A reliability study on the ACLEW corpora

Experimental set-up

For studying the reliability of human annotators, and get a sense of their level of agreement, we asked to a second person to annonate a 1-mn long chunk from the daylong recording of each child. The purpose being to compare metrics obtained by these two annotators. This reliability study has been performed on SOD, WAR, ROS, TSE, ROW and BER corpus. We have 10 children by sub-corpora, for a total of 60 chunks of 1 minute.

We mapped all the labels into :

- CHI : for the key-child, the one wearing the recording device
- OCH : for other chidren
- MAL : for male speakers
- FEM : for female speakers
- OVL : for overlap
- SIL : for silence

Performances metrics

Identification Error Rate as a overall performance measure

One might want compare the level agreement of the two annotators as a function of the identification error rate. As a reminder, the identification error rate is computed as follow :

$$\text{identification error rate} = \frac{\text{false alarm} + \text{miss} + \text{confusion}}{\text{total}}$$

where :

false alarm is the duration of non-speech incorrectly classified as speech

miss is the duration of speech incorrectly classified as non-speech

confusion is the duration of speaker confusion (agreements on the fact that there's speech, but disagreement on the talker identity).

speech is the duration of speaker confusion (agreements on the fact that there's speech, but disagreement on the talker identity).

The two annotators obtained an identification error rate of 59.09% shared amongst a false alarm rate of 19.13%, a miss rate of 22.4% and a confusion of 17.56%.

Here's the per corpora identification error rate :

corpora	ider
ROW	44.73892
BER	52.10606
WAR	54.10426
ALL	59.09289
SOD	60.39397
ROS	71.00388
TSE	72.14068

Best cases

The three best cases, for which the agreement was the highest were :

	ider%	total	correct	correct%	fa	fa%	miss	miss%	conf	conf%	co
WAR_3528_006660_006720.rttm	11.40	40.01	36.12	90.28	0.67	1.67	3.89	9.72	0	0	W
WAR_4995_026700_026760.rttm	11.45	46.53	43.68	93.87	2.48	5.33	2.85	6.13	0	0	W
WAR_9398_005100_005160.rttm	11.48	14.89	14.20	95.37	1.02	6.85	0.69	4.63	0	0	W

Worst cases

The three worst cases, for which the disagreement was the highest were :

	ider%	total	correct	correct%	fa	fa%	miss	miss%	conf	conf%
WAR_1130_023040_023100.rttm	921.74	4.83	1.89	39.13	41.58	860.87	0.19	3.93	2.75	56.94
TSE_0643_020364_020424.rttm	521.55	6.96	6.03	86.64	35.37	508.19	0.90	12.93	0.03	0.43
ROS_1299_004320_004380.rttm	153.48	33.88	6.72	19.83	24.84	73.32	0.25	0.74	26.91	79.43

Detection Error Rate as a per-class performance measure

One can have a look at the per-class detection error rate defined as :

$$\text{detection error rate} = \frac{\text{false alarm} + \text{miss}}{\text{total}}$$

	ALL	BER	ROW	WAR	TSE	ROS	SOD
CHI	20.40425	31.49816	24.74762	8.927915	28.18321	38.19307	14.44508
FEM	28.71480	21.16044	32.99284	22.006243	24.01249	49.93739	23.35978
MAL	35.92883	19.15982	21.95915	34.973005	60.27861	45.32833	100.00000
OCH	43.37580	74.30380	27.84415	28.211749	48.86864	34.15829	61.59200
OVL	56.07320	64.01310	32.90025	71.106030	72.94763	17.16997	70.76886
ELE	64.71095	48.72972	17.44501	81.967213	NaN	66.80583	100.00000

With no surprise, there's a high disagreement for classes such as the OVL one for which it is harder to tell when it starts and it ends exactly. The highest agreement is obtained for the CHI class for which the two annotators obtainted a detection error rate of 45.1%

Best agreement for the CHI class

The three best cases, for which the agreement on the CHI class was the highest were :

	deter%	total	fa	fa%	miss	miss%
BER_6035_030360_030420.rttm	0	0	0	NA	0	NA
BER_7758_034320_034380.rttm	0	0	0	NA	0	NA
ROS_3510_004740_004800.rttm	0	0	0	NA	0	NA

Worst agreement for the CHI class

The three worst cases, for which the disagreement on the CHI class was the highest were :

	deter%	total	fa	fa%	miss	miss%
SOD_3634_036120_036180.rttm	7466.67	0.03	2.24	7466.67	0.00	0
TSE_7220_030589_030649.rttm	1650.00	0.12	1.86	1550.00	0.12	100
BER_1196_034740_034800.rttm	817.27	1.39	11.36	817.27	0.00	0

Precision/Recall as a per-class performance measure

As illustrated by the two tables shown above, the detection error rate (like the identification error rate) can be tricky to interpret when little speech is contained in the chunk. Indeed, in that particular case, the denominator is close to 0 (or equal to 0 if there's no speech), hence pumping up the measure. One might be more familiar with metrics such as the precision and the recall defined as :

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn}$$

where : tp is the duration of true positive (e.g. speech classified as speech)

fp is the duration of false positive (e.g. non-speech classified as speech)

fn is the duration of false negative (e.g speech classified as non-speech)

class	precision	recall
CHI	76.32	79.60
MAL	71.39	64.07
FEM	70.80	71.29
OCH	62.21	56.62
ELE	48.85	35.29
OVL	35.58	43.93

```
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
```

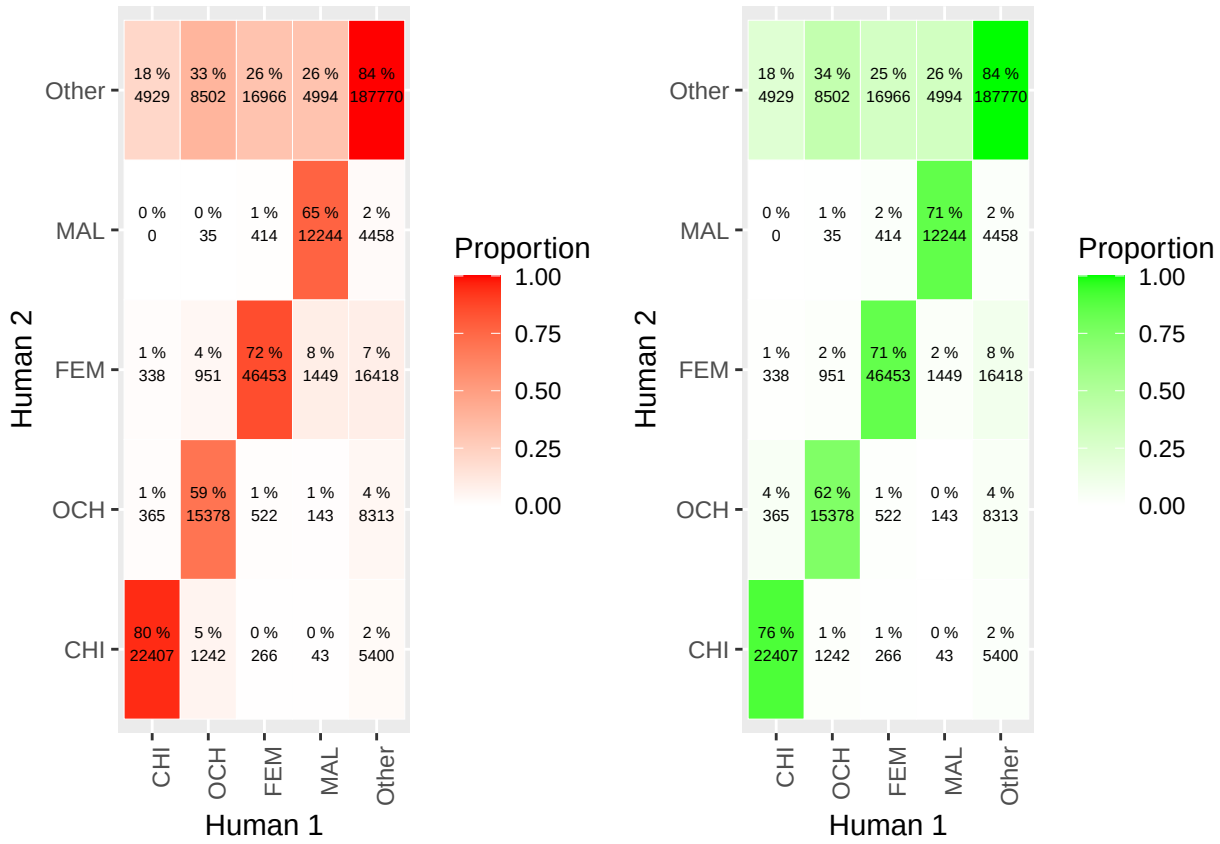


Figure 1: Precision (left) and recall (right) confusion matrices on all of the corpus

```
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
```

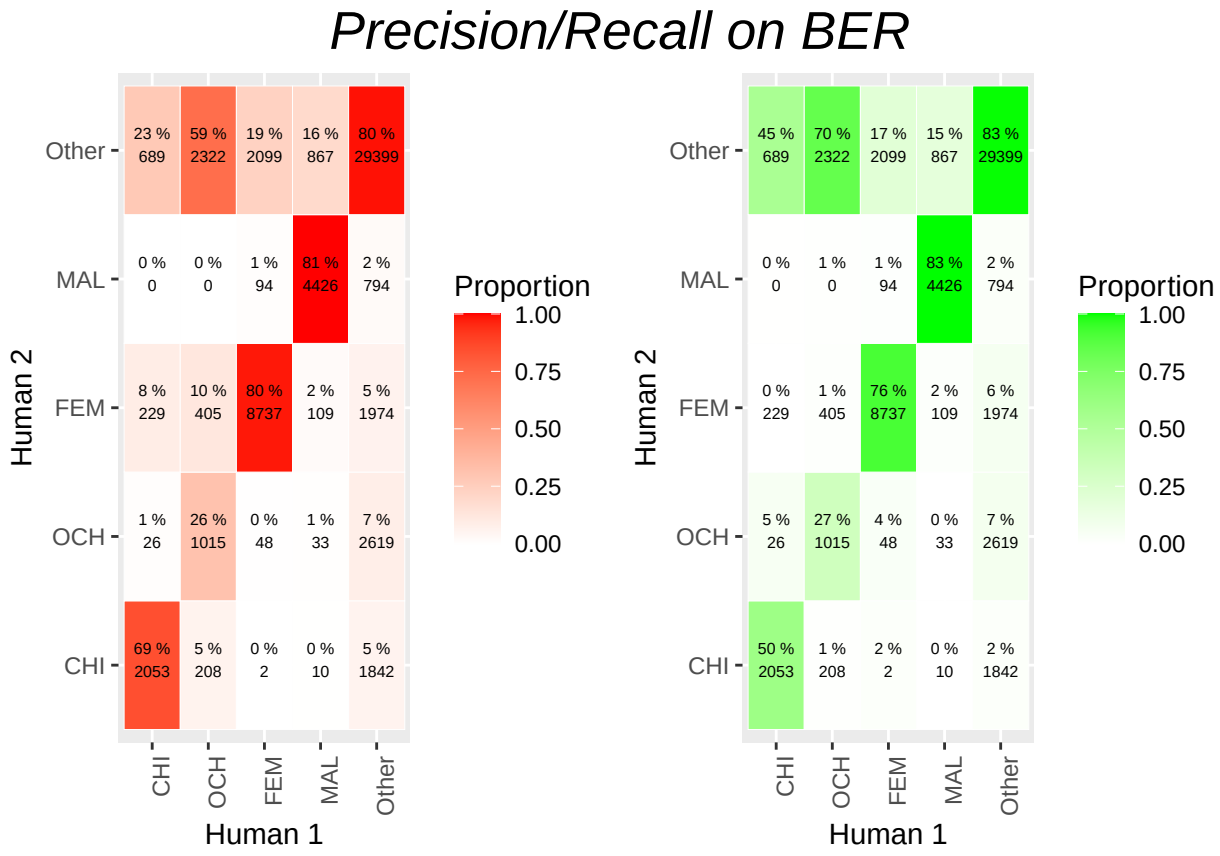


Figure 2: Precision (left) and recall (right) confusion matrices per corpora

```
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
```

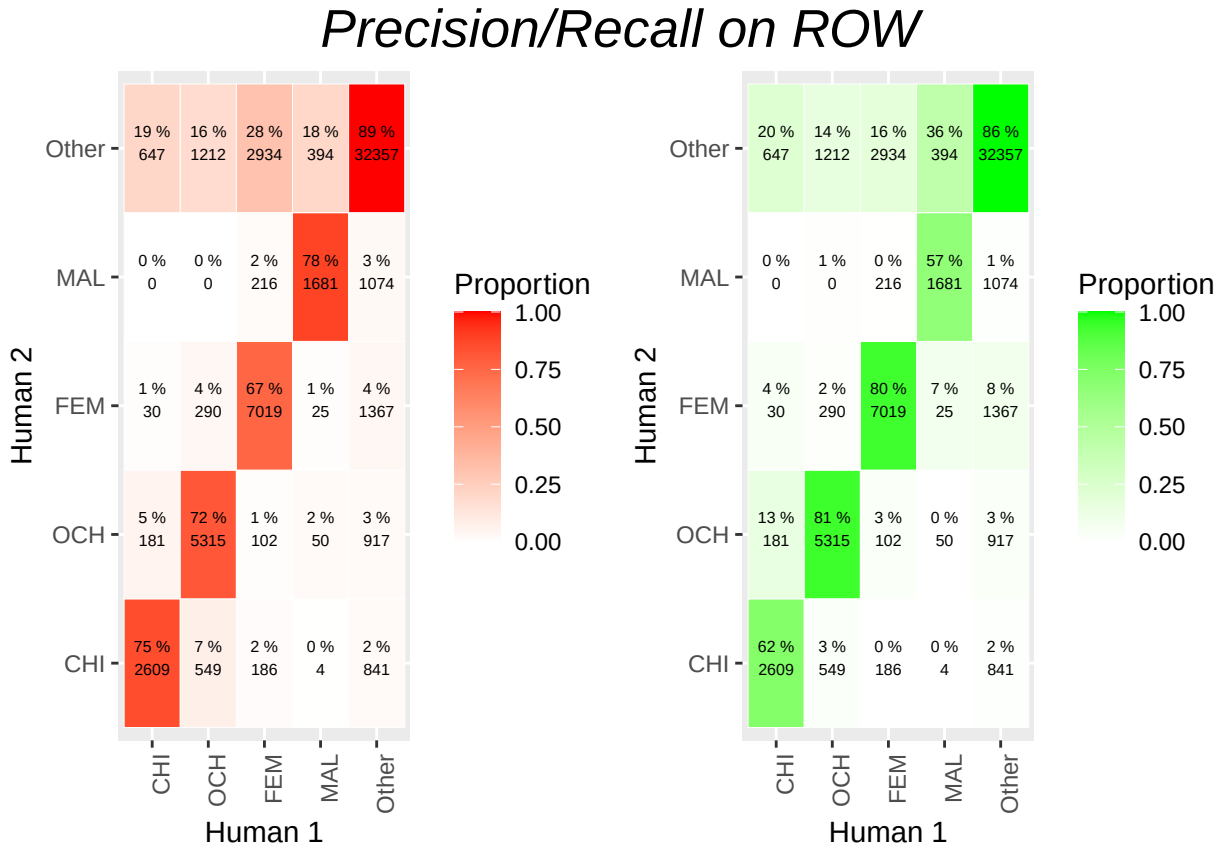


Figure 3: Precision (left) and recall (right) confusion matrices per corpora

```
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
```

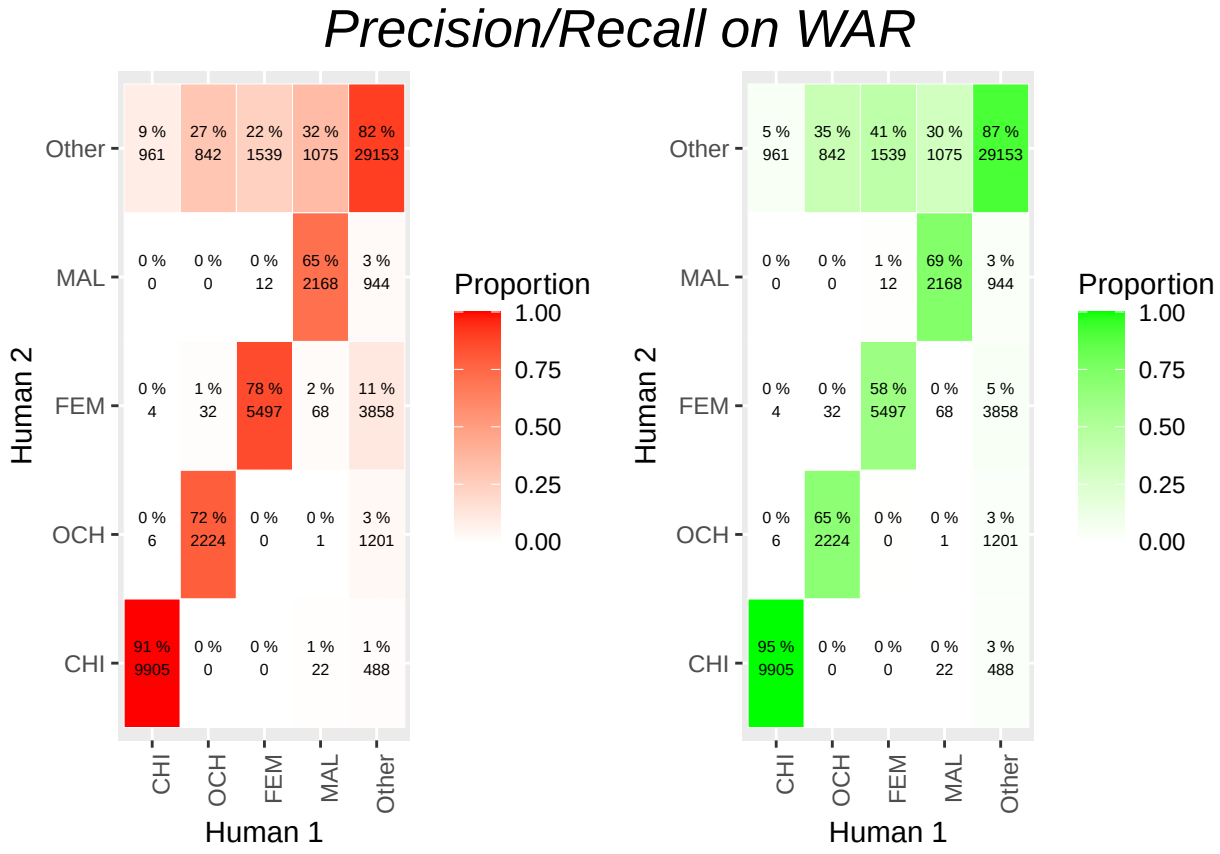


Figure 4: Precision (left) and recall (right) confusion matrices per corpora

```
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
```

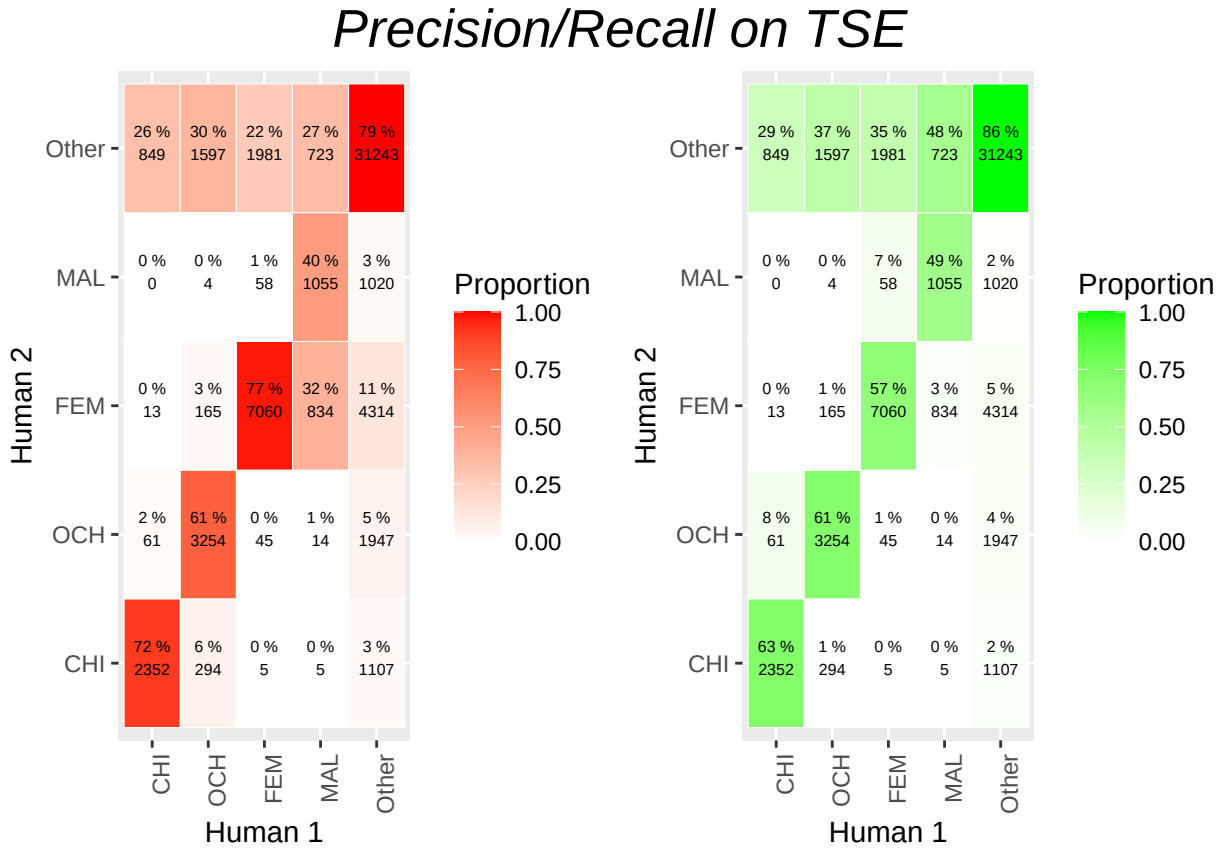


Figure 5: Precision (left) and recall (right) confusion matrices per corpora

```
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
```

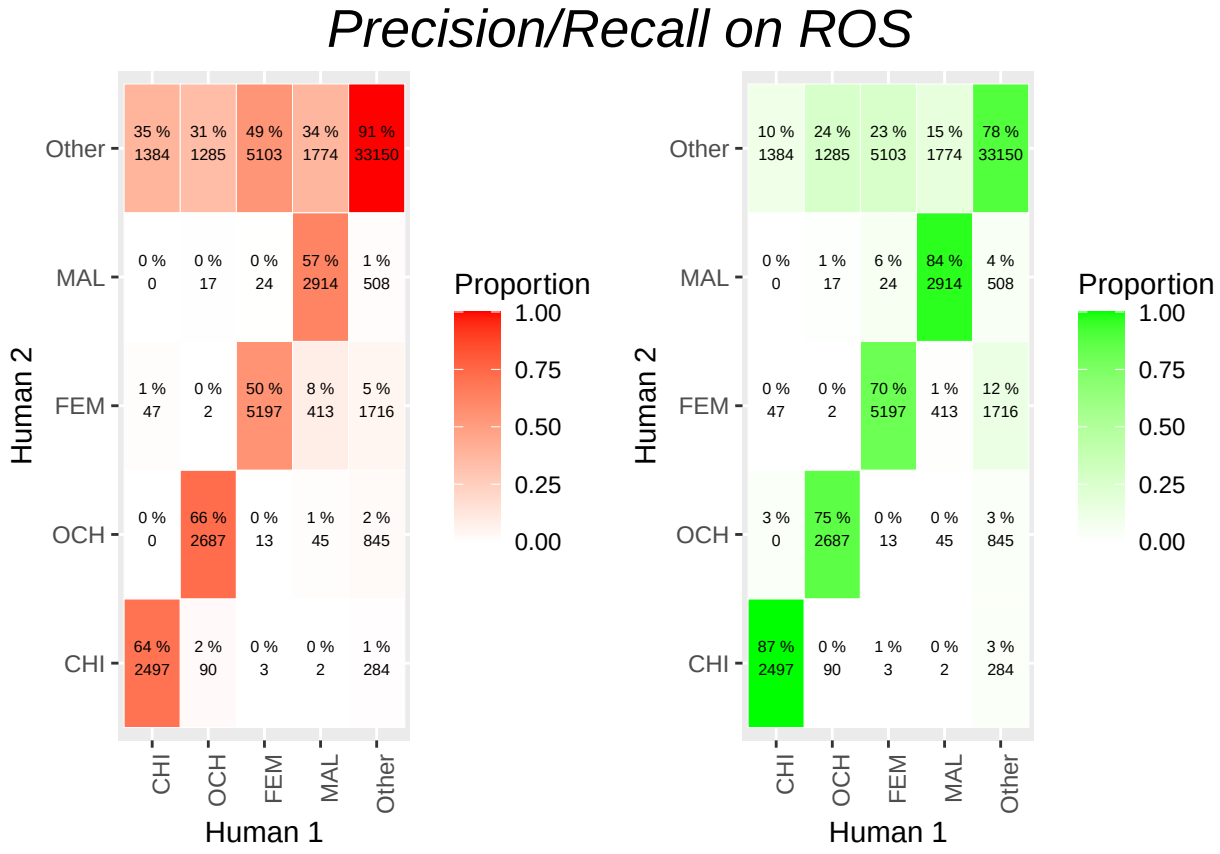


Figure 6: Precision (left) and recall (right) confusion matrices per corpora

```
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
```

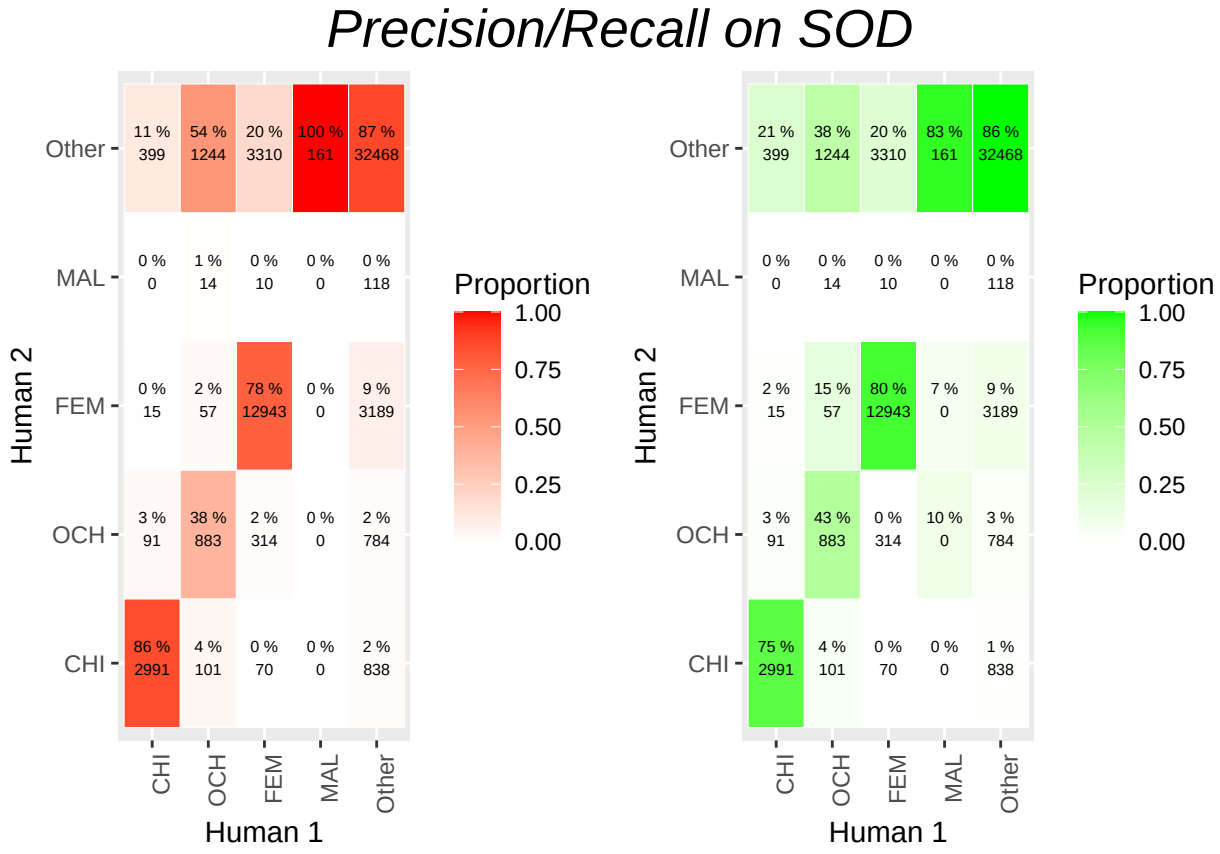


Figure 7: Precision (left) and recall (right) confusion matrices per corpora

```
## Warning: Use of `stall$pr` is discouraged. Use `pr` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$rec` is discouraged. Use `rec` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
## Warning: Use of `stall$n` is discouraged. Use `n` instead.
## Warning: Use of `stall$LENA` is discouraged. Use `LENA` instead.
## Warning: Use of `stall$human` is discouraged. Use `human` instead.
```

Cohen's kappa

As an other measure of the level of agreement between the two annotators, we propose the use of the Cohen's kappa measure, defined as follow :

$$\kappa = \frac{\Pr(\alpha) - \Pr(e)}{1 - \Pr(e)}$$

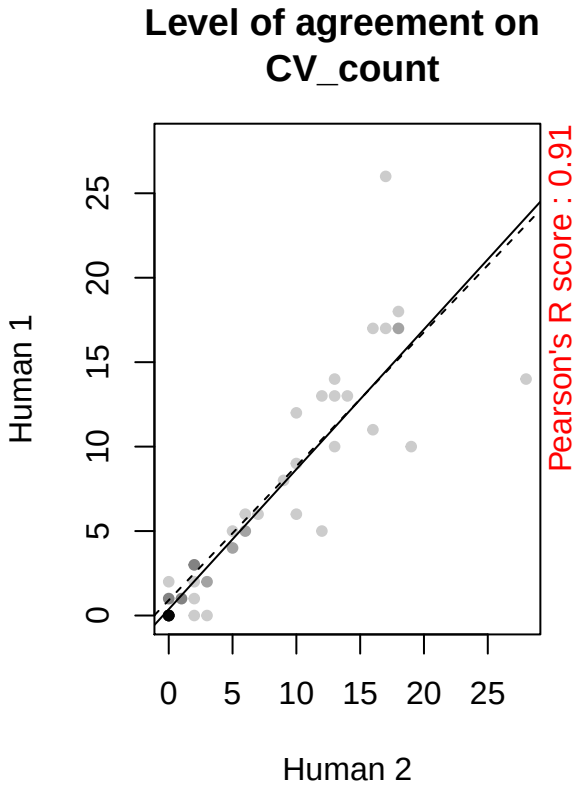
where: $\Pr(\alpha)$ is the relative agreement between the annotators. $\Pr(\alpha)$ is the probability of a random agreement on a given frame.

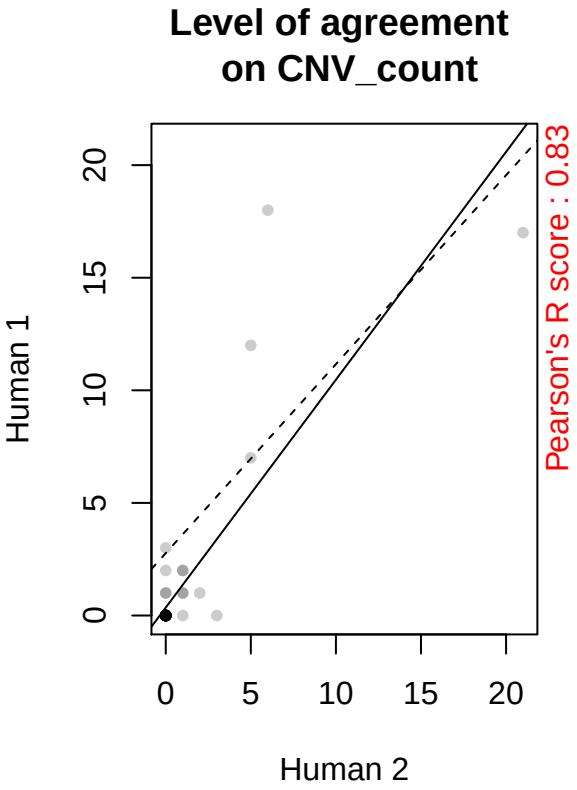
If both annotators fully agree, $\kappa = 1$, if they fully disagree (or agree randomly), $\kappa = 0$

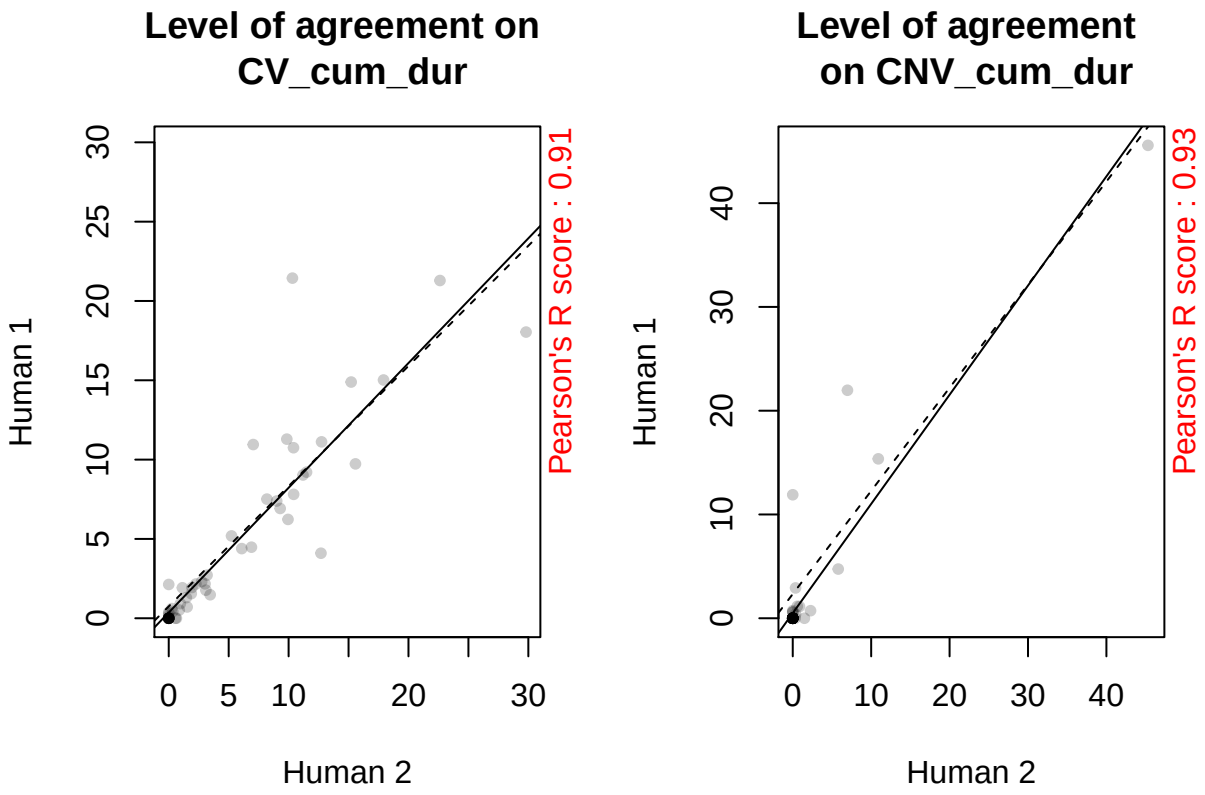
corpora	n_obs	kappa	weighted_kappa
WAR	60000	0.6988889	0.7806605
ROW	60000	0.6801391	0.7056961
SOD	60000	0.6627016	0.6595857
all	360000	0.6312279	0.6380880
ROS	60000	0.5767226	0.5864284
TSE	60000	0.5485063	0.5474113
BER	60000	0.5937692	0.4417776

Vocalizations and turn-taking

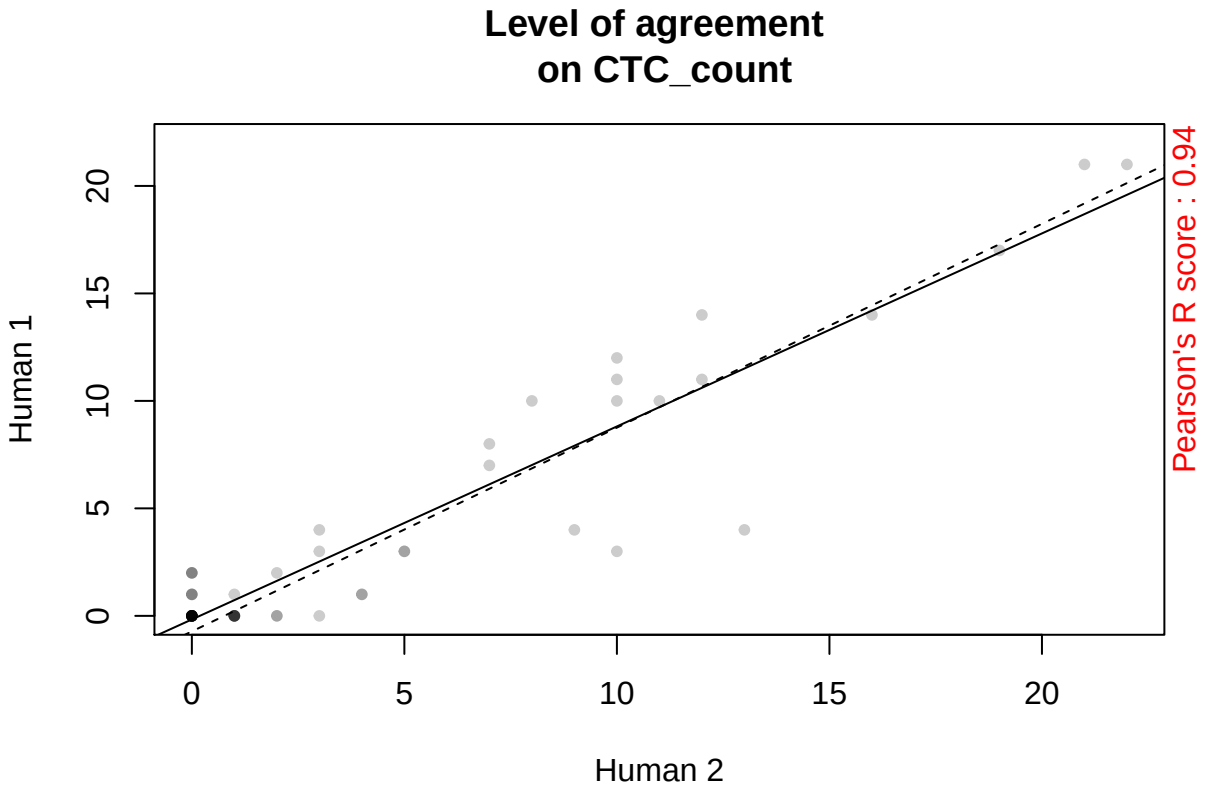
For this study, we considered only the children that have been annotated as vcm or lex, for which vocalizations were classified as C, N, W, L, U, or Y. That led us to remove 7 children from the study, for a total of 53 children.






Level of agreement on CV_cum_dur
Human 1
Human 2
Pearson's R score : 0.91
0
5
10
15
20
25
30

Level of agreement on CNV_cum_dur
Human 1
Human 2
Pearson's R score : 0.93
0
10
20
30
40


Level of agreement on CTC_count
Human 1
Human 2
Pearson's R score : 0.94
0
5
10
15
20