Standard names analyses and tooling: aims and preliminary work

Sadie Bartholomew
National Centre for Atmospheric Science & University of Reading

Building on related work by Jonathan Gregory (NCAS, UoR & MOHC)

2021 CF Workshop, CF Standard Names session 21st September 2021







Outline

- The set of all CF Standard Names forms a very interesting data set from numerous perspectives!
 - Can be seen (largely) as a list of the precise and systematicallyformed names of physical geophysical quantities in active usage
 - Grammar: both the phrases that make up each name (lexicon) and their syntax (rules for assembling the phrases) are informative
- Can we pick out patterns across the latest table? What can this analysis of the full set of names tell us?
- Ultimately, we hope analysis of the names can allow us to develop tools to streamline the proposal and/or acceptance process e.g. a bot to make suggestions

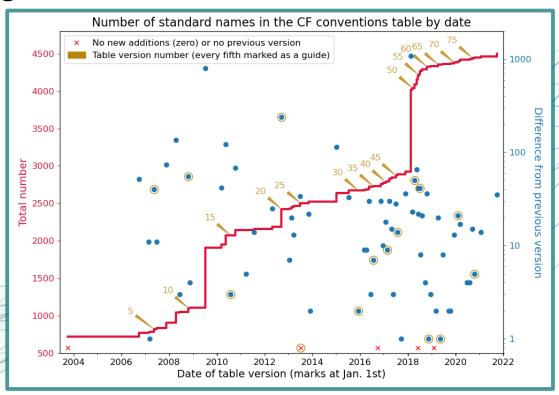






History

Table v.78 released today has 4495 Standard Names! That is 35 more than v.77released back in January '21.









Background

- Several years ago, Jonathan conducted grammatical study & analyses for table v.14 (released ~2010); we'd like to take it forward with the current, much larger table, and extend it.
- For example, let's discuss a tiny sub-sample of the table:
 - downward_northward_momentum_flux_in_air
 - downward_northward_momentum_flux_in_air_due_to_diffusion
 - downward_water_vapor_flux_in_air_due_to_diffusion
 - downwelling_longwave_flux_in_air
 - downwelling_longwave_flux_in_air_assuming_clear_sky
 - downwelling_longwave_radiance_in_air
 - downwelling_radiance_per_unit_wavelength_in_air
 - downwelling_radiative_flux_per_unit_wavelength_in_air







Analysis methods

- Such analysis comes under the umbrella of linguistics
- Luckily there are numerous mature and powerful tools for doing this kind of analysis (i.e. computational linguistics) with our language of choice, Python
- Libraries that can help analyse the names include NLTK (Natural Language Toolkit), SpaCy, Pattern, TextBlob
- I have started with the lighter tool TextBlob (see <u>textblob.readthedocs.io/en/dev/</u>) which wraps NLTK with simpler syntax, though will move to a more comprehensive and powerful tool next







 Initial analysis was on the lexical side (looking at the underlying phrases) and focused on finding the socalled n-grams, contiguous sequences of words

• e.g. a single word is n=1, a unigram, a two-word phrase is n=2, or a bigram, etcration due surface radioactivity surface upward density and itrogen compounds in a real fraction depth area fraction depth and real fraction depth area fraction

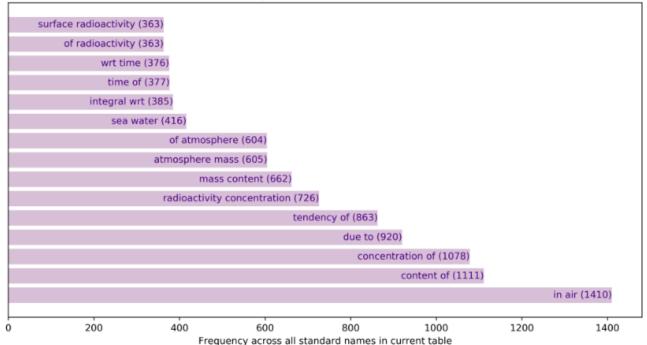
```
introgen monoxide example precised and are fraction depth and are fraction depth and are fraction depth are fraction deposition deposition depth are fraction deposition deposition depth are fraction deposition deposition depth are fraction d
```







15 most common n-grams of size 2 for the CF Standard Names

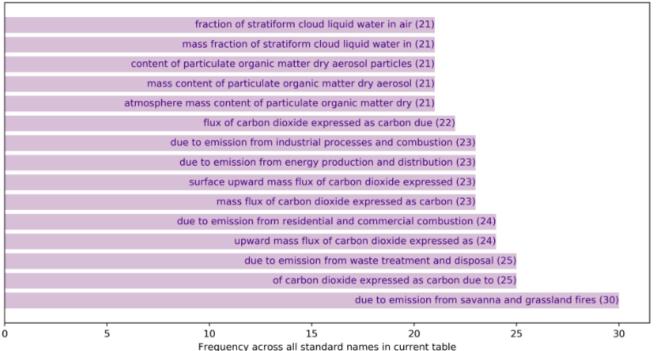








15 most common n-grams of size 8 for the CF Standard Names











15 most common n-grams of size 18 for the CF Standard Names

depth at shallowest local minimum in vertical profile of mole concentration of dissolved molecular oxygen in sea water (1) concentration of colored dissolved organic matter in sea water expressed as equivalent mass fraction of quinine sulfate dihydrate (1 to x angle of rotation from east to y angle of rotation from solar azimuth to platform azimuth (1) east to x angle of rotation from east to y angle of rotation from solar azimuth to platform (1) from east to x angle of rotation from east to y angle of rotation from solar azimuth to (1) rotation from east to x angle of rotation from east to y angle of rotation from solar azimuth (1) of rotation from east to x angle of rotation from east to y angle of rotation from solar (1) angle of rotation from east to x angle of rotation from east to y angle of rotation from (1) upward mass flux of carbon dioxide expressed as carbon due to anthropogenic land use or land cover change (2) surface upward mass flux of carbon dioxide expressed as carbon due to anthropogenic land use or land cover (2) mole concentration of particulate organic matter expressed as carbon in sea water due to net primary production by (5) of mole concentration of particulate organic matter expressed as carbon in sea water due to net primary production (6) tendency of mole concentration of particulate organic matter expressed as carbon in sea water due to net primary (6) of atmosphere mass content of particulate organic matter dry aerosol particles expressed as carbon due to emission from (9) tendency of atmosphere mass content of particulate organic matter dry aerosol particles expressed as carbon due to emission (10) 10 Frequency across all standard names in current table









```
residential and commercial combustion (24)
                                                                                 Note the colours of the bars
 carbon monoxide (25)
                                                                                 and their labels do not have
 waste treatment and disposal (25)
                                                                                 any meaning; they are plotted
  of carbon (26
  product of (26)
                                                                                 in alternating colours to make it
  optical thickness (26
                                                                                 easier to distinguish neighbouring
  convective cloud (26)
                                                                                 bars and their labels.
  air pressure (26)
   nitrogen compounds nitrogen (28)
I longwave flux (29)
     sea salt (29)
     variance spectral density (29) 
forest fires (30)
      net downward (30)
      savanna and grassland fires (30)
eddy advection (31)
       surface downwelling (31)
       radiative flux (31)
       nitrogen monoxide (31)
       surface snow (31)
       swell wave (32)
         production and (33)
         area fraction (33)
         surface downward (34)
         clear sky (36)
            ocean mole (38
              water vapor (39)
              land ice (40)
              atmosphere layer (40)
               air temperature (42
                surface upward (43)
                matter dry (43)
carbon in (45)
                  shortwave flux (46)
 heat flux (46)
                      wet deposition (52)
       per unit (56)
                         carbon dioxide (59)
      liquid water (61)
                           stratiform cloud (63)
particulate organic (63)
                             dry deposition (68)
                                      sea ice (93)
                                                           moles of (201)
                                       aerosol particles (242)
                                                                    emission from (268)
                                                expressed as (296)
                                                                            fraction of (357)
                                              surface radioactivity (363)
                                                                              time of (377)
                                                                                sea water (416)
                                                                                                     tendency of (863
                                                                                       due to (920)
                                                                                       content of (1111)
                                                                                  500 600 700 800 900
                           70 80 90
                                                                                                      1000
```

Frequency across all standard names in current table (log scale)

Some preliminary results

• In this case, we look for most common *n*-gram of any *n*, but removing it once found









Thanks for listening. Any questions?

Do you have any ideas about tools that you think could streamline the standard names proposal process? If so, please email *sadie.bartholomew@ncas.ac.uk*. I'd be interested to hear your thoughts!

- Quick links related to the talk:
 - Jonathan's original work on 'Parsing CF standard names': www.met.rdg.ac.uk/~jonathan/CF_metadata/14.1/
 - Repository where I am storing code for the analyses: github.com/sadielbartholomew/cf-standard-names-linguistics
 - Dedicated webpage to display key results from the above: sadielbartholomew.github.io/cf-standard-names-linguistics/
 - A relevant open issue regarding Standard Name tooling: github.com/cf-convention/discuss/issues/88





