AnnoTize: A Flexible Annotation Tool for Documents with Mathematical Formulae

Lukas Panzer **Jan Frederik Schaefer**

FAU Erlangen-Nürnberg/KWARC

MathUI Workshop
CICM Conference
Cambridge, UK
September 7, 2023

Natural Language Processing and Mathematical Language

- Natural language processing has benefitted from a long tradition of annotation tasks and benchmarks
- STEM documents pose problems: formulae, tables, . . . *not really unicode strings*
- Why care?
  $\rightsquigarrow$ Semantic services

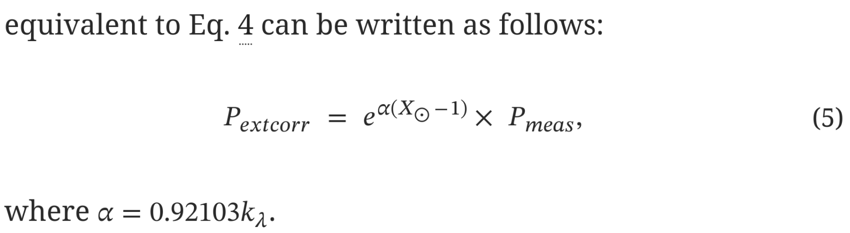
Motivation: semantic services

- 🔍 $\boldsymbol{1.5\,\mathrm{eV}}$
    - ↗ $1.43 \pm 0.9\,\mathrm{eV}$
    - ↗ $2.4 \cdot 10^{-19}\,J$

- 🔍 $\boldsymbol{\sum_{k=-\infty}^{\infty} \exp(-\pi k^2)}$
    - ↗ $\sum_{n=-\infty}^{\infty} e^{-\pi n^2} = \ldots$

Motivation: semantic services

Example from [Kri22]



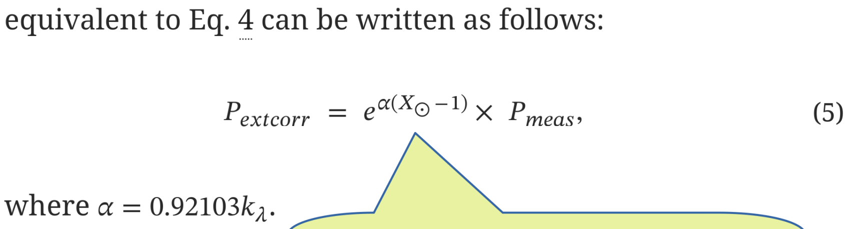
equivalent to Eq. 4 can be written as follows:

$$P_{extcorr} = e^{\alpha(X_{\odot} - 1)} \times P_{meas}, \tag{5}$$

where $\alpha = 0.92103k_{\lambda}$.

Motivation: semantic services

Example from [Kri22]



equivalent to Eq. 4 can be written as follows:

$$P_{extcorr} = e^{\alpha(X_{\odot} - 1)} \times P_{meas}, \qquad (5)$$

where $\alpha = 0.92103k_{\lambda}$.

“the Sun’s airmass”

Amount of air in direction of sun
If relative: Divided by amount of air at zenit

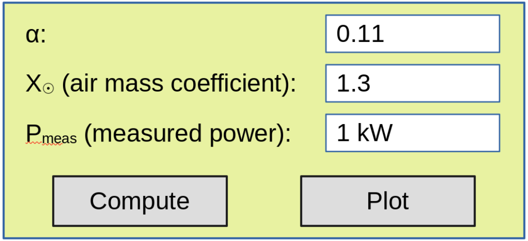
Motivation: semantic services

Example from [Kri22]

equivalent to Eq. 4 can be written as follows:

$$P_{extcorr} = e^{\alpha(X_{\odot} - 1)} \times P_{meas}, \quad (5)$$

where $\alpha = 0.92103 k_{\lambda}$.



α: 0.11

$X_{\odot}$ (air mass coefficient): 1.3

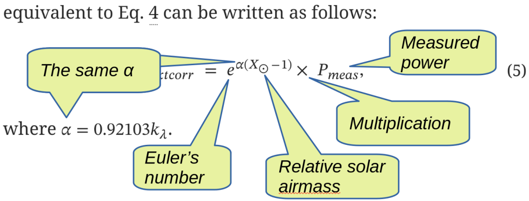
P_{meas} (measured power): 1 kW

Motivation: semantic services

For all those services

we need semantic annotations!

(full formalization not necessary)



equivalent to Eq. 4 can be written as follows:

$$_{tcorr} = e^{\alpha(X_{\odot}-1)} \times P_{meas}, \qquad (5)$$

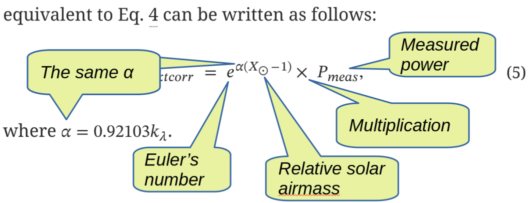
where $\alpha = 0.92103k_{\lambda}$.

Motivation: semantic services

For all those services

we need semantic annotations!

(full formalization not necessary)



equivalent to Eq. 4 can be written as follows:

$$_{tcorr} = e^{\alpha(X_{\odot}-1)} \times P_{meas}, \qquad (5)$$

where $\alpha = 0.92103k_{\lambda}$.

Authors don't provide them ⇝ We have to infer them

AnnoTize

We will need manual annotations

for evaluation and possibly training

Formulae prevent us from using the standard tools

no reasonable plaintext representation

⇝ We present **AnnoTize**, a flexible annotation tool for math documents

https://github.com/rezakul/AnnoTize

Conclusion

AnnoTize

- is an annotation tool for HTML documents with a particular focus on formula support
- supports a wide range of annotation types

  ABoSpec files for new types of declarations
- makes the annotation process more efficient with templates

References I

[Kri22] Kevin Krisciunas. *Including Atmospheric Extinction in a Performance Evaluation of a Fixed Grid of Solar Panels*. 2022. arXiv: 2107.02876 [astro-ph.IM].