A Development Outlook for CLARIN's Northernmost Center

Steven Coats
University of Oulu, Finland
steven.coats@oulu.fi
CLARIN Conference, Barcelona
October 17th, 2024

Outline

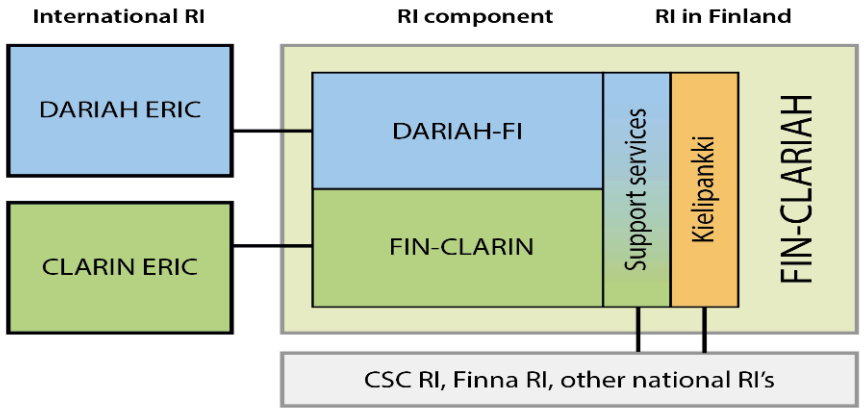
1. Background
2. Related CLARIN Centers: Fin-CLARIAH, CKCMC
3. Oulu CLARIN Center
    - Legal contexts
    - CoANZSE Audio
4. Plans for development
5. Summary

Slides for the presentation are on my homepage at https://cc.oulu.fi/~scoats

Background

- CoANZSE Audio website created in autumn 2023
- To implement Shibboleth login, certification as a CLARIN center was necessary
- Why not expand the center to incorporate new resources?
- Focus on multimedia corpora of computer-mediated communication (from, e.g., YouTube, Twitch, TikTok, 𝕏, etc.)

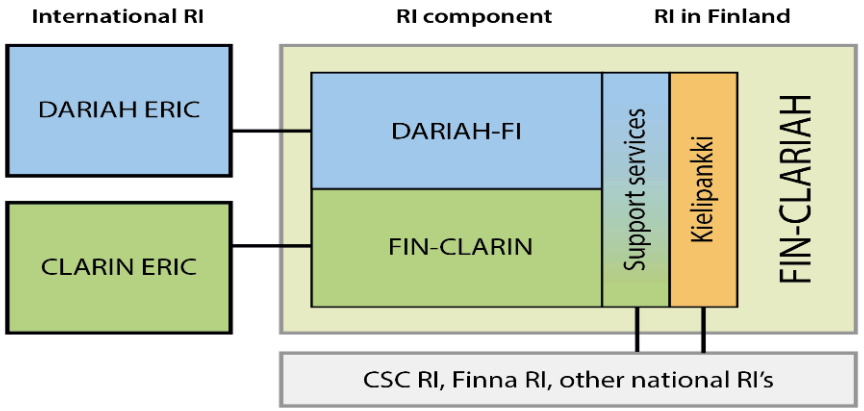
Related CLARIN Centers: Fin-CLARIAH



Kielipankki (Language Bank of Finland)

- Wide range of language resources, tools, and services
- Most resources hosted by Finland's Centre for Scientific Computing, many accessible through Korp
- Focus on Finland's national and official languages of Finnish, Swedish, Sámi languages, Finno-Ugric languages, as well as English and other languages
- As of 2024, social-media-platform-sourced multimedia content is not a primary focus

Related CLARIN Centers: Fin-CLARIAH

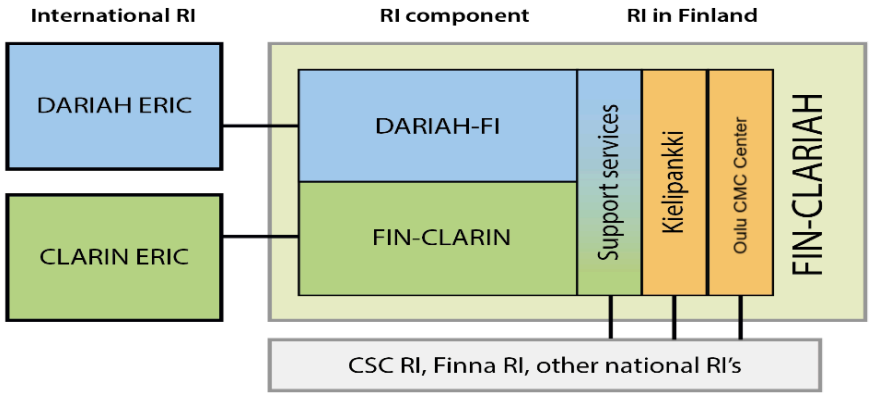


FIN-CLARIN

KIELIPANKKI
The Language Bank of Finland

Kielipankki (Language Bank of Finland)

- Wide range of language resources, tools, and services
- Most resources hosted by Finland's Centre for Scientific Computing, many accessible through Korp
- Focus on Finland's national and official languages of Finnish, Swedish, Sámi languages, Finno-Ugric languages, as well as English and other languages
- As of 2024, social-media-platform-sourced multimedia content is not a primary focus



Related CLARIN Centers: CKCMC

https://cmc-corpora.org/ckcmc

- Site to disseminate information about CMC (computer-mediated communication) resources, technologies, and community activities
- Proceedings of CMC-Corpora conference series
- Training opportunities
- Expertise in TEI-XML for CMC, among other topics
  - https://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

Doesn't host corpora *per se*

Oulu CLARIN Center: Legal contexts

- U.S. Code Title 17, § 107: Fair use
- Australian Copyright Act of 1968, New Zealand Copyright Act of 1994: Fair dealing
- EU 2019/790: Directive on Copyright in the Digital Single Market authorizes text and data mining of material under copyright for purposes of research and education (European Council 2019)
- Codified in Finnish law in 2023
    - Section 13b of the *Tekijänoikeuslaki* updated
    - "Research organizations and cultural heritage institutions that have legal access to the work may make copies of it for text and data mining in scientific research and keep them for scientific research, including for later verification of research results, provided that copies of the work are only available to those entitled to do so." (3.3.2023/263)

Oulu CLARIN Center: CoANZSE Audio (Coats 2022, 2024a, 2024b)

- *Corpus of Australian and New Zealand Spoken English*
- 57k ASR transcripts from YouTube channels of regional and local councils in Australia and New Zealand
- Can be used for research in grammar (Morin & Coats 2023), lexis, phonetics, pragmatics (Thaler & Elsweiler 2023), discourse, etc.
- Many recordings are meetings: advantages in terms of representativeness and comparability
- Speaker place of residence (cf. videos collected based on place-name search alone)
- Topical contents and communicative contexts comparable

Example video

WebVTT file

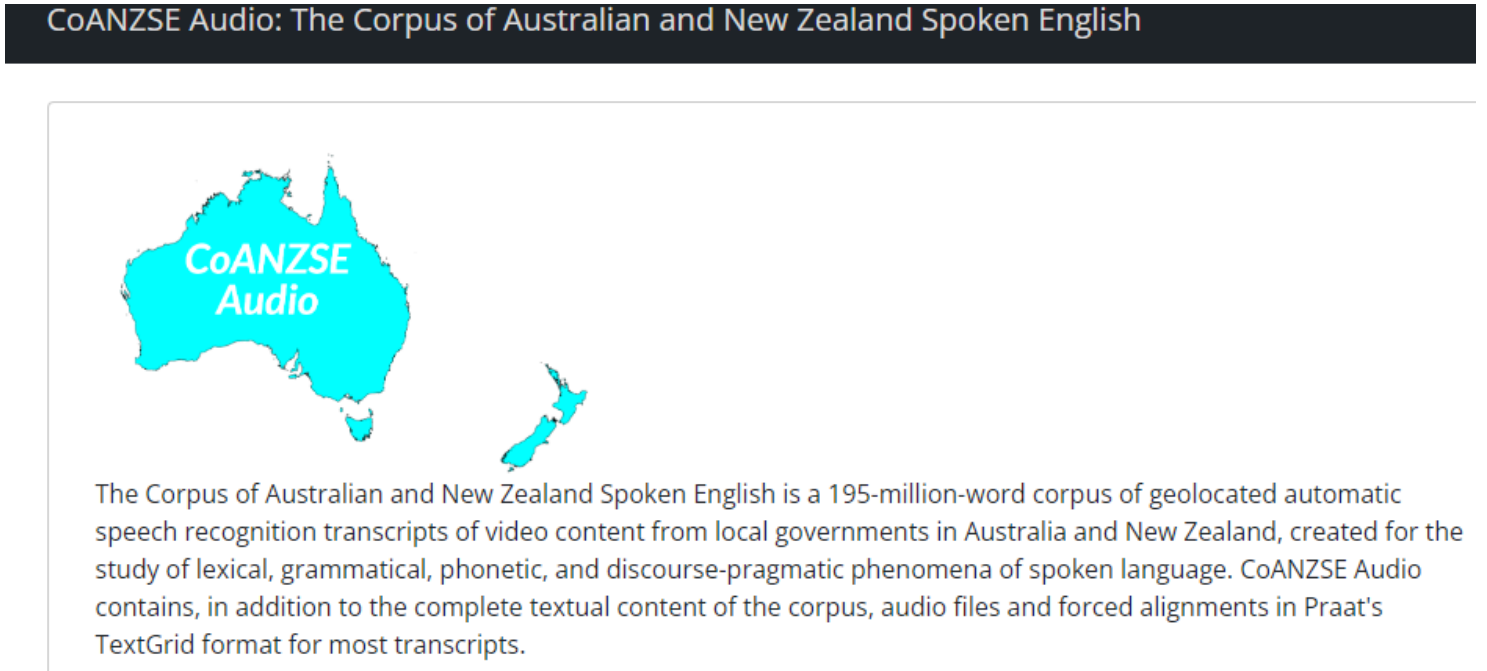
```
WEBVTT
Kind: captions
Language: en

00:00:01.160 --> 00:00:06.550 align:start position:0%

[Music]

00:00:06.550 --> 00:00:06.560 align:start position:0%
[Music]


00:00:06.560 --> 00:00:08.150 align:start position:0%
[Music]
uh<00:00:06.960><c> welcome</c>

00:00:08.150 --> 00:00:08.160 align:start position:0%
uh welcome


00:00:08.160 --> 00:00:10.950 align:start position:0%
uh welcome
i'd<00:00:08.320><c> like</c><00:00:08.480><c> to</c><00:00:08.639><c> open</c><00:00:08.880><c> the</c><00:00:09.040><c> meeting</c><00:00:09.360><c> at</c><00:00:09.519><

00:00:10.950 --> 00:00:10.960 align:start position:0%
i'd like to open the meeting at 9 12 a.m


00:00:10.960 --> 00:00:13.190 align:start position:0%
i'd like to open the meeting at 9 12 a.m
thank<00:00:11.200><c> you</c><00:00:11.280><c> for</c><00:00:11.440><c> your</c><00:00:11.599><c> attendance</c>
```

CoANZSE Data format

	country	state	name	channel_name	channel_url	video_title	video_id	upload_date	video_length	text_pos	location	latlong	nr_words
0	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Road Resurfacing Video	zVr6S5XkJ28	20181127	146.120	g_NNP_2.75 'day_XX_2.75 my_PRP$_3.75 name_NN_4.53 is_VBZ_4.74 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	433
1	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Weather update 5pm 1 March 2022 - Mayor Matt Gould	p4MjirCc1oU	20220301	181.959	hi_UH_0.64 guys_NNS_0.96 i_PRP_1.439 'm_VBP_1.439 just_RB_1.76 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	620
2	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Transport Capital Works Video	DXlkVTcmeho	20180417	140.450	council_NNP_0.53 is_VBZ_1.53 placing_VBG_1.65 is_VBZ_2.07 2018-19_CD_2.57 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	347
3	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	Council Meeting Wrap Up February 2022	2NhuhF2fBu8	20220224	107.840	g_NNP_0.399 'day_NNP_0.399 guys_NNS_0.799 and_CC_1.12 welcome_JJ_1.199 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	341
4	AUS	NSW	Wollondilly Shire Council	Wollondilly Shire	https://www.youtube.com/c/wollondillyshire	CITY DEAL 4 March 2018	4-cv69ZcwVs	20180305	130.159	[Music]_XX_0.85 it_PRP_2.27 's_VBZ_2.27 a_DT_3.27 fantastic_JJ_3.36 ...	62/64 Menangle St, Picton NSW 2571, Australia	(-34.1700078, 150.612913)	420

CoANZSE Audio: https://coanzse.org

- Searchable online CoAZNSE, including audio and forced alignment files
- Powered by BlackLab (De Does et al. 2017), developed at the Dutch Language Institute
- "Under the hood": Apache Lucene
- Accessible via CLARIN's Shibboleth authentication
- Backend: CSC server (Pouta service)



CoANZSE size by country/state/territory

Location	nr_channels	nr_videos	nr_words	video_length (h)	nr_audio_files
Australian Capital Territory	8	650	915,542	111.79	41,752
New South Wales	114	9,741	27,580,773	3,428.87	1,299,949
Northern Territory	11	289	315,300	48.72	6,628
Queensland	58	7,356	19,988,051	2,642.75	950,084
South Australia	50	3,537	13,856,275	1,716.72	643,866
Tasmania	21	1,260	5,086,867	636.99	240,453
Victoria	78	12,138	35,304,943	4,205.40	1,624,830
Western Australia	68	3,815	8,422,484	1,063.78	386,898
New Zealand	74	18,029	84,058,661	10,175.80	3,926,216
Total	**482**	**56,815**	**195,528,896**	**24,030.82**	**9,122,676**

Plans for Oulu center development

Resources to be made available

- CoNASE, CoBISE, CoGS
- Pipelines for data collection, processing, annotation
- Twitch transcript and live chat corpus
- Singapore Podcast Corpus

Potential problems

- More personnel needed
- Funding/grants needed
- Computational resources needed

Thanks for your attention!

References

Coats, S. (2022). CoANZSE: The Corpus of Australian and New Zealand Spoken English: A new resource of naturalistic speech transcripts. In P. Parameswaran, J. Biggs & D. Powers (Eds.), *Proceedings of the the 20th Annual Workshop of the Australasian Language Technology Association*, 1–5.

Coats, S. (2024a). Building a searchable online corpus of Australian and New Zealand aligned speech. Australian Journal of Linguistics.

Coats, S. (2024b). CoANZSE Audio: Creation of an online corpus for linguistic and phonetic snalysis of Australian and New Zealand Englishes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3407-3412.

De Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), CLARIN in the Low Countries (pp. 245–257). Ubiquity Press. https://doi.org/10.5334/bbi.20

European Council (2019). Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790

Morin, C., & Coats, S. (2023). Double modals in Australian and New Zealand English. *World Englishes*. https://doi.org/10.1111/weng.12639

Thaler, M., & Elsweiler, C. (2023). The role of gender in the realisation of apologies in local council meetings: A variational pragmatic approach in British and New Zealand English. *Zeitschrift für Anglistik und Amerikanistik*, 71(3), 217–239.