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

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

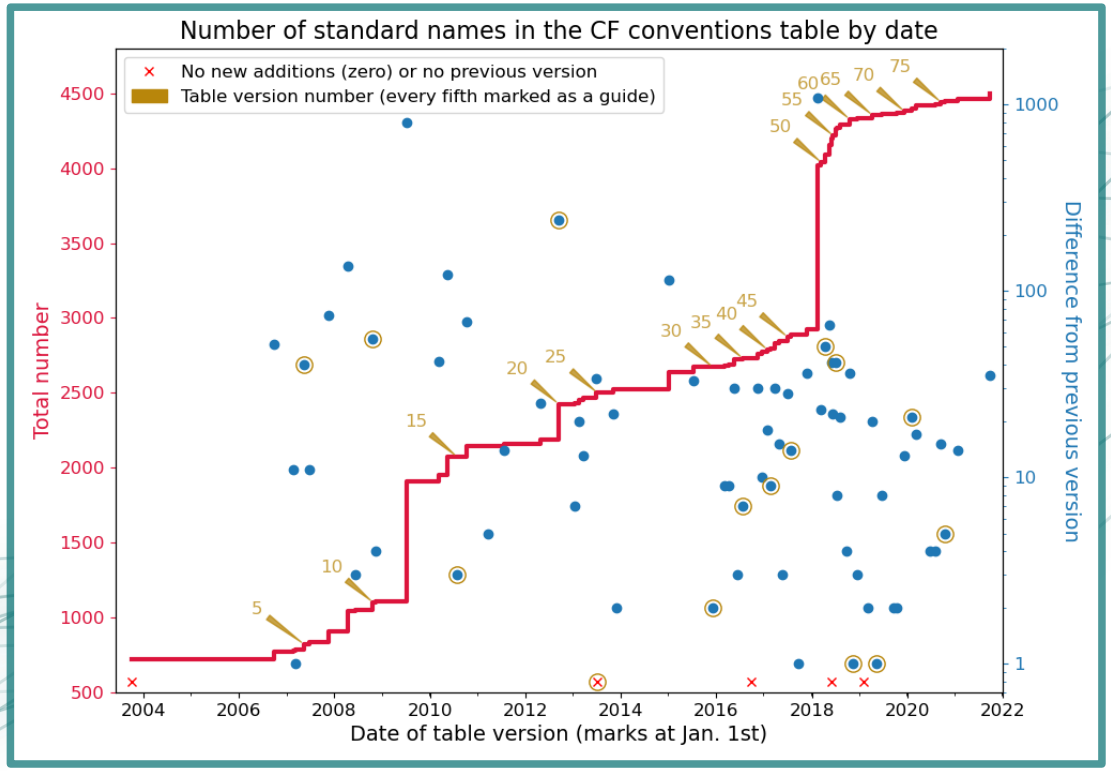
Outline

- 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 systematically-formed 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

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
University of Reading
NCAS work on CF is supported by:
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

History

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



National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

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

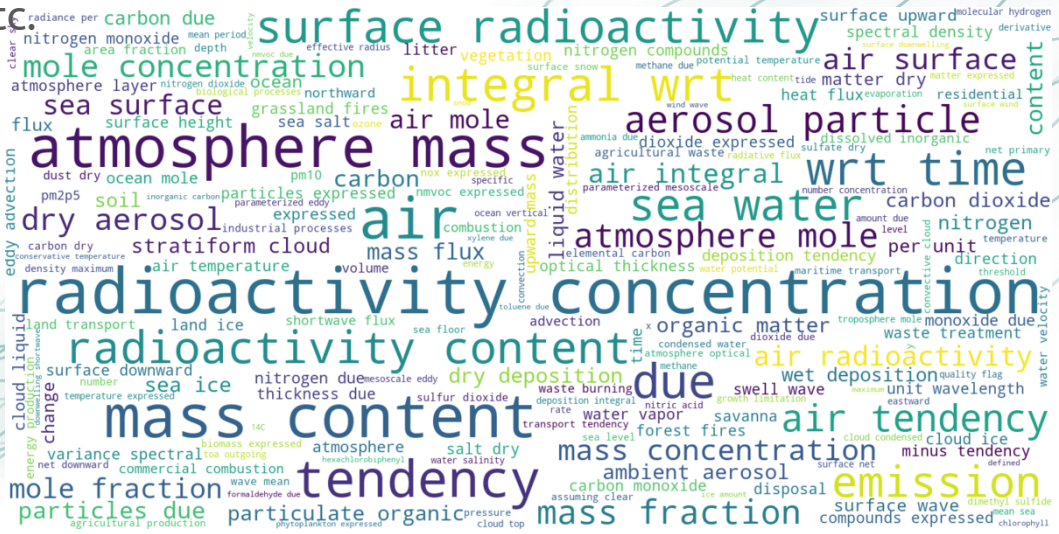
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
University of Reading
NCAS work on CF is supported by:
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

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 textblob.readthedocs.io/en/dev/) which wraps NLTK with simpler syntax, though will move to a more comprehensive and powerful tool next

Some preliminary results

- Initial analysis was on the *lexical* side (looking at the underlying *phrases*) and focused on finding the so-called **n-grams, contiguous sequences of words**
  - e.g. a single word is *n=1*, a *uni*gram, a two-word phrase is *n=2*, or a *bi*gram, etc.



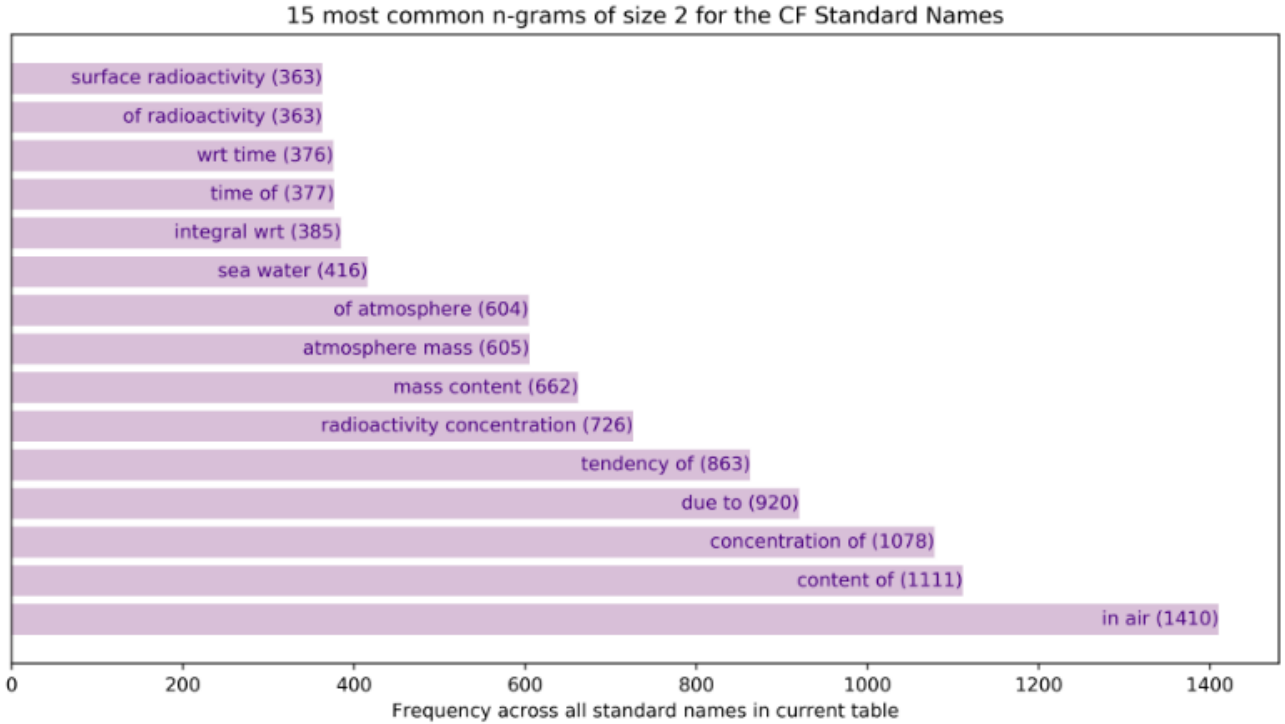
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Some preliminary results



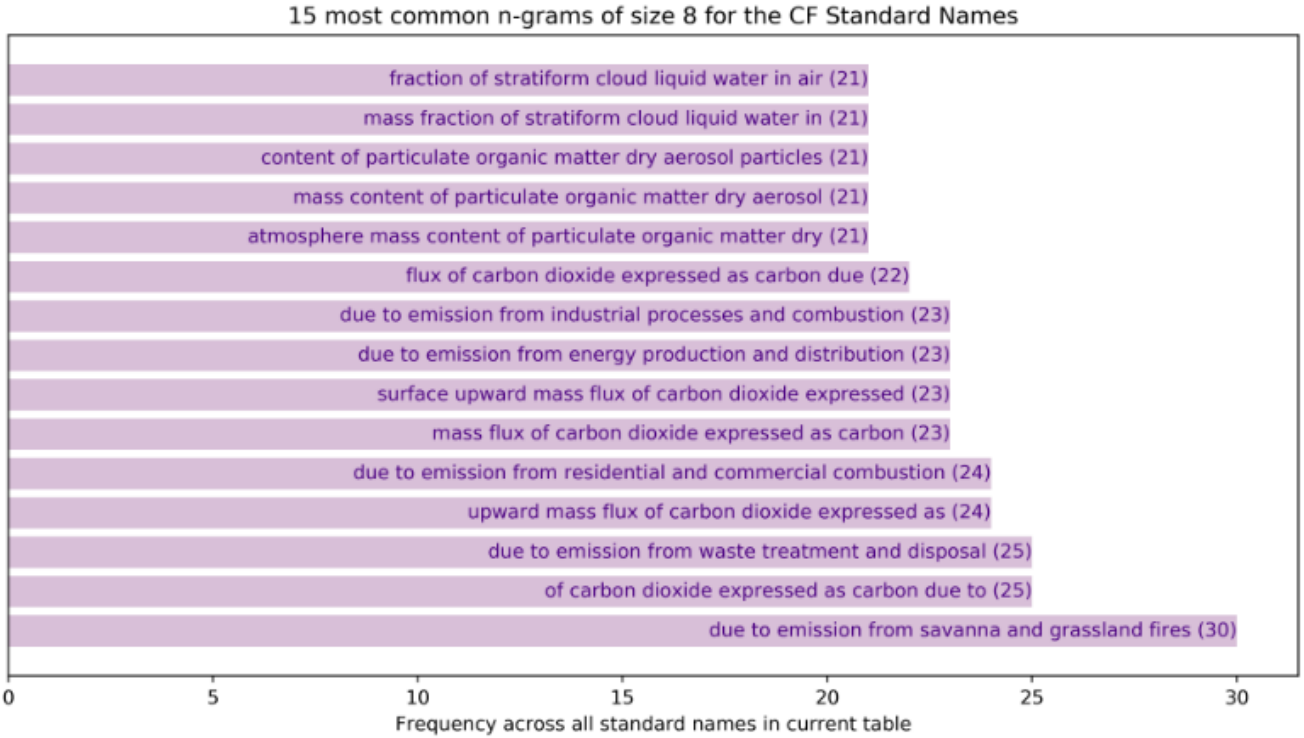
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Some preliminary results



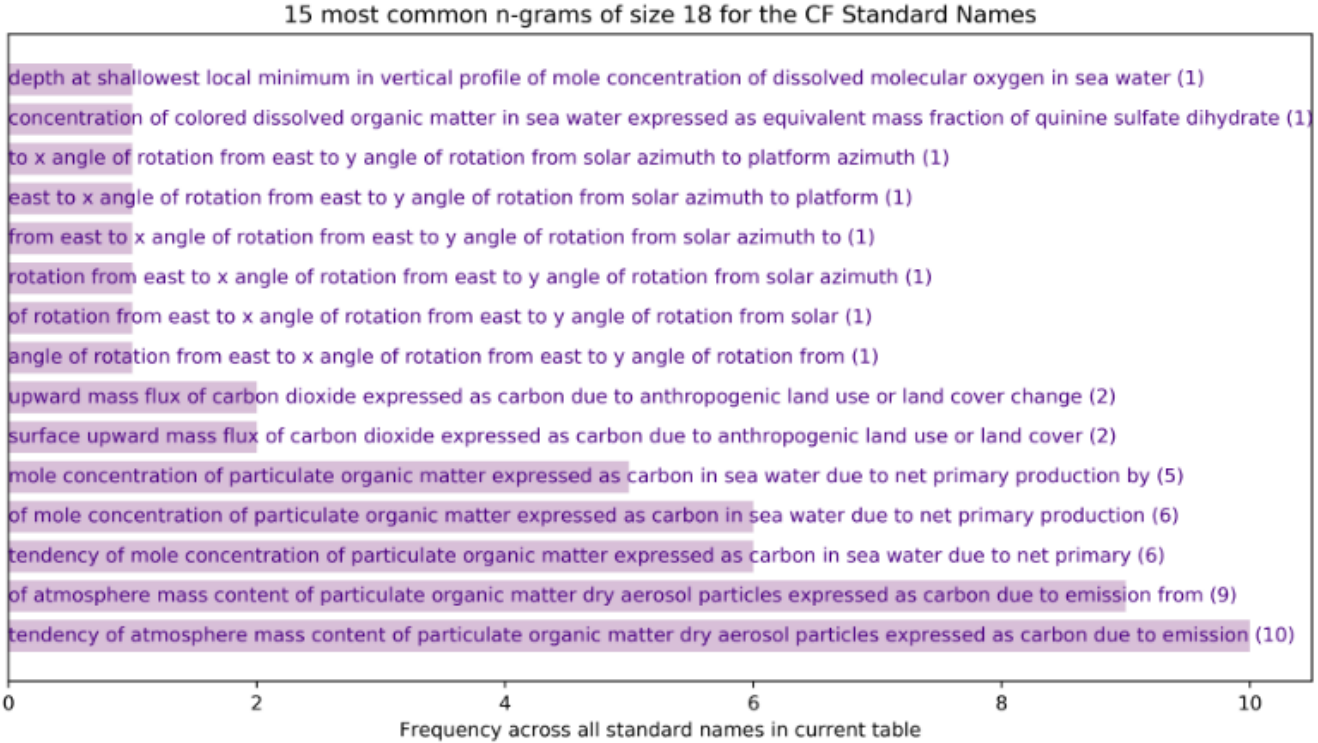
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Some preliminary results

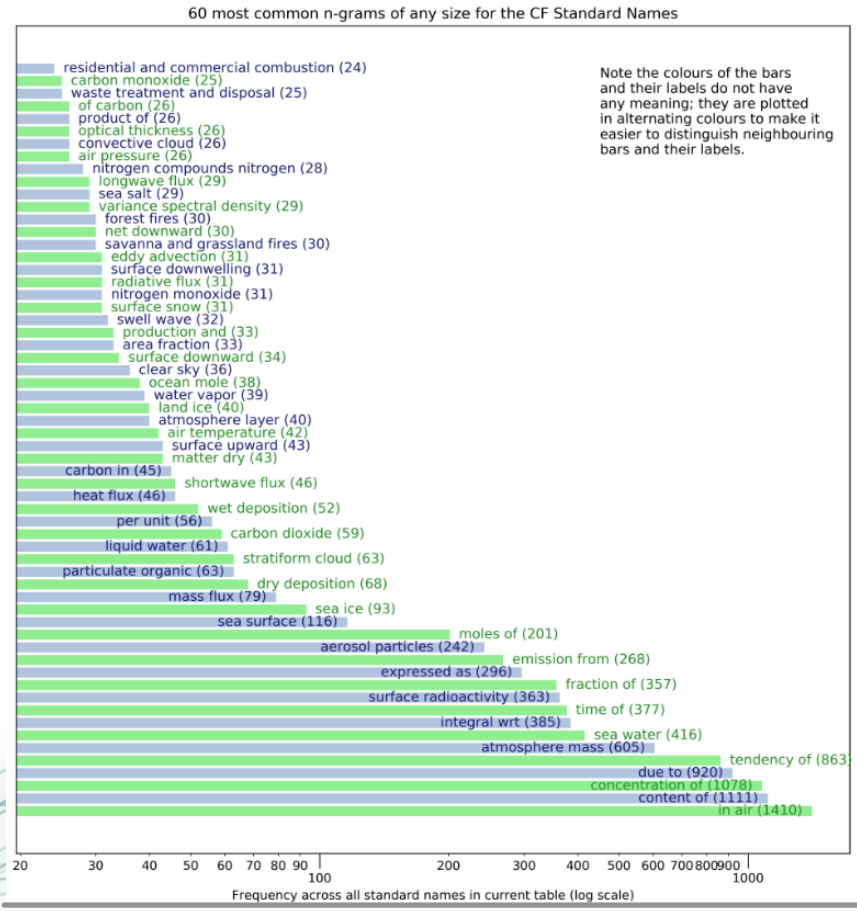


National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING



Some preliminary results

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

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
University of Reading
NCAS work on CF is supported by:
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

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

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NCAS work on CF is supported by:

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING