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 question is how well we can study the former matter, given the current structure of the corpus and what the corpus currently lacks in this respect. Thirdly, we provide some conclusions on how collectors' genders could influence the quantity of the collected material.

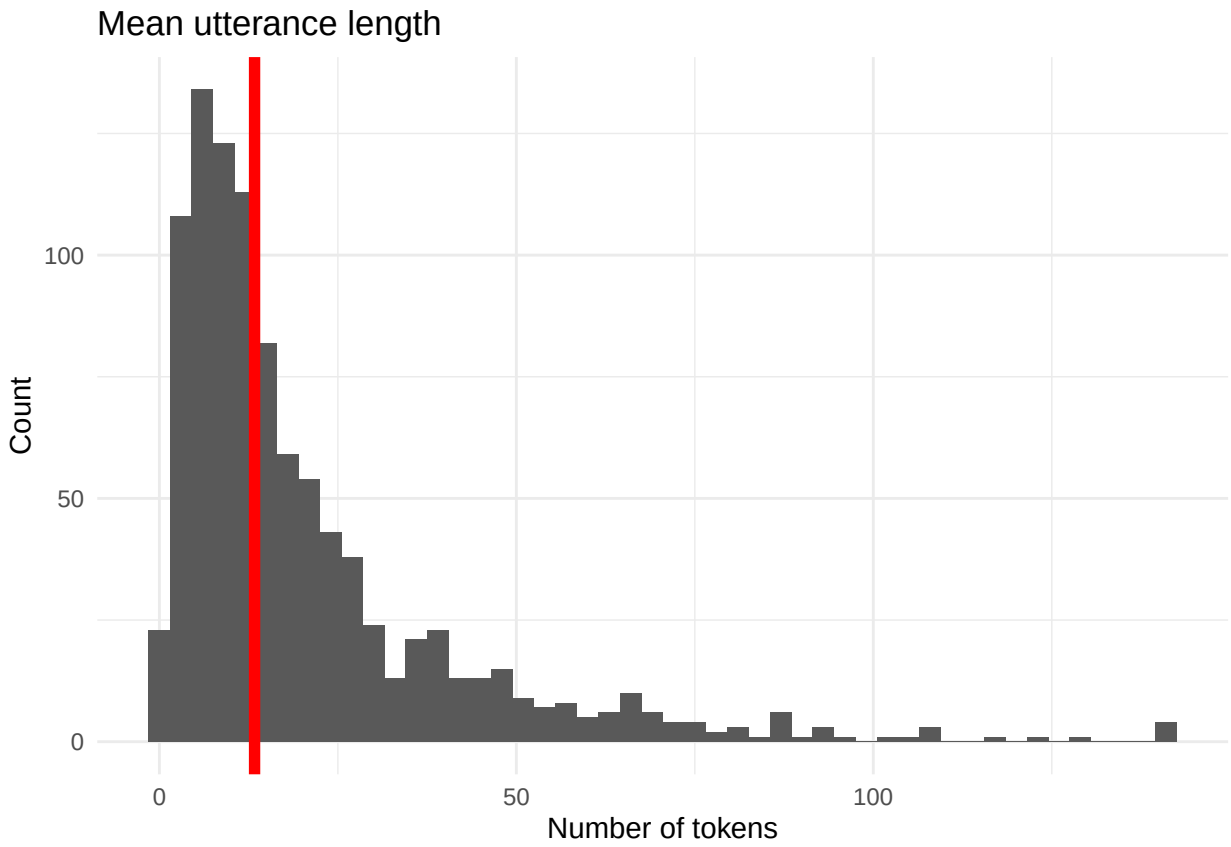
Introduction

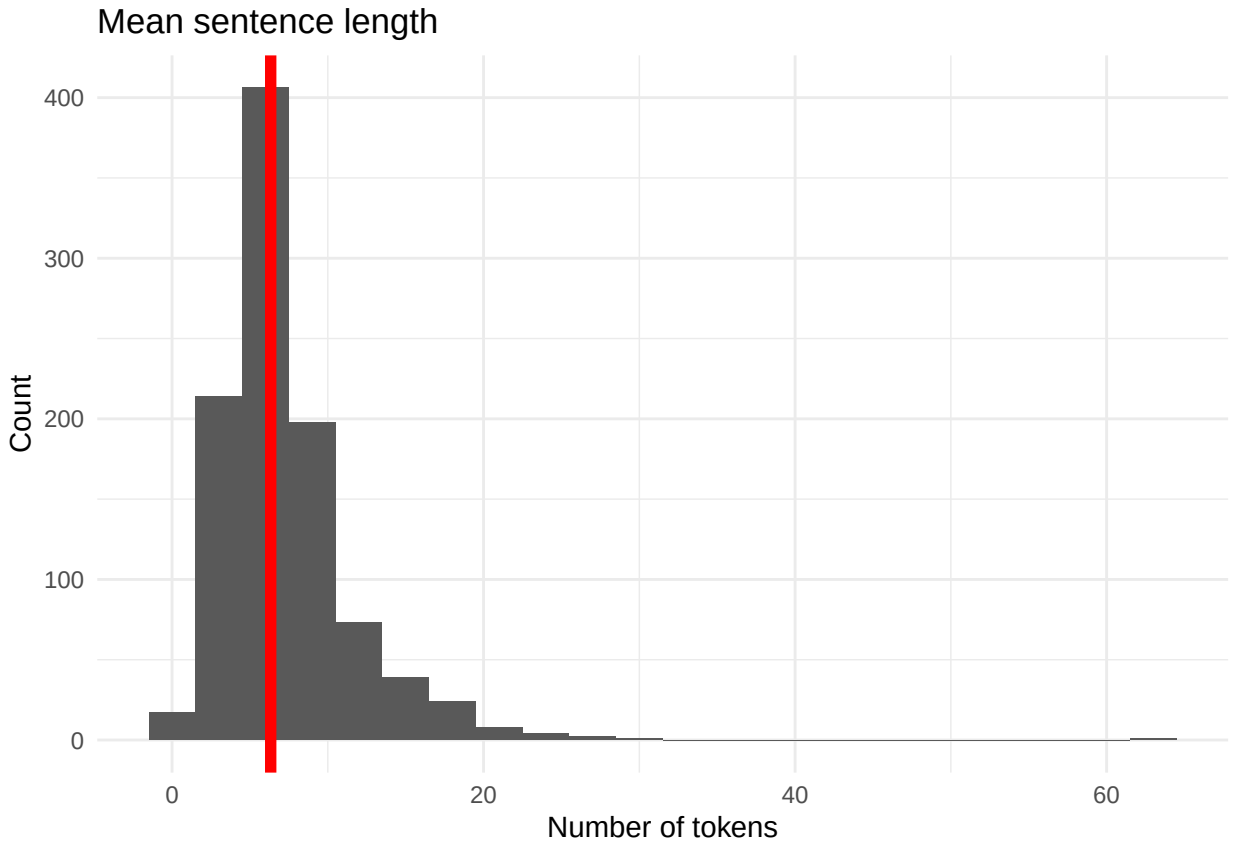
Mean utterance length (MLU) and mean sentence length (MSU) are both traditional variables that have been excessively studied in relation to sociolinguistic factors. What makes a study of such kind especially interesting is the stereotype than women generally talk more than men do. This motivated a lot of researchers to address the matter on the material of different languages (see Daniel & Zelenkov 2012, Tannen 1990). 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 M. A. Daniel 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 archive 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 considered the span between two questions or remarks a single utterance and thus the mean number of tokens inside those spans was viewed as the MLU for each entry. The mean sentence length on the other hand was calculated after all the interviewers' utterances were excluded.

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





In the histograms above the red line shows the median value. Filtering the dataset, we exclude entries in which the mean sentence length exceeds 40 or the mean utterance length exceeds 120, as those can be viewed as outliers that might affect the comparison results. This step will also make both distributions more or less symmetrical around the median, which in its turn will simplify the upcoming tests.

```
dataset <- dataset %>%
  filter(inf_gender1 == "f" | inf_gender1 == "m") %>%
  filter(mean_sl <= 40 & mean_ul < 120)
```

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 figures below).
- 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.

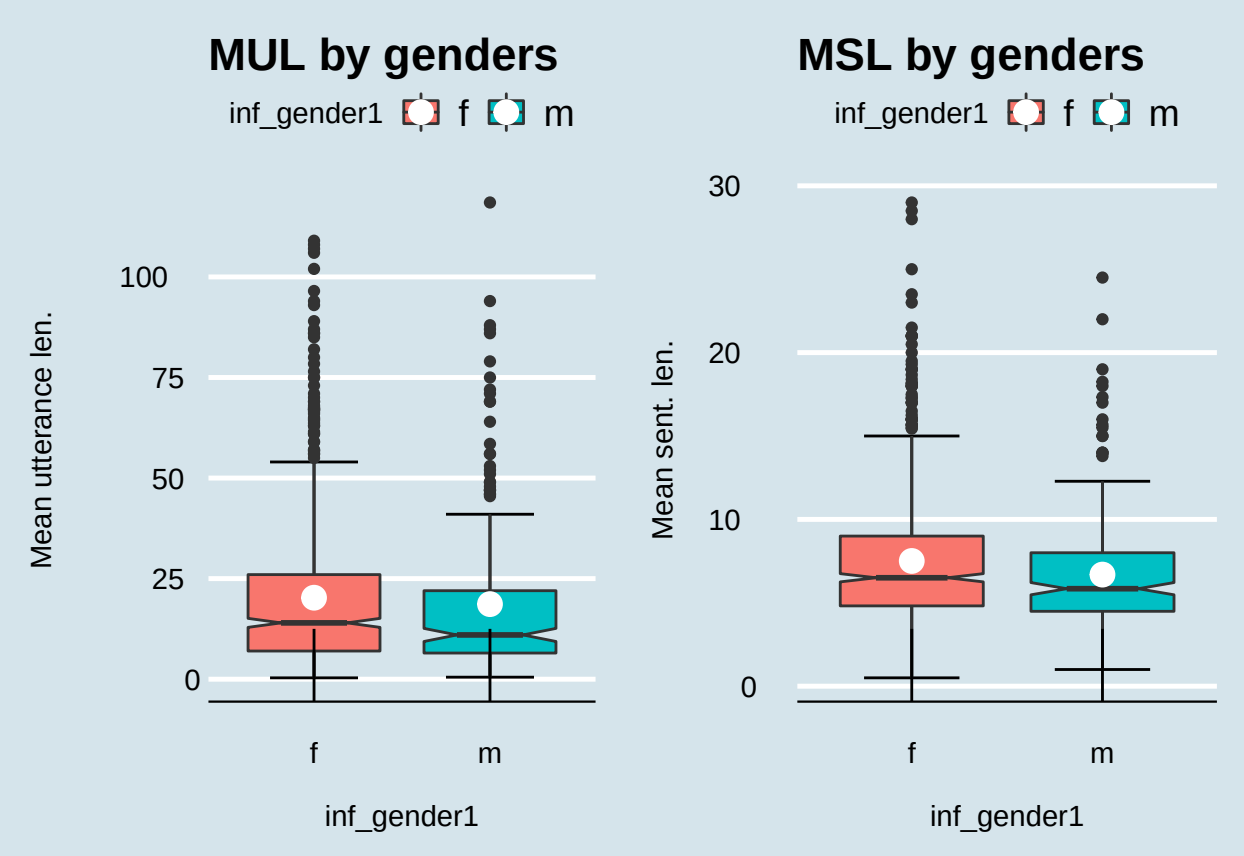
The relevant variables that are present in the dataset are:

- 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 specialists that informers are much more eager to share their knowledge about magic and other forbidden subjects with the interviewers that they perceive as equal to themselves.

Experiments

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 run the corresponding tests. 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        718   228    88968. 0.0477 *
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_ul f      m       0.0644   718   228 small
```

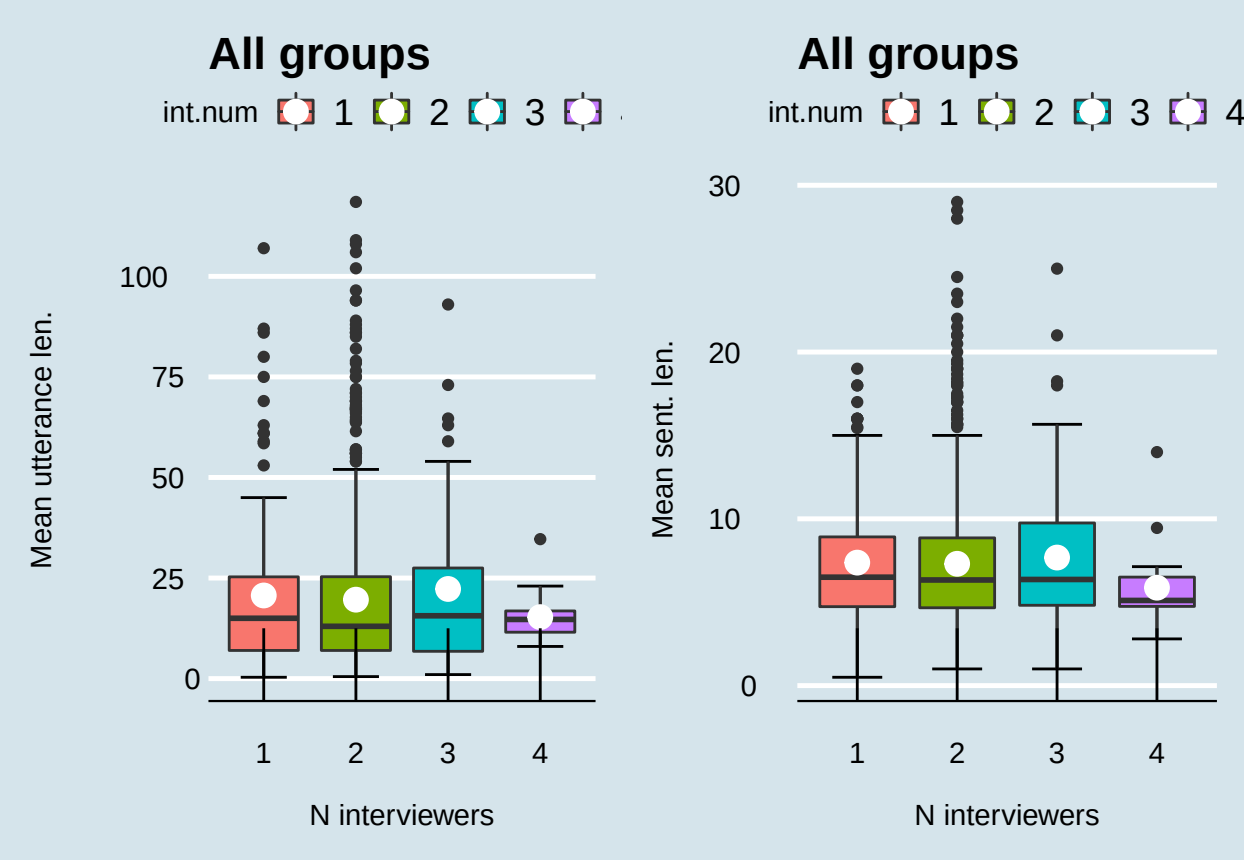
MSL

```
## # 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        718   228    92590. 0.00281 **
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_sl f      m       0.0971   718   228 small
```

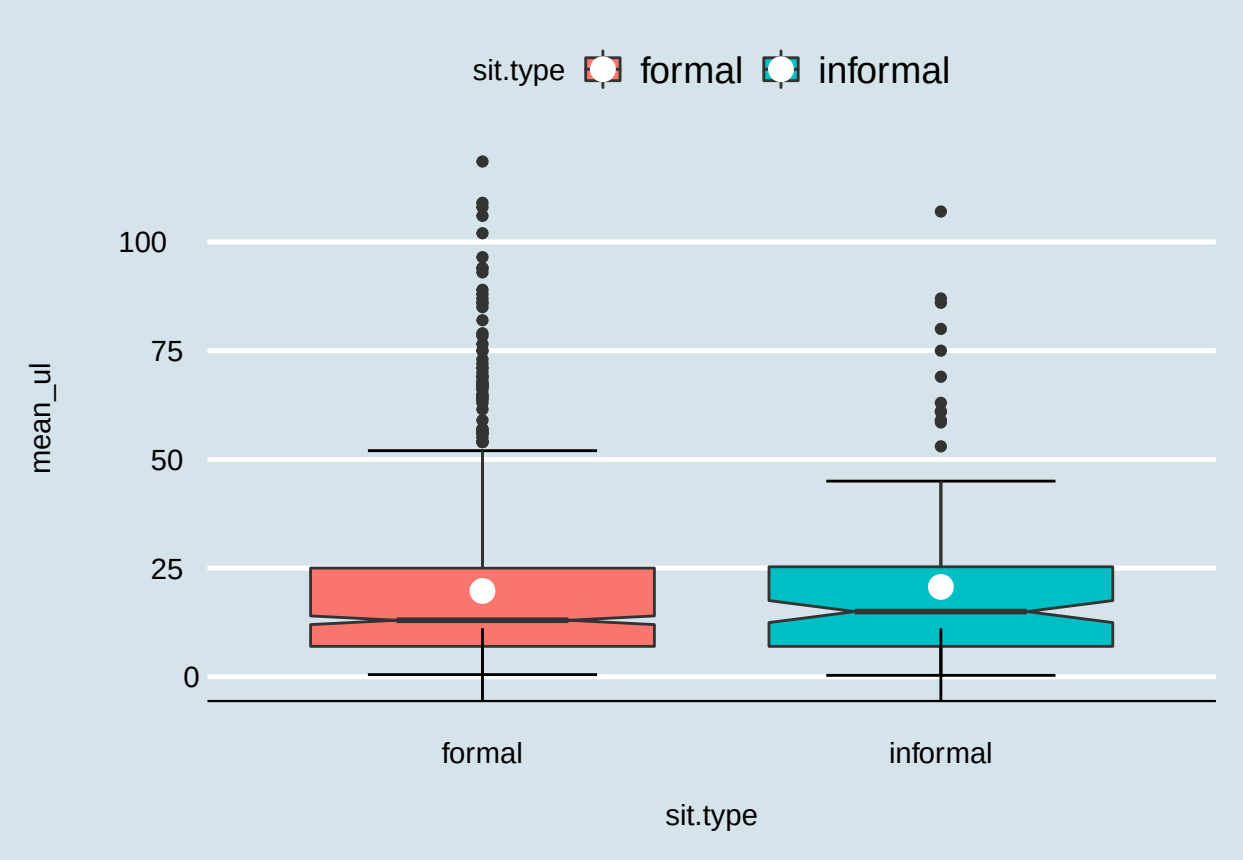
Our interpretation of the tests above depends on what alpha we choose for testing the hypotheses. As long as we pick an alpha of 0.001, like M. A. Daniel did in his research, the difference of means in both of the tests has no statistical significance. However, if we choose an alpha of 0.05 that is generally acknowledged as sufficient for the social sciences, the results appear to be significant. Whichever fashion we go for, the r-statistic still implies that the effect size is quite small in both of the tests. This fact allows us to conclude that according to the data that we have, the common assumption that women speak more is not supported statistically. Having determined this conclusion, 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 corresponding corpus entries. Firstly, we can compare the subgroups determined by the exact number of interviewers.

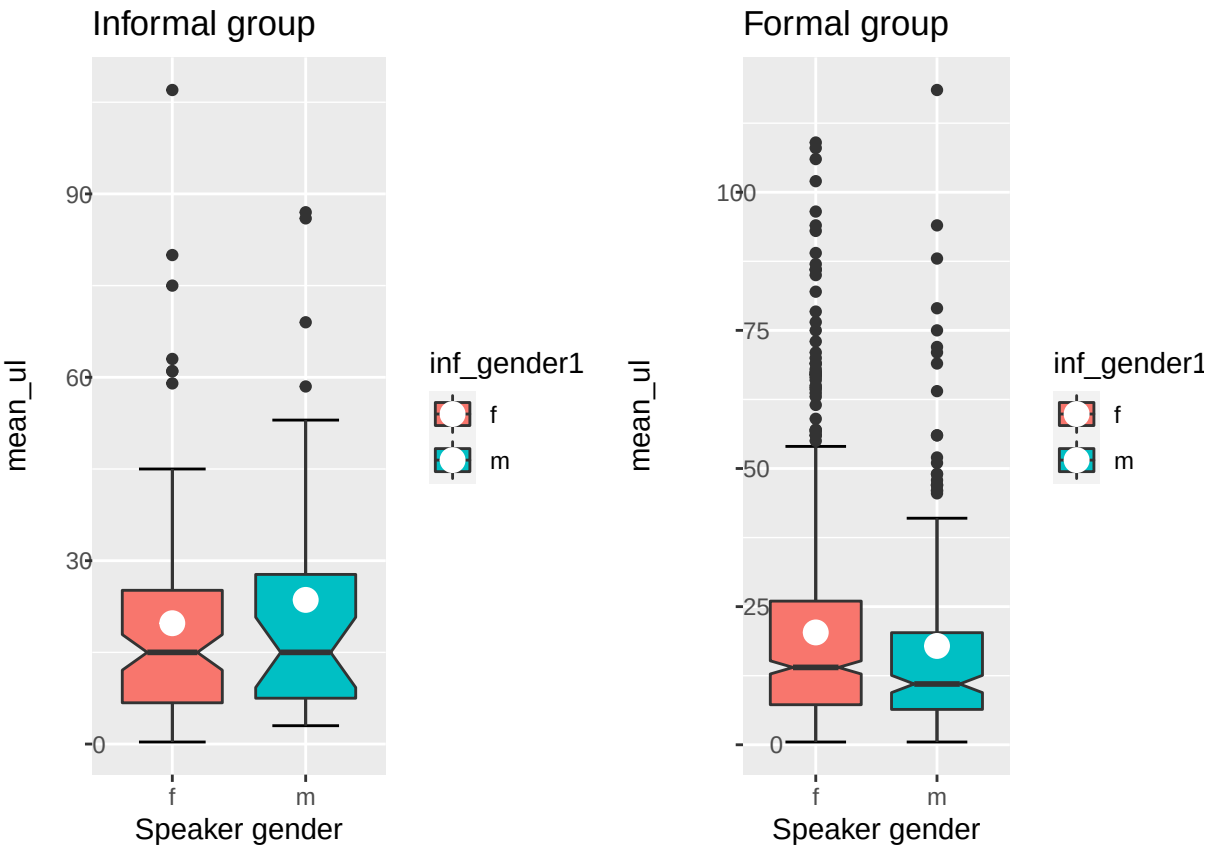


Then we may also look at the difference between the formal (>2 interviewers) and the informal subgroups. The sample sizes are 815 for formal and 131 for informal types.

```
## # A tibble: 2 x 2
##   sit.type     n
##   <chr>    <int>
## 1 formal     815
## 2 informal   131
```



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 graph for the informal group suggests that no differences are present. The graph 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. Running the same comparison tests inside the formal subgroup, 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        618   197    67602. 0.0194 *
```

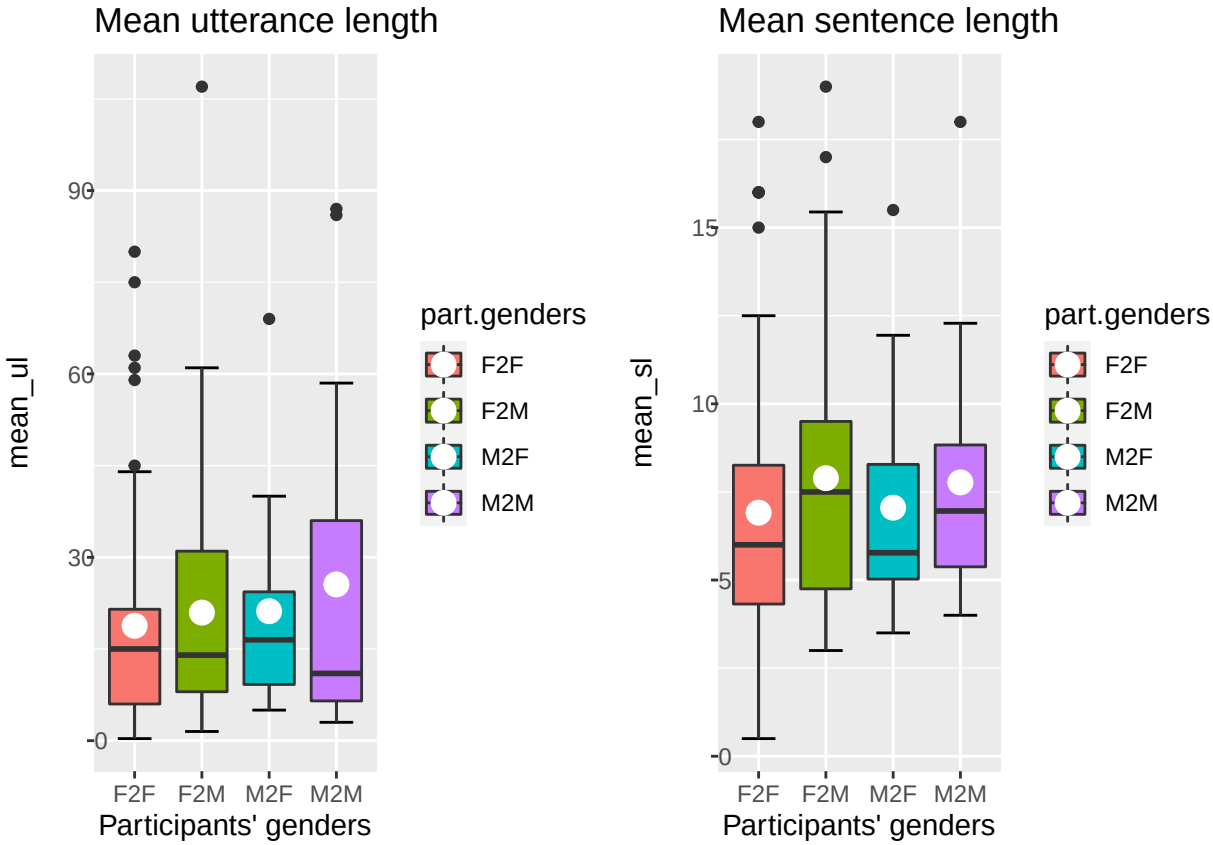
```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_ul f      m       0.0819   618   197 small
```

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        618   197    70514. 0.000804 ***
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 mean_sl f      m        0.117   618   197 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 appears to be quite notable. If we try to interpret these distinctions in terms of some real-world tendencies, we might suppose that female speakers are more eager to adapt their speaking patterns to the specific communicative situation. 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. So far the tendencies 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.



MUL

```
## # A tibble: 4 x 3
##   part.genders mean_MUL     n
##   <chr>           <dbl> <int>
## 1 F2F              18.8    55
## 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              6.91    55
## 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 an ANOVA with TukeyHSD to check this out. The reason, why we chose TukeyHSD is because this post-hoc test has less requirements than a pairwise t.test.

MUL

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## part.genders   3    607   202.4   0.494  0.687
## Residuals    127  52029   409.7
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mean_ul ~ part.genders, data = data_single)
##
## $part.genders
##              diff        lwr      upr     p adj
## F2M-F2F 2.1515373  -8.440298 12.74337 0.9519426
## M2F-F2F 2.3809234 -13.392937 18.15478 0.9793308
## M2M-F2F 6.7672403  -7.855179 21.38966 0.6249252
## M2F-F2M 0.2293861 -15.896164 16.35494 0.9999817
## M2M-F2M 4.6157030 -10.385425 19.61683 0.8538180
## M2M-M2F 4.3863170 -14.631084 23.40372 0.9317390
```

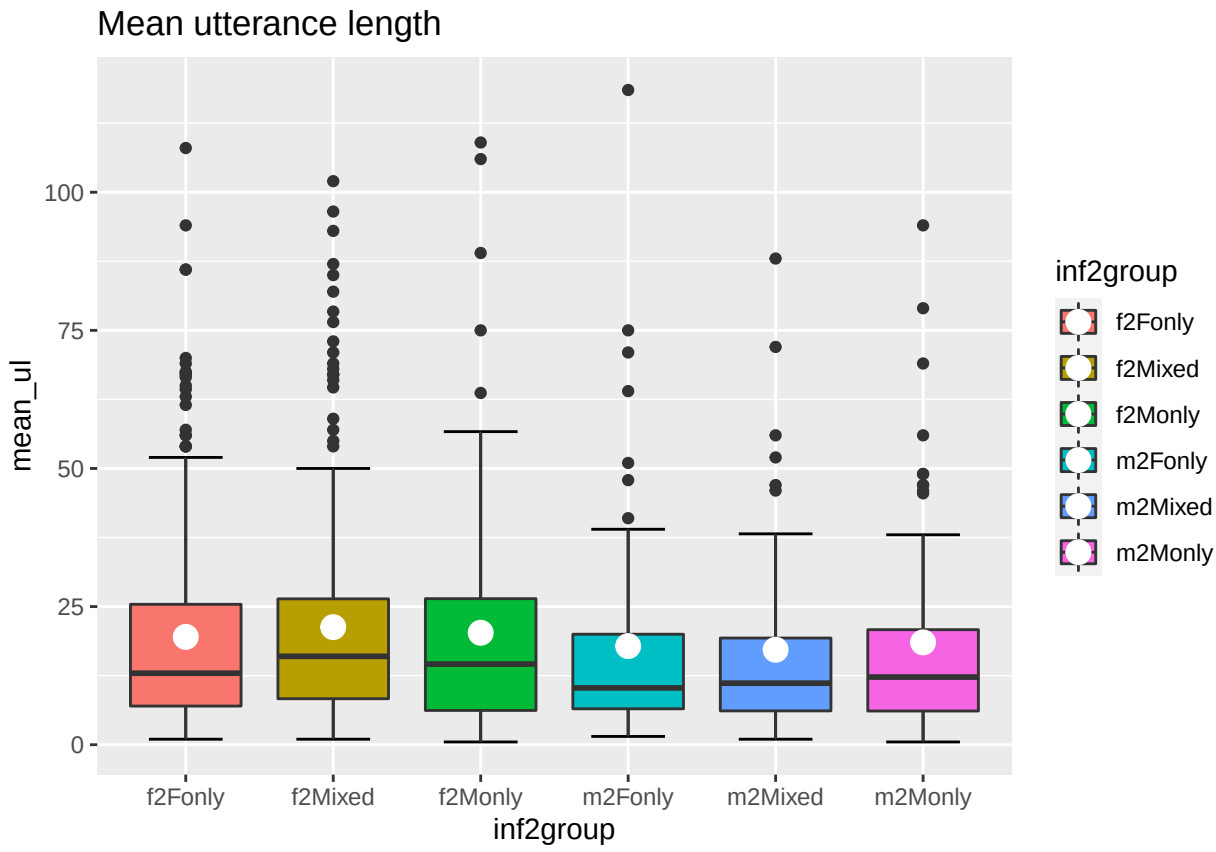
MSL

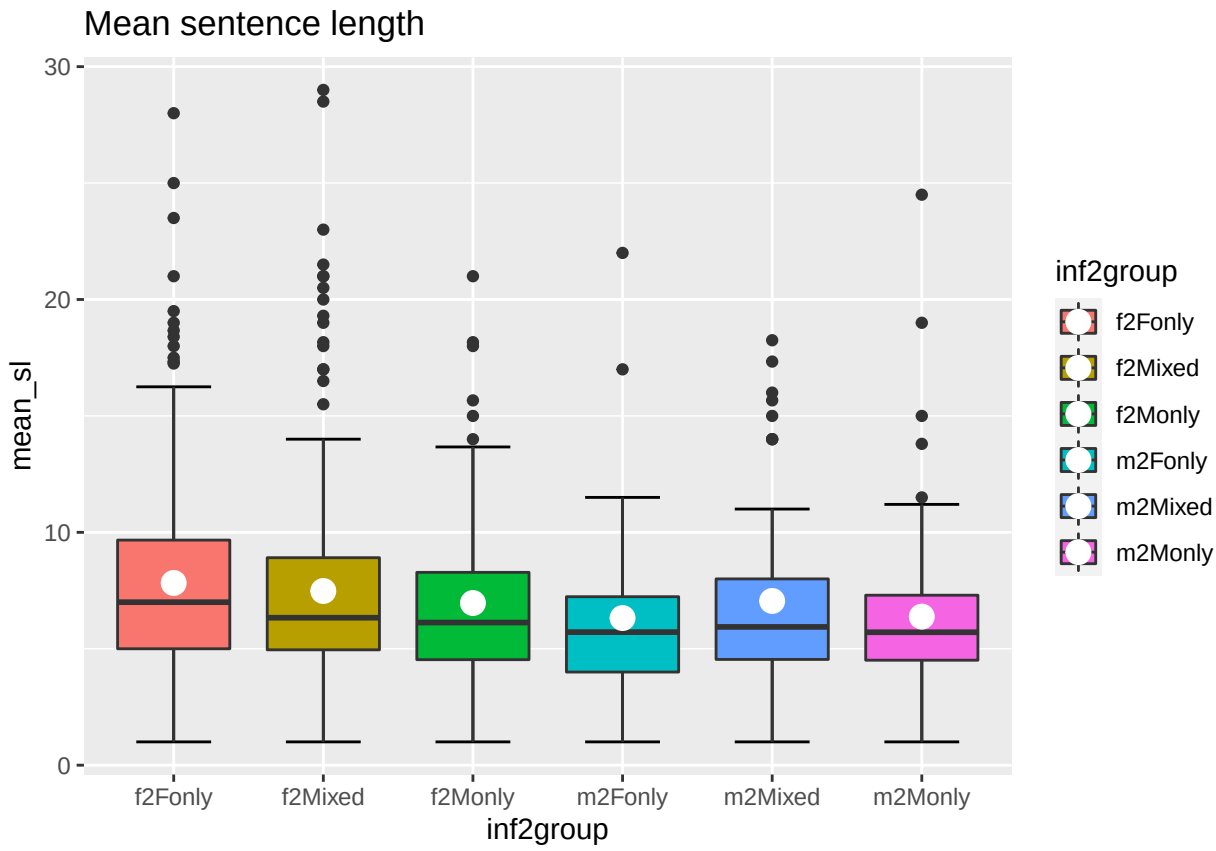
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## part.genders   3   27.9   9.311   0.612  0.608
## Residuals    127 1931.3  15.207
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mean_sl ~ part.genders, data = data_single)
##
## $part.genders
##                diff       lwr      upr     p adj
## F2M-F2F  0.9797111 -1.060959 3.020381 0.5964374
## M2F-F2F  0.1421809 -2.896881 3.181243 0.9993509
## M2M-F2F  0.8606398 -1.956580 3.677860 0.8564542
## M2F-F2M -0.8375301 -3.944350 2.269290 0.8961965
## M2M-F2M -0.1190713 -3.009255 2.771113 0.9995562
## M2M-M2F  0.7184589 -2.945518 4.382436 0.9564613
```

As the reports suggest, the slight difference that exists is of no statistical significance.

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.


Mean utterance length
mean_ul
100
75
50
25
0
f2Fonly
f2Mixed
f2Monly
m2Fonly
m2Mixed
m2Monly
inf2group
inf2group
f2Fonly
f2Mixed
f2Monly
m2Fonly
m2Mixed
m2Monly


Mean sentence length
mean_sl
30
20
10
0
f2Fonly
f2Mixed
f2Monly
m2Fonly
m2Mixed
m2Monly
inf2group
inf2group
f2Fonly
f2Mixed
f2Monly
m2Fonly
m2Mixed
m2Monly

Although the graphs show no drastic difference between the distributions, the means are obviously not equal, which is why we may compare them using the ANOVA test and utilizing the Tukey Honest Significant differences as a post-hoc comparison.

MUL

```
## # A tibble: 6 x 3
##   inf2group   MUL   num
##   <chr>     <dbl> <int>
## 1 f2Fonly    19.5   268
## 2 f2Mixed    21.3   235
## 3 f2Monly    20.3   115
## 4 m2Fonly    17.9    65
## 5 m2Mixed    17.2    62
## 6 m2Monly    18.5    70
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## inf2group     5   1344   268.7   0.732    0.6
## Residuals   809 296958   367.1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mean_ul ~ inf2group, data = grouped)
##
## $inf2group
##                         diff        lwr       upr     p adj
## f2Mixed-f2Fonly  1.7661861  -3.124834  6.657206 0.9073024
## f2Monly-f2Fonly  0.7429578  -5.358026  6.843941 0.9993301
## m2Fonly-f2Fonly -1.6623699  -9.229214  5.904475 0.9889789
## m2Mixed-f2Fonly -2.3325552 -10.045327  5.380217 0.9549011
## m2Monly-f2Fonly -0.9933902  -8.339522  6.352741 0.9988865
## f2Monly-f2Mixed -1.0232283  -7.251506  5.205049 0.9971660
## m2Fonly-f2Mixed -3.4285560 -11.098405  4.241293 0.7976608
## m2Mixed-f2Mixed -4.0987414 -11.912594  3.715111 0.6654218
## m2Monly-f2Mixed -2.7595763 -10.211763  4.692611 0.8978159
## m2Fonly-f2Monly -2.4053277 -10.898060  6.087404 0.9658914
## m2Mixed-f2Monly -3.0755131 -11.698518  5.547492 0.9116867
## m2Monly-f2Monly -1.7363480 -10.033035  6.560339 0.9911808
## m2Mixed-m2Fonly -0.6701854 -10.385707  9.045336 0.9999591
## m2Monly-m2Fonly  0.6689797  -8.758117 10.096076 0.9999529
## m2Monly-m2Mixed  1.3391651  -8.205460 10.883790 0.9986695
```

MSL

```
## # A tibble: 6 x 3
##   inf2group   MSL   num
##   <chr>     <dbl> <int>
## 1 f2Fonly    7.83   268
## 2 f2Mixed    7.48   235
## 3 f2Monly    6.96   115
## 4 m2Fonly    6.32    65
## 5 m2Mixed    7.05    62
## 6 m2Monly    6.38    70
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
```

```
## inf2group     5    220   44.02   2.647  0.022 *
## Residuals   809  13455   16.63
## ---
## 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 ~ inf2group, data = grouped)
##
## $inf2group
##                          diff       lwr       upr     p adj
## f2Mixed-f2Fonly -0.34446198 -1.385567 0.6966430 0.9345818
## f2Monly-f2Fonly -0.86078667 -2.159445 0.4378718 0.4069028
## m2Fonly-f2Fonly -1.50655330 -3.117236 0.1041291 0.0820599
## m2Mixed-f2Fonly -0.77218640 -2.413931 0.8695582 0.7605133
## m2Monly-f2Fonly -1.44473834 -3.008440 0.1189630 0.0891772
## f2Monly-f2Mixed -0.51632469 -1.842079 0.8094296 0.8762718
## m2Fonly-f2Mixed -1.16209132 -2.794699 0.4705166 0.3242764
## m2Mixed-f2Mixed -0.42772442 -2.090985 1.2355363 0.9776147
## m2Monly-f2Mixed -1.10027636 -2.686553 0.4860000 0.3539132
## m2Fonly-f2Monly -0.64576663 -2.453534 1.1620007 0.9111472
## m2Mixed-f2Monly  0.08860026 -1.746897 1.9240975 0.9999931
## m2Monly-f2Monly -0.58395168 -2.349989 1.1820853 0.9347467
## m2Mixed-m2Fonly  0.73436689 -1.333684 2.8024177 0.9131974
## m2Monly-m2Fonly  0.06181496 -1.944842 2.0684716 0.9999993
## m2Monly-m2Mixed -0.67255194 -2.704226 1.3591218 0.9344439
```

Once again, the difference seems to be too small to be considered. Although the ANOVA shows that not all groups have a similar mean MSL, the post-hoc test proves that the differences are not significant. However, we can state that the difference between the female-to-female group and two other groups, namely male-to-female and male-to-male, are on the verge of statistical significance. Of course, the latter fact does not allow us to make any strict conclusions.

Discussion

- It may seem from the experiments above that we failed to achieve any significant result, as we did not find any notable tendencies. However, the main goal of the study was not to confirm, but to test a common assumption about the communicative behavior of certain genders. In this sense the experiment can be called a success, as our study, made on a pragmatically uniform corpus, proves that the gender-based differences are not statistically significant.
- 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. We would need some markup for speaker's age and for the age of the interviewers to study this question in detail.

References

- Daniel, M. A., Zelenkov, Yu. G. (2012). 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