variable: geomean_feat_err <dbl>
```

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)")
```

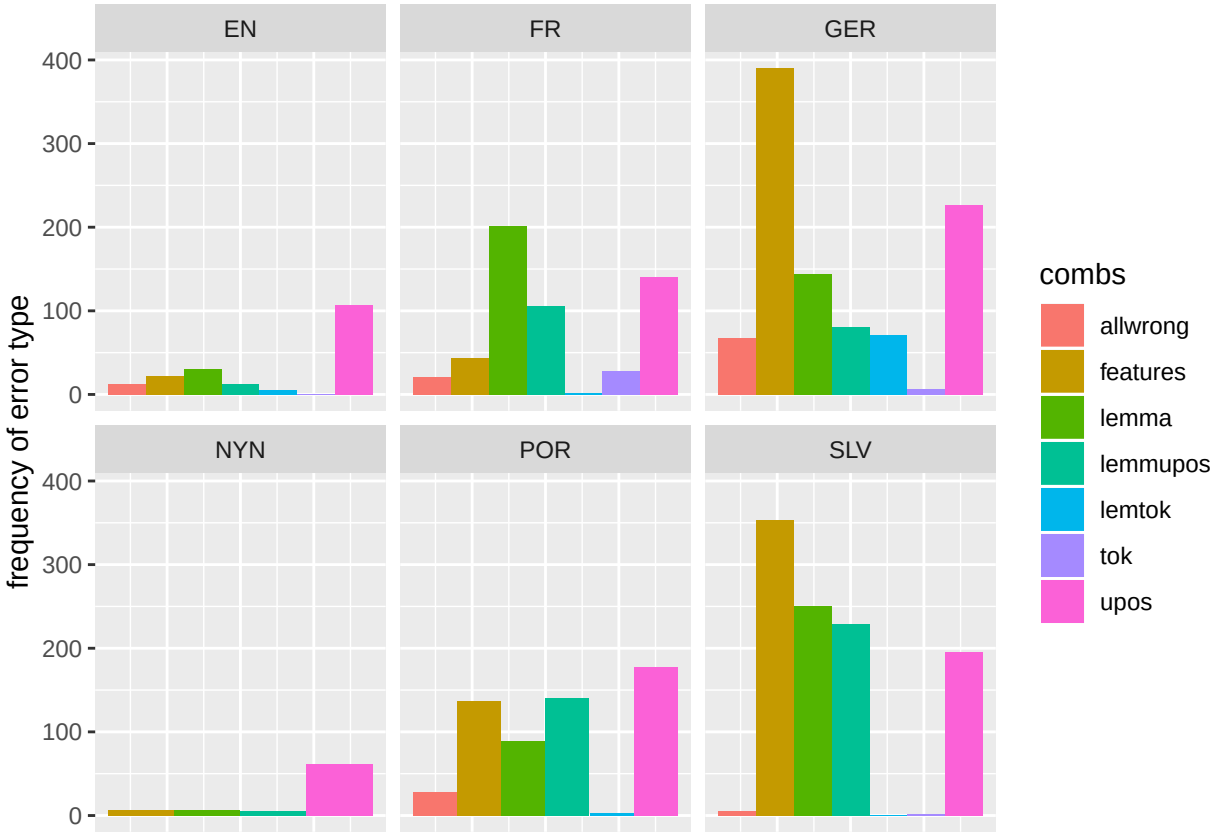


Comparison of languages according to different combinations of error types

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")
```

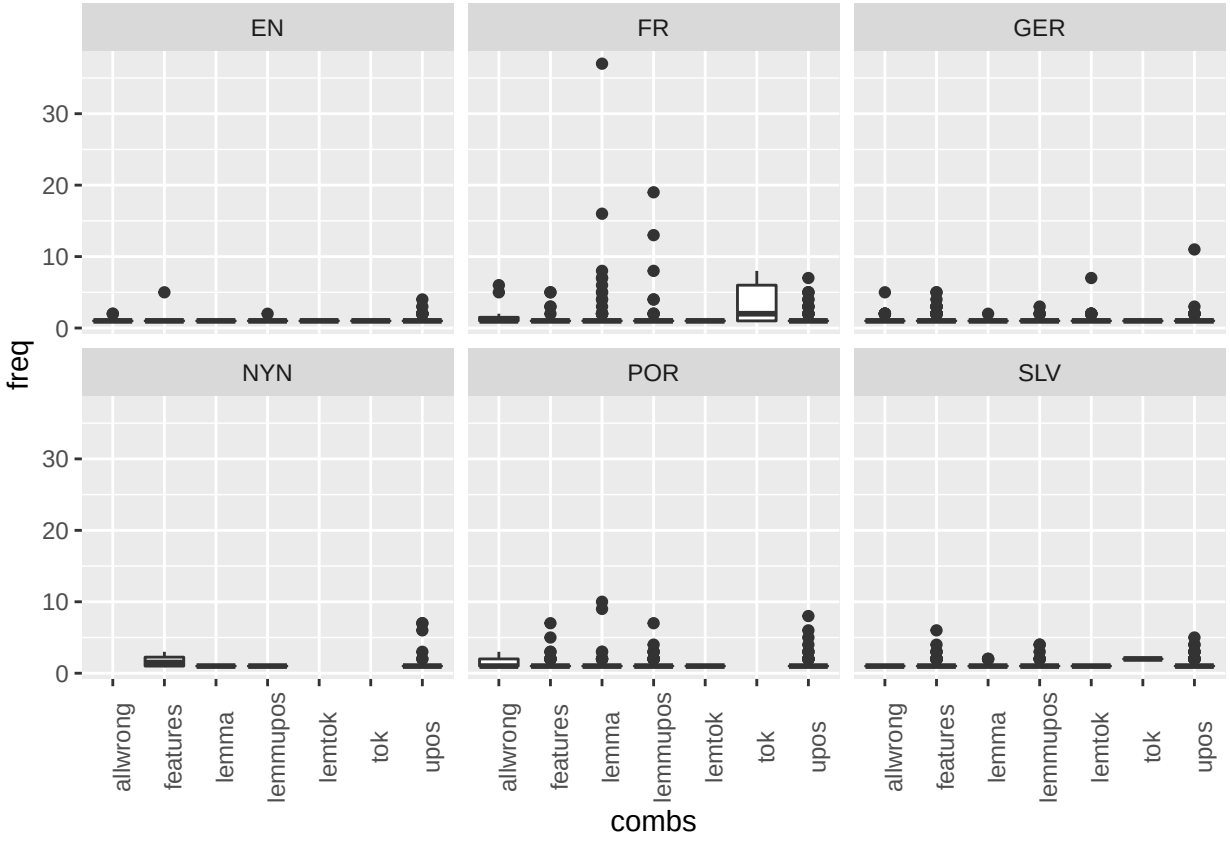


Distribution of errors across lemmas

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.

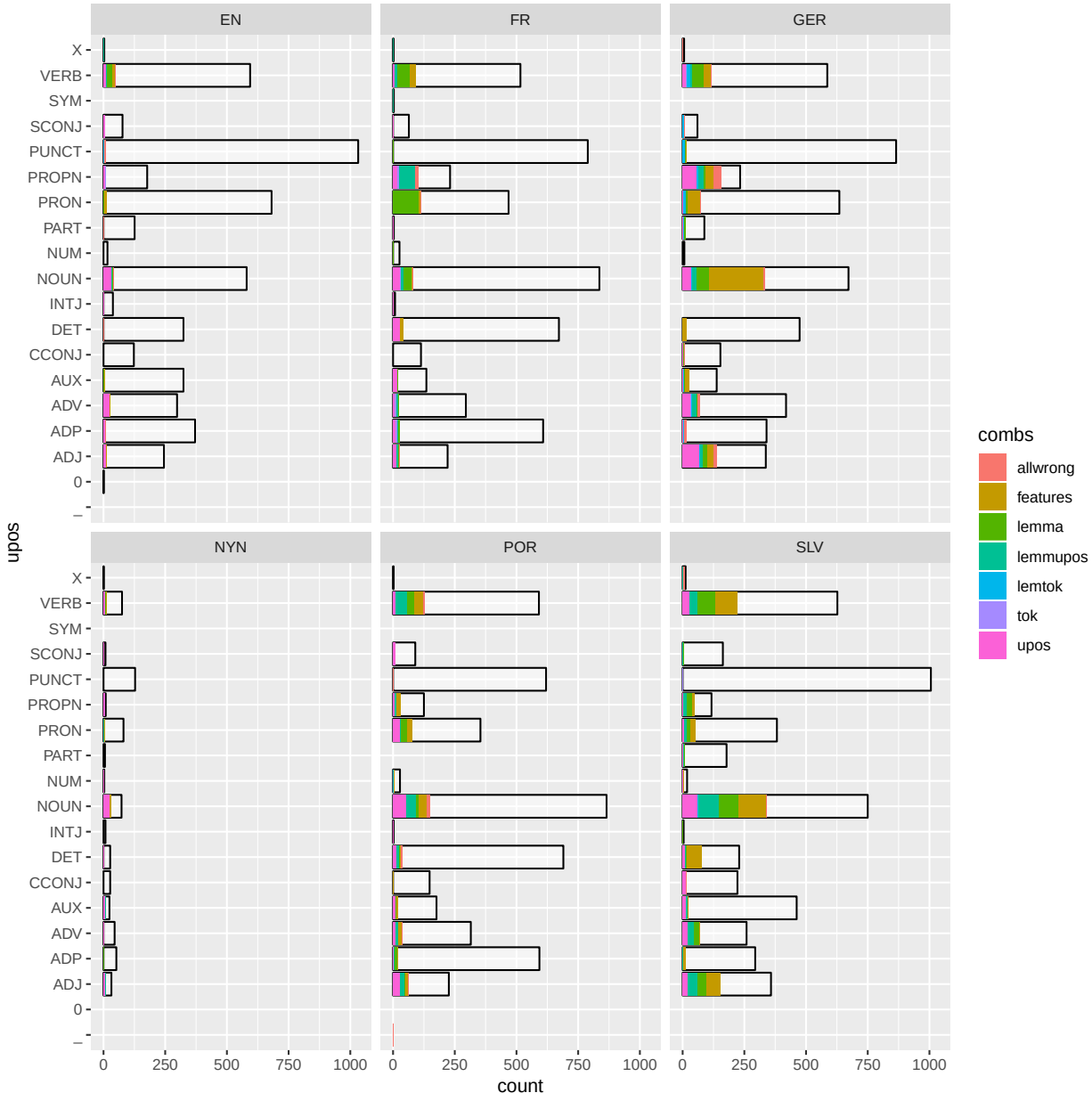
```
freqs <- select(err_freqs, language, combs, freq) %>% na.omit()
freqs$combs <- factor(freqs$combs)
levels(freqs$combs)[2] <- "features"

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



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.

```
ggplot(altok_df) +
  geom_bar(aes(x = upos), stat="count", alpha = 0.5, color = "black",
           fill = "white") +
  geom_bar(err_freqs[complete.cases(err_freqs),], mapping = aes(x = upos, fill = combs, y = freq),
           stat = "identity") +
  facet_wrap(~ language) +
  coord_flip()
```



Word clouds for individual languages

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

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.

English

EN - Across parts of speech

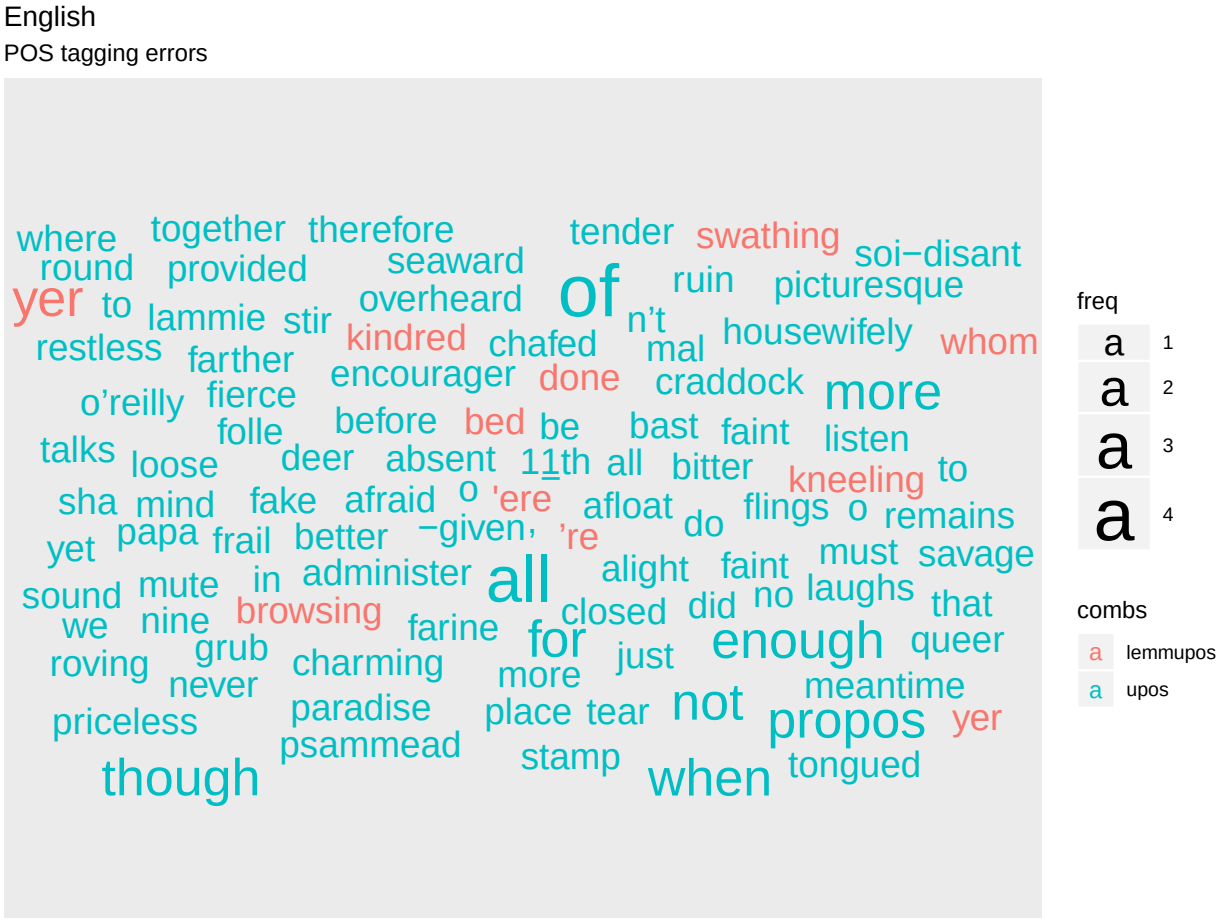
```
i <- 1
df <- dplyr::filter(err_freqs, language == vars[i]) %>% na.omit()
set.seed(123)
ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
  geom_text_wordcloud(show.legend = TRUE, shape = "square") +
  scale_size_area(max_size = 12) +
  ggtitle(label = titles[i], subtitle = "All error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


```
i <- 1
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


```
i <- 1
df <- dplyr::filter(err_freqs, language == vars[i], combs %in%
                      c("lemmupos", "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 = "POS tagging errors")
```



Function to extract different combinations of errors across all POS.

```
plot_errcombs_acrossPOS <- function(language, df = err_freqs, combs_vector,
                                    max_size = 12,
                                    shape = "square", show.legend = TRUE,
                                    label = "Language Sample",
                                    subtitle = "POS tagging errors") {
df <- dplyr::filter(df, language == language, combs %in%
                      combs_vector) %>% 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 = label, subtitle = subtitle) +
  facet_wrap(~ combs)

}
```

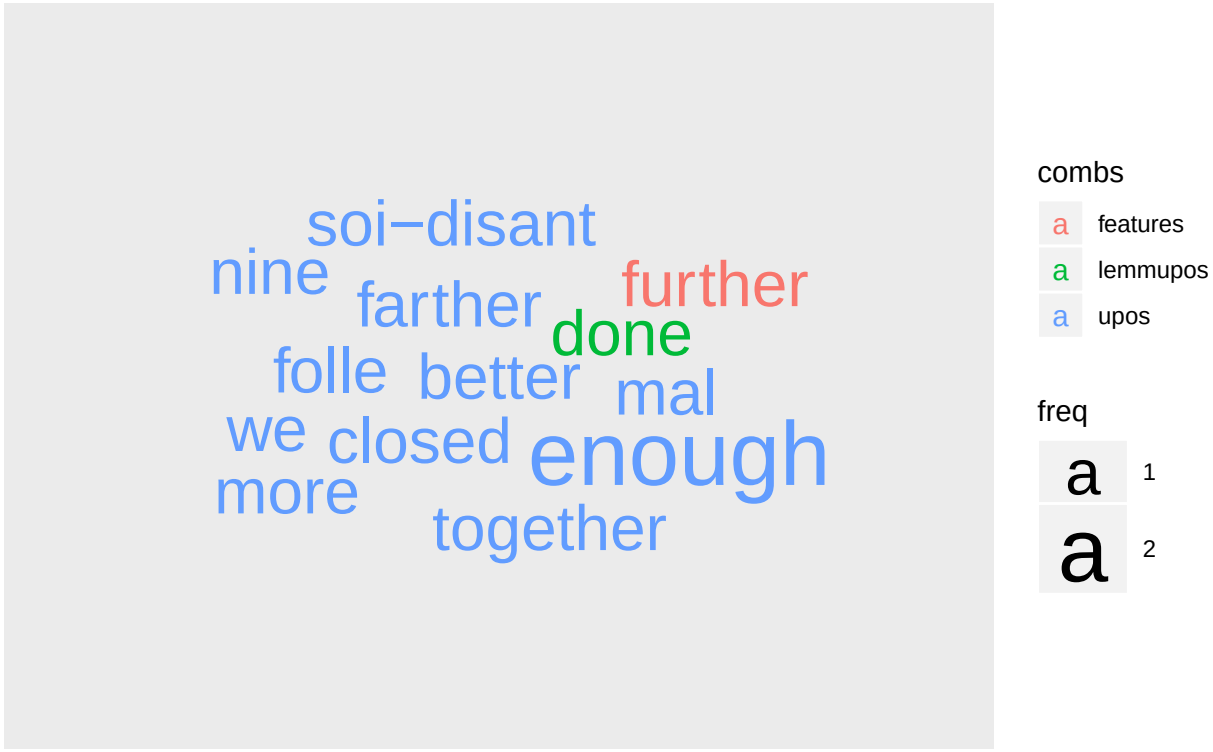
EN - Individual parts of speech

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], "\""))
```



Again, we extract a function to shorten the code.

```
plot_errcombs_POSwise <- function(lang, combs_vec = c(
  "allwrong", "features", "lemma", "lemtok", "lemupos", "tok","upos"),
  upos_vec = c("PUNCT", "ADV", "PRON", "AUX", "NOUN", "VERB", "PROPN", "DET",
               "ADP", "ADJ", "SCONJ", "INTJ", "PART", "X", "NUM", "SYM", "CCONJ",
```

```
                "_" ),
  max_size = 12, shape = "square", show.legend = TRUE,
  label = "Error combination in selected POS",
  subtitle = paste0("Errors in POS \"", upos_vec, "\"")) {
  df <- dplyr::filter(err_freqs, language == lang, upos %in% upos_vec,
                      combs %in% combs_vec) %>%
    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 = label, subtitle = subtitle)
}
```

```
i <- 1
y <- 3

plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
```

English

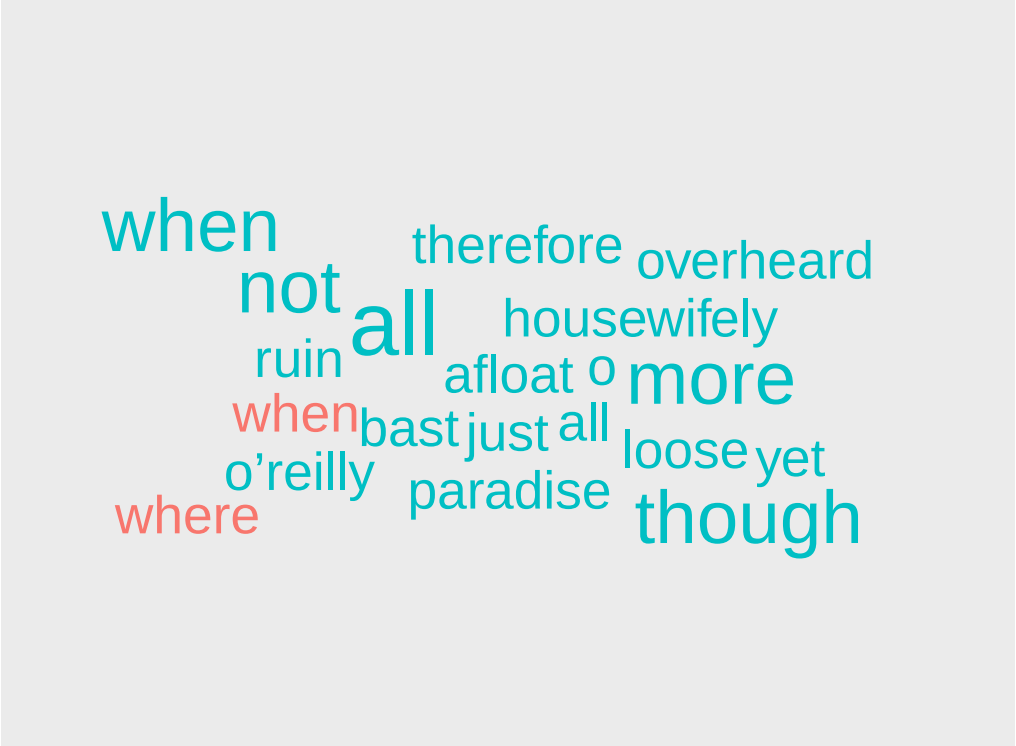
Errors in POS "ADP"


```
# 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])
```

English

Errors in POS "ADV"

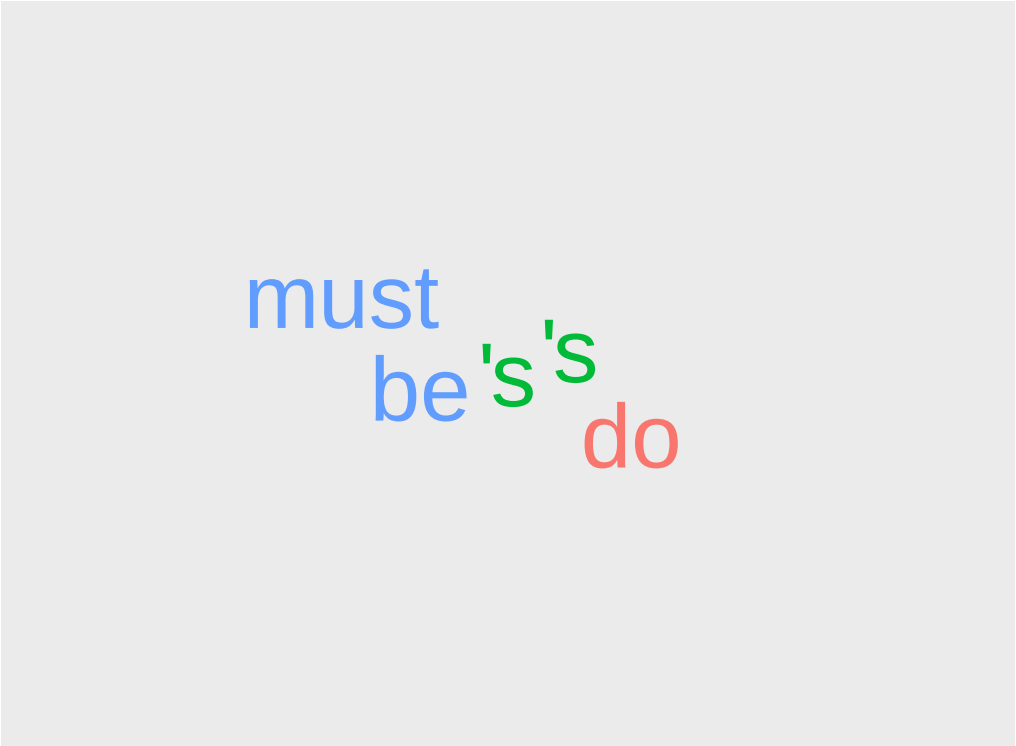



```
# df <- dplyr::filter(err_freqs, language == "EN", upos == ups[y]) %>% na.omit()
# set.seed(123)
# ggplot(df, aes(label = lower_form, color = combs, size = freq)) +
#   geom_text_wordcloud(show.legend = TRUE, shape = "square") +
#   scale_size_area(max_size = 12) +
#   ggtitle(label = titles[i],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

```
i <- 1
y <- 5
plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
```

English

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 = paste0("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])
```

English

Errors in POS "DET"


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

English

Errors in POS "INTJ"


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

English

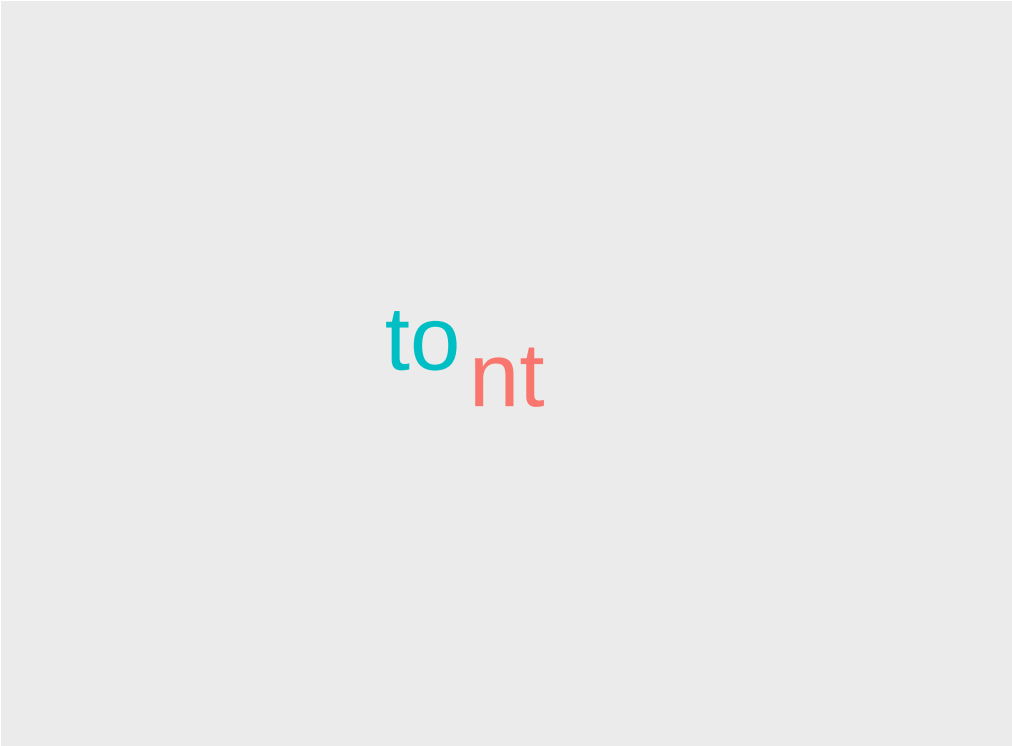
Errors in POS "NOUN"


```
#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])
```

English

Errors in POS "PART"




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

English

Errors in POS "PRON"


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

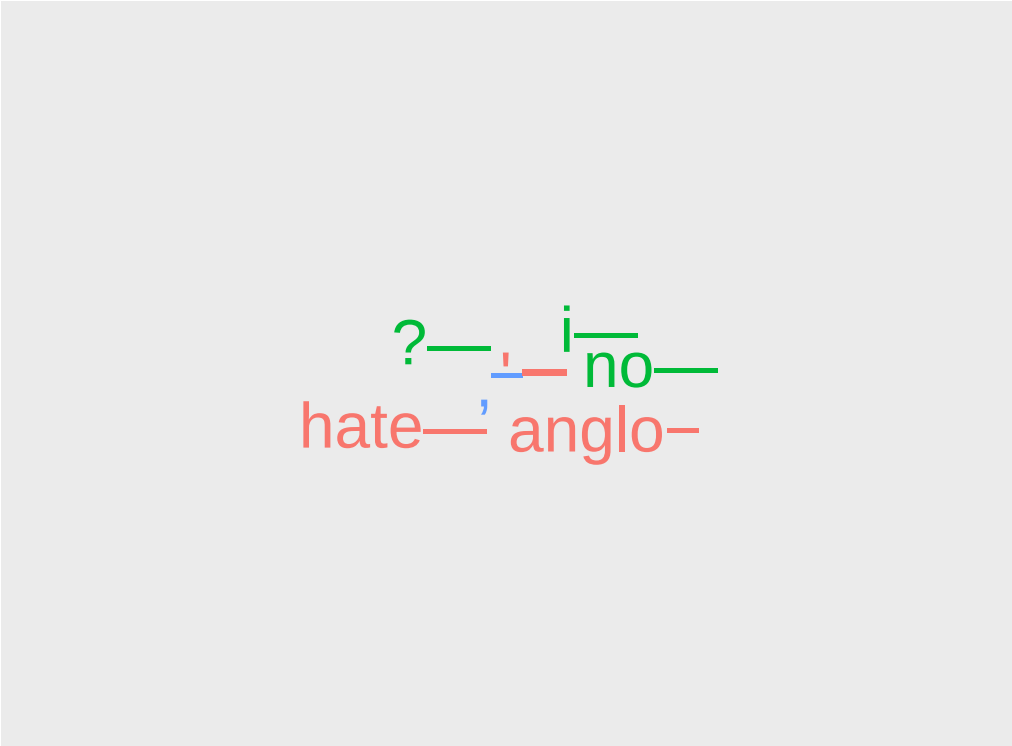
English

Errors in POS "PROPN"


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

English

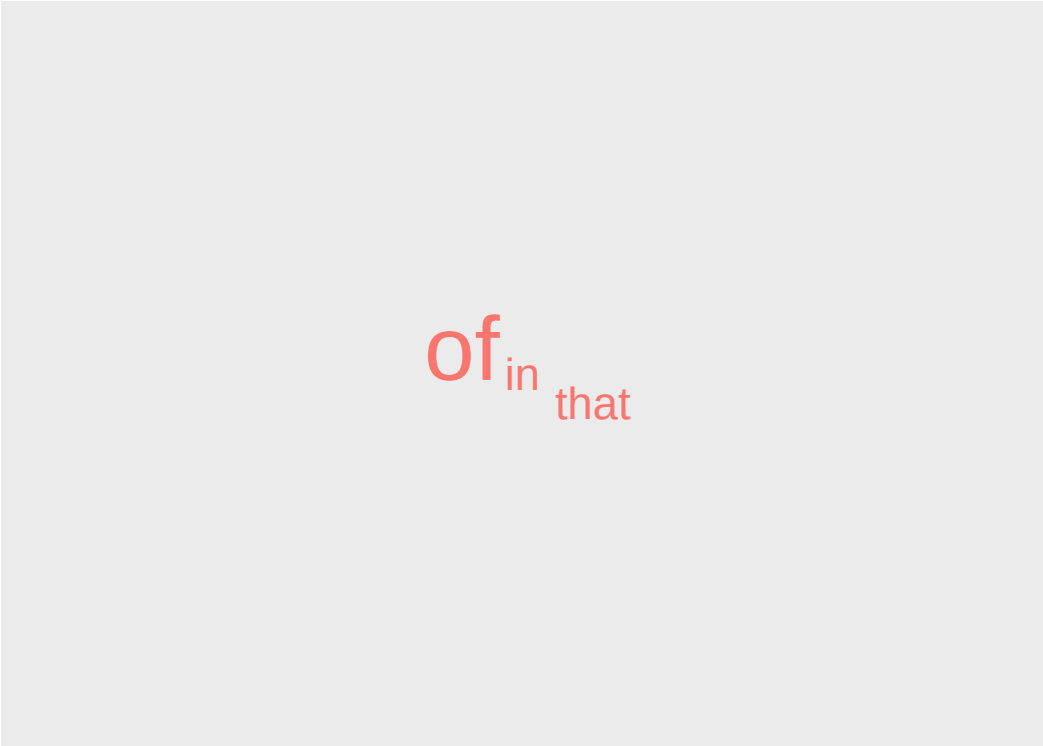
Errors in POS "PUNCT"




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

English

Errors in POS "SCONJ"




```
# 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])
```

English

Errors in POS "VERB"

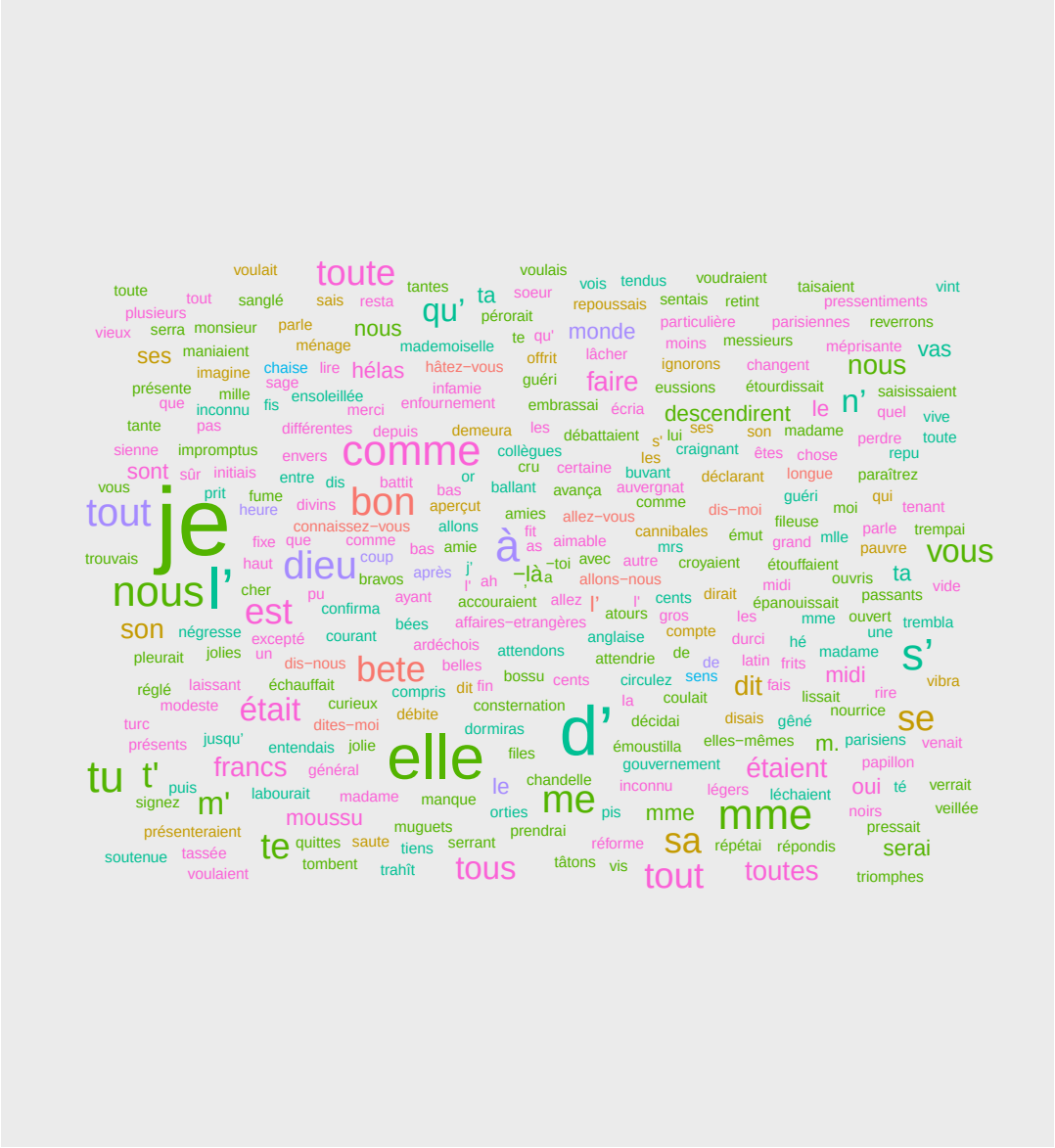

```
# No errors in English X
# i <- 1
# y <- 18
# plot_errcombs_POSwise(lang = vars[i], upos_vec = ups[y], label = titles[i])
```

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

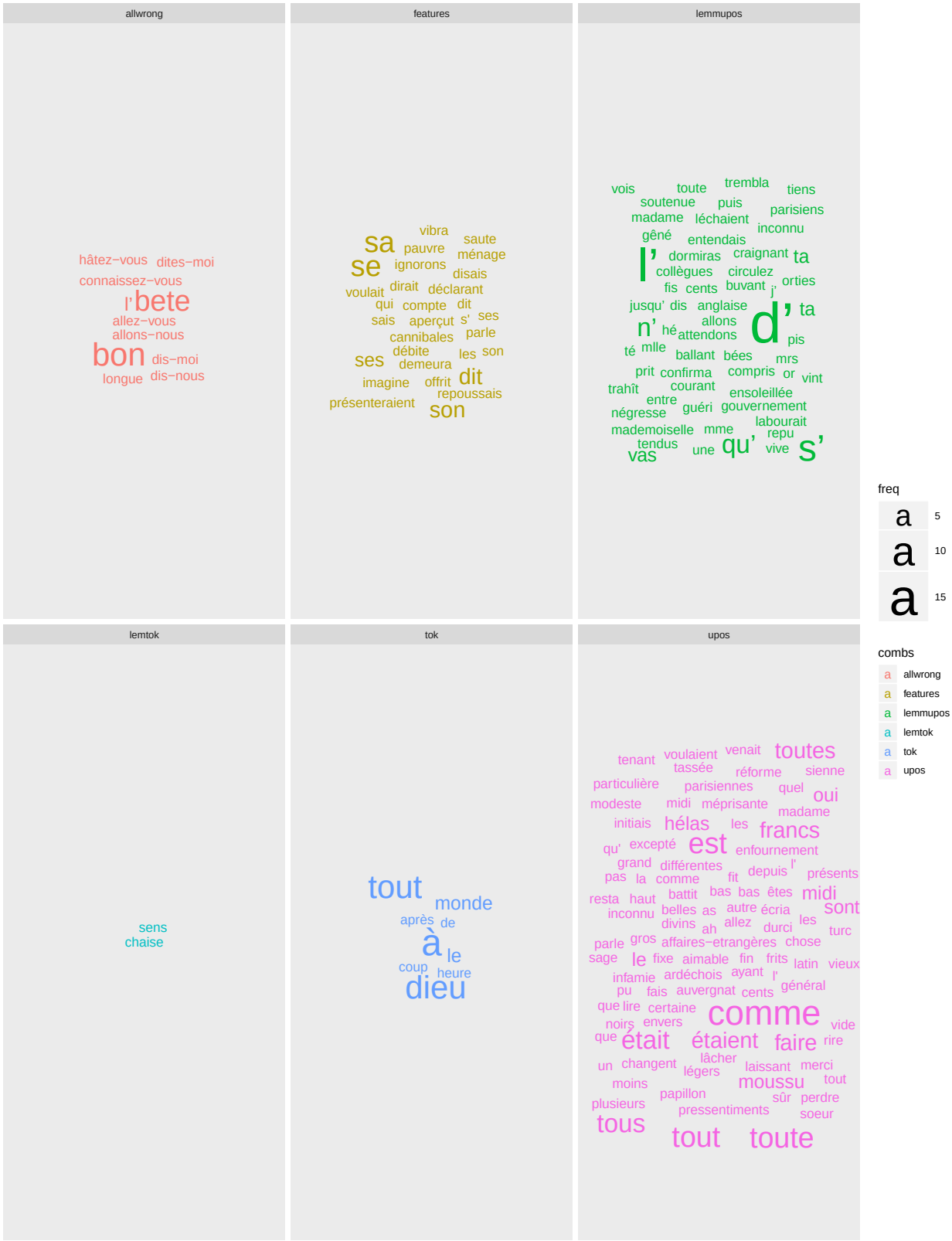
All error types




```
i <- 2
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("allwrong", "features", "lemmupos","lemtok",
                                 "tok", "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 = 18) +
  ggtitle(label = titles[i], subtitle = "Selected error types") +
  facet_wrap(~ combs)
```

French

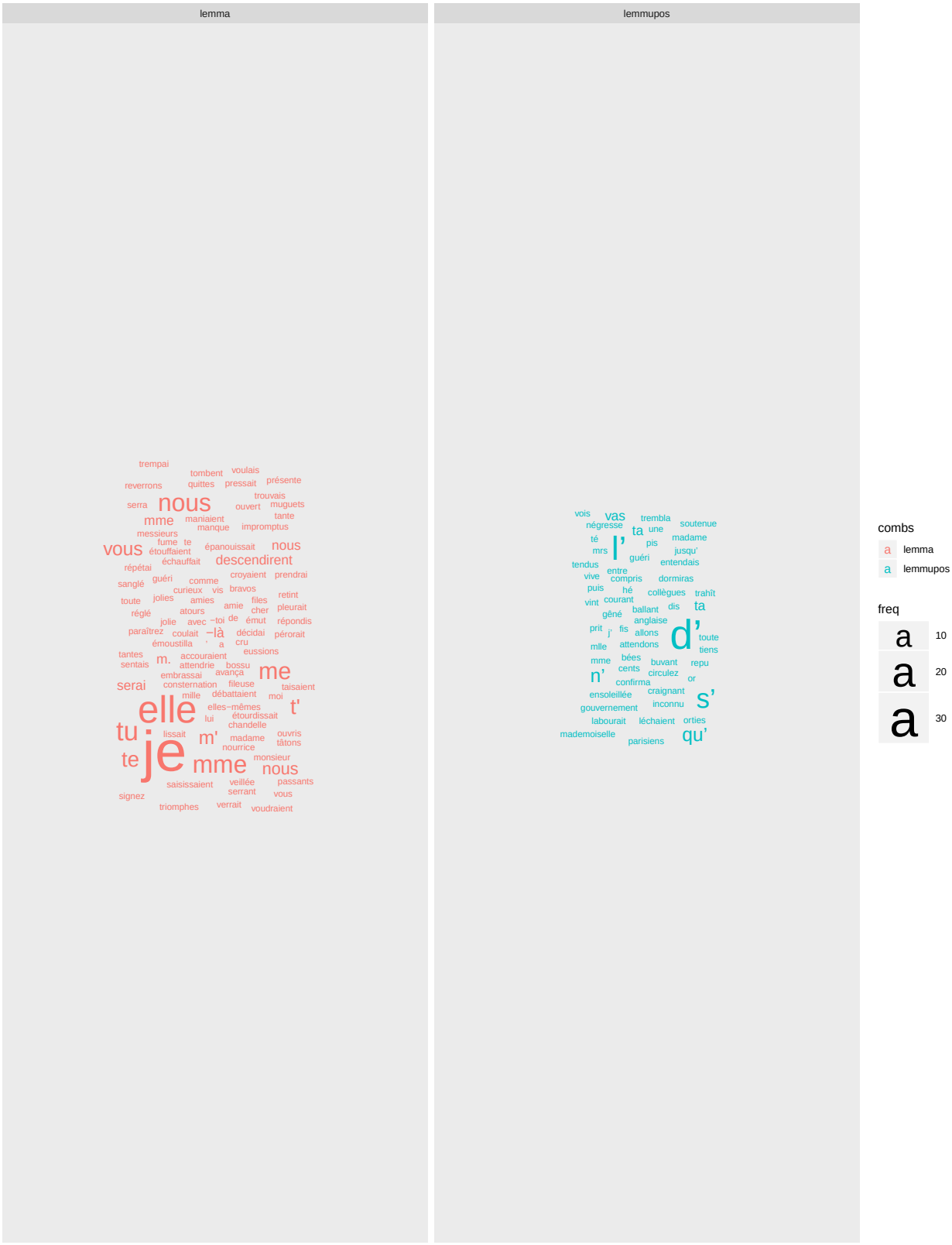
Selected error types



```r
i <- 2
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("lemma", "lemmupos")) %>%
  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 = 18) +
  ggtitle(label = titles[i], subtitle = "Lemmatization errors") +
  facet_wrap(~ combs)
```

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])
```


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

French

Errors in POS "ADP"

combs

a	lemma
a	tok
a	upos

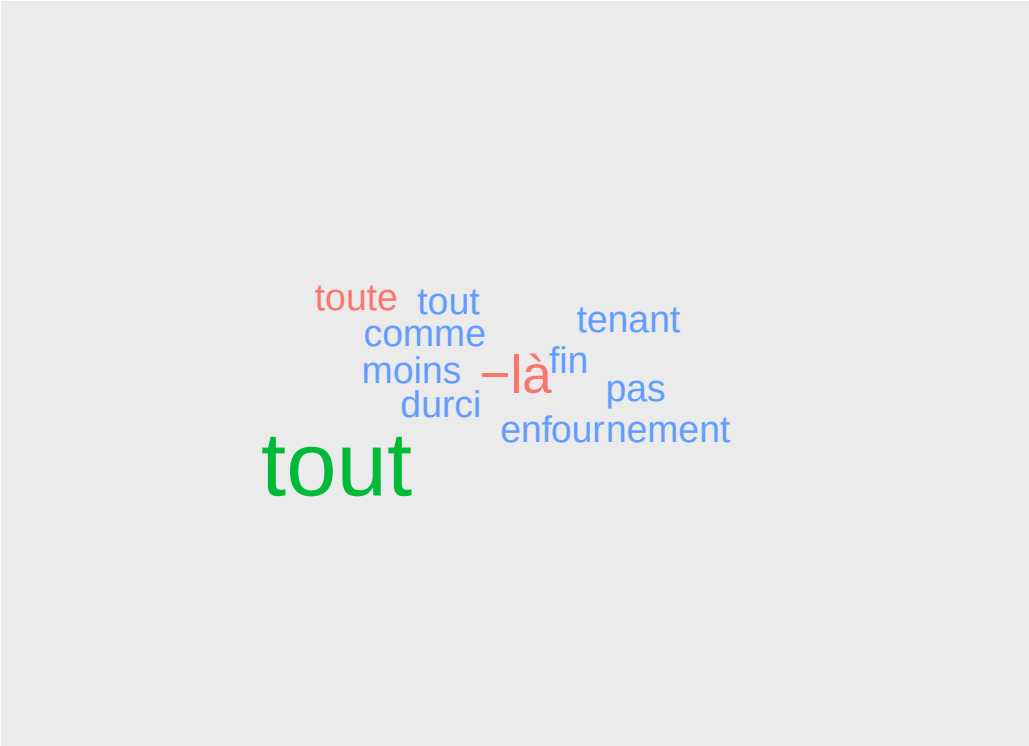
freq

a	1
a	2
a	3
a	4
a	5
a	6
a	7
a	8

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

French

Errors in POS "ADV"




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

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])
```

French

Errors in POS "DET"


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

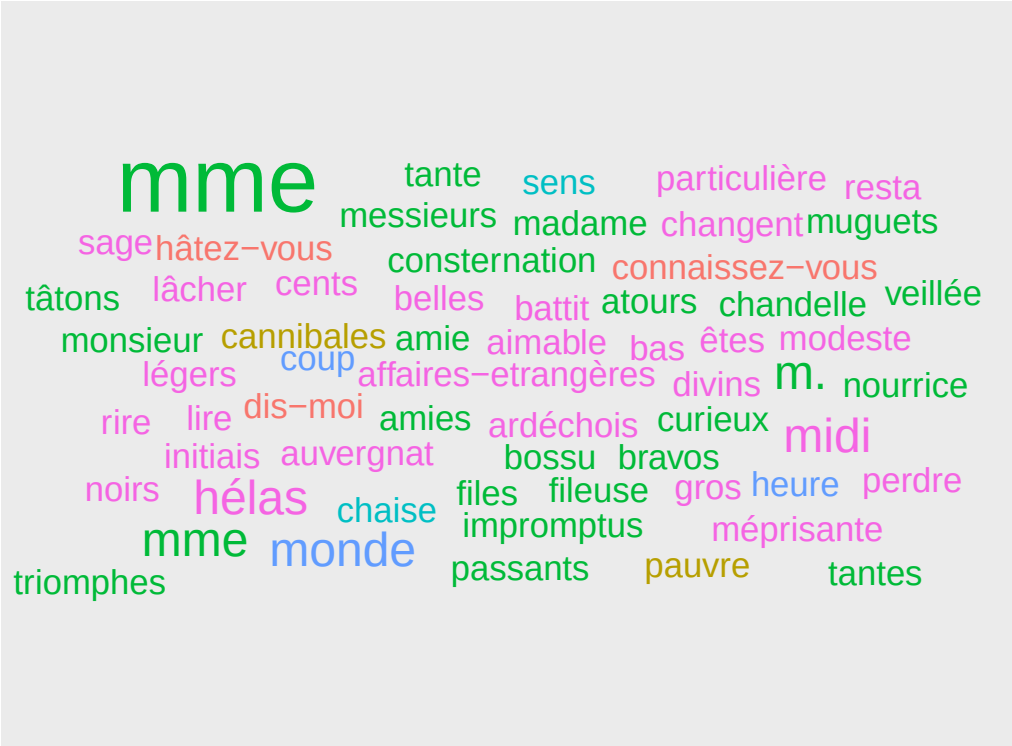
French

Errors in POS "INTJ"


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

French

Errors in POS "NOUN"



combs

a	allwrong
a	features
a	lemma
a	lemtok
a	tok
a	upos

freq

a	1
a	2
a	3
a	4
a	5
a	6

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

French

Errors in POS "NUM"


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

French

Errors in POS "PART"


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

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])
```

French

Errors in POS "PROPN"


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

French

Errors in POS "PUNCT"


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

French

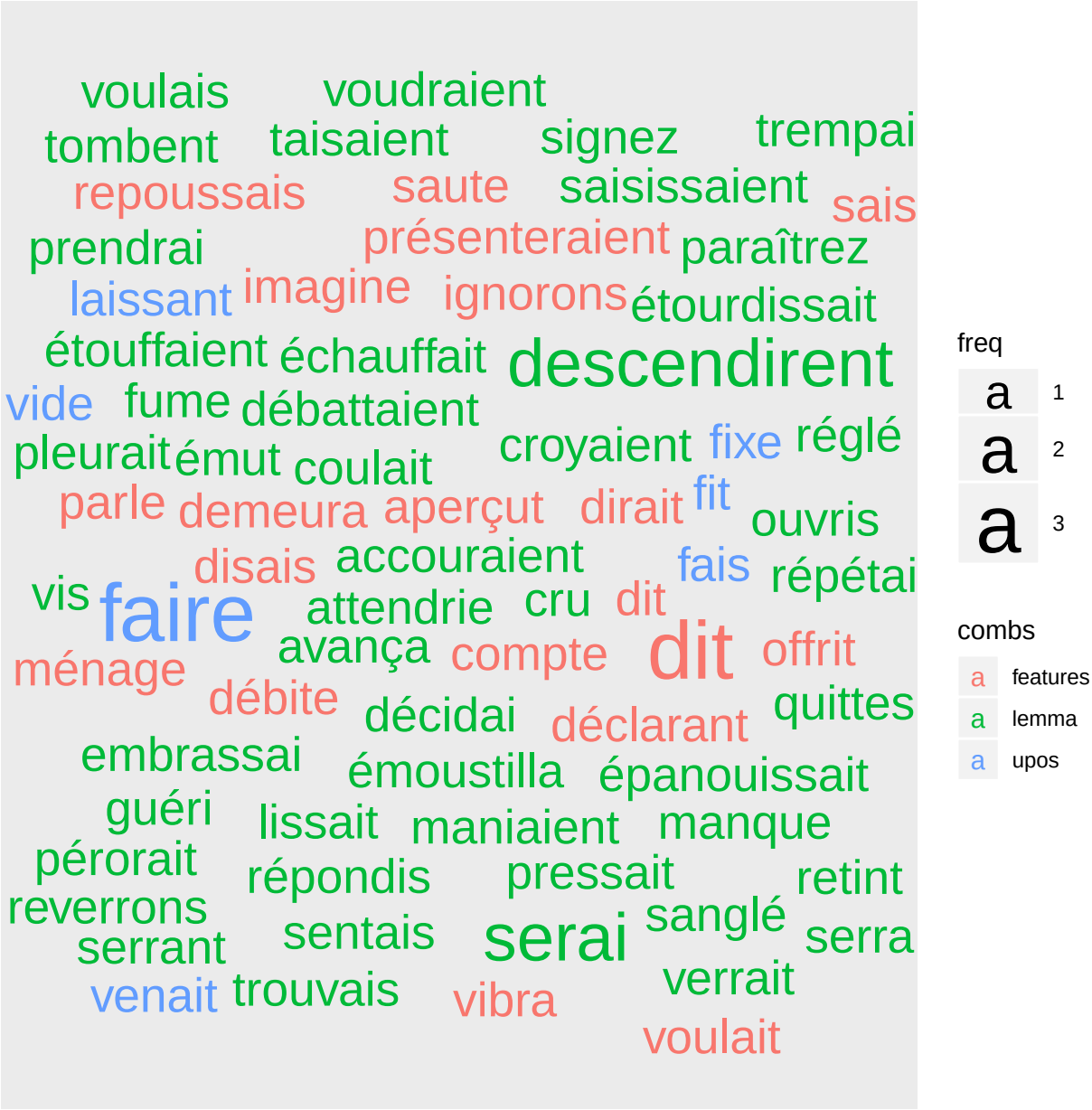
Errors in POS "SCONJ"


```
# 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])
```

French

Errors in POS "VERB"




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

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

All error types

```
i <- 3
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("allwrong", "lemmupos","lemtok",
                                 "tok")) %>%
  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 I") +
  facet_wrap(~ combs)
```

German

Selected error types I

```r
i <- 3
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("features", "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 II") +
  facet_wrap(~ combs)
```

German

Selected error types II

```r
i <- 3
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("lemma", "lemmupos")) %>%
  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 = 1:3) +
  ggtitle(label = titles[i], subtitle = "Lemmatization errors") +
  facet_wrap(~ combs)
```

German

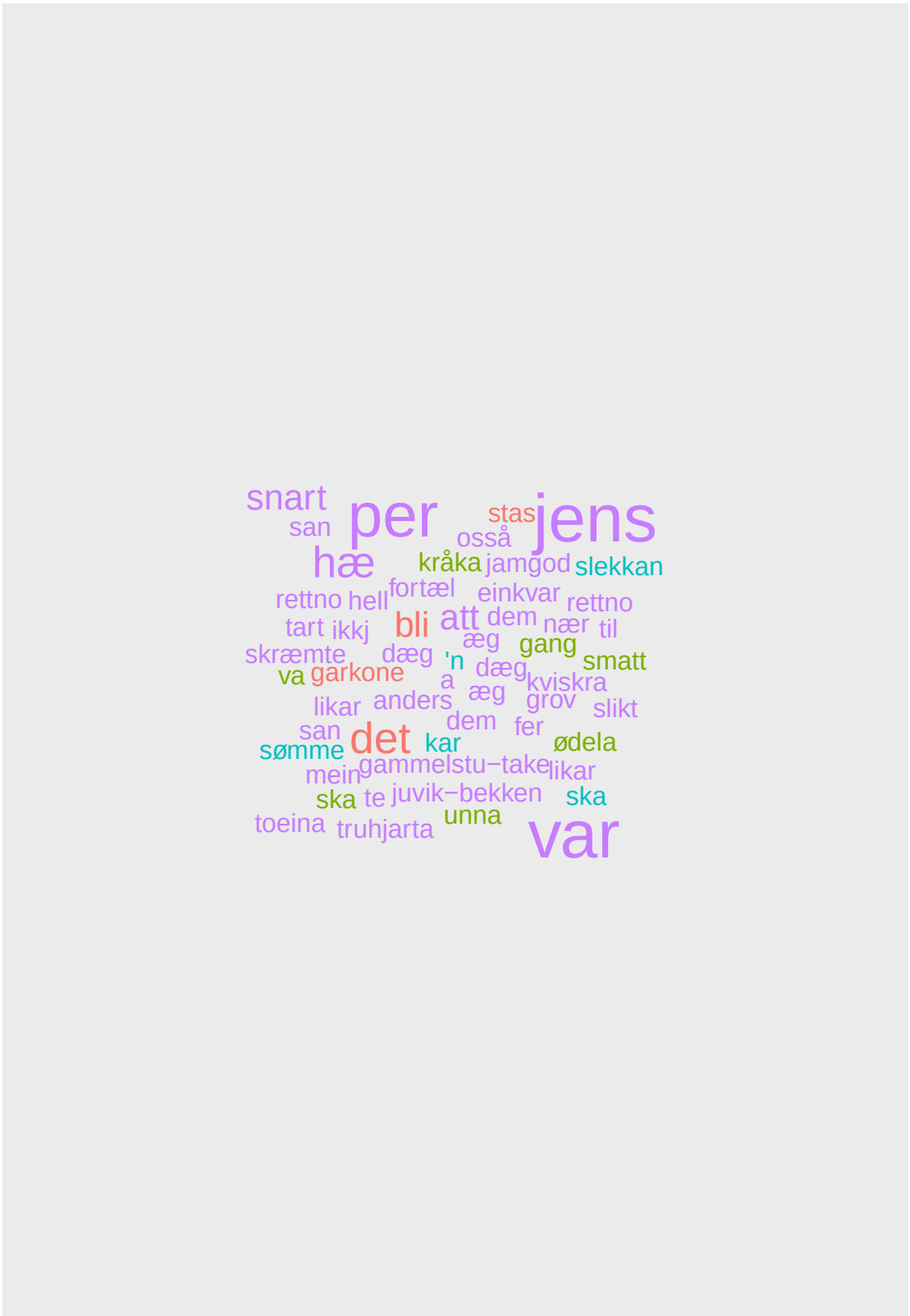
Lemmatization errors

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

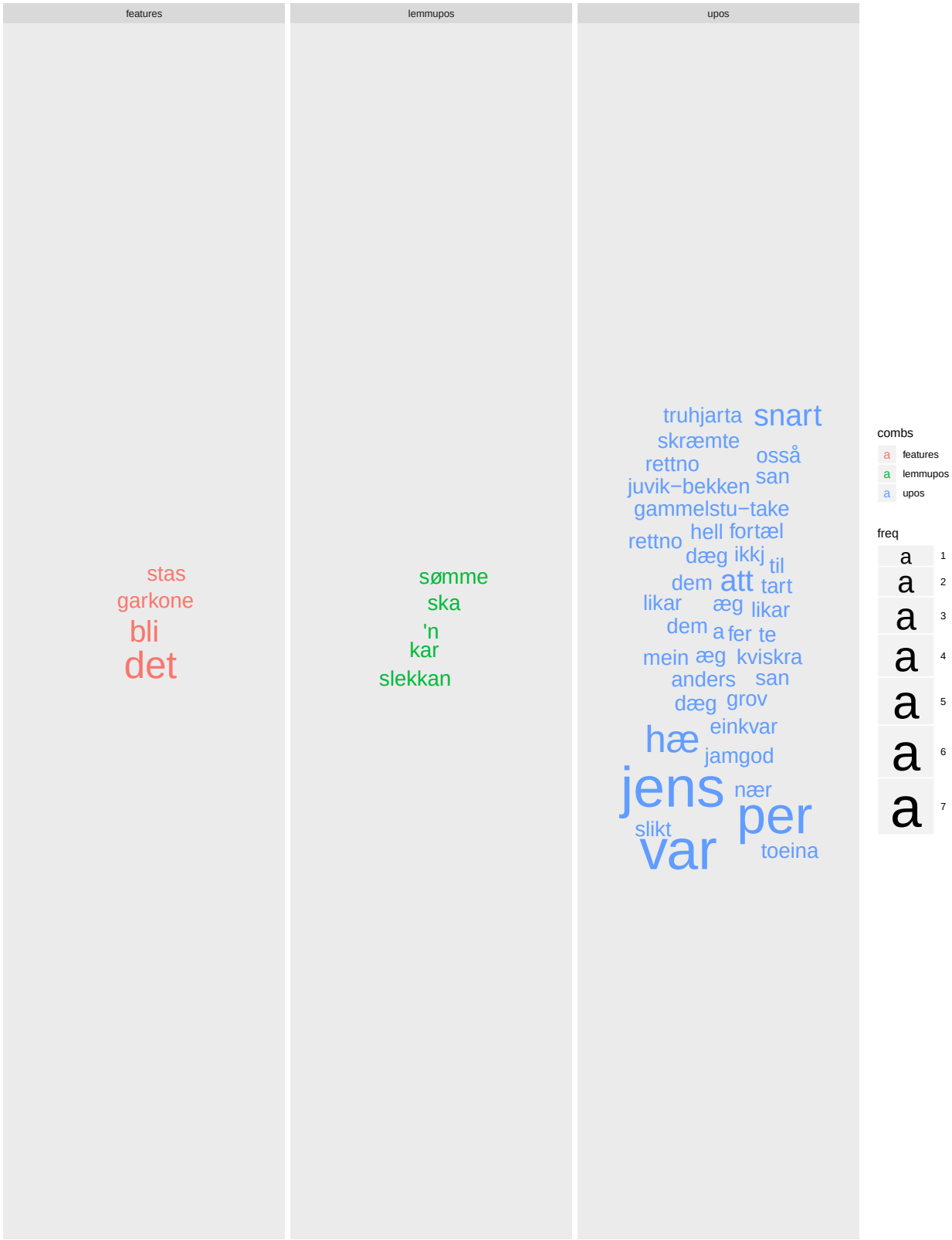
All error types



```
i <- 4
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("allwrong", "features", "lemmupos","lemtok",
                                 "tok", "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 = 18) +
  ggtitle(label = titles[i], subtitle = "Selected error types") +
  facet_wrap(~ combs)
```

Nynorsk

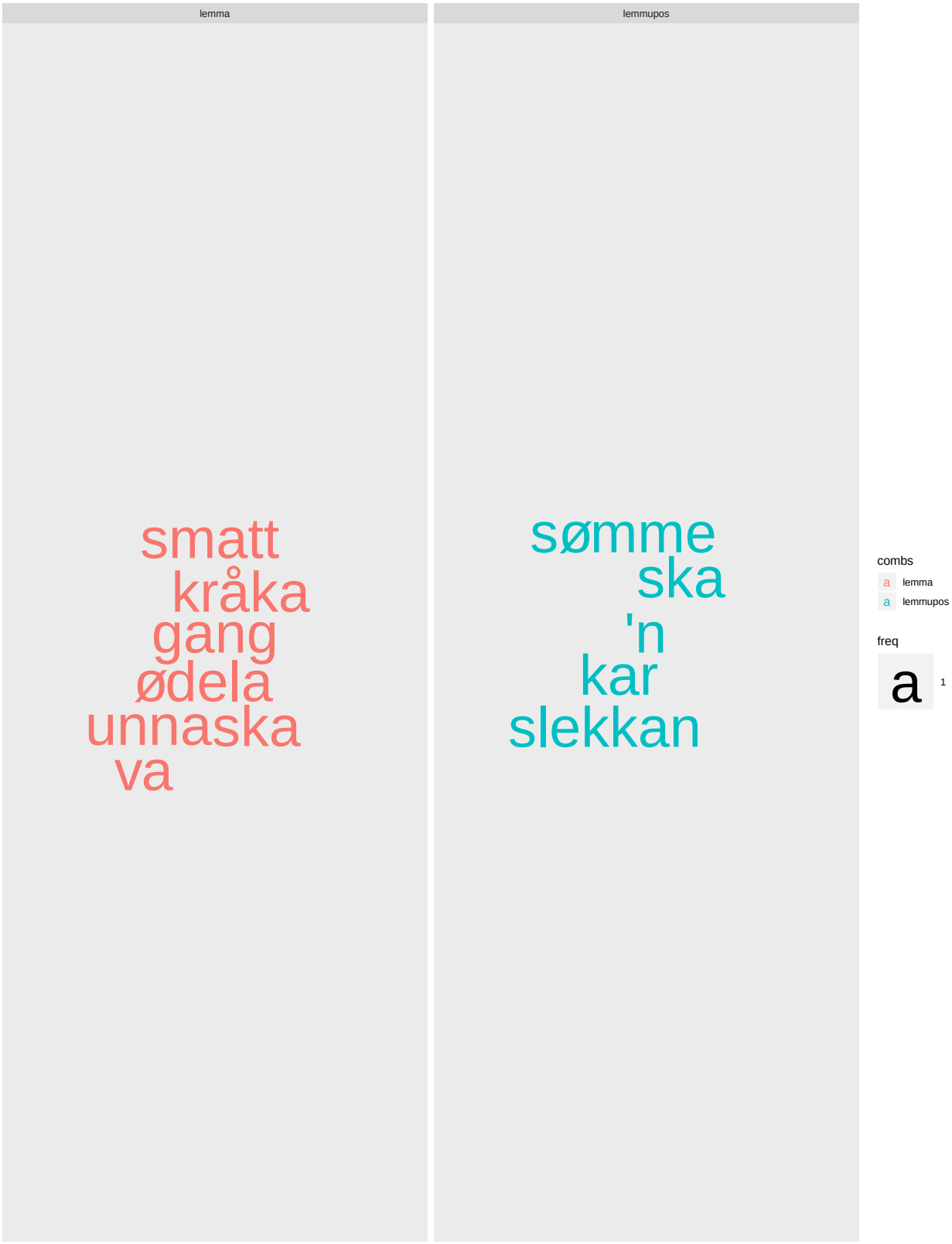
Selected error types



```r
i <- 4
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("lemma", "lemmupos")) %>%
  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 = 18) +
  ggtitle(label = titles[i], subtitle = "Lemmatization errors") +
  facet_wrap(~ combs)
```

Nynorsk

Lemmatization errors



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

All error types

```
i <- 5
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("allwrong", "features", "lemmupos","lemtok",
                                 "tok", "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") +
  facet_wrap(~ combs)
```

Portuguese

Selected error types

```r
i <- 5
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("lemma", "lemmupos")) %>%
  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 = 18) +
  ggtitle(label = titles[i], subtitle = "Lemmatization errors") +
  facet_wrap(~ combs)
```

Portuguese

Lemmatization errors

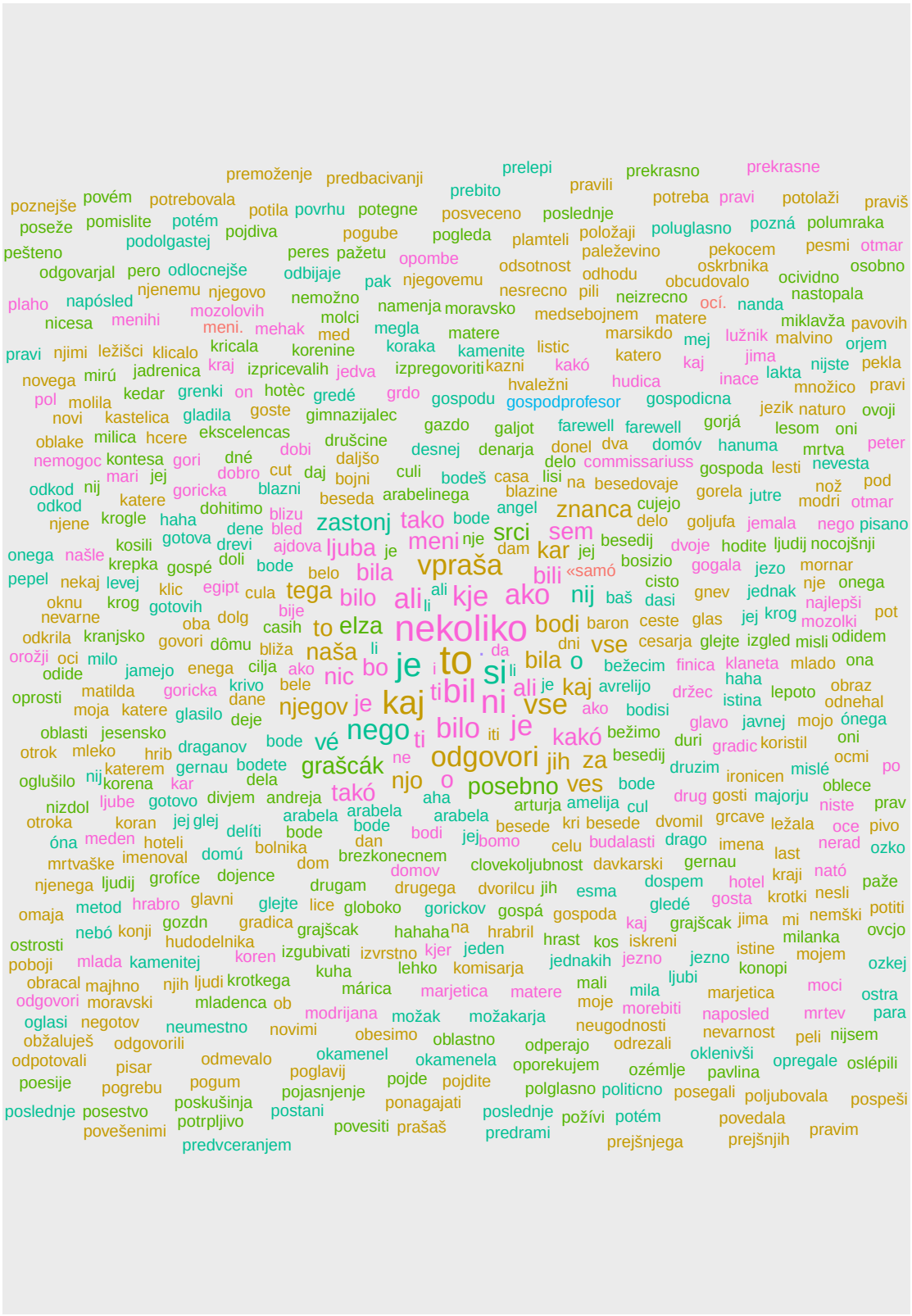
Slovene

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

All error types



```
i <- 6
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("allwrong", "lemtok",
                                 "tok", "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 I") +
  facet_wrap(~ combs)
```

Slovene

Selected error types I

```r
i <- 6
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("lemma")) %>%
  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 = 1:2) +
  ggtitle(label = titles[i], subtitle = "Lemmatization errors")
```

Slovene

Lemmatization errors

```r
i <- 6
df <- dplyr::filter(err_freqs, language == vars[i],
                    combs %in% c("lemmupos")) %>%
  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 = 1:2) +
  ggtitle(label = titles[i], subtitle = "Combined lemmatization and POS errors")
```

Slovene

Combined lemmatization and POS errors

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

vs. our annot. samples as a proxy for recall