MUL in field interviews

Daniil Ignatiev

4 June 2021

MUL & MSL in field interviews

Abstract

The paper explores several linguistic questions, making use of the RSUH folklore archive. The first of them is how traditional sociolinguistic variables, namely the mean utterance length and the mean sentence, length relate to the gender of the speaker and to the gender of the addressee. The second goal is to scrutinize, how well we can study the former matter, given the current structure of the corpus, and to find out, what the corpus currently lacks in this respect.

Introduction

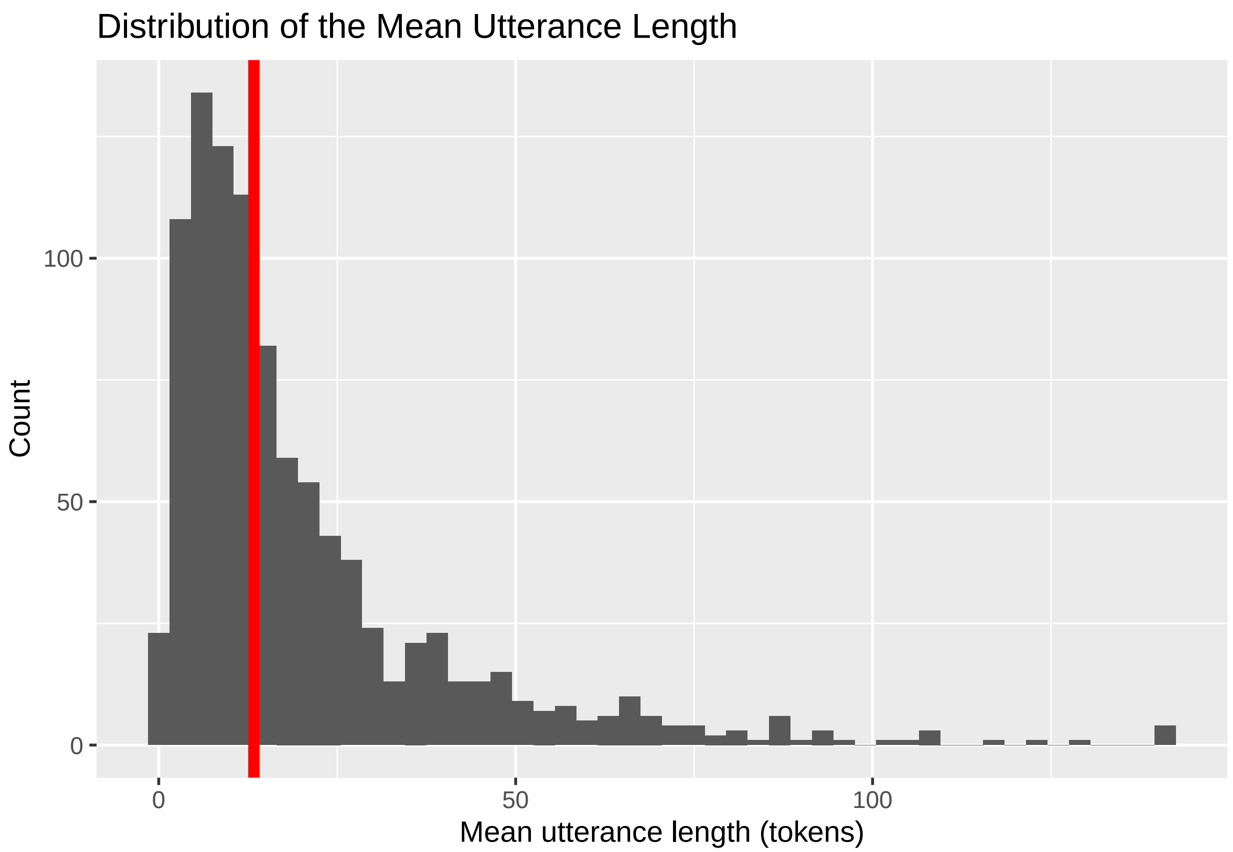
Mean utterance length (MUL) and mean sentence length (MSL) are both traditional variables that have been excessively studied in relation to sociolinguistic factors, including gender. What makes a study of the possible difference between genders especially interesting, is the stereotype than women generally talk more than men do. This motivated a lot of researchers to address the matter on the material of different languages (see Daniel & Zelenkov 2012, Tannen 1990). **In our experiments we continue this trend and seek to test statistically, whether men or women speak in longer passages by comparing the mean utterance length. Additionally, we measure the length of sentences to see, if respondents of a certain gender tend to employ lengthier sentences in their speech**.

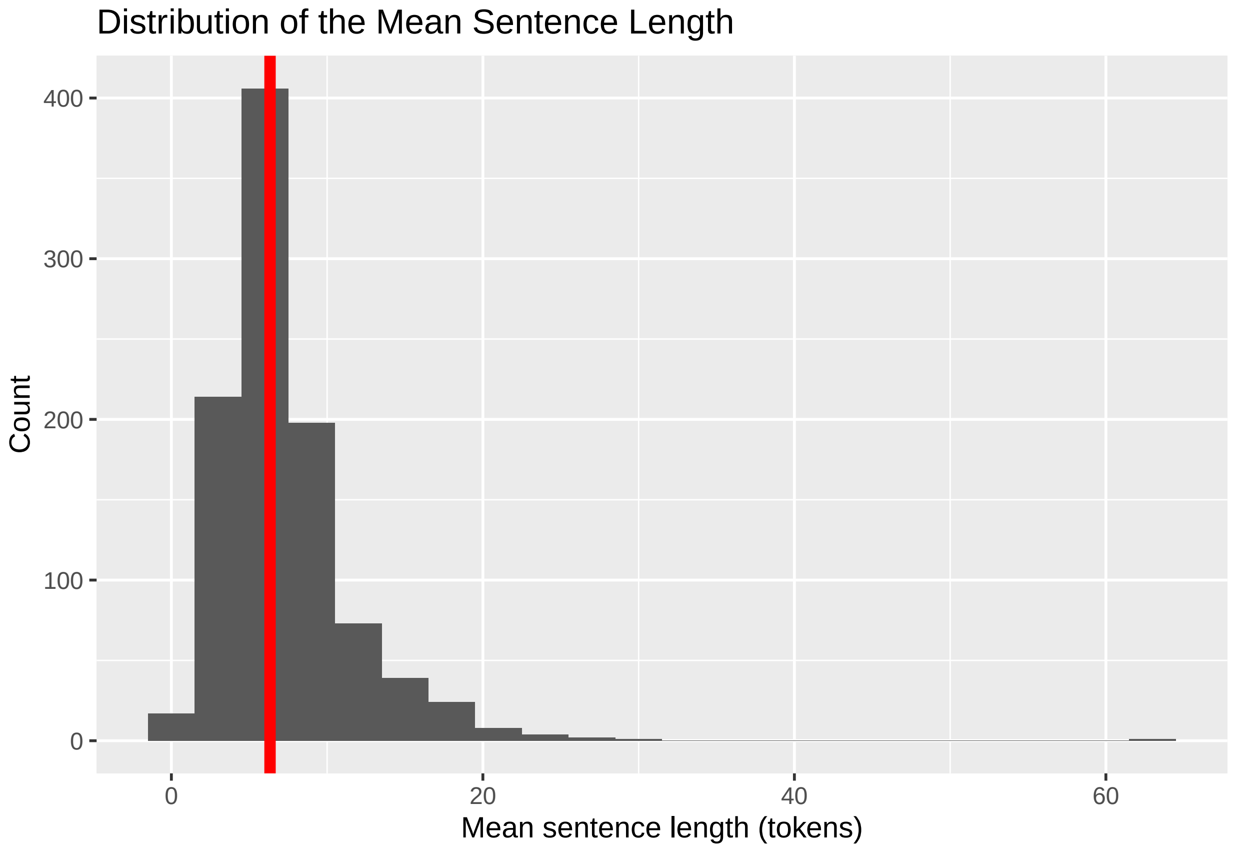
Despite the fact, that the problem is generally well-studied, we still raise it thanks to the specific properties of our corpus that offer a new perspective on the old research object. If we compare our study to the one Daniel and Zelenkov performed on the data from the Russian National Corpus, we can see one important advantage of our dataset, namely the pragmatic uniformity. The data available in the Oral Sub-corpus of the RNC comes from very different sources, including purely artificial ones, like films or plays, and therefore this data unifies very different examples of speech in terms of pragmatics, which a researcher can hardly account for. Our dataset on the other hand consists of transcribed field recordings and interviews that more or less belong to a same type of communicative situations. This fact makes the statistical hypotheses we intend to test a lot more convincing.

Preparation

The dataset was previously extracted from the RSUH folklore archive (in the process of publication) and prepared using Python scripts. The original archive consists of some 24000 entries, each of which contains one or several answers to interviewers' questions by different speakers. All the questions or interviewers' remarks are enclosed in square brackets, which makes it easy to filter them by using regular expressions. When calculating the mean utterance length, we viewed the span between the end of the previous previous question and the beginning of the next question as a single utterance and thus the mean number of tokens (e.g. words) inside those spans was viewed as the MUL for each entry. Thus, each entry may potentially contain one or several utterances and in the former case the MUL is equal to the number of tokens in the only utterance. The mean sentence length on the other hand was calculated after splitting the respondents' answers by punctuation marks (".", "!", "…", "?") by counting the mean number of tokens inside the remaining spans.

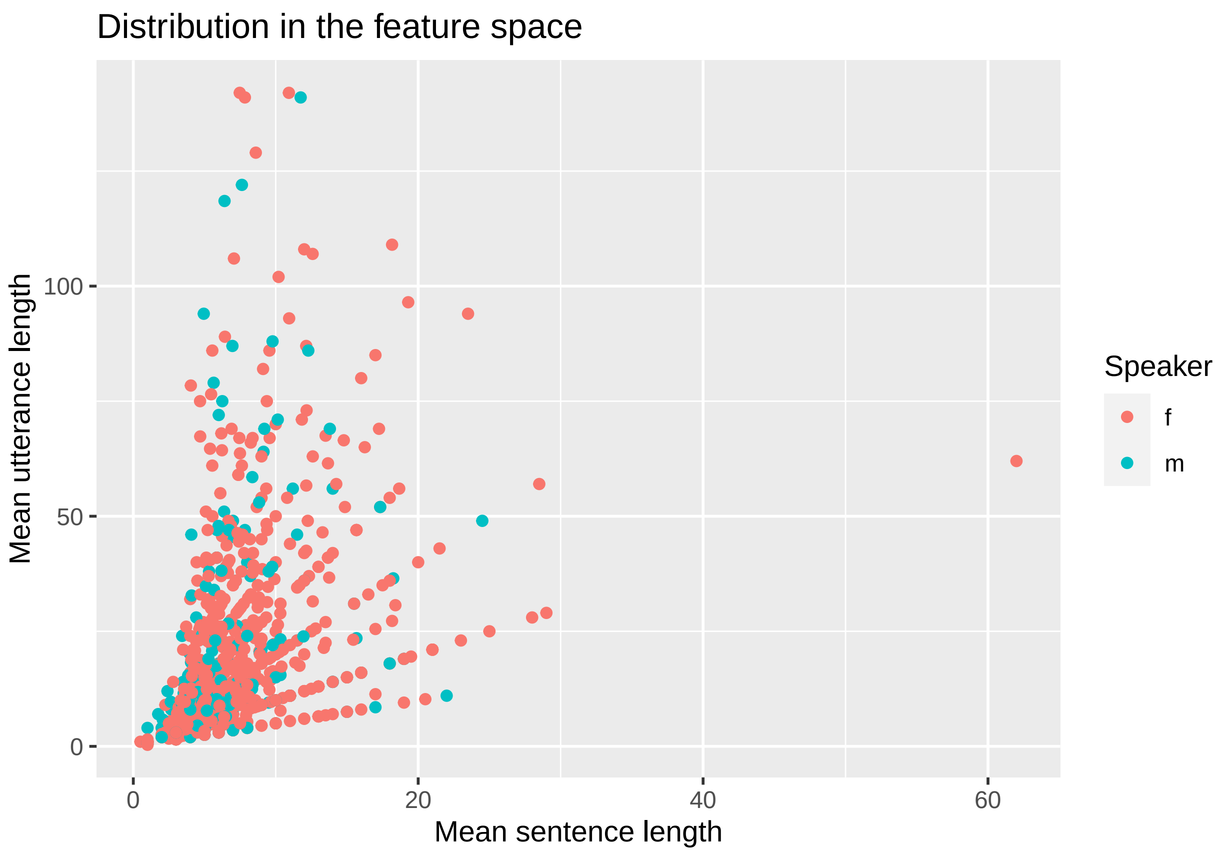
```
dataset <- read.csv("https://raw.githubusercontent.com/ruthenian8/int_mul/master/uq_infs_genders.csv")
```





In the two histograms above, the red line shows the median value of the MUL and MSL parameters.

To explore the properties of the dataset, we may also take a look at how different speakers are distributed in the feature space with mean sentence length serving as an x-axis and mean utterance length as an y-axis.



The only entry with mean sentence length more than 60 will be excluded, as it can be viewed as an outlier that is likely to affect the computations.

The three factors that generally need to be taken care of when comparing parameters like the MUL are independence of observations, normal distribution of data and homosedasticity (equivalence of variance). The first of this constraints was taken care of, when we selected our examples from the archive, as we included only one randomly selected entry from each of the speakers. This means that none of the examples in the dataset was influenced by another. The two other factors are commonly viewed as less restrictive for several reasons.

- Firstly, since the impact that a non-normal distribution of data would make on a t-test can easily be avoided by resorting to Wilcoxon's test instead. This possibility greatly alleviates the construction of our dataset, since the graph clearly suggests that the distribution of data is quite skewed and even resembles Poisson's distribution (see the figures above).
- Secondly, while we still intend to check the homosedasticity of the data, using the Wilcoxon's test or the version of the t-test, known as Welch's independent sample t-test reduces the influence of this factor.
  These two assumptions lead us to the conclusion that the current state of the dataset is acceptable for comparing mean values inside any groups of choice.

Another assumption that should be accounted for is that the distributions should be symmetrical around the median for successfully running the Wilcoxon's test. This was almost the case before filtering out the outliers, as the graphs suggest. We can hope that after the cleanup has been performed, the data finally meets this requirement.

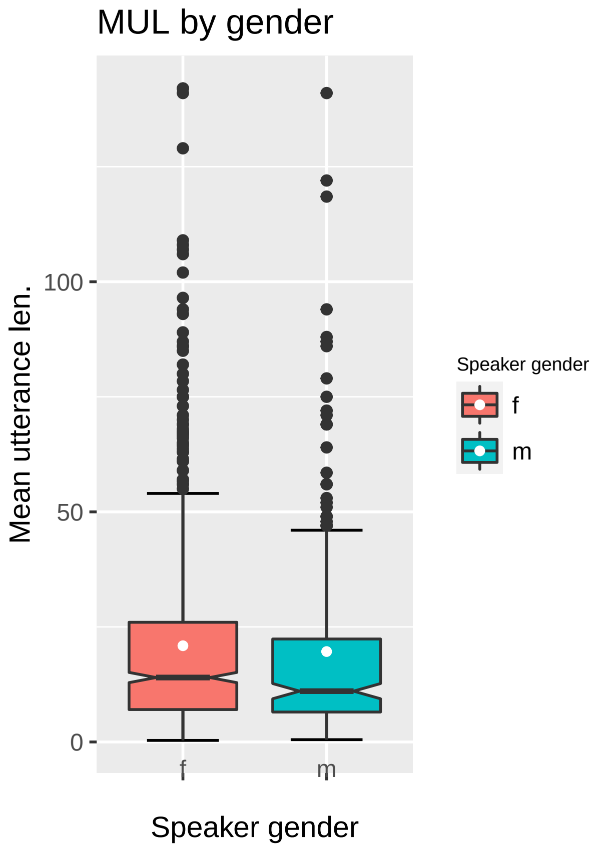
The relevant variables that are present in the dataset are:

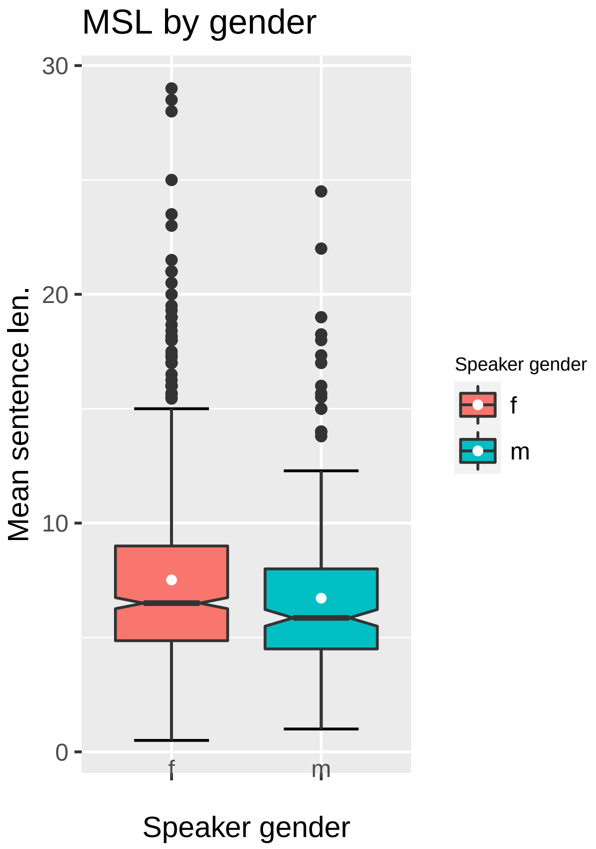
- text: includes a full version of the text for each of the entries.
- mean_ul: mean utterance length, calculated in the fashion described above
- mean_sl: mean sentence length,
- inf_gender1: the gender of the speaker
- sob_gender1: the gender of the first interviewer
- sob_gender2: the gender of the second interviewer (if present, ”” if absent)
- sob_gender3 &
- sob_gender4: genders of the third and the fourth interviewer

The four latter parameters allow us to separate the entries in several groups depending on the gender of the speakers. Thus, we can compare the cases, in which the speaker has a conversation either with a single interviewer of a certain gender or with a team of interviewers, that can be either diverse or uniform in terms of gender. This aspect of the situation can possibly influence the speaker, determining, whether they wish to share their knowledge. For instance, it was noted by the scholars that respondents are much more eager to share their knowledge about magic and other forbidden subjects with the interviewers that they perceive as equal to themselves.

Dataset analysis

After the requirements have been accounted for, we can proceed to the comparison of groups. First of all, we are going to compare speaker genders overall, without making any further distinctions.





In both cases, the third quartile boundary is higher for the female group, although we may assume that this tendency is due to the unequal sample sizes, as the female group includes more entries. The white points that mark the mean of the distribution suggest that little difference is present, but we are still going to check the difference. Firstly, we check the homogenity of variance using Levene's test.

MUL

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.0783 0.7797
##       950
```

MSL

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  3.7483 0.05316 .
##       950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the case of the MUL, the equivalence of variances is asserted by the test. As for the MSL, the tests suggests that the variances may differ, but the p-value and the f-statistic do not confirm the presence of a striking difference. This means that Wilcoxon's test is likely to show correct results. After running the test itself, we also calculate the corresponding r-statistic so as to determine the effect size.

MUL

```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic     p p.signif
## * <chr>   <chr>  <chr>  <int> <int>     <dbl> <dbl> <chr>
## 1 mean_ul f      m        722   230     89887 0.059 ns
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_ul f      m       0.0612   722   230 small
```

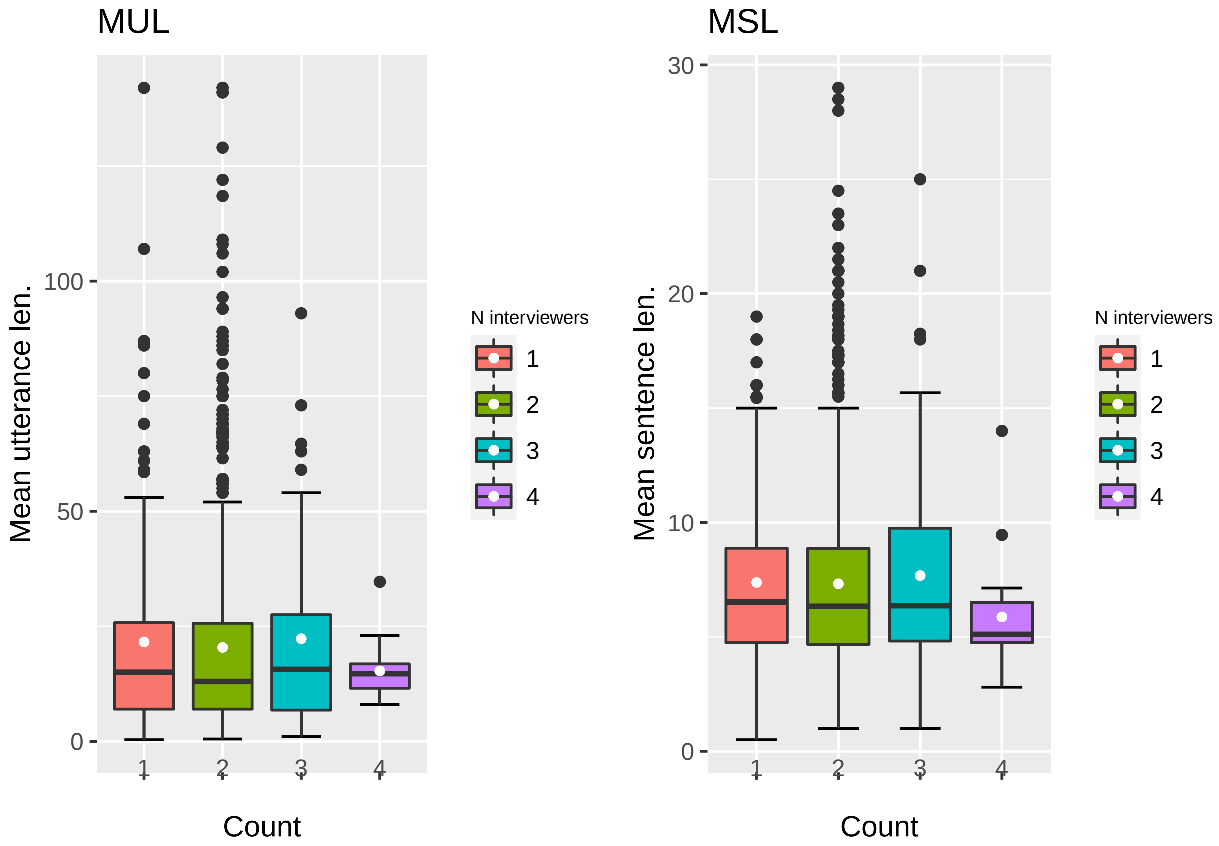
MSL

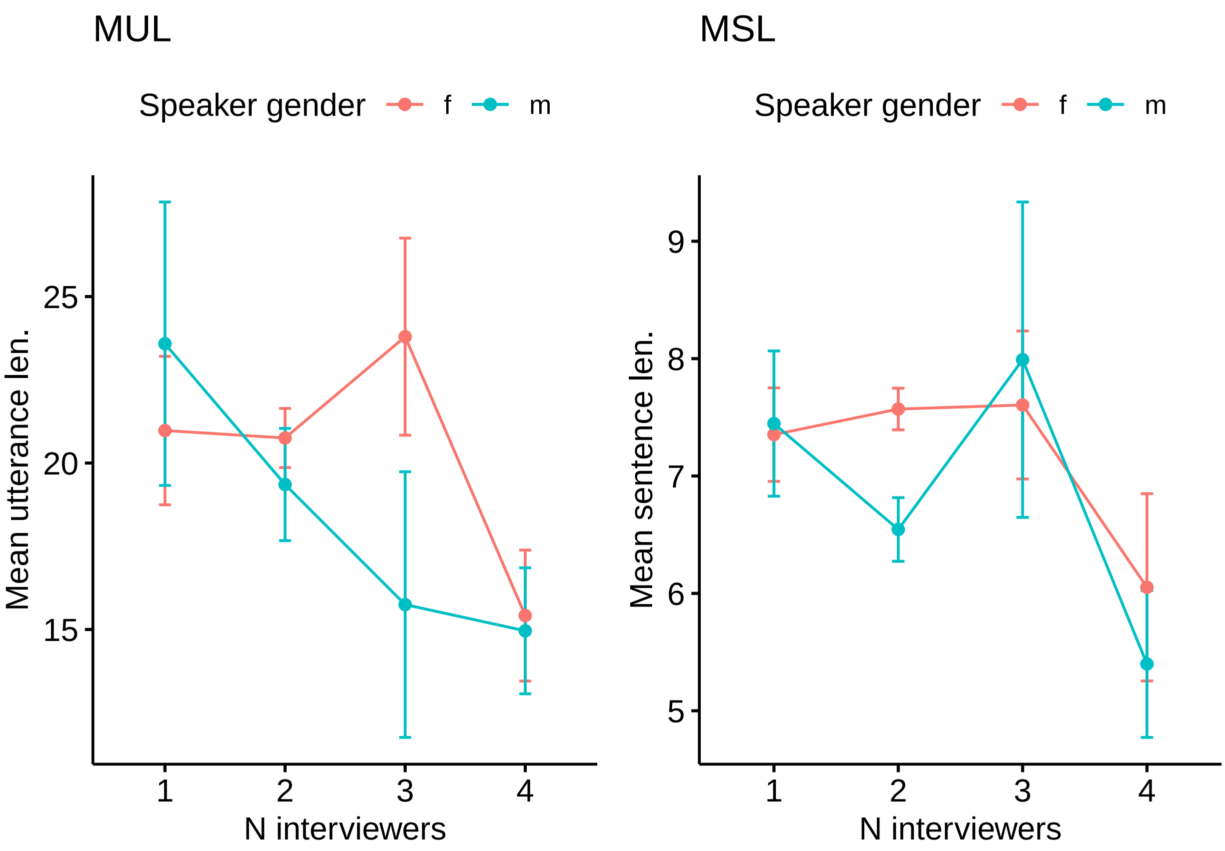
```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic       p p.signif
## * <chr>   <chr>  <chr>  <int> <int>     <dbl>   <dbl> <chr>
## 1 mean_sl f      m        722   230    93674. 0.00337 **
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_sl f      m       0.0950   722   230 small
```

Our interpretation of the tests above depends on what alpha we choose for testing the hypotheses. As long as we pick an alpha of 0.001, like Daniel and Zelenkov did in their research, the difference of means in both of the tests has no statistical significance. However, if we choose an alpha of 0.05 that is generally acknowledged as sufficient for the social sciences, both results appear to be significant, especially in the case of the sentence length. Whichever fashion we go for, the r-statistic still implies that the effect size is quite small in both of the tests. Thus, the difference in the utterance length is on the verge of statistical significance, but does not appear to be large, whereas the sentence length is likely to differ, being larger in the case of female respondents. We may conclude that **according to the data that we have, the common assumption that women speak more is hardly supported statistically. On the other hand, this very impression can potentially be explained by the fact that they tend to use longer sentences (at least, in the case of personal interviews)**. Having addressed the main matter in question, we can now try to split the groups into smaller fractions, so as to see, if any patterns can be spotted in this manner.

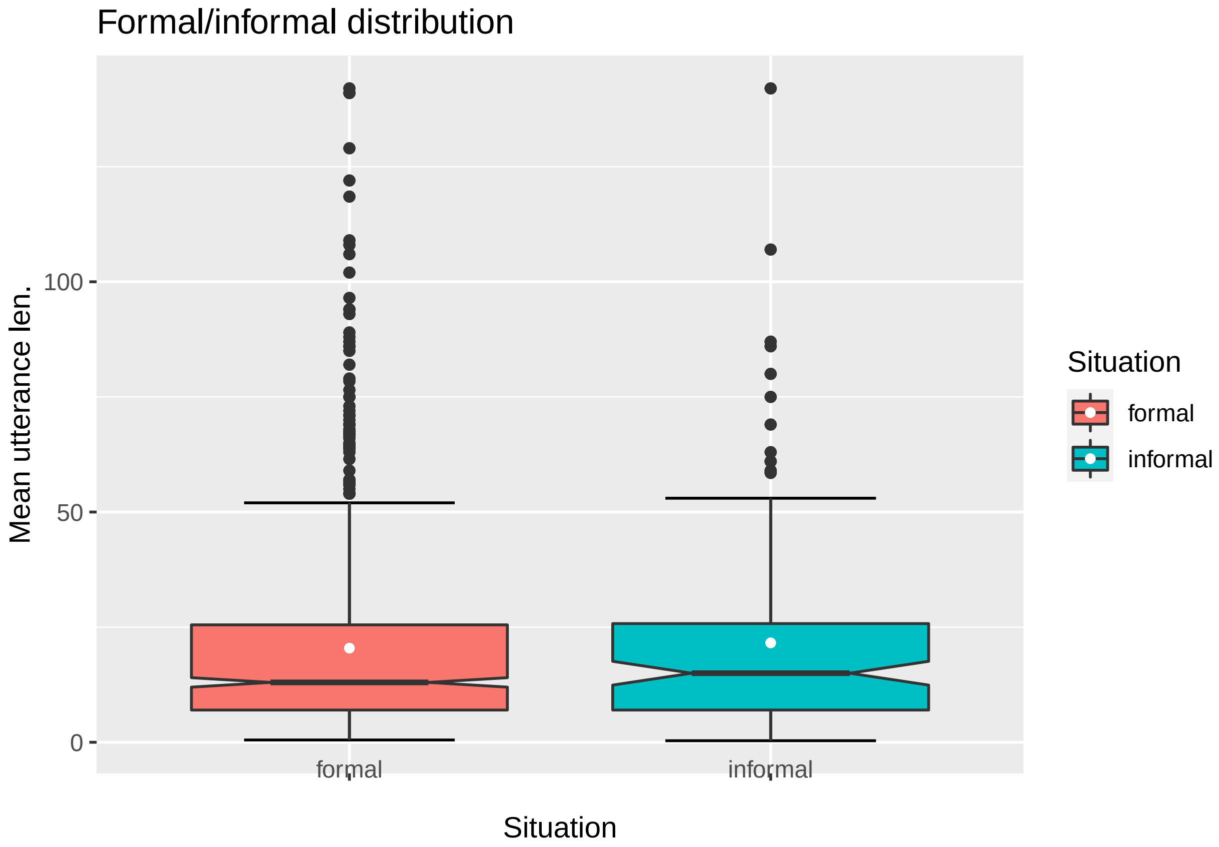
Although we stated at the beginning, that the corpus that we are working with is generally uniform in terms of pragmatics, speakers' perception of the communicative situation can still differ, depending on the number of interviewers they converse with. As long as only one interviewer participates in the dialogue, the situation can possibly be viewed as a private talk, which does not force the informant to change his everyday speaking patterns. **Meanwhile a conversation with multiple interviewers (normally from 2 to 4) is more likely to be perceived as a formal situation that requires the informant to make respective adjustments to their manner of speaking.** Therefore it would be interesting to compare the "formal" and "informal" corpus entries. Firstly, we can compare the subgroups determined by the exact number of interviewers.



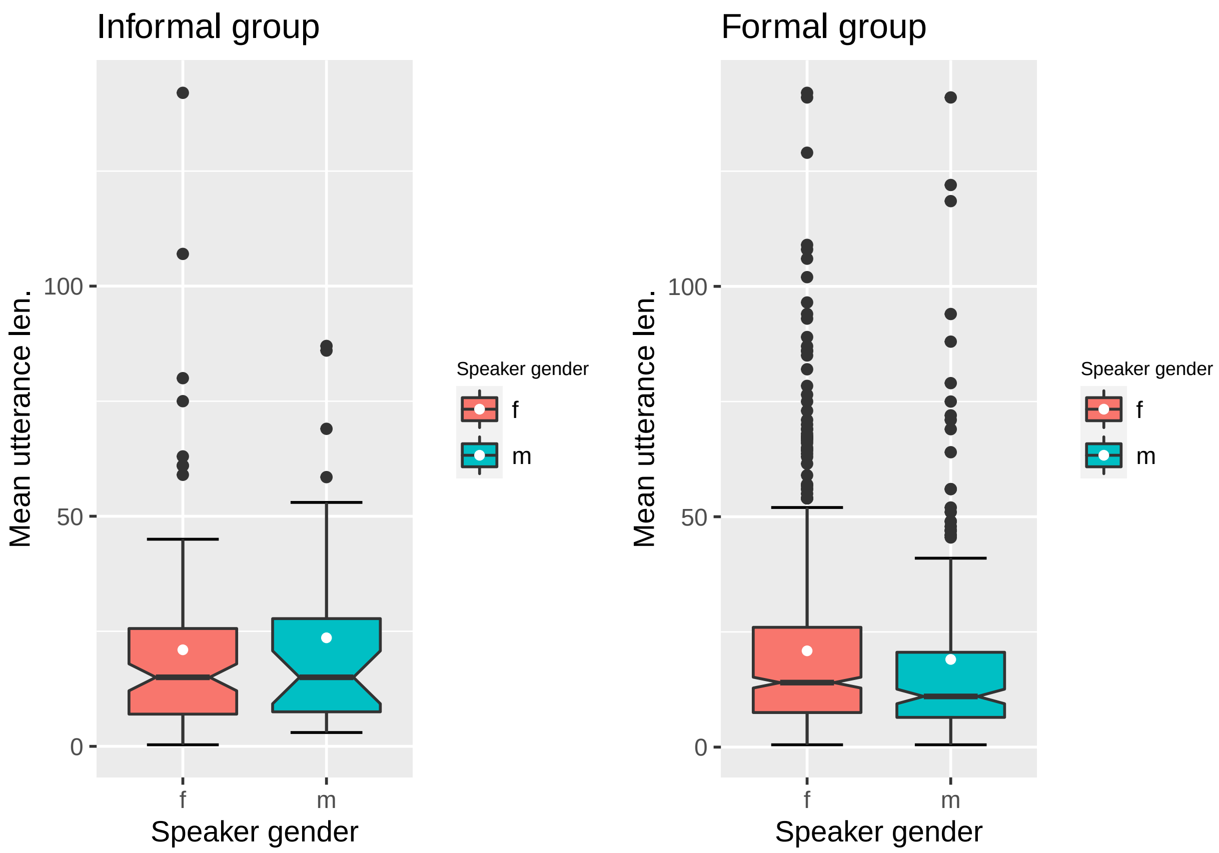


Then we may also look at the difference between the formal (>2 interviewers) and the informal subgroups. The sample sizes are 815 for formal and 131 for informal types.

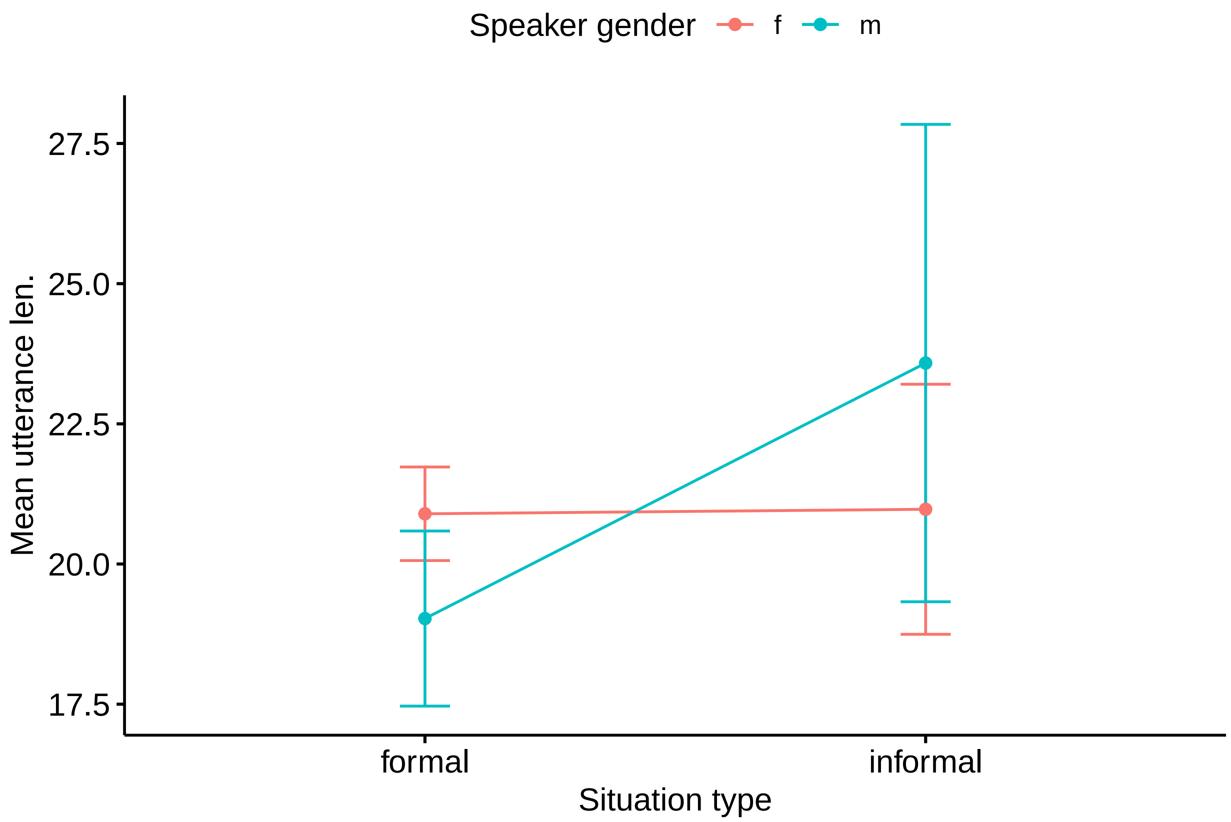
```
## # A tibble: 2 x 2
##   sit.type     n
## * <chr>    <int>
## 1 formal     820
## 2 informal   132
```



What we can see from the graph is that the assumption we made above is presumably false, as the means and the medians of the two distributions are roughly the same. The notches also clearly intersect, which implies the lack of statistically significant differences between the medians. Still, it can also be noted that the whisker ends are different and include higher values in the case of the formal situation type. The other difference is that the plot for the formal type includes many more outliers with high MUL values, although this fact can be explained by the difference of sample sizes. However, what would be interesting to test is whether gender-based distinctions exist in the two new groups. Firstly, we will create plots for both the formal and the informal case.



The box plot for the informal group suggests that little difference is present, since the quartile borders of the boxes and the medians are positioned on similar levels. The box plot for the formal group on the other hand shows that the quartile borders and the whiskers cover a narrower range and find themselves lower in the case of the male speaker group. The notches of the boxes are also differently positioned, which hints at the difference of medians.



The line plot also shows that the data is distributed differently inside the two groups, as the intersection of intervals is much smaller inside the formal group. Running the same comparison tests inside the formal group, we can see, that the significance between speaker genders is more notable in this particular case.

MUL

```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic      p p.signif
## * <chr>   <chr>  <chr>  <int> <int>     <dbl>  <dbl> <chr>
## 1 mean_ul f      m        621   199     68197 0.0275 *
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_ul f      m       0.0770   621   199 small
```

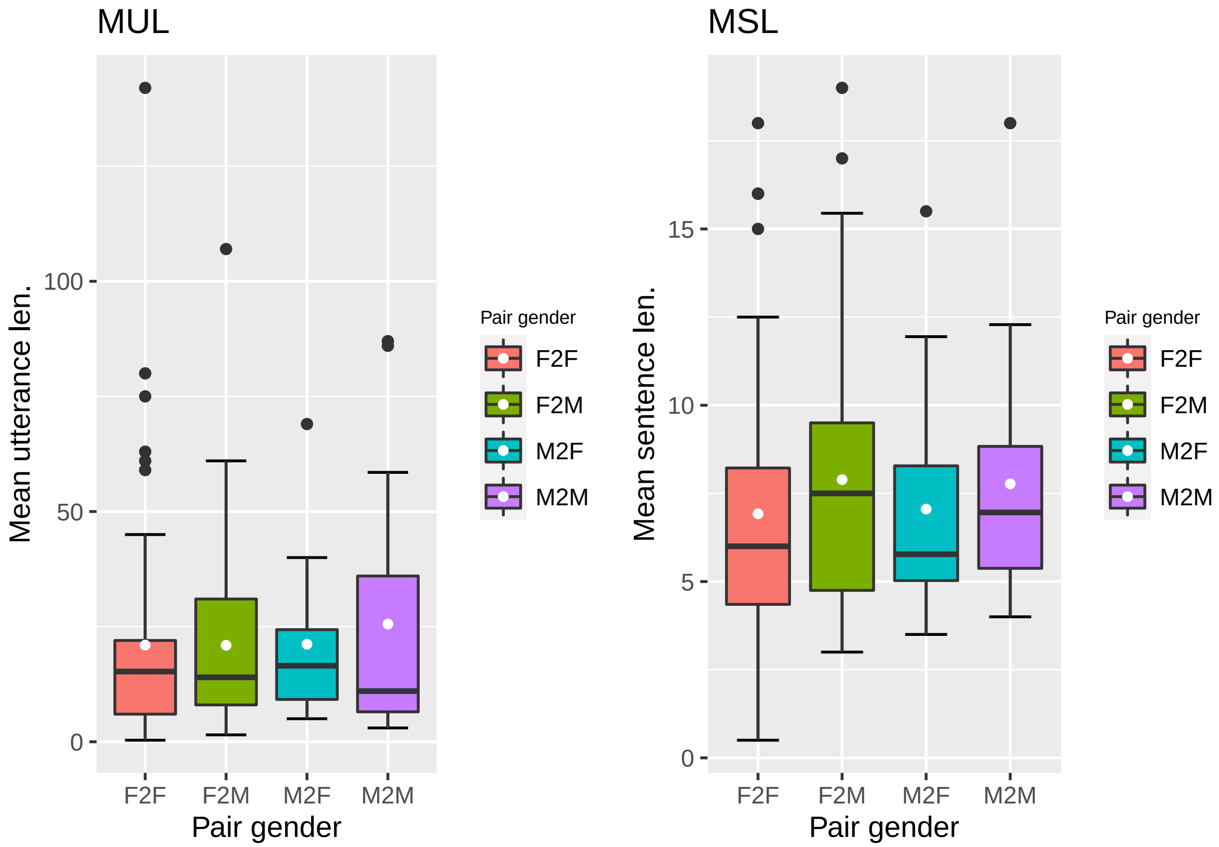
MSL

```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic       p p.signif
## * <chr>   <chr>  <chr>  <int> <int>     <dbl>   <dbl> <chr>
## 1 mean_sl f      m        621   199     71312 0.00106 **
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_sl f      m        0.114   621   199 small
```

While the difference of the mean utterance lengths still does not have to be taken too seriously, as the p-value shows a high probability of a type-1 error, while the r-statistic value is small, the difference of mean sentence lengths remains quite notable. The effect size for this comparison turned out to be even larger than the one obtained in the initial experiment (0.114 vs 0.0950).

If we try to interpret these distinctions in terms of some real-world tendencies, we may suppose that **female speakers are more eager to adapt their speaking patterns to the specifics of the communicative situation and thus end up using longer sentences**. It may be either due to differences in the perception of the situation (members of one group view it as responsible, while others do not) or due to some other factor. This tendency in its turn can presumably contribute to the general impression that women speak more than men do. So far, the trends that we observed were hardly significant, but still informative. Nevertheless, the dataset can be fractured even more, if we take into account the gender of the interviewers. At first, we can take a look at how the distributions look, when there is only one interviewer. In this case we create a separate variable for distinct speaker-interviewer pairs: "F2M" means that a female respondent is talking to a male interviewer etc.



MUL

```
## # A tibble: 4 x 3
##   part.genders mean_MUL     n
## * <chr>           <dbl> <int>
## 1 F2F              21.0    56
## 2 F2M              20.9    45
## 3 M2F              21.2    14
## 4 M2M              25.6    17
```

MSL

```
## # A tibble: 4 x 3
##   part.genders mean_MSL     n
## * <chr>           <dbl> <int>
## 1 F2F              6.92    56
## 2 F2M              7.89    45
## 3 M2F              7.05    14
## 4 M2M              7.77    17
```

Judging by the graphs, we can conclude that no real difference between the means is present, although both the MUL and the MSL seem to be slightly higher, when the interviewer is male. Just to be sure, we may apply a two-way anova with respondent's gender and interviewer's gender as interacting variables to check this out. We favored TukeyHSD for the post-hoc testing, since it has less requirements than a pairwise t.test.

MUL

```
##                           Df Sum Sq Mean Sq F value Pr(>F)
## inf_gender1               1    161   161.3   0.308  0.580
## sob_gender1               1     32    32.3   0.062  0.804
## inf_gender1:sob_gender1   1    115   115.5   0.221  0.639
## Residuals               128  66937   522.9
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mean_ul ~ inf_gender1 * sob_gender1, data = data_single)
##
## $inf_gender1
##        diff       lwr      upr     p adj
## m-f 2.60789 -6.682793 11.89857 0.5795823
##
## $sob_gender1
##          diff       lwr      upr     p adj
## m-f 0.9874238 -6.903799 8.878647 0.8048485
##
## $`inf_gender1:sob_gender1`
##                diff       lwr      upr     p adj
## m:f-f:f  0.18087811 -17.60646 17.96821 0.9999933
## f:m-f:f -0.04850797 -11.96588 11.86886 0.9999996
## m:m-f:f  4.56719506 -11.91682 21.05121 0.8884740
## f:m-m:f -0.22938608 -18.44634 17.98756 0.9999873
## m:m-m:f  4.38631695 -17.09754 25.87018 0.9512710
## m:m-f:m  4.61570303 -12.33099 21.56240 0.8933747
```

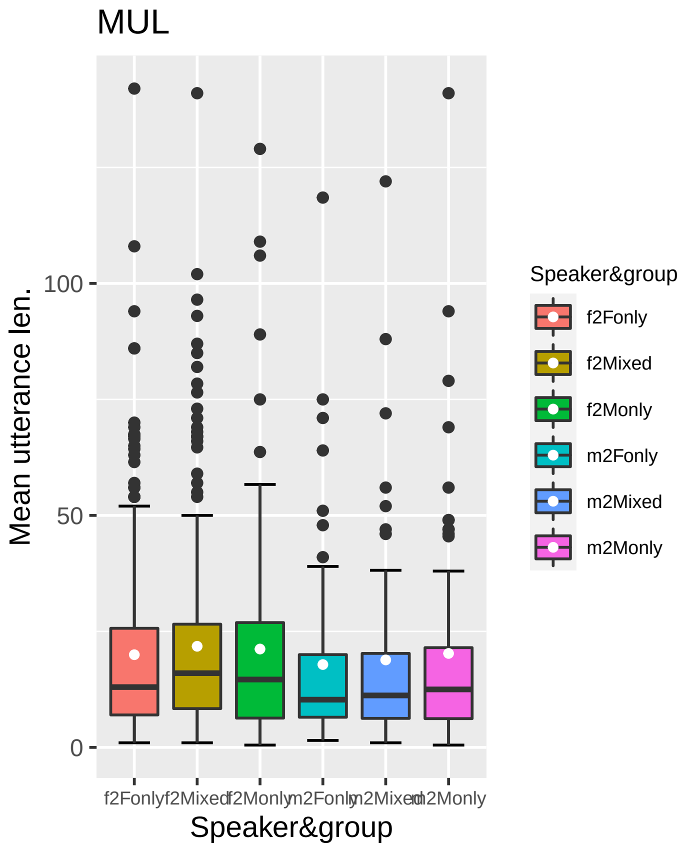
MSL

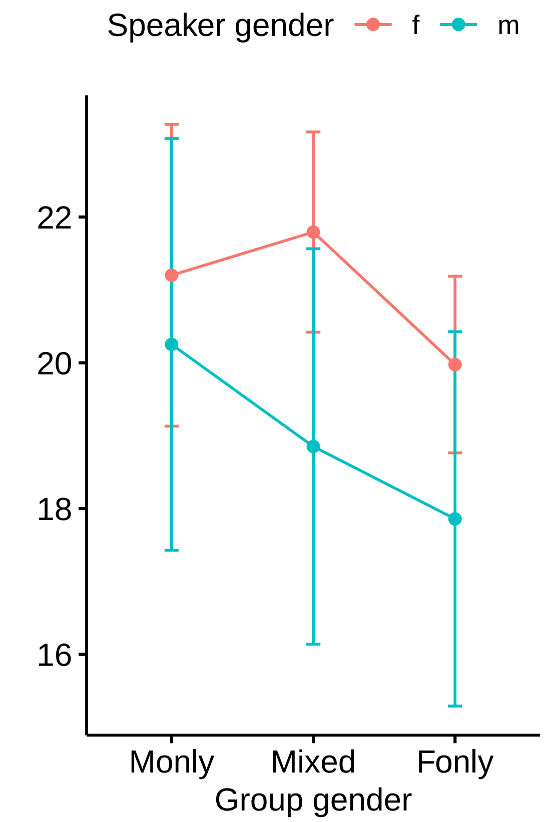
```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## inf_gender1               1    0.2    0.21   0.014  0.906
## sob_gender1               1   27.1   27.05   1.793  0.183
## inf_gender1:sob_gender1   1    0.4    0.37   0.025  0.876
## Residuals               128 1931.6   15.09
```

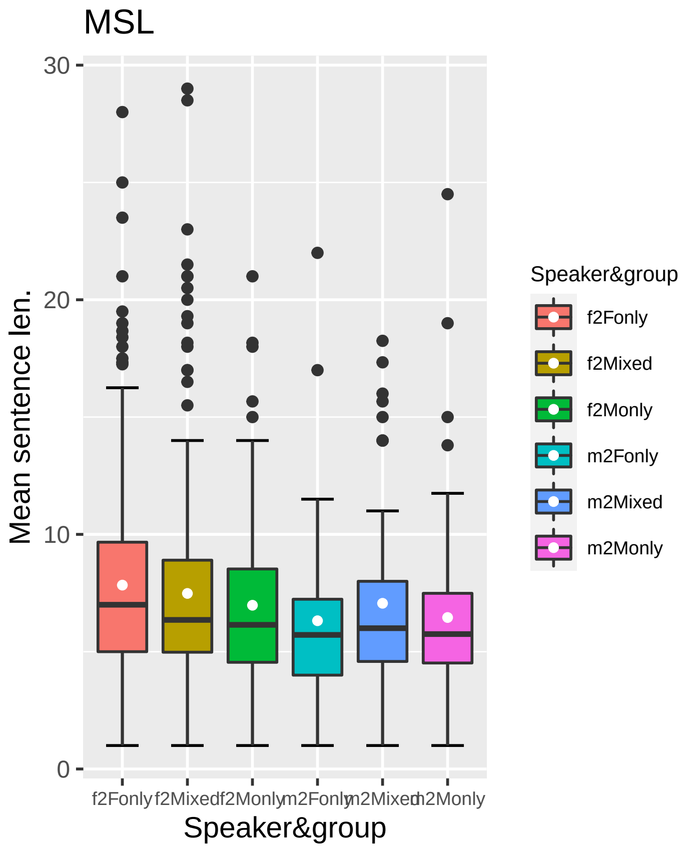
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mean_sl ~ inf_gender1 * sob_gender1, data = data_single)
##
## $inf_gender1
##           diff       lwr      upr     p adj
## m-f 0.09409455 -1.484152 1.672341 0.9062781
##
## $sob_gender1
##          diff       lwr      upr     p adj
## m-f 0.9036012 -0.436913 2.244115 0.1846497
##
## $`inf_gender1:sob_gender1`
##               diff       lwr      upr     p adj
## m:f-f:f  0.1321260 -2.889481 3.153733 0.9994698
## f:m-f:f  0.9696561 -1.054796 2.994108 0.5983584
## m:m-f:f  0.8505849 -1.949621 3.650791 0.8585574
## f:m-m:f  0.8375301 -2.257057 3.932118 0.8951414
## m:m-m:f  0.7184589 -2.931092 4.368010 0.9559911
## m:m-f:m -0.1190713 -2.997876 2.759733 0.9995511
```

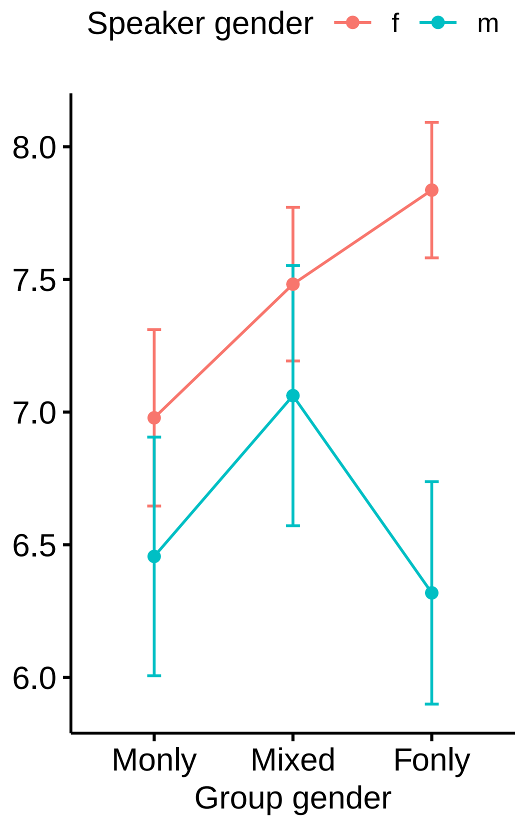
As the reports suggest, the slight difference that exists is of no statistical significance. This fact supports our assumption, that the difference between the genders reveals itself mainly in formal situations with respect to both the MUL and the MSL. On the other hand, this very test demonstrates, that the interaction between the speaker gender and the interviewer gender is not significant. In a sense, this can be viewed as good news, because it appears that both males and females are equally open to a talk with any interviewer.

We may also look at the situations, in which there are two or more interviewers. For this comparison we introduce a variable that has values “Fonly” and “Monly” if the interviewer group is composed solely of females or males respectively. All other composition types are marked as “Mixed”. After combining this value with the gender of the speaker we are left with 6 distinct dialogue types.


MUL
Mean utterance len.
0
50
100
f2Fonly
f2Mixed
f2Monly
m2Fonly
m2Mixed
m2Monly
Speaker&group
Speaker&group
f2Fonly
f2Mixed
f2Monly
m2Fonly
m2Mixed
m2Monly


Speaker gender
f
m
16
18
20
22
Monly
Mixed
Fonly
Group gender





Although the graphs show no drastic difference between the distributions, the means are obviously not equal, which is why we may compare them using a two-way anova test and the Tukey Honest Significant differences post-hoc test. Thus the gender of the group and respondent's gender will be additionally tested for interaction.

MUL

```
## # A tibble: 6 x 3
##   inf2group   MUL   num
## * <chr>     <dbl> <int>
## 1 f2Fonly    20.0   269
## 2 f2Mixed    21.8   236
## 3 f2Monly    21.2   116
## 4 m2Fonly    17.9    65
## 5 m2Mixed    18.9    63
## 6 m2Monly    20.3    71
```

```
##                            Df Sum Sq Mean Sq F value Pr(>F)
## inf_gender1                 1    526   526.3   1.179  0.278
## group.gender                2    533   266.5   0.597  0.551
## inf_gender1:group.gender    2     93    46.6   0.104  0.901
## Residuals                 814 363495   446.6
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = mean_ul ~ inf_gender1 * group.gender, data = grouped)
##
## $inf_gender1
##          diff       lwr      upr     p adj
## m-f -1.868816 -5.247644 1.510012 0.2779499
##
## $group.gender
##                    diff       lwr      upr     p adj
## Mixed-Fonly  1.64037435 -2.309915 5.590664 0.5928884
## Monly-Fonly  1.62278437 -2.908911 6.154480 0.6777302
## Monly-Mixed -0.01758998 -4.643505 4.608325 0.9999561
##
## $`inf_gender1:group.gender`
##                       diff        lwr       upr     p adj
## m:Fonly-f:Fonly -2.1176835 -10.460536  6.225169 0.9788709
## f:Mixed-f:Fonly  1.8181317  -3.565653  7.201917 0.9289367
## m:Mixed-f:Fonly -1.1241876  -9.573022  7.324647 0.9989703
## f:Monly-f:Fonly  1.2250270  -5.479982  7.930036 0.9953158
## m:Monly-f:Fonly  0.2762646  -7.777674  8.330203 0.9999987
## f:Mixed-m:Fonly  3.9358152  -4.519791 12.391421 0.7685114
## m:Mixed-m:Fonly  0.9934959  -9.678661 11.665653 0.9998202
## f:Monly-m:Fonly  3.3427105  -6.009791 12.695212 0.9109622
## m:Monly-m:Fonly  2.3939481  -7.968390 12.756286 0.9861456
## m:Mixed-f:Mixed -2.9423194 -11.502511  5.617872 0.9237194
## f:Monly-f:Mixed -0.5931047  -7.437901  6.251691 0.9998737
## m:Monly-f:Mixed -1.5418672  -9.712547  6.628812 0.9945484
## f:Monly-m:Mixed  2.3492146  -7.097948 11.796378 0.9807156
## m:Monly-m:Mixed  1.4004522  -9.047402 11.848307 0.9989329
## m:Monly-f:Monly -0.9487624 -10.044476  8.146952 0.9996856
```

MSL

```
## # A tibble: 6 x 3
##   inf2group   MSL   num
## * <chr>     <dbl> <int>
## 1 f2Fonly    7.84   269
## 2 f2Mixed    7.48   236
## 3 f2Monly    6.98   116
## 4 m2Fonly    6.32    65
## 5 m2Mixed    7.06    63
## 6 m2Monly    6.46    71
```

```
##                           Df Sum Sq Mean Sq F value  Pr(>F)
## inf_gender1                1    133  132.78   8.008 0.00477 **
## group.gender               2     44   21.92   1.322 0.26722
## inf_gender1:group.gender   2     37   18.64   1.124 0.32546
## Residuals                814  13496   16.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = mean_sl ~ inf_gender1 * group.gender, data = grouped)
##
## $inf_gender1
##           diff       lwr        upr     p adj
## m-f -0.9386458 -1.589707 -0.2875851 0.0047705
##
## $group.gender
##                   diff       lwr       upr     p adj
## Mixed-Fonly -0.1324713 -0.893646 0.6287034 0.9120872
## Monly-Fonly -0.5874098 -1.460615 0.2857951 0.2550796
## Monly-Mixed -0.4549385 -1.346298 0.4364213 0.4544535
##
## $`inf_gender1:group.gender`
##                        diff       lwr       upr     p adj
## m:Fonly-f:Fonly -1.51806982 -3.125640 0.0895006 0.0768005
## f:Mixed-f:Fonly -0.35448417 -1.391877 0.6829084 0.9254925
## m:Mixed-f:Fonly -0.77462272 -2.402615 0.8533691 0.7514386
## f:Monly-f:Fonly -0.85820273 -2.150180 0.4337744 0.4043890
## m:Monly-f:Fonly -1.38062662 -2.932527 0.1712734 0.1135300
## f:Mixed-m:Fonly  1.16358565 -0.465711 2.7928823 0.3205315
## m:Mixed-m:Fonly  0.74344710 -1.312953 2.7998472 0.9068871
## f:Monly-m:Fonly  0.65986709 -1.142251 2.4619850 0.9021646
## m:Monly-m:Fonly  0.13744320 -1.859259 2.1341449 0.9999595
## m:Mixed-f:Mixed -0.42013855 -2.069588 1.2293105 0.9785434
## f:Monly-f:Mixed -0.50371856 -1.822631 0.8151938 0.8850446
## m:Monly-f:Mixed -1.02614245 -2.600537 0.5482522 0.4266852
## f:Monly-m:Mixed -0.08358001 -1.903938 1.7367781 0.9999946
## m:Monly-m:Mixed -0.60600390 -2.619184 1.4071758 0.9557834
## m:Monly-f:Monly -0.52242390 -2.275062 1.2302141 0.9575789
```

Both the anova and the post-hoc test demonstrate that the gender of the speaker (inf_gender) is the most influential variable. Neither its interaction with the gender of the group, nor the latter variable itself impact the MUL and the MSL in a notable manner. On the other hand, this result shows that the importance of the speaker gender should not be underestimated with respect to the mean sentence length, since the presence of many other variables does not diminish its effect.

Discussion

- The results show that the difference between genders in terms of speech parameters is likely to exist. When it comes to the mean utterance length, the tests lead us to the conclusion, that it generally tends to be larger for female speakers. Nevertheless, the effect size is small, so this the distance between the genders is not critical. On the other hand, the study reveals that **female speakers tend to speak in longer sentences, especially when confronted by multiple interviewers**, e. g. in formal situations. This trend is much more evident and we suppose that it could be the reason behind the stereotype that women speak more.
- The study also showed some disadvantages of the corpus, **like the lack of an age variable or like the lack of balance in terms of speaker genders**. These weak points can potentially be taken into account, when collecting new data.
- As for the interviewer team composition in terms of gender, the study shows that the effect that the corresponding variables make is hardly noticeable - at least as far as the mean utterance length and the mean sentence length are concerned. As was stated above, this is good news, as it means that **both male and female interviewers can collect texts with equal success**.

References

- Daniel, M. A., Zelenkov, Yu. G. (2012). NKRJa kak instrument sociolingvisticheskih issledovanij. Epizod IV. Pol govorjashhego i dlina repliki [NRC as an intrument for sociolinguistic research]. Episode IV: speaker gender and utterance length. In: Computational linguistics and intellectual technologies. Proceedings of “Dialogue-2012”. 11. Moscow: RSUH. pp. 112-121.
- Tannen, Deborah (1990). You just don’t understand. Women and men in conversation. NY: Ballantine Books