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 55.57% shared amongst a false alarm rate of 19.85%, a miss rate of 18.97% and a confusion of 16.75%.

Here's the per corpora identification error rate :

corpora	ider
BER	41.05884
ROW	44.29150
SOD	49.11254
WAR	54.26772
ALL	55.57232
ROS	71.43684
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
ROW_9801_000000_000001.rttm	0.00	0.00	0.00	NA	0.00	NA	0.00	NA	0	NA	R
WAR_9398_005100_005160.rttm	11.93	15.01	14.22	94.74	1.00	6.66	0.79	5.26	0	0	W
WAR_3528_006660_006720.rttm	11.95	40.26	36.12	89.72	0.67	1.66	4.14	10.28	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	21.06832	28.445883	20.98293	12.25221	28.18321	39.13366	13.80426
FEM	29.10666	21.441932	35.15565	22.72983	24.01249	50.71766	23.06536
MAL	36.56353	19.737533	21.77419	35.19952	60.27861	47.57974	100.00000
OCH	43.24036	66.388175	29.27194	30.66376	48.86864	33.25165	59.52865
ELE	49.63453	6.909967	17.61876	82.54005	NaN	66.80583	NaN
OVL	50.02050	40.797218	36.26136	71.31915	72.94763	16.28559	64.08200

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 40.27%

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%
TSE_7220_030589_030649.rttm	1650.00	0.12	1.86	1550.00	0.12	100.00
ROW_2745_020220_020280.rttm	338.71	0.31	1.05	338.71	0.00	0.00
SOD_1499_024300_024360.rttm	226.12	1.34	2.05	152.99	0.98	73.13

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	80.44	78.93
FEM	72.10	70.89
MAL	72.05	63.44
OCH	64.50	56.76
ELE	50.35	50.37
OVL	35.40	49.98

```
## Warning: package 'scales' was built under R version 3.4.4
```

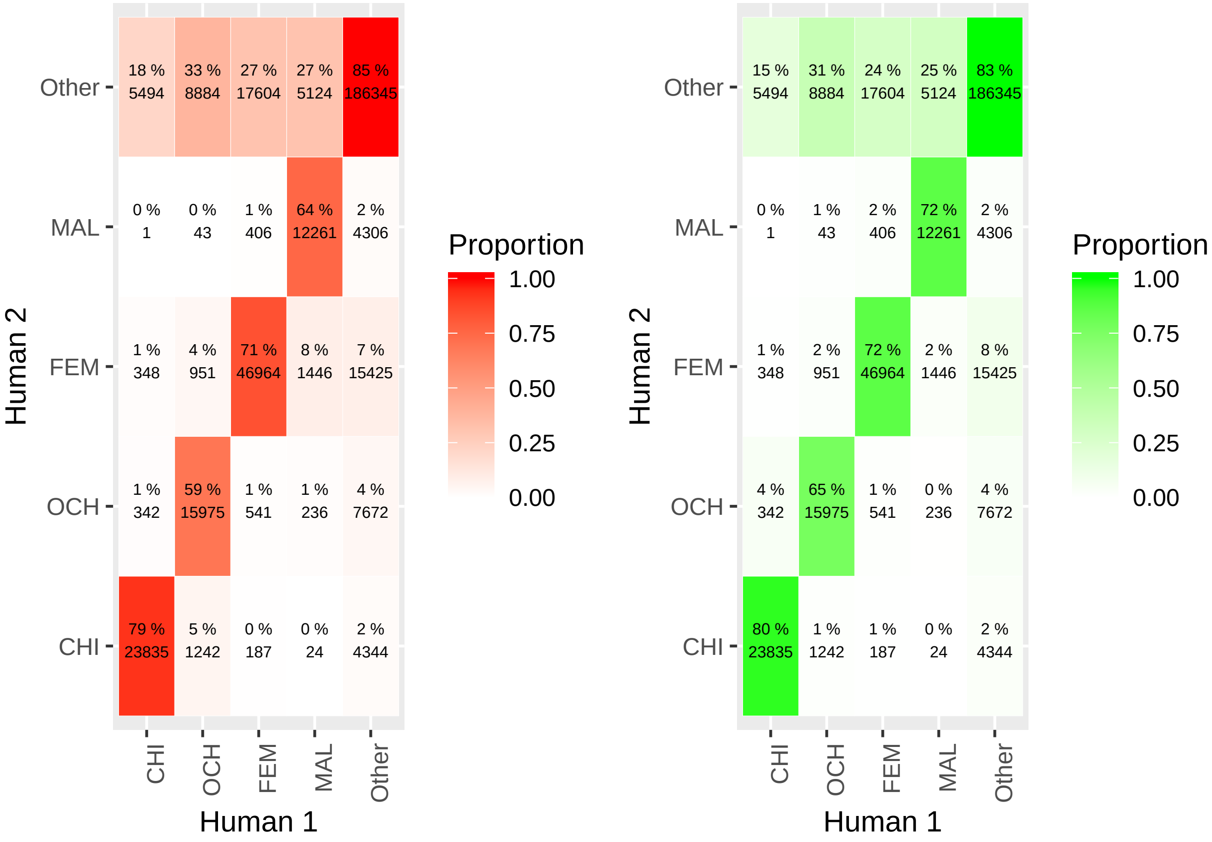
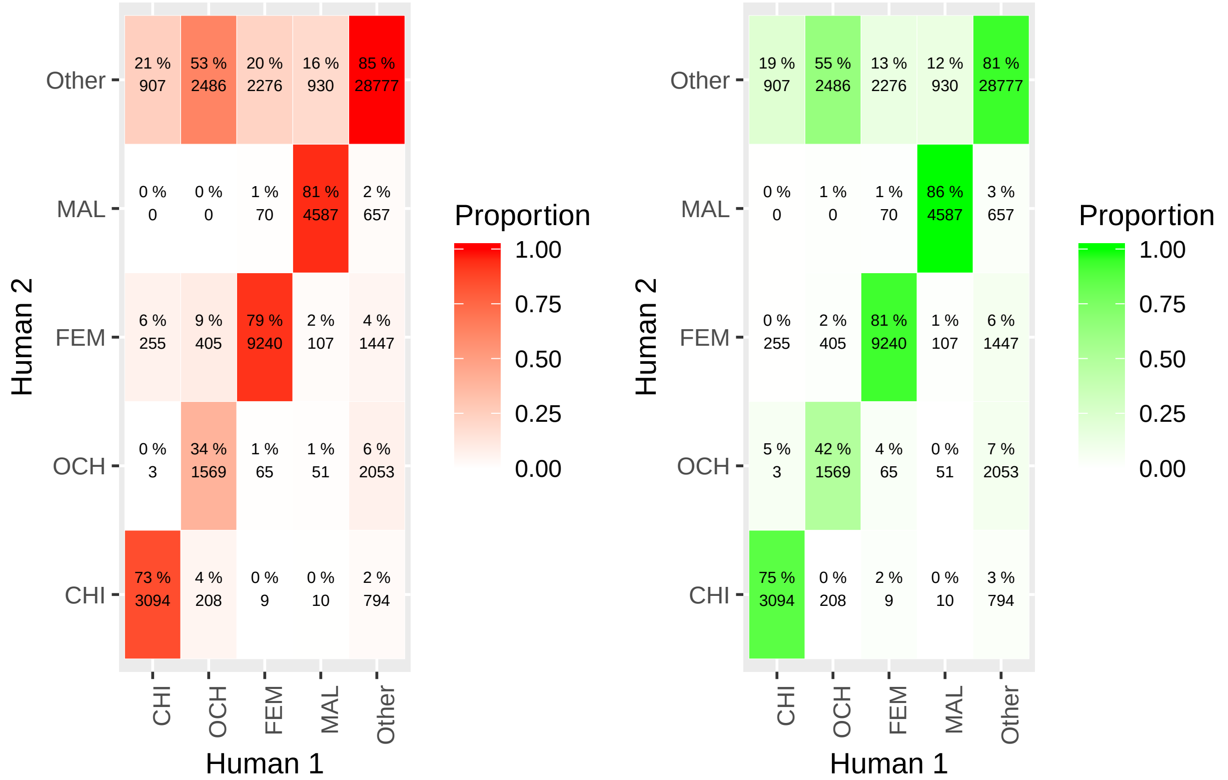
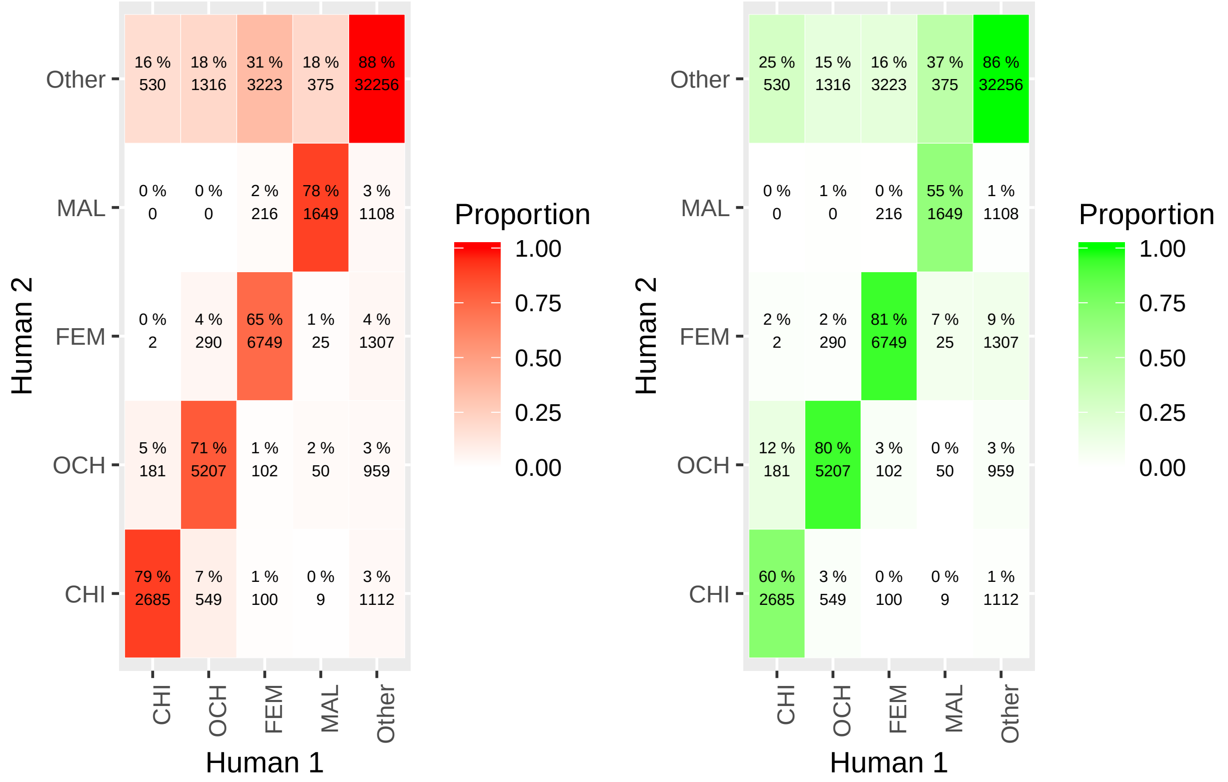


Figure 1: Precision (left) and recall (right) confusion matrices on all of the corpus

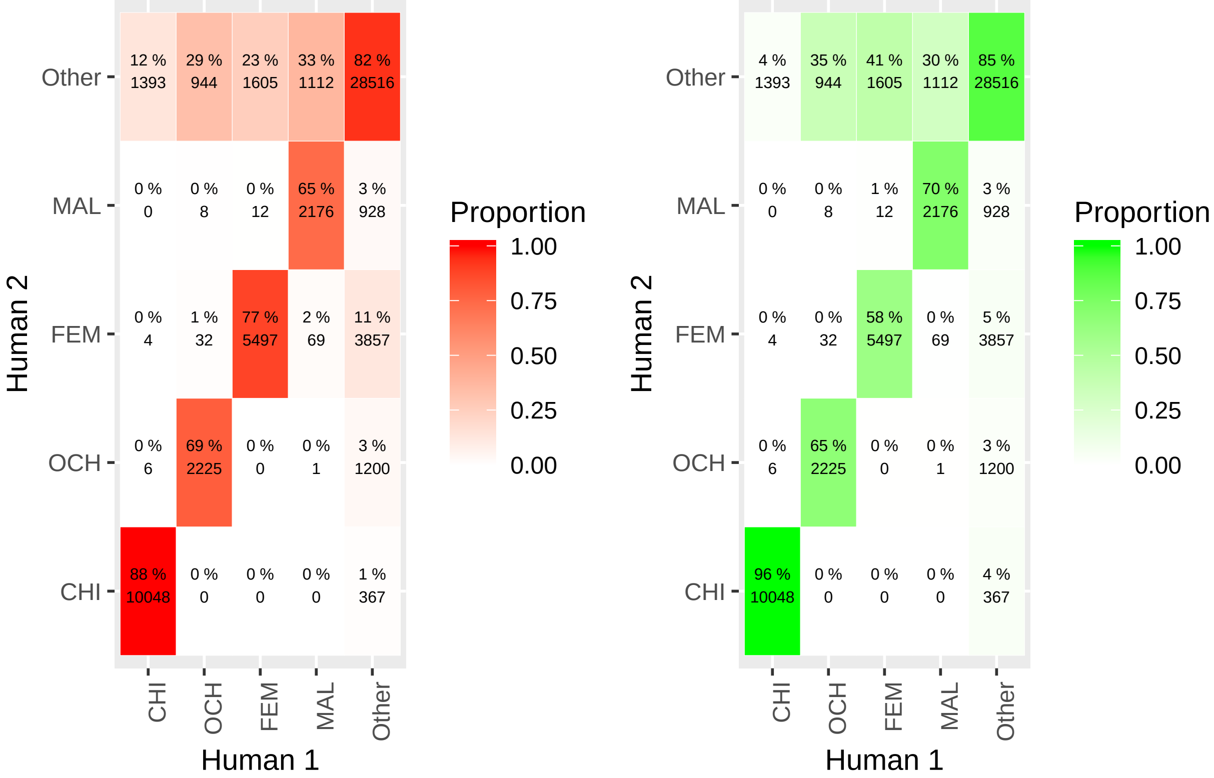
Precision/Recall on BER



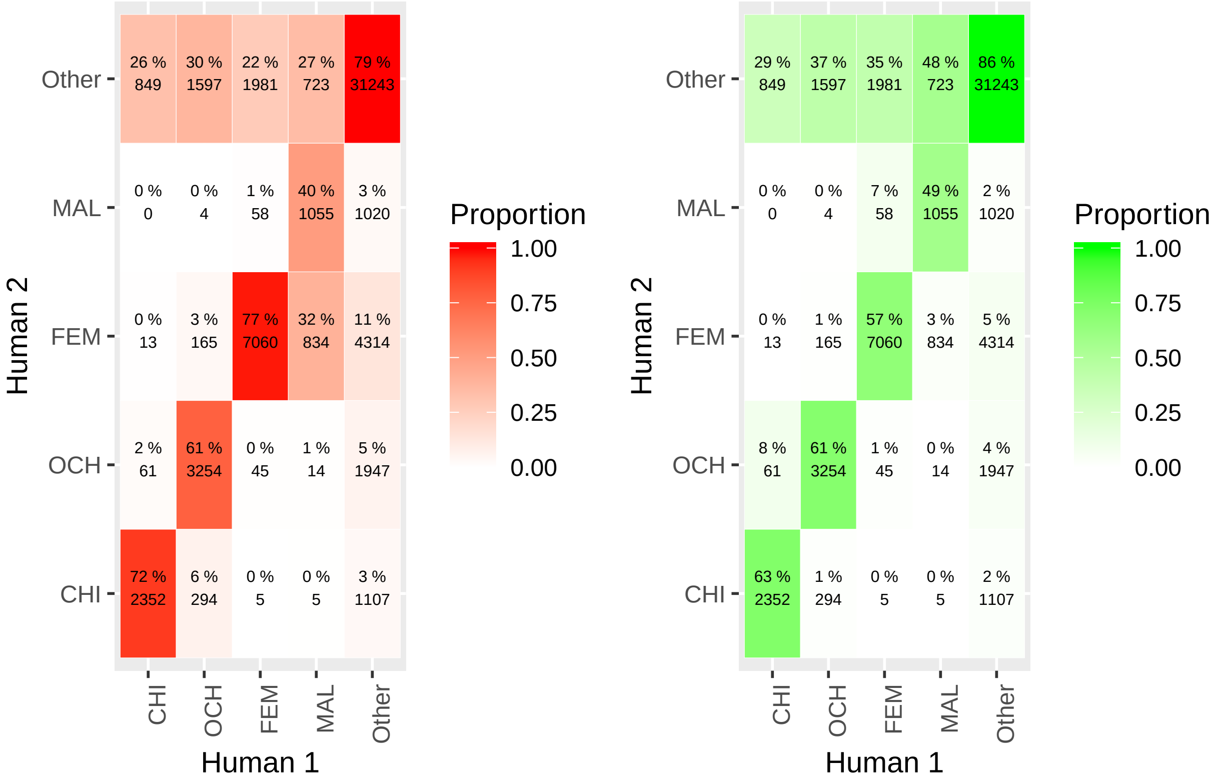
Precision/Recall on ROW



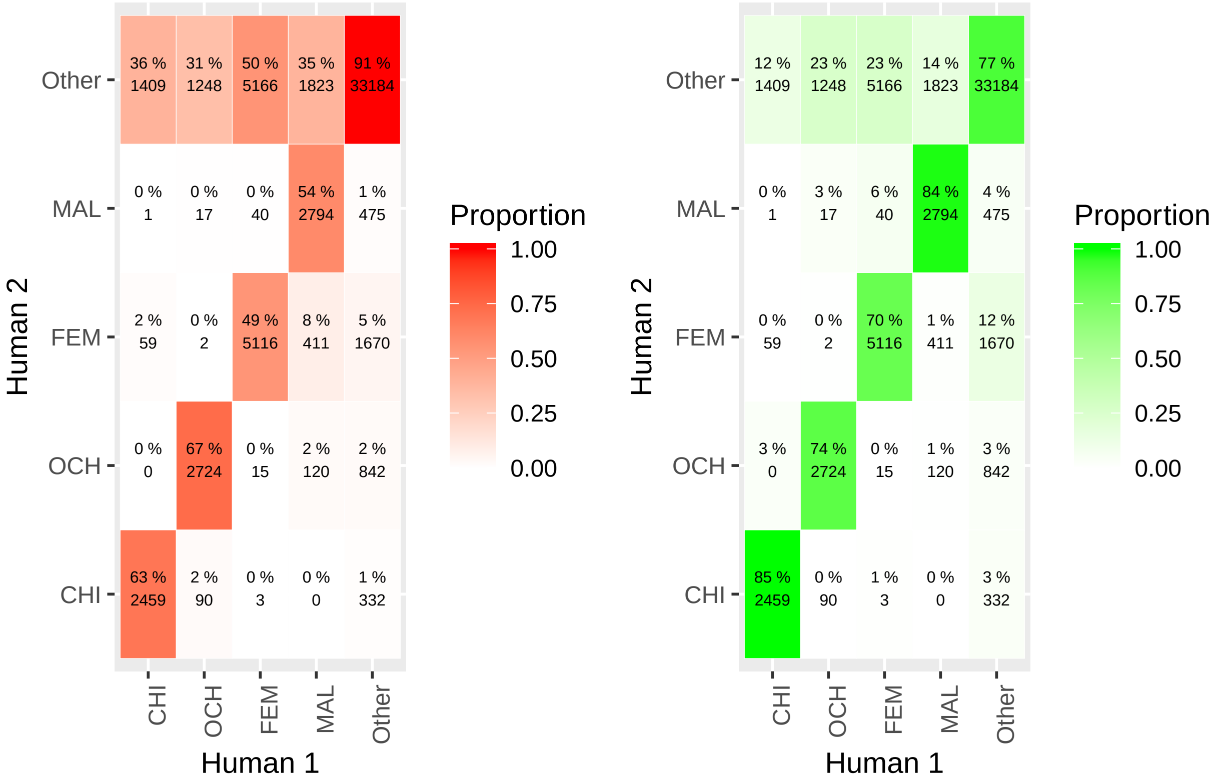
Precision/Recall on WAR



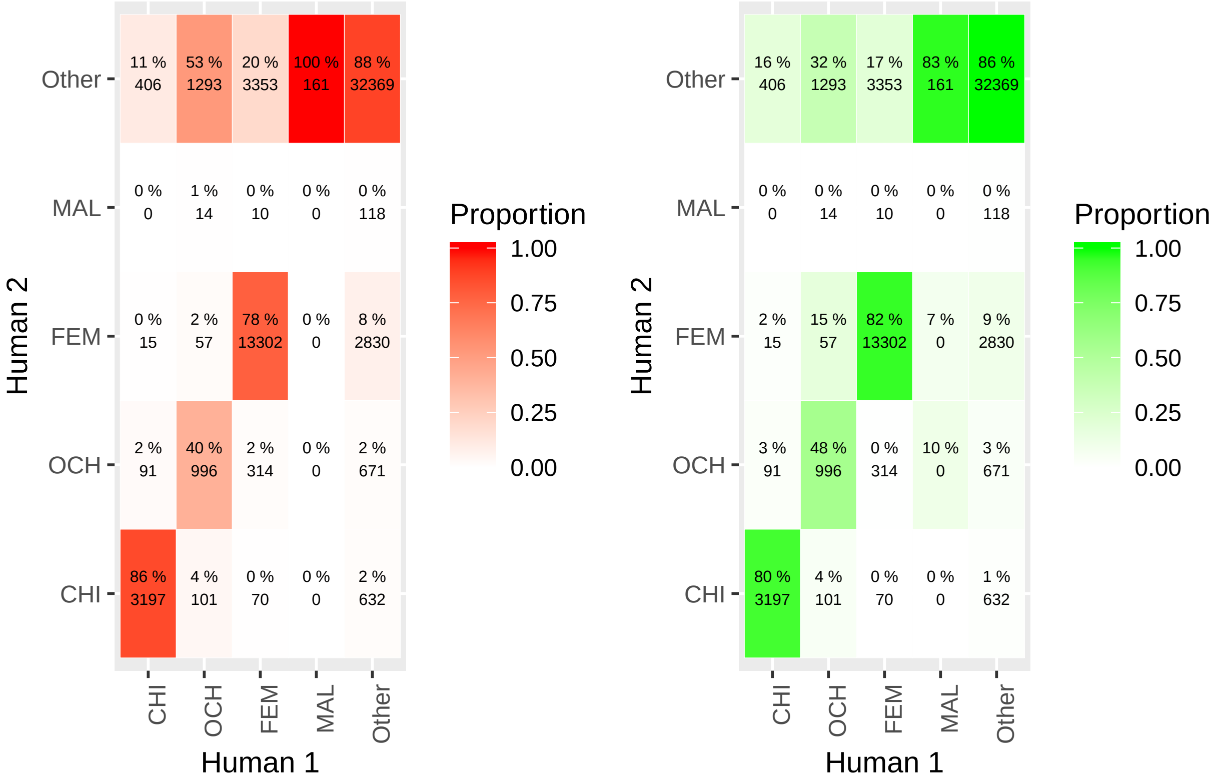
Precision/Recall on TSE



Precision/Recall on ROS



Precision/Recall on SOD



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.

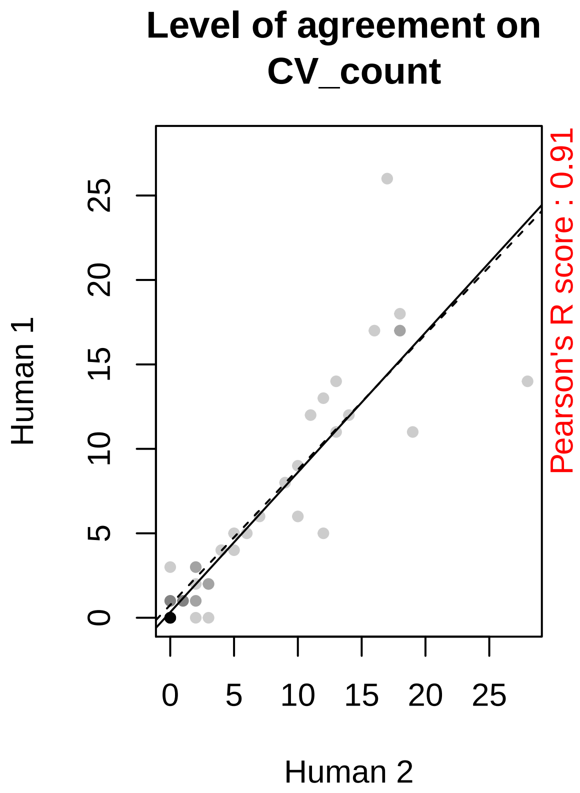
If both annotators fully agree, $\kappa = 1$, if they fully disagree (or agree randomly), $\kappa = 0$

```
## Warning: package 'psych' was built under R version 3.4.4
```

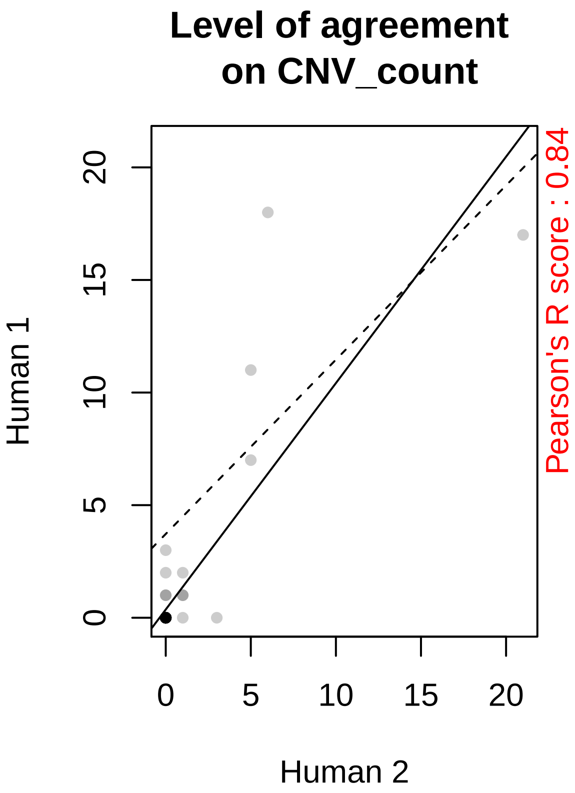
corpora	n_obs	kappa	weighted_kappa
WAR	60000	0.6883621	0.7634267
SOD	60000	0.6844869	0.6899259
ROW	60000	0.6661583	0.6879248
all	360000	0.6403963	0.6531616
ROS	60000	0.5707022	0.5808867
BER	60000	0.6540080	0.5688731
TSE	60000	0.5485063	0.5474113

Vocalizations and turn-taking

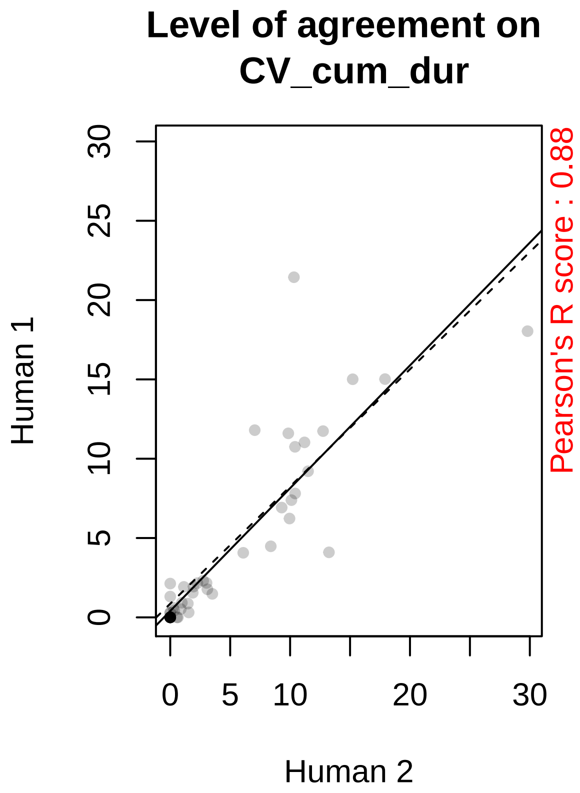
For this study, we considered only the children that have been annotated as vcm or lex, for which vocalizations were classified as C, N, W, L, U, or Y. That led us to remove 7 children from the study, for a total of 53 children.



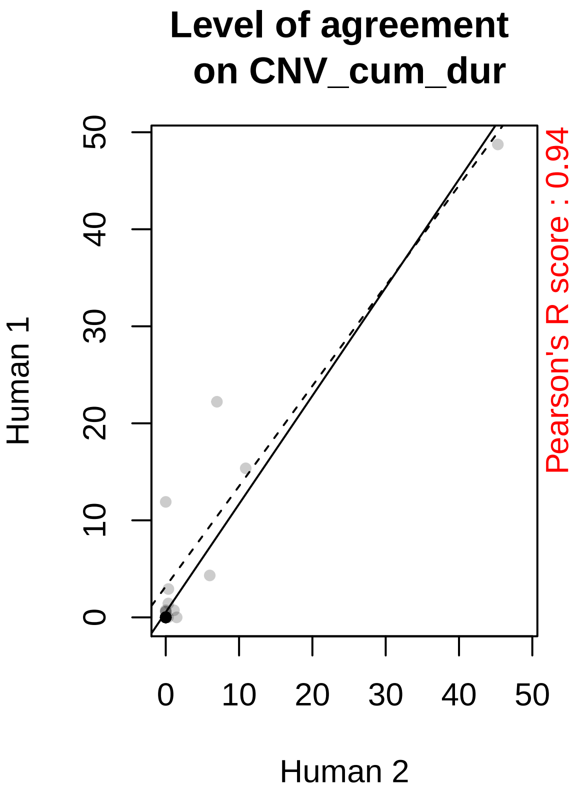
Level of agreement on CV_count



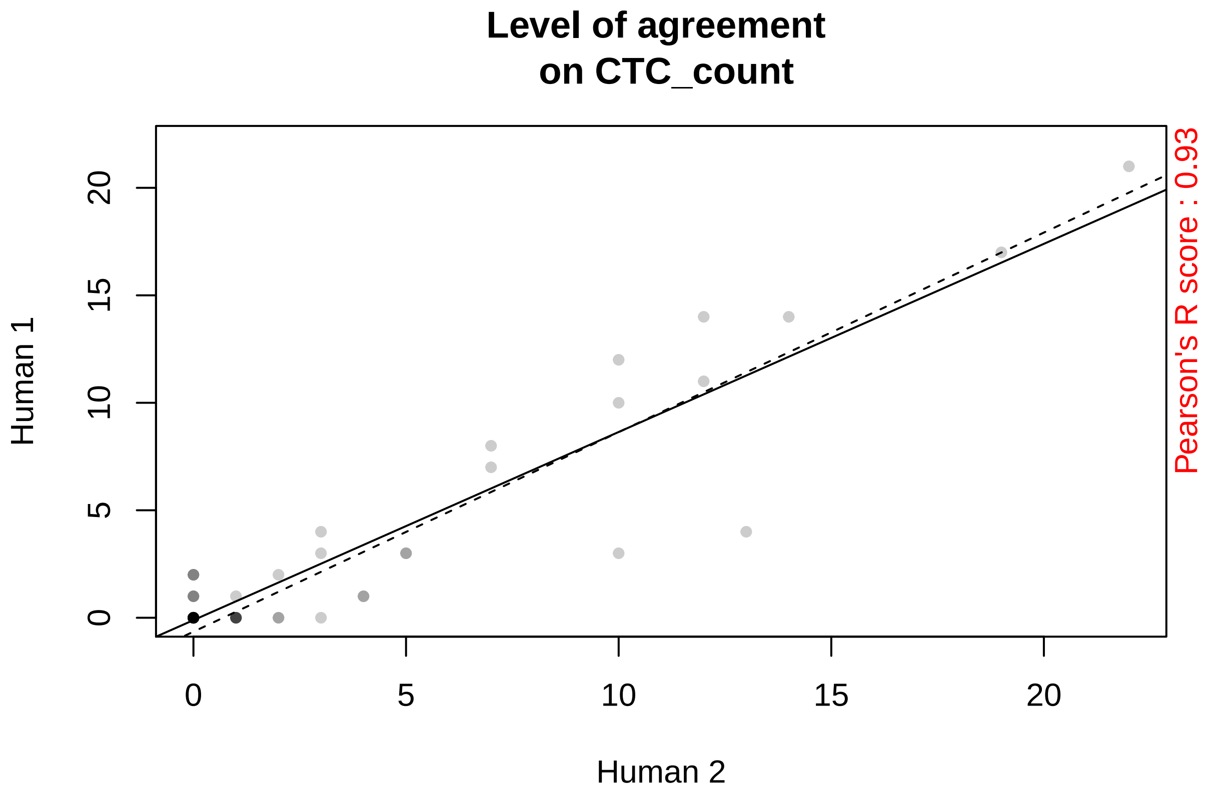
Level of agreement on CNV_count



Level of agreement on CV_cum_dur



Level of agreement on CNV_cum_dur


Level of agreement
on CTC_count
Human 1
Human 2
0
5
10
15
20
0
5
10
15
20
Pearson's R score : 0.93