Towards an explorer tool for visualisation of grammatical patterns in the CF Standard Names through decomposition into *n*-grams

Keywords: metadata, geophysics, natural language processing (NLP), (computational) linguistics, data visualisation, network science
American Geophysical Union Fall Meeting 2023 • IN31C: CF and NetCDF: 30 Years of Wide-Open Science

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Sadie L. Bartholomew
National Centre for Atmospheric Science and University of Reading (UK)

University of Reading

Context | Assembling geophysical metadata

The Standard Names of the CF Conventions [1] are metadata identifiers for geophysical quantities, designed to be assigned to the CF `standard_name` attribute to describe what a given data variable *means*. An open community process sustains the set of names, with much effort and care taken to foster cross-domain application, and around two decades of by-request extension has seen it diversify from ~400 to ~4700 names. Name composition is informed by guidelines outlining a general structure which permits contextual qualifications:

`[<surface>] [<component>] <base_quantity> [at_<surface>] [in_<medium>] [due_to_<process>] [assuming_<condition>]`

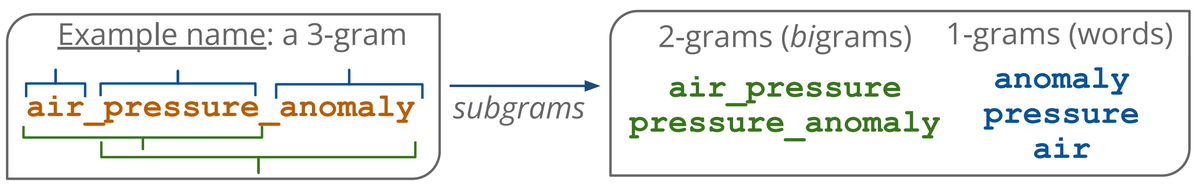
with an example being:

`downward_water_vapor_flux_in_air_due_to_diffusion`,

however other than these loose patterns, addition of new names is regulated largely through moderation by volunteers. Furthermore, the CF Conventions website only sets out the names in listed (tabular) format. With deeper insight into composition of the names and their interconnections, and/or alternative non-linear forms of presentation, the process of exploring existing names and/or proposing or approving new ones could be streamlined. Ideally, new tooling can be forged based on up-to-date deconstruction. Though some grammatical analysis has been conducted towards this aim [2], it is over a decade out of date.

Method | Dissecting the names into *n*-grams

The current release of the CF Standard Name Table, Version 83 [1], consisting of 4667 names, was studied with methodology that centered on the *n*-gram [3], a contiguous sequence of $n \in \mathbb{N}$ items, which is used extensively within natural language processing (NLP) especially for probabilistic analysis, e.g. to facilitate auto-completion of words in text applications. For the names, the natural atomic item is a word, ignoring underscore delimination. To demonstrate, all *subgram*s contained by one small standard name are:



The first aim was to, in NLP terms, treat the set of all names as a *text corpus* and perform *word tokenisation* to find all *n*-grams occurring with at least cutoff recurrence *c* for $n \geq 1$. Then, link names to all subgrams, in turn with decreasing *n*, to highlight shared *n*-grams. The Python code used the `TextBlob` and `NetworkX` libraries and is shared on GitHub [4].

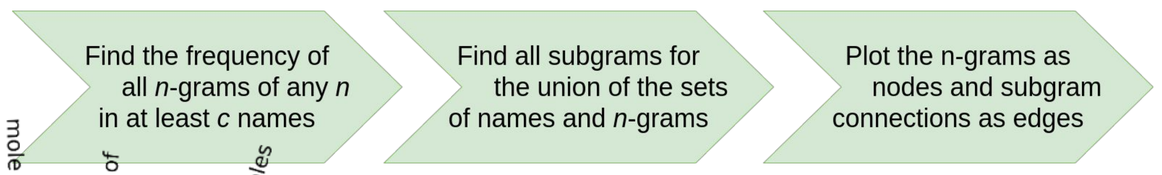


Figure 1: The core steps of the computational algorithm.

Results | Networks linking names via shared *n*-grams

Key figures from the *n*-gram frequency counting stage (call it frequency *f*) were that:

- '`in air`' was the most common *n*-gram and 2-gram, having *f* = 1456;
- '`radioactivity concentration of`' was the most common 3-gram (*f* = 724);
- the largest *n* for which an *n*-gram recurred was *n* = 20 ('`tendency of mole concentration of particulate organic matter expressed as carbon in sea water due to net primary production by`') with *f* = 5.

Plots detailing the 15 most common *n*-grams across the full range $1 \leq n \leq 20$ have been shared online [4] for further reference.

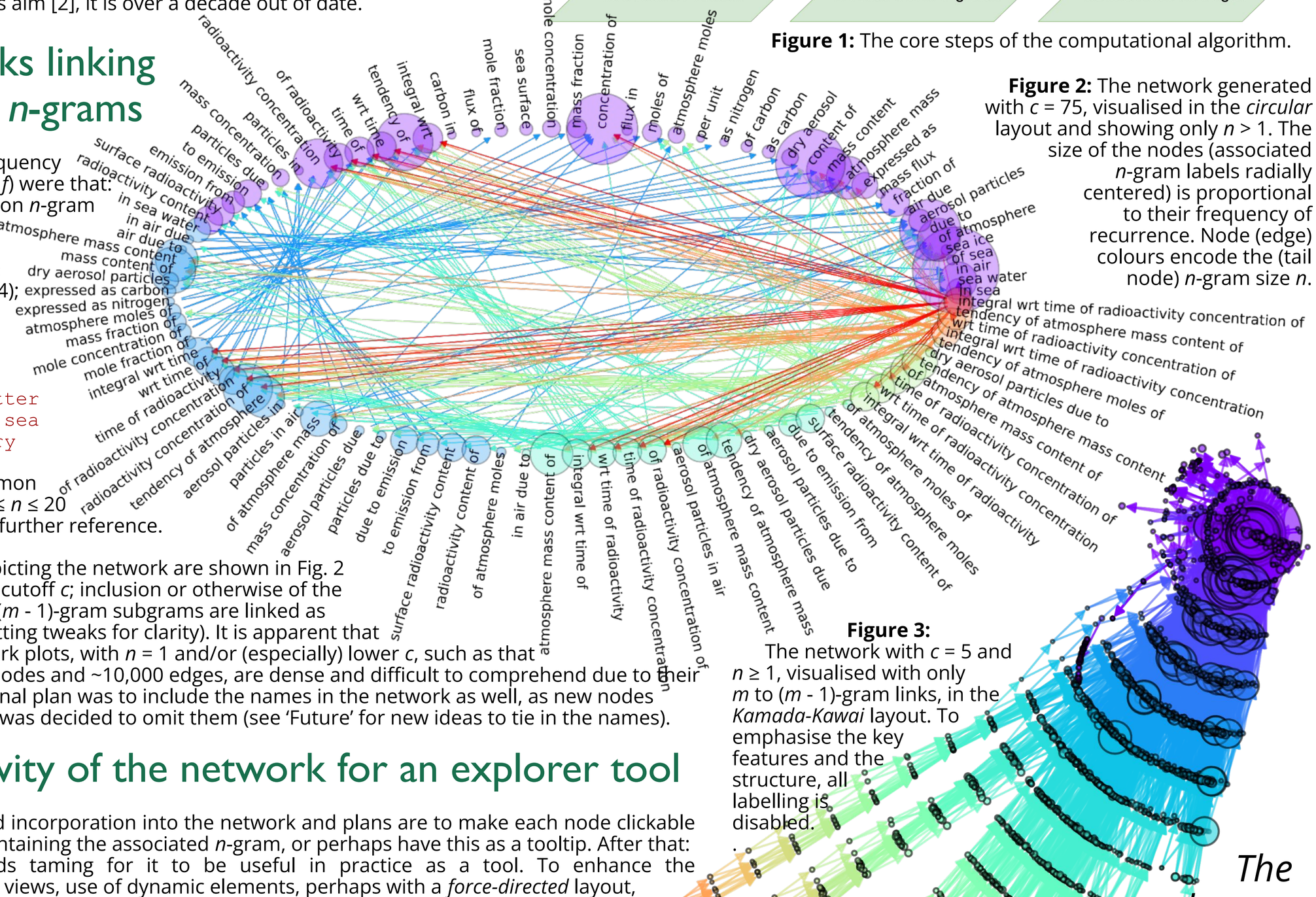


Figure 2: The network generated with *c* = 75, visualised in the *circular* layout and showing only *n* > 1. The size of the nodes (associated *n*-gram labels radially centered) is proportional to their frequency of recurrence. Node (edge) colours encode the (tail node) *n*-gram size *n*.

Examples of plots produced depicting the network are shown in Fig. 2 and Fig. 3, which vary by layout; cutoff *c*; inclusion or otherwise of the case of *n* = 1; and whether only (*m* - 1)-gram subgrams are linked as opposed to all (as well as formatting tweaks for clarity). It is apparent that the more-comprehensive network plots, with *n* = 1 and/or (especially) lower *c*, such as that in Fig. 3, which portrays ~5000 nodes and ~10,000 edges, are dense and difficult to comprehend due to their size. Therefore, though the original plan was to include the names in the network as well, as new nodes linked to any subgram nodes, it was decided to omit them (see 'Future' for new ideas to tie in the names).

Figure 3: The network with *c* = 5 and $n \geq 1$, visualised with only *m* to (*m* - 1)-gram links, in the *Kamada-Kawai* layout. To emphasise the key features and the structure, all labelling is disabled.

The subgram network needs taming for it to be useful in practice

Future | Interactivity of the network for an explorer tool

The names themselves still need incorporation into the network and plans are to make each node clickable to produce a list of all names containing the associated *n*-gram, or perhaps have this as a tooltip. After that:

1. The subgram network needs taming for it to be useful in practice as a tool. To enhance the comprehensibility of network views, use of dynamic elements, perhaps with a *force-directed* layout, may help [5]. The highlighting of subgram pathways from a given node of interest is another idea.
2. Map this systematic top-down *n*-gram analysis to external bottom-up work to deduce a *lexicon* of *phrase type*s and a *syntax*, e.g. [2]. Is a given *n*-gram a phrase type, or just coincidental?

Acknowledgements

This work has been partially supported by the Horizon 2020 programme through IS-ENES3. The author would like to thank Jonathan Gregory for discussions prompting this work based upon the intention to update and/or extend [2] and NCAS-CMS for affording some time to investigate this topic.

↘ Take-home points

- ➔ The CF Standard Names identify geophysical quantities. More insight into their composition and interconnections will benefit both those who maintain and who consult the table of names, to discover, or propose new, names.
- ➔ One approach to such evaluation to use word tokenisation (splitting) into *n*-grams, sequences of $n \in \mathbb{N}$ words.
- ➔ Code has been developed in Python using the `TextBlob` and the `NetworkX` libraries to determine the recurrence frequency *f* for all possible *n*-grams, optionally parameterised by a cutoff *c* where $f \geq c$ for inclusion, and plot this in various layouts as a network of nodes, proportional in size to *f*, connected through all successive subgrams.
- ➔ The intention is to make the network useful through reduction or interactivity and share it as an exploration tool.

↖ References

With all websites last accessed on 2023-11-29T15:07:46Z (UTC):

1. CF Conventions (2023), *CF Standard Name Table,* Version 83, <http://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html>
2. Gregory, J. M. (2010), *Parsing CF standard names*, Version 14.1, <http://www.met.rdg.ac.uk/~jonathan/CF_metadata/14.1/>
3. Jurafsky, D. and Martin, J.H. (2009), *Chapter 4: N-grams* from *Speech and Language Processing*, Second Edition, Pearson Prentice Hall
4. Bartholomew, S. L. (2023), *Lexical and semantic analysis, and visualisation, of the CF Conventions Standard Names*, <https://github.com/sadielbartholomew/cf-standard-names-linguistics>
5. Barabási, A. L. and Pósfai, M. (2016), *Network Science*, Cambridge University Press