TG4 content

SVN path: Corpora/TG4

**Scripts**

The ‘scripts’ directory contains the scripts I used to download the episodes:

‘descriptions.tsv’ contains item descriptions for each episode/movie, with the following fields:

Programme name; Series; Episode number; Category; Description

The description is usually bilingual, with English first.

‘videos.tsv’ contains the video-specific information: the first four fields are the same as in ‘descriptions.tsv’, the final field is the URI (web address) of the video, which may be either a single video file, or an m3u playlist containing a list of video pieces.

‘wayback-video.txt’ contains the video (or video pieces, as extracted from the m3u) that I was able to save in the Internet Archive’s wayback machine (in case it was desirable to have the original video), while ‘wayback-subs.txt’ contains the corresponding subtitle URIs.

The script ‘vtt-patcher.pl’ converts vtt subtitles to Audacity labels, with the ability to merge segments, split segments, replace time signatures, and perform time-specific corrections.

The ‘unsorted-march’ directory contains fragments from the first attempt at downloading from TG4, where the video lacks corresponding subtitles, or vice versa. It may still be possible to recover more of the data, but in any case, the subtitles are useful for expanding the language model.

‘lm-data’ contains data from the language model, either crawled from the internet (lm-data/corpuscrawler) or from books/book samples (lm-data/ga-books)

**Issues**

All of the TG4 content has multiple speakers, and multiple dialects.

Lots of cross talk, crowd talk (in animated shows), background noises and music.

Child speech is usually learner speech from Gaelscoil students (my original use case was a pronunciation training tool, so I wanted to collect as many samples as possible, even if it meant I would have to transcribe them myself).

Where there are subtitles for programmes with child speech, such as ‘Whiz sa Chistin’, it has typically been corrected for errors: ‘spraoim peil’ transcribed as ‘imrím peil’, etc.

The emphasis in the animated TV shows seems to be on the energy of the actors’ performances, and pronunciation can sometimes suffer as a result (‘agus ansin’ pronounced as ‘agatiun’; ‘chuile dhuine’ pronounced as ‘chola dhona’, etc.)

There is a high instance of English words throughout: approximately 25% of unique words are English (this number is comparable to the web corpus of Irish that I collected for the language model, though).

There are also a number of semi-English words: waxáil, etc.

Ros na Rún has unique issues, because the actors are reciting lines from memory: the actors will quite often rephrase things in an equivalent way, phrases will be repeated, etc.

There are some minor issues with the text in the subtitles (they clearly did not use a spell checker), and the timings are intended to give people with hearing difficulties a chance to read them, so are not always in sync with the audio.

**Kids’ animated movies**

amhrannamara <https://www.cic.ie/en/general/other/the-song-of-the-sea>

niko https://en.wikipedia.org/wiki/The\_Flight\_Before\_Christmas\_(2008\_film)

**Kids’ animated TV series**

(Saol faoi Shráid is Irish-produced, and features puppets rather than animation, but I include it here)

astroblast https://en.wikipedia.org/wiki/Astroblast!

catahata https://en.wikipedia.org/wiki/The\_Cat\_in\_the\_Hat\_Knows\_a\_Lot\_About\_That!

dinotrain <https://en.wikipedia.org/wiki/Dinosaur_Train>

dora https://en.wikipedia.org/wiki/Dora\_the\_Explorer

garfield <https://ga.wikipedia.org/wiki/The_Garfield_Show>

gearoidnagaisce https://en.wikipedia.org/wiki/Inspector\_Gadget\_(2015\_TV\_series)

harveybeaks https://en.wikipedia.org/wiki/Harvey\_Beaks

olivia https://en.wikipedia.org/wiki/Olivia\_(TV\_series)

qpootle https://en.wikipedia.org/wiki/Q\_Pootle\_5

saolfaoishraid https://en.wikipedia.org/wiki/Saol\_faoi\_Shr%C3%A1id

spongebob https://en.wikipedia.org/wiki/SpongeBob\_SquarePants

toirbeir https://en.wikipedia.org/wiki/We%27re\_Going\_on\_a\_Bear\_Hunt#Television\_adaptation

wallaceandgromit https://en.wikipedia.org/wiki/Wallace\_and\_Gromit

**Other TG4**

**Kids programming**

lurgan2k17 <https://www.tg4.ie/ga/clair/lurgan-2k17/>

Mixture of music and speech; single episode only (other episodes were only subtitled in English). Teenagers.

wac <http://old.tg4.ie/ie/programmes/cula4/programmes/wac.html>

Mixture of speech and dubbed animal footage. Mostly teenagers.

whizsachistin https://www.cula4.com/ga/clair/whiz-sa-chistin/

Children’s cookery programme. Pre-teens/teenagers.

**Adult programming**

bealoideasbeo https://www.tg4.ie/ga/clair/bealoideas-beo/

rosnarun https://www.tg4.ie/ga/clair/ros-na-run/

**Non-TG4**

Corpora/teanglann

Contains the pronunciations from teanglann.ie

Corpora/TG4/CloIarChonnacht

This directory contains all of the audio I was able to download from Cló Iar-Chonnacht; no text (as yet), but it may be worth contacting them

Corpora/TG4/RnaG

Contains a dramatic adaptation of ‘Cré na Cille’, along with some fragments of the book that I was able to find online. As a dramatic adaptation, it diverges from the text quite a bit. It may be possible to do a ‘semi-forced alignment’ (i.e., recognition using the book text as the language model).

Corpora/TG4/coislife

This directory contained the MP3s from Cois Life’s Soundcloud; I’ve removed the files that overlap with the audiobooks. What remains are some of the manual alignments and corrections I had made, though it seems I failed to add all of them.

Corpora/cdroms/

Contains the extracted content from 90s-era CDROMs I found around the lab.

Corpora/cdroms/coisceim

Introduction to Irish: short phrases in 3 dialects, with transcription

Corpora/cdroms/dunchaochain

Mayo placenames with pronunciations

Corpora/TG4/no-subtitles-kids

This directory contains shorter programmes with younger children speaking, mostly Gaelscoil students. None of these programmes had subtitles, though I did transcribe some of the shorter pieces, to have examples of learner speech.