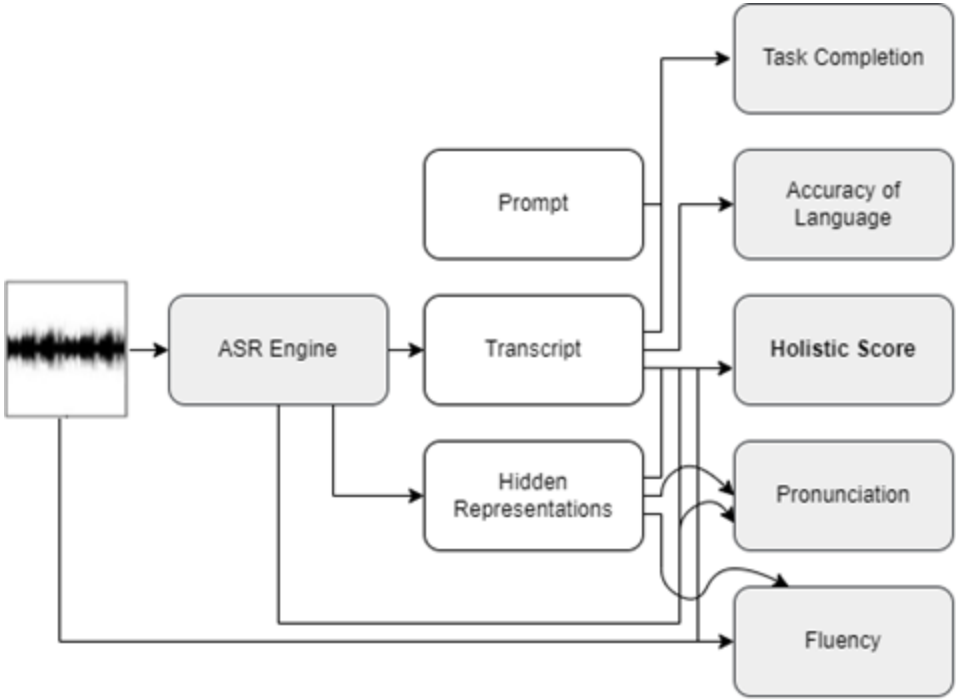
Deep Learning Methods in L2 low-resource speech recognition and assessment

Yaroslav Getman

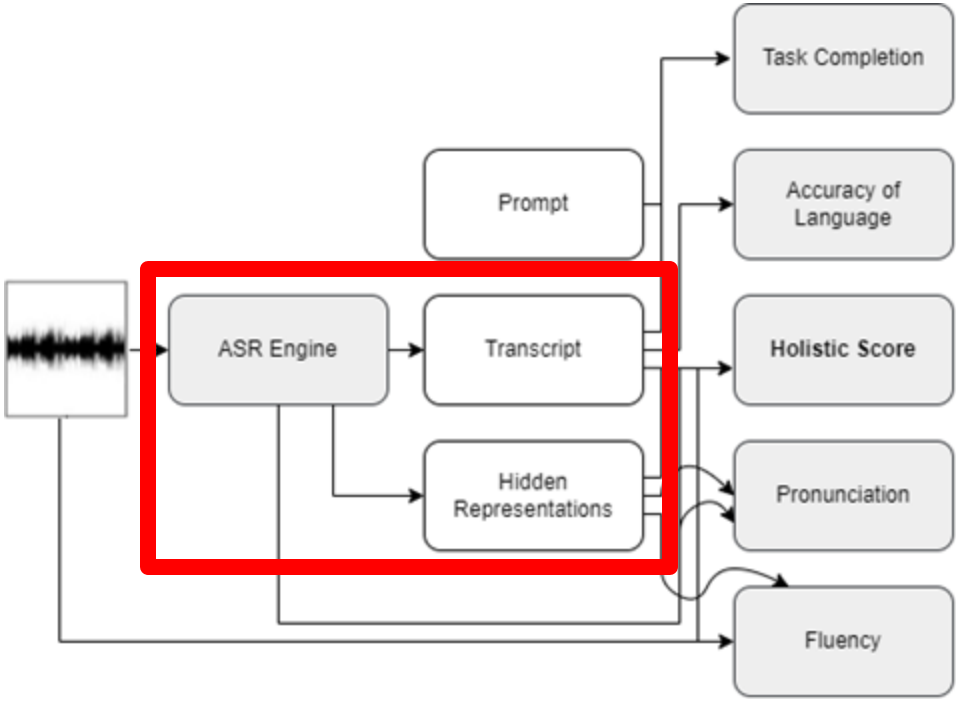
ASA

- Individual grade for each dimension + overall (holistic) score
- DNN classifiers



ASA

- Individual grade for each dimension + overall (holistic) score
- DNN classifiers

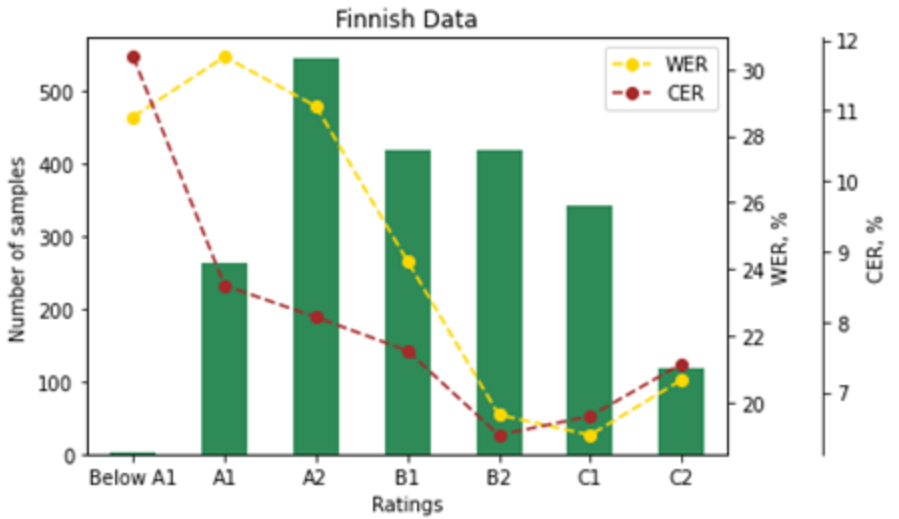


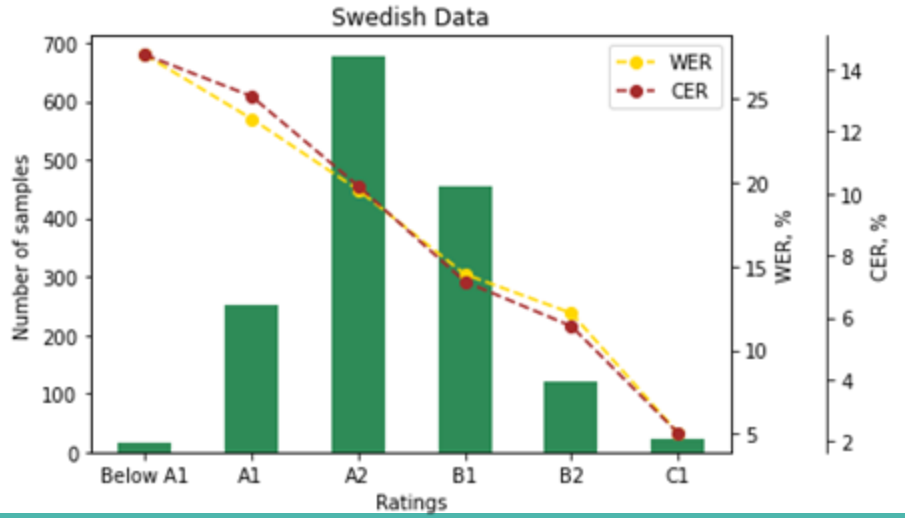
ASR

- Architecture: **wav2vec 2.0**
- Pre-trained on large amount of unlabeled speech
  - Learns deep acoustic representations
- Can perform ASR after fine-tuning on labeled data
- Suits for low-resource ASR
- Easy to adapt to the target domain
  - Fine-tune first on L1 data, then on L2 speech
- No language model

ASR

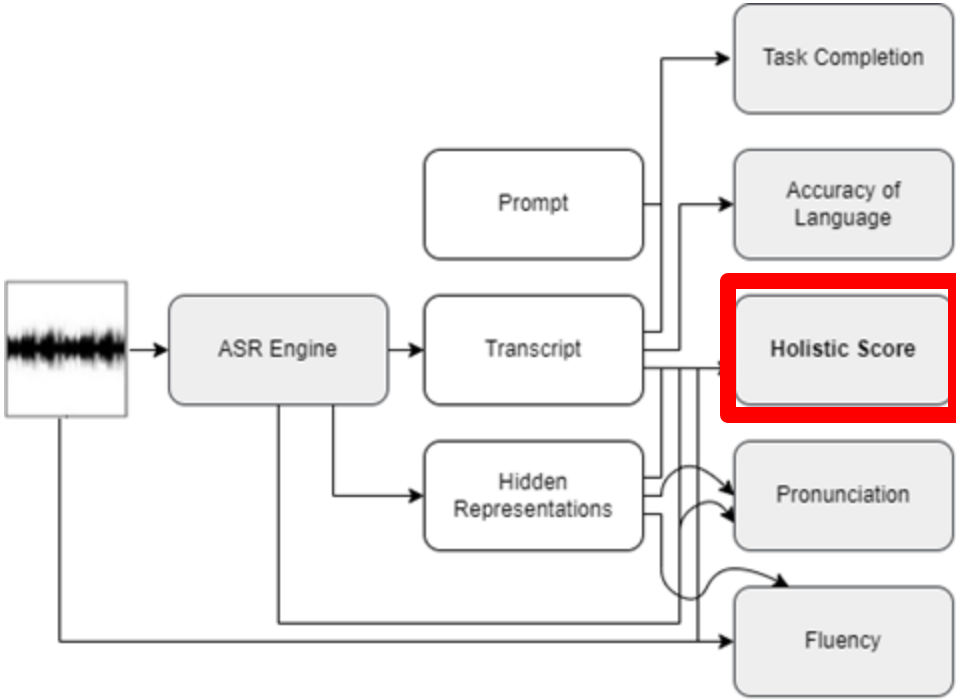
- **Finnish**: 308 speakers / 2112 samples / **14** hours
- **Swedish**: 178 speakers / 1542 samples / **5.5** hours
- 4-fold cross-validation
- Finnish: **21.89%/7.06%** WER/CER
- Swedish: **17.71%/9.08%** WER/CER



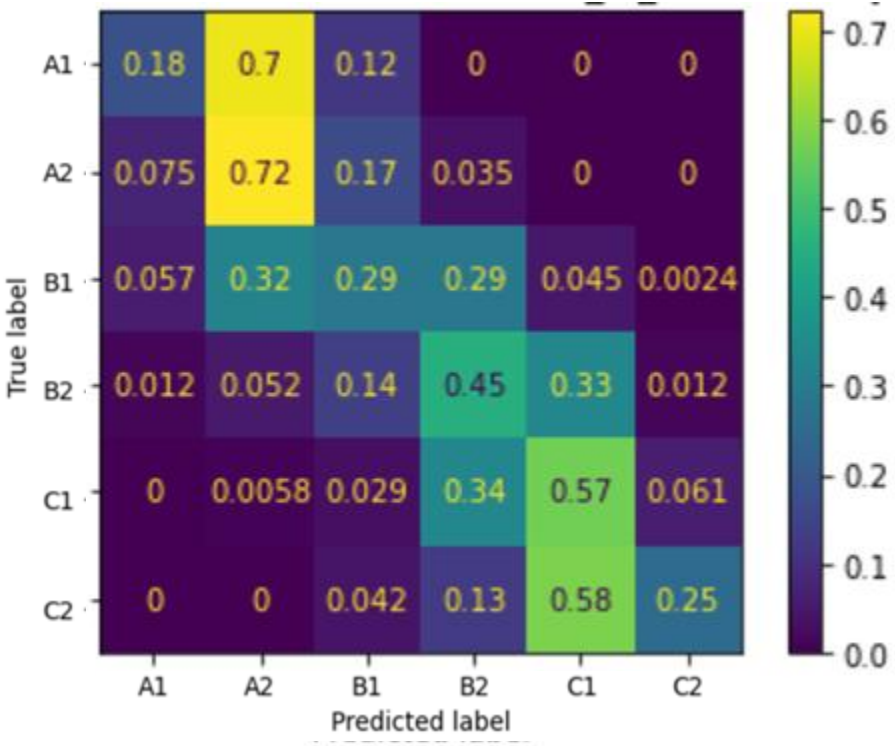


Holistic Score

- 4 holistic levels for Swedish and 6 levels for Finnish
- Manually extracted features + deep acoustic representations from the middle layer of wav2vec 2.0
- Classifier: 6-layer DNN

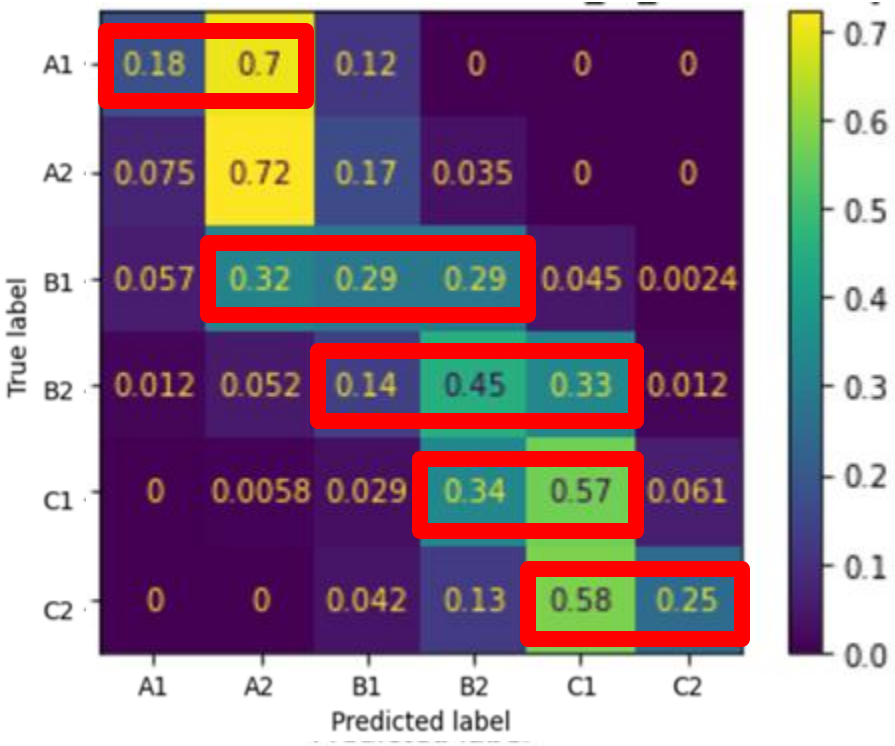


Finnish **Holistic**



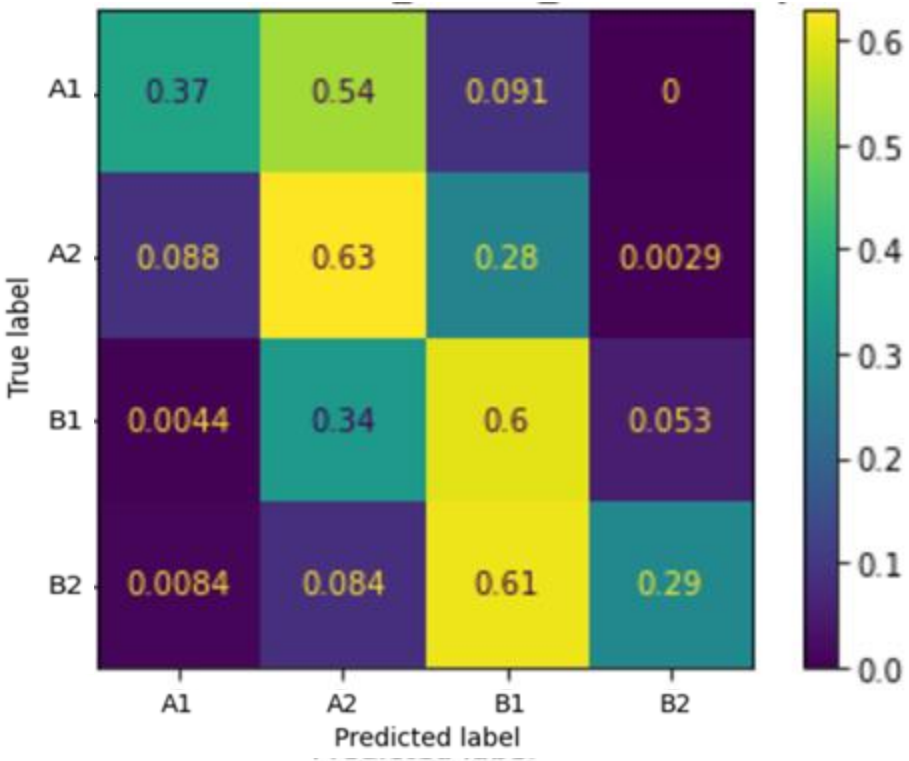
- **46%** accuracy / **39%** F1

Finnish **Holistic**



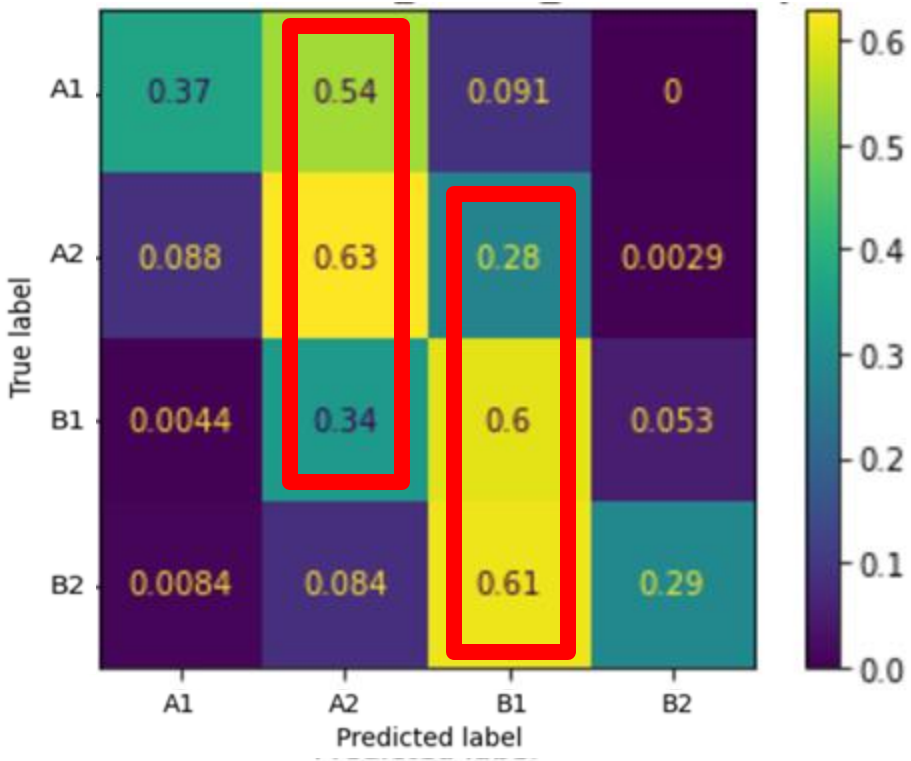
- **46%** accuracy / **39%** F1
- Confusion between the neighboring classes

Swedish **Holistic**



- **55%** accuracy / **50%** F1

Swedish **Holistic**



- **55%** accuracy / **50%** F1
- Tends to rate as intermediate levels (A2 and B1)
- More beginner level samples are being collected

Full Results. **Finnish**

Criterion and	human-to-human			machine-to-human					
Range of Classes	**Kappa**	**Correlation**	**MAE**	**Kappa**	**Correlation**	**MAE**	**P, %**	**R, %**	**F1, %**
Holistic (**2-7**)	0.732	0.751	0.782	0.807	0.803	0.612	46.85	39.95	39.05
Fluency (**2-4**)	0.393	0.392	0.575	0.507	0.522	0.359	63.95	55.67	57.62
Pronunciation (**2-4**)	0.513	0.531	0.445	0.583	0.612	0.269	66.51	54.62	55.18
Lex.-Gram. (**1-3**)	0.576	0.580	0.404	0.529	0.546	0.265	47.56	49.14	48.18
Task Achiev. (**1-3**)	0.340	0.298	0.410	0.365	0.390	0.318	54.52	44.09	46.41

Full Results. **Finnish**

Criterion and	human-to-human			machine-to-human					
Range of Classes	Kappa	Correlation	MAE	Kappa	Correlation	MAE	P, %	R, %	F1, %
Holistic (**2-7**)	0.732	0.751	0.782	0.807	0.803	0.612	46.85	39.95	39.05
Fluency (**2-4**)	0.393	0.392	0.575	0.507	0.522	0.359	63.95	55.67	57.62
Pronunciation (**2-4**)	0.513	0.531	0.445	0.583	0.612	0.269	66.51	54.62	55.18
Lex.-Gram. (**1-3**)	0.576	0.580	0.404	0.529	0.546	0.265	47.56	49.14	48.18
Task Achiev. (**1-3**)	0.340	0.298	0.410	0.365	0.390	0.318	54.52	44.09	46.41

Full Results. **Swedish**

Criterion and	human-to-human			machine-to-human					
Range of Classes	Kappa	Correlation	MAE	Kappa	Correlation	MAE	P, %	R, %	F1, %
Holistic (**2-5**)	0.496	0.490	0.613	0.524	0.524	0.461	56.13	47.33	49.76
Fluency (**1-3**)	0.498	0.490	0.425	0.560	0.574	0.305	63.06	59.41	60.53
Pronunciation (**2-3**)	0.162	0.162	0.419	0.276	0.290	0.343	66.97	67.53	66.85
Lex.-Gram. (**1-3**)	0.427	0.435	0.516	0.246	0.259	0.460	47.10	42.84	43.33
Task Achiev. (**1-3**)	0.376	0.371	0.621	0.650	0.714	0.320	60.97	59.51	59.92

Full Results. **Swedish**

Criterion and	human-to-human			machine-to-human					
Range of Classes	Kappa	Correlation	MAE	Kappa	Correlation	MAE	P, %	R, %	F1, %
Holistic (**2-5**)	0.496	0.490	0.613	0.524	0.524	0.461	56.13	47.33	49.76
Fluency (**1-3**)	0.498	0.490	0.425	0.560	0.574	0.305	63.06	59.41	60.53
Pronunciation (**2-3**)	0.162	0.162	0.419	0.276	0.290	0.343	66.97	67.53	66.85
Lex.-Gram. (**1-3**)	0.427	0.435	0.516	0.246	0.259	0.460	47.10	42.84	43.33
Task Achiev. (**1-3**)	0.376	0.371	0.621	0.650	0.714	0.320	60.97	59.51	59.92

Conclusions and Future Work

- L2 ASR + speech rating system for L2 Finnish and Finland Swedish
- 4 systems assessing different aspects of L2 speakers' proficiency
- Some systems have low agreement with human raters
  - The degree of human-to-human agreement is often even lower
- Model compression
- Updating L2 Swedish models

References

- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.
- Al-Ghezi, R., Getman, Y., Rouhe, A., Hildén, R., & Kurimo, M. (2021). Self-supervised end-to-end ASR for low resource L2 Swedish. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. ISCA.
- Al-Ghezi, R., Getman, Y., Voskoboinik, E., Singh, M., & Kurimo, M. (2023, January). Automatic Rating of Spontaneous Speech for Low-Resource Languages. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 339-345). IEEE.

ASR (extra)

Reference	ASR Output	WER, %	CER, %
L2 Swedish ASR			
när den här fri	när den här ledig le ledigheten fri	75.00	141.67
okej först gå gå till norragatan	okej först gå öö vad det mm gå till norragatan	66.67	37.04
aa kan kan jag har mot list	kan kan jag ha de mot list	42.86	19.05
ööm ha en bra födelsedag	öömha en bra födelsedag ningen	60.00	30.00
L2 Finnish ASR			
moi maria hauska tutustua minun nimeni on minulle kuuluu hyvää öö vähä väsyttää viclä mutta entäs itse	moi marja hauska tutustuaa minun nimeni on belttama koko minulle kuuluu hyvää öö vähä väsyyttää tiällä mutta entäs itse	35.29	20.93
se kuuluu tähän vuodenaikaan oletko levännyt tarpeeksi ja yrittänyt ottaa lääkkeitä	ja kuuluu tähän vuoden aikaan oletko hävennyt tarpeeksi ja yrittänyt ottaa lääkkeitä	36.36	6.85
öö tänään meillä on o olet ollut koulussa meillä oli ensin psykologiaa ja me aloitettiin yhden projektin öö sitten meillä oli ruokailu ja nyt meil on tämä suomen juttu	öö tänään millä on o olet ollut koulussa me olin sin sykologiaa ja me aloitettiin projekkin öa sitte meil oli ruokailla ja nyt meil on me suomen jutt	44.83	16.55