

## **Atypicality score v0 (speaker×dataset)**

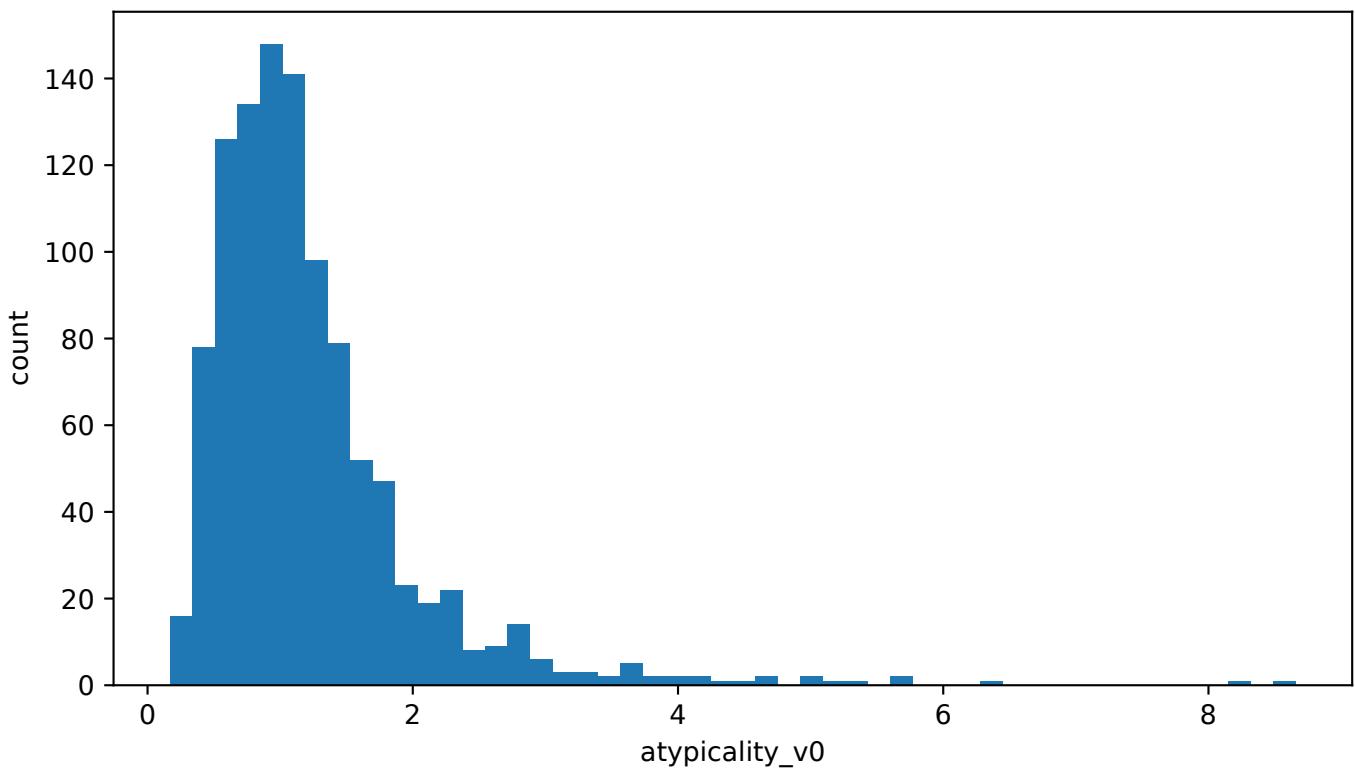
- rows: 1,072
- datasets: cejc, csj
- score: RMS(z) within dataset
- scaler: robust\_z (median/MAD) if generated as default

Output columns:

dataset, speaker\_id, role, n\_rows, atypicality\_v0, top\_contrib\_json, is\_outlier\_p99

Note: v0 is a deviation score for pragmatic feature distribution, not a diagnosis.

## Distribution: cejc



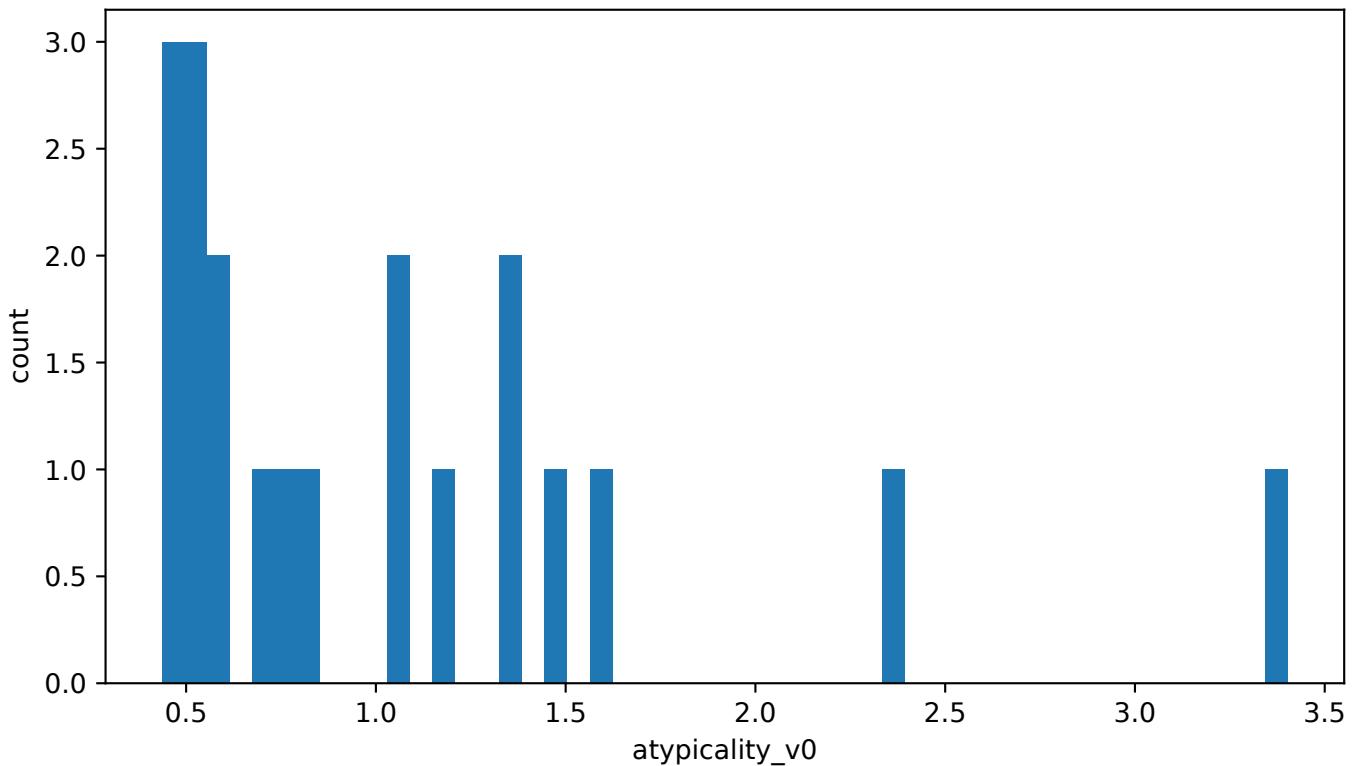
```
{  
    "n": 1052,  
    "mean": 1.22889,  
    "std": 0.814379,  
    "p50": 1.04811,  
    "p90": 2.06733,  
    "p95": 2.70066,  
    "p99": 4.54623,  
    "max": 8.65921  
}
```

Next: see the next page for Top-N speakers table.

## Top 30 speakers: cejc

speaker_id	atypicality_v0	is_outlier_p99
T007_007:IC06	8.65921	True
K010_003a:IC01	8.26943	True
W001_001a:IC01	6.37322	True
T014_012:IC04	5.68601	True
K010_005b:IC01	5.6591	True
K010_005a:IC01	5.34351	True
T004_016:IC06	5.13067	True
K004_022:IC02	4.99705	True
T018_022:IC02	4.93058	True
K008_010:IC02	4.64753	True
T023_017:IC04	4.62465	True
C002_013b:IC02	4.47089	False
T013_020:IC02	4.26397	False
W010_001b:IC02	4.11999	False
T014_012:IC03	4.07883	False
T001_009:IC04	4.00611	False
T007_006:IC04	3.97857	False
T021_011b:IC01	3.86268	False
K009_009a:IC01	3.7983	False
W010_001a:IC02	3.7075	False
T003_006:IC05	3.67235	False
T003_007:IC01	3.63265	False
T004_011:IC02	3.62823	False
T023_009:IC03	3.62691	False
T013_013:IC01	3.54356	False
T021_011a:IC01	3.43897	False
T018_021:IC01	3.31404	False
K010_011a:IC03	3.31032	False
K007_008:IC02	3.24203	False
T006_001:IC02	3.21916	False

## Distribution: csj



```
{  
    "n": 20,  
    "mean": 1.07145,  
    "std": 0.724135,  
    "p50": 0.795342,  
    "p90": 1.66979,  
    "p95": 2.42852,  
    "p99": 3.20808,  
    "max": 3.40297  
}
```

Next: see the next page for Top-N speakers table.

## Top 30 speakers: csj

speaker_id	atypicality_v0	is_outlier_p99
D01M0047:R	3.40297	True
D01M0019:L	2.37723	False
D04F0050:R	1.59119	False
D04M0052:R	1.48723	False
D04F0022:L	1.37915	False
D01F0023:R	1.36233	False
D02F0015:L	1.20663	False
D04F0022:R	1.07346	False
D01F0023:L	1.05696	False
D01F0030:R	0.827911	False
D02F0015:R	0.762773	False
D04F0044:R	0.717695	False
D01M0047:L	0.577275	False
D04F0044:L	0.574594	False
D02M0028:L	0.552641	False
D01F0030:L	0.550735	False
D04F0050:L	0.534966	False
D01M0009:R	0.4851	False
D04M0010:R	0.47214	False
D04M0052:L	0.436071	False

## **How to use v0 (next)**

- 1) Pick outliers (p99 or top-N) per dataset.
- 2) For each outlier, inspect `top_contrib_json` (features with largest  $|z|$ ).
- 3) Then go to `analysis/v1/gold=v13/examples` and sample representative turns.
- 4) (Phase3-2) LLM labeling: add functional labels (repair/question/backchannel/topic-shift etc.)