A thorough evaluation of the Language Environment Analysis (LENATM) system

many[1]

[1]

Author Note

Correspondence concerning this article should be addressed to many, . E-mail:

Abstract

waiting

A thorough evaluation of the Language Environment Analysis (LENATM) system

Brief introduction to LENA(R) products.

Previous validation work.

Present work.

Methods

Corpora.

Processing.

LENA classification accuracy.

Speech and talker segmentation metrics.

Precision and recall.

CVC and CTC evaluation.

AWC evaluation.

Results

Before starting, we provide some general observations based on the human annotation. Silence is extremely common, constituting 79% of the frames. In fact, 45% of clips contained no speech by any of the human speaker types (according to the human annotators). As for speakers, female adults make up 11% of the frames, the child contributes to 4% of the frames, whereas male adult voices, other child voices, and electronic voices are found in only 1% of the frames each. Overlap makes up the remaining 3% of the frames. The following consequences ensue: if frame-based accuracy is sought, a system that classifies every frame as silence would be 79% correct. This is of course not what we want, but it indicates that systems adapted to this kind of speech should tend to have low "false alarm" rates, i.e. a preference for being very conservative as to when there is speech. If the system does say there is speech, then it had better say that this speech comes from female adults, who provide a great majority of the speech. In second place, it should be key child. Given that male adults

and other children are rare, a system that makes a lot of mistakes in these categories may still have a good global performance, because these categories are extremely rare.

LENA classification accuracy: False alarms, misses, confusion. Our first analysis is based on standard speech technology metrics, which put errors in the perspective of how much speech there is. That is, if 10 frames are wrong in a file where there are 100 frames with speech, this is a much smaller problem than if 10 frames are wrong in a file where there is 1 frame with speech. In other words, these metrics should be considered relative error metrics. One problem, however, emerges when there is no speech whatsoever in a given file. In the speech technology literature, this is never discussed, because most researchers working on this are basing their analyses on files that have been selected to contain speech (e.g., recorded in a meeting, or during a phone conversation). We still wanted to take into account clips with no speech inside because it is key for our research goals: We need systems that can deal well with long stretches of silence, because we want to measure how much speech children hear. Indeed, as mentioned above, 45% of our clips had no speech whatsoever. In these cases, the false alarm, miss, and confusion rates are all undefined, because the denominator is zero. In all likelihood, this leads to an overestimation of LENA's performance, because potential false alarms in these files are not counted against the system. It also occurred that there was just a little speech; in this case, the denominator is very small, and therefore the ratio for these two metrics ended up being a very large number. To avoid such outliers having an undue impact on our report, we present medians (rather than means).

There were, a priori, several ways of analyzing the data:

- collapsing near and far together (i.e., CHN and CHF were mapped onto a single CH category)
- treating the near and far categories separately (i.e., CHN and CHF are both treated as "speakers", but not the same one)
- not considering TV as a speaker category, since it is conceptually not identical to the

electronic voices detected by ACLEW human annotators; in this case, the gold annotations should also map electronic voices to non-speech or silence

- not considering OLN as a speaker category, since it is not conceptually identical to the overlap derived from humans' annotating different speaker categories.

We thought the most informative decision would be to report on several of these settings, albeit briefly. We start with the situation that yields the best LENA performance: Electronic voices in the gold annotation are mapped onto silence, so that the categories found in the human annotation are FEM, MAL, CHI, OCH, and overlap; in the LENA annotation, only CHN, FAN, MAN, and CXN are considered speakers (with all far categories, TVN, and OLN all mapped onto silence). In this setting, LENA's false alarm (i.e., saying that someone was speaking when they were not) had a median of 12%, whereas the miss rate had a median of 49%. The confusion rate, as mentioned above, is only calculated for the correctly detected speech (i.e., not the speech that was missed, which counts towards the miss rate, nor the speech that was falsely identified, which is considered in the false alarm). The confusion rate was very low, with a median of 10%. These three metrics can be added together into a single "diarization error rate". The median diarization error rate over the clips that had some speech was 79%.

If electronic voices in the gold annotation are still mapped onto silence and in the LENA annotation, CHN, FAN, MAN, CXN as well as OLN are considered speakers (with all far categories as well as TVN mapped onto silence), so that the human categories considered were CHI, FEM, MAL, OCH, and overlap; and the LENA categories considered were CHN, FAN, MAN, CXN, and OLN. In this setting, LENA's false alarm, missed, and confusion rate medians were 33%, 21%, and 28% respectively, for a total median diarization error of 82%. Performance likely degrades because OLN is not picking up the same regions as the overlapping speech found in the human annotations.

Next, we allowed the electronic voices segmented by humans, and TVN among the LENA speaker categories, to be considered during the evaluation (rather than mapping them

all to non-speech or silence), so that the human categories considered were CHI, FEM, MAL, OCH, overlap, and electronic; and the LENA categories considered were CHN, FAN, MAN, CXN, OLN, and TVN. LENA's false alarm, missed, and confusion rate medians were 40%, 18%, and 30% respectively, for a total median diarization error of 88%. Performance likely degrades because TVN is not picking up the electronic speech segmented by ACLEW annotators.

Finally, we declared the maximum possible number of categories: The human categories considered were still CHI, FEM, MAL, OCH, overlap, and electronic; but the LENA categories considered were CHN, FAN, MAN, CXN, OLN, TVN, CHF, FAF, MAF, CXF, OLF, TVF. LENA's false alarm, missed, and confusion rate medians were 74%, 7%, and 41% respectively, for a total median diarization error of 122%. Performance degrades because everything is treated as speech, leading to huge apparent false alarm rates.

LENA classification accuracy: Precision and recall. By now, we have established that the best performance (when "far" labels such as CHF and OLF are mapped onto silence, as are TVN and OLN), the overall relative diarization error rate is about 79%, due mainly to missing speech (49%), with false alarms (12%) and confusion between talker categories (10%) constituting a relatively small proportion of errors. However, this metric may not capture what our readers are interested in, for two reasons. First, this metric gives more importance to correctly classifying segments as speech versus non-speech (False alarms + misses) than confusing talkers (confusion). Second, many LENA adopters use the system not to make decisions on the sections labeled as non-speech, but rather on sections labeled as speech, and particularly those labeled adults and key child. The metrics above do not give more importance to these two categories, and do not give us insight on the patterns of error made by the system. Looking at precision of speech categories is crucial for users who interpret LENA's estimated quantity of adult speech or key child speech, as low precision means that some of what LENA called e.g. key child was not in fact the key child, and thus it is providing overestimates. Looking at recall may be most interesting for adopters who

intend to employ LENA as a first-pass annotation: the lower the recall, the more is missed by the system and thus cannot be retrieved (because the system labeled it as something else, which will not be inspected given the original filter). Recall also impacts quantity estimates, since it indicates how much was missed of that category.

Therefore, this subsection shows confusion matrices, containing information on precision and recall, for each key category. For this analysis, we collapsed over all human annotations that contained overlap between two speakers into a category called "overlap". Please remember that this category is not defined the same way as the LENA overlap category. For LENA, overlap between any two categories falls within overlap – i.e., CHN+TV would be counted towards overlap; whereas for us, only overlap between two talker categories (e.g., key child and female adult) counts as overlap.

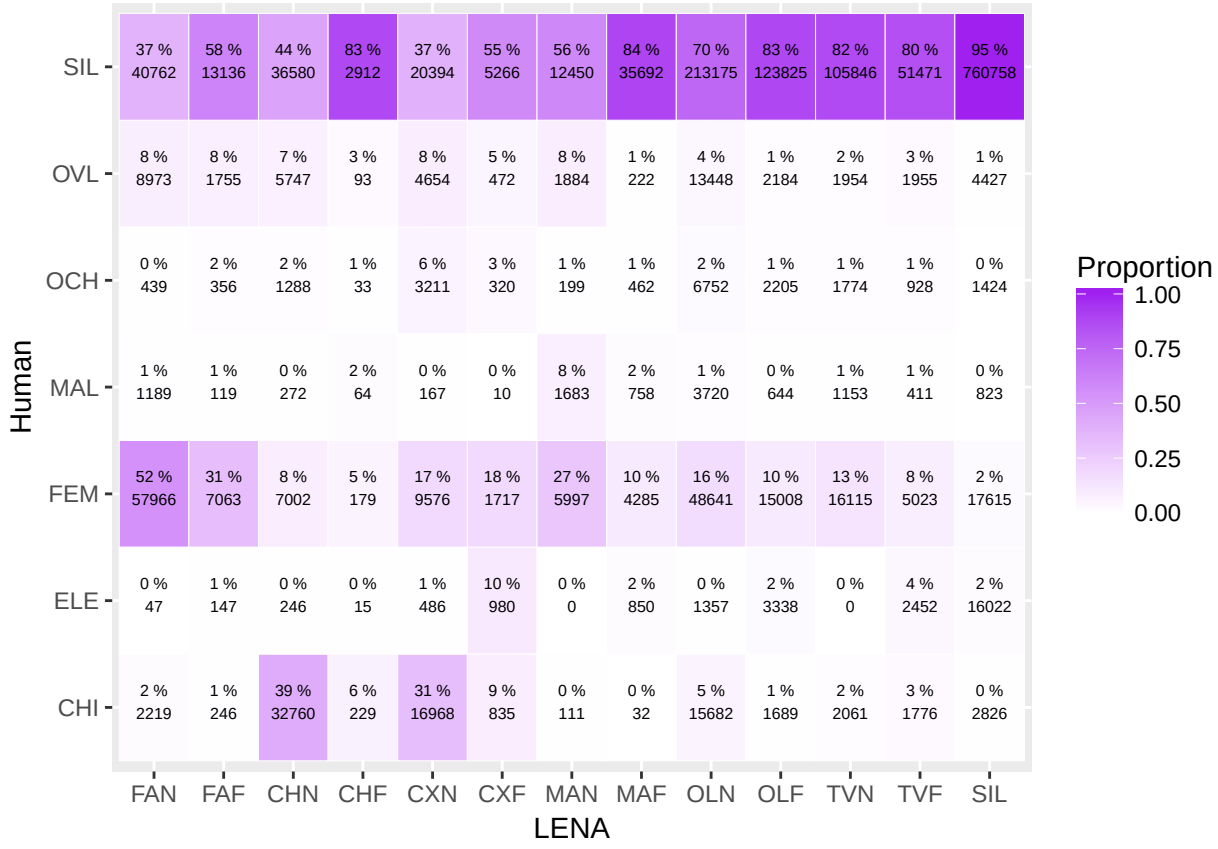

Figure 1

We start by explaining how to interpret one cell in Figure (precision): Focus on the

crossing of the human category FEM and the LENA category FAN; when LENA tags a given frame as FAN, this corresponds to a frame tagged as being a female adult by the human 52% of the time. This category, as mentioned above, is the most common speaker category in the audio, so that over 57k frames (representing 52% of the frames tagged as FAN by LENA) were tagged as being female adult by both the human and LENA. The remaining 2, 0, 1, 8, 0, and 37% of frames that LENA tagged as FAN were actually other categories according to our human coders: 37% were silence, 8% were in regions of overlap between speakers or between a speaker and an electronic voice, and 3% were due to confusions with other speaker tags. Inspection of the rest of the confusion matrix shows that, other than silence, this is the most precise LENA tag.

Precision for CHN comes in secondplace, at 39%; thus, fewer than half of the frames labeled as being the key child are, in fact, the key child. The majority of the frames, LENA incorrectly tagged as being the key child are actually silence (or rather, lack of speech) according to the human annotator (44%), with the remaining errors being due to confusion with other categories: About 8% of them are actually a female adult; 2% are another child; and 7% are regions of overlap across speakers, according to our human coders.

MAN and CXN score similarly, 8 and 6% respectively, meaning that less than a tenth of the areas LENA tagged as being these speakers actually correspond to them. As with the key child, most errors are due to LENA tagging silent frames as these categories. However, in this case confusion with other speaker tags is far from negligible. In fact, the most common speaker tag in the human annotation among the regions that LENA tagged as being MAN were actually female adult speech (27%); and, for CXN, it was not uncommon to find a CXN tag for a frame human listeners identified as a female adult (17%) or the key child (6%). In a nutshell, this suggests extreme caution before undertaking any analyses that rely on the precision of MAN and CXN, since most of what is being tagged as such is silence or other speakers.

Another observation is that the “far” tags of the speaker categories do tend to more

frequently correspond to what humans tagged as silence (74%) than the "near" tags (54%), and thus it is reasonable to exclude them from consideration. The relatively high proportion of near LENA tags that correspond to regions that humans labeled as silence could be partially due to the fact that the LENA system, in order to process a daylong recording quickly, does not make judgments on small frames independently, but rather imposes a minimum duration for all speaker categories, padding with silence in order to achieve it. Thus, any key child utterance that is shorter than .6 secs will contain as much silence as needed to achieve this minimum (and more for the other talker categories). Our system of annotation, whereby human annotators had no access whatsoever to the LENA tags, puts us in an ideal situation to assess the impact of this design decision, because any annotation that starts from the LENA segmentation should bias the human annotator to ignore such short interstitial silences to a greater extent than if they have no access to their tags whatsoever.

These analyses shed light on the extent to which we can trust the LENA tags to contain what the name indicates. We now move on to recall, which indicates a complementary perspective: how much of the original annotations were captured by LENA.

Again, we start with an example to facilitate the interpretation of this figure: The best performance for a talker category this time is CHN: Nearly half of the original frames humans tagged as being uttered by the key child were captured by the LENA under the CHN tag. Among the remaining regions that humans labeled as being the key child, 22% was captured by LENA's CXN category and 20% by its OLN tag, with the remainder spread out across several categories. This result can be taken to suggest that an analysis pipeline that uses the LENA system to capture the key child's vocalizations by extracting only CHN regions will get nearly half of the key child's speech. Where additional human vetting is occuring in the pipeline, such researchers may consider additionally pulling out segments labeled as CXN, since this category actually contains a further 22% of the key child's speech. Moreover, as we saw above, over a third of these LENA tags corresponds to the key child, which means that human coders who are re-coding these regions could filter out the two thirds that do not.

Figure 2

Many colleagues also use the LENA as a first pass to capture female adult speech via their FAN label. Only 30 of the female adult speech can be captured this way. Unlike the case of the key child, missed female speech is classified into many of the other categories, and thus there may not exist an easy solution (i.e., one would have to pull out all examples of many other categories to get at least half of the original female adult). However, if the hope is to capture as much of the female speech as possible, perhaps a solution may be to also pull out OLN regions, since these capture a further 25% of the original female adult speech and, out of the OLN tags, 16% are indeed female adults (meaning that human annotators re-coding these regions need to filter out 4 out of 5 clips, on average).

For the remaining two near speaker labels (MAN, CXN), recall averaged 15%, meaning that less than a quarter of male adult and other child speech is being captured by LENA. In fact, most of these speakers' contributions are being tagged by the LENA as OLN (mean

across MAN and CXN 34%) or silence (mean across MAN and CXN 7%), although the remaining sizable proportion of misses is actually distributed across many categories.

Finally, as with precision, the "far" categories show worse performance than the "near" ones. It is always the case that a higher percentage of frames is "captured" by the near rather than the far labels. For instance, out of all frames attributed to the key child by the human annotator, 42% were picked up by the LENA CHN label versus 0% by the LENA CHF label. This result can be used to argue why, when sampling LENA daylong files using the LENA software, users need not take into account the "F" categories.

Child Vocalization Counts (CVC) accuracy. Given the inaccuracy of far LENA tags, and in order to follow the LENA system procedure, we only counted vocalizations attributed to CHN and ignored those attributed to CHF. As shown in Figure (CVC), there is a strong association between clip-level counts estimated via the LENA system and those found in the human annotations: the Pearson correlation between the two was $r = 0.71$ ($p = 0.00$) when all clips were taken into account, and $r = 0.77$ ($p = 0.00$) when only clips with some child speech (i.e., excluding clips with 0 counts in both LENA and human annotations) were considered. This suggests that the LENA system captures differences in terms of number of child vocalizations across clips well.

However, users need more: They also interpret the absolute number of vocalizations found by LENA. Therefore, it is important to also bear in mind the absolute error rate and the relative error rate. The absolute error rate tells us, given a LENA estimate, how close the actual number may be. The relative error rate puts this number in relation to the actual number of vocalizations tagged by the human coder. For instance, imagine that we find that LENA errs by 10 vocalizations according to the absolute error rate; this means that, on average across short clips like the ones used here, the numbers by LENA would be off by 10 vocalizations. We may think this number is small; by using the relative error rate, we can check whether it is small relative to the actual number: An error of 10 vocalizations would seem less problematic if there are 100 vocalizations on average (LENA would be just 10%

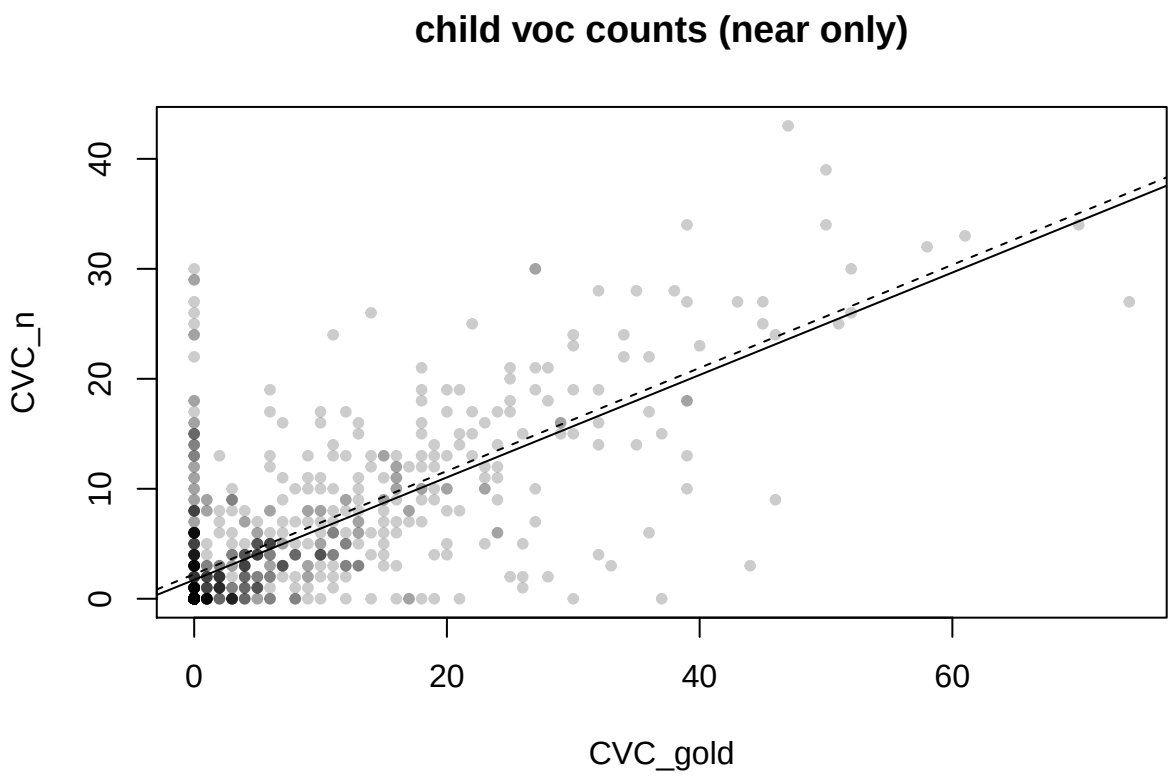

Figure 3

off) than if there are 10 (LENA would be doubling the number of vocalizations).

The absolute error rate ranged from -47 to 30, with a mean of -1.63 and a median of 0. As for relative error rates, these require the number in the denominator to be non-null. For this analysis, therefore, we need to remove the 460 clips in which the human annotator said there were no child vocalizations whatsoever. When we do this, the mean relative error rate ranged from -100 to 800, with a mean of -20.17 and a median of -44.44

Conversational Turn Counts (CTC) accuracy. Again, we only considered "near" speaker categories in the turn count, and applied the same rule the LENA does, where a turn can be from the key child to an adult or vice versa, and should happen within 5 seconds to be counted. The association between clip-level LENA and human CTC was weaker than that found for CVC: the Pearson correlation between the two was r = 0.55 (p = 0.00) when all clips were taken into account, and r = 0.47 (p = 0.00) when only clips with some child speech (i.e., excluding 322 clips with 0 counts in both LENA and human

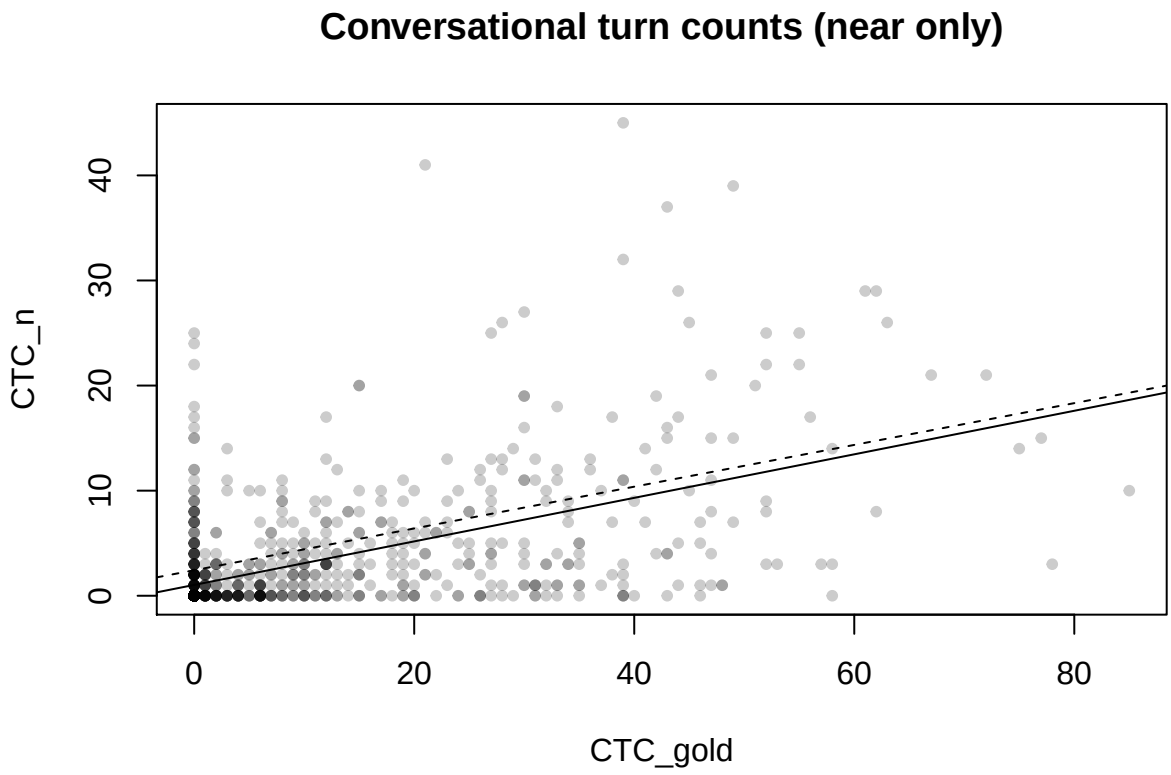

Figure 4

annotations) were considered. The absolute error rate ranged from -75 to 25, with a mean of -6.82 and a median of 0. As for relative error rates, these require the number in the denominator to be non-null. For this analysis, therefore, we need to remove the 427 clips in which the human annotator said there were no child-adult or adult-child turns whatsoever. When we do this, the mean relative error rate ranged from -100 to 366.67, with a mean of -64.64 and a median of -81.82

Adult Word Counts accuracy. One child in the SOD corpus was learning French. We have included this child to increase power, but results without this one child are nearly identical. The association between clip-level LENA and human AWC was strong: the Pearson correlation between the two was r=0.75 (p=0.00) when all clips were taken into account, and r=0.69 (p=0.00) when only clips with some child speech (i.e., excluding 309 clips with 0 counts in both LENA and human annotations) were considered. This suggests that the LENA system captures differences in terms of number of child vocalizations across

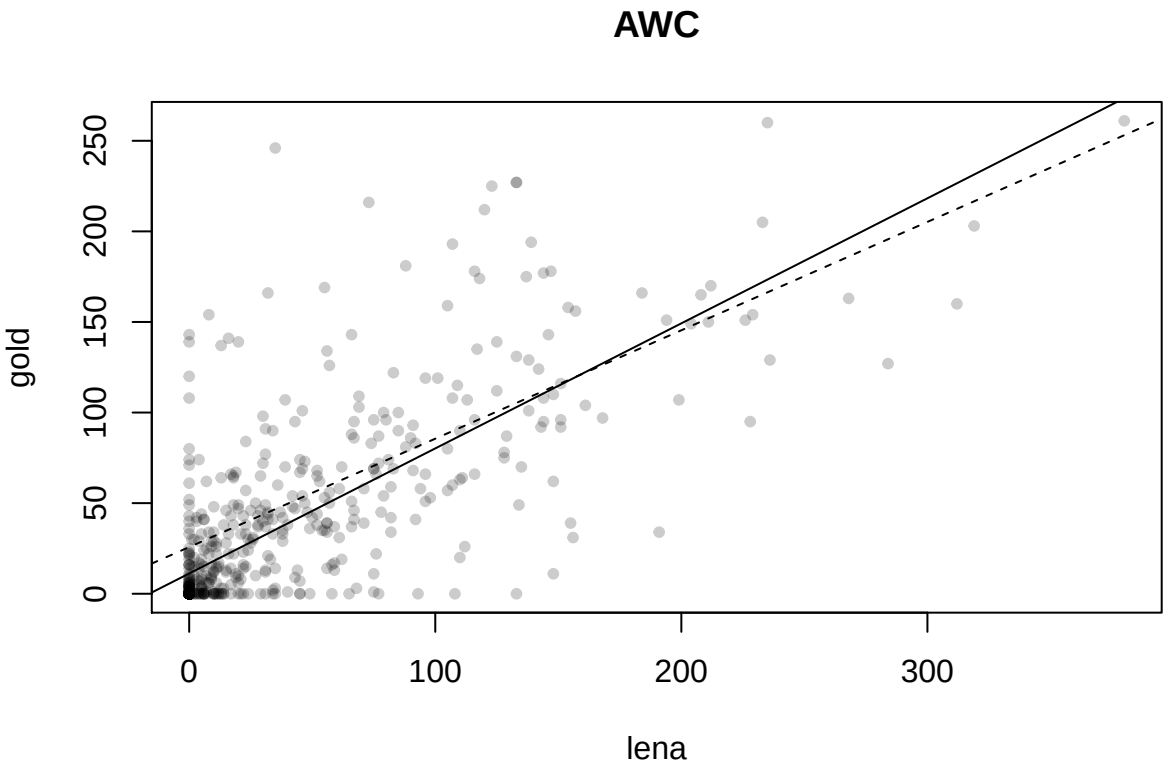

Figure 5

clips well. The absolute error rate ranged from -211 to 157, with a mean of -0.20 and a median of 0. As for relative error rates, these require the number in the denominator to be non-null. For this analysis, therefore, we need to remove the 361 clips in which the human annotator said there were no child-adult or adult-child turns whatsoever. When we do this, the mean relative error rate ranged from -100 to 7400, with a mean of 55.04 and a median of -17.78

Effects of age and differences across corpora. The preceding sections include results that are wholesale, over all corpora. However, we have reason to believe that performance could be higher for the corpora collected in North America (BER, WAR, SOD) than those collected in other English-speaking countries (L05) or non-English speaking populations (TSI). Additionally, our age ranges are wide, and in the case of TSI children, some of the children are older than the oldest children in the LENA training set. To assess whether accuracy varies as a function of corpora and child age, we fit mixed models as

follows.

We predicted false alarm, miss, and confusion rates (when all "F" categories, TV, and overlap were mapped onto silence, which yielded the best results in Section XX) from corpus, child age, and the interaction as fixed effects, child ID as random effect, on clips where there was some speech according to the human annotator. We followed up with an Analysis of Variance (type 2) to assess significance. In none of these analyses was corpus, child age, or their interaction significant.

For CVC, we fit a mixed model where CVC according to the human was predicted from CVC according to LENA, in interaction with corpus and age, as fixed factors; with child ID as random effect. An Analysis of Variance (type 2) found a triple interaction, suggesting that the predicted value of LENA with respect to human CVC depended on both the corpus and the child age; and a two-way interaction between CVC by LENA and corpus. To investigate these further, we fit a model where CVC according to the human was predicted from CVC according to LENA in interaction with age (as fixed factors, with child ID as random) within each corpus separately. This revealed a significant interaction between LENA CVC and age for BER (indicating that the predictive value of LENA CVC increased with child age), whereas for the other four corpora this interaction was not significant, nor was the main effect of age, and only the LENA CVC emerged as a significant predictor of variance in child vocalization counts derived from human annotation.1

For CTC, we fit a mixed model where CTC according to the human was predicted from CTC according to LENA, in interaction with corpus and age, as fixed factors; with child ID as random effect. An Analysis of Variance (type 2) found a two-way interaction between CTC by LENA and corpus. To investigate this further, we fit the same regressions within each corpus separately.2 These follow-up analyses revealed that CTC by LENA varied in its strength of prediction of human-tagged CTC across corpora (BER estimate = 1.16, SE of estimate = 0.16, t = 7.41; L05 (estimate = 1.62, SE of estimate = 0.21, t = 11.74) TSI estimate = 0.94, SE of estimate = 0.08, t = 11.74; SOD estimate = 0.96, SE of estimate =

0.22, t = 4.46; WAR estimate = 1.68, SE of estimate = 0.14, t = 12.22).

Discussion

Acknowledgments

References