

Developing Lexical Resources for Controlled Authoring Purposes

Kara Warburton

City University of Hong Kong

kara@termologic.com

Agenda

- A few concepts and assertions
- Lexical resources for CA
- Examples of use
- Challenges of using an existing termbase
- Conclusion

Controlled authoring (CA)

- CA is "the process of applying a set of predefined style, grammar, punctuation rules and approved terminology to content (documentation or software) during its development" (Ó Broin, 2009).
- CA is increasingly being adopted by companies as well as by public institutions as a way to improve content and decrease production and translation costs
- Active vs passive CA

Terminology for authoring

- Passive authoring (“pull”)
 - spreadsheets
 - Word files/glossaries
 - Lookup Web sites
- Active authoring (“push”)
 - “term checker” running in conjunction style checker and spell checker
 - different formats depending on the tool – Acrolinx, HyperSTE (Tedopres), crossAuthor (Across)

Key objectives for CA

- Minimize synonyms and variants
- Minimize homonyms
- Manage abbreviations
- Protect non-translatable text

Lexical resources for CA

- Differ from “terminology”
 - include general lexicon expressions
 - more non-nominal units
- Are not “lexicographical”
 - Require the concept-oriented approach to manage synsets
 - ISO standards: 16642, 30042.

Usage status values

- Preferred – the “best” term
- Prohibited – not allowed, ever
- *Constrained* – allowed in some contexts
 - What are those contexts? Subject field? Product? Grammatical category?
 - Usage notes to help
- *Admitted* – allowed but not preferred

Ter·mo·lo·gic

The screenshot displays the termweb application interface. At the top left is the 'termweb' logo. Below it, the 'Dictionary' is set to 'MiniT Test' and the view is 'All sections, No additional dictionaries'. The 'Source' is 'English' and the 'Filter' is '<No filter>'. A search bar contains the word 'perform'. On the left, a list of terms is shown, with 'perform' at the top. The main content area shows three search results for 'perform' in English. The first result (Term ID 2181-3) includes a sample sentence, usage status (Constrained), part of speech (Verb), register (Neutral), and a usage note. The second result (Term ID 2181-1) is for the word 'finish' with similar metadata. The third result (Term ID 2181-2) is for the word 'complete' with usage status (Rejected), part of speech (Verb), and register (Neutral).

termweb

Dictionary: MiniT Test All sections, No additional dictionaries < Change

Source English Filter <No filter>

perform Search

Rikstermbanken
 Domain search

perform
performed
period
period of time
permit
permitted
permutation
person
personal computer
personnel
pi
PI
picture
pie chart
Piepel
Pillai's test
pinpoint
place
place
placeholder
placement
Plackett-Burman design
Plant Flag
plot
plot
plotted point
notted points

New Edit Copy Delete Admin

Target Custom Custom view... View condensed Settings

Concept ID 2181

Term ID 2181-3

English
perform
Sample sentence Minitab requires at least 10 observations to perform this analysis; however, you should have at least 25 observations for an adequate study.
Usage status Constrained
Part of speech Verb
Register Neutral
Usage note Use "perform" only to mean "run a command in Minitab".

Term ID 2181-1

English
finish
Usage status Constrained
Part of speech Verb
Register Neutral
Usage note Use "finish" in the sense of completing something, of bringing it to an end.

Term ID 2181-2

English
complete
Usage status Rejected
Part of speech Verb
Register Neutral

Ter·mo·lo·gic

termweb

Dictionary:

Source: Filter:

Rikstermbanken
 Domain search

- perform
- performed
- period
- period of time
- permit
- permitted
- permutation
- person
- personal computer
- personnel
- pi
- PI
- picture
- pie chart
- Piepel
- Pillai's test
- pinpoint
- place
- place
- placeholder
- placement
- Plackett-Burman design
- Plant Flag
- plot
- plot
- plotted point
- plotted points
- plotter
- PLS
- PLS coefficient plot

Term ID 2181-4
English
conduct
Usage status Rejected
Part of speech Verb
Register Neutral
Usage note Simplified Technical English not approved word

Term ID 2181-5
English
do
Usage status Preferred
Part of speech Verb
Register Neutral

Term ID 2181-6
English
accomplish
Usage status Rejected
Part of speech Verb
Register Neutral
Usage note Simplified Technical English not approved word

Term ID 2181-7
English
carry out
Usage status Rejected
Part of speech Verb
Register Neutral
Other term type Idiom

Ter·mo·lo·gic

Variant of Term: 'hostname'
Status: Use with caution
Part of speech: noun
Replace with : host name
Edit Flag
Ignore Flag

section

title **Hostname, a use-with-caution noun** title

p Enter the **hostname**. Enter the host name. p

section

section

title **Dumpfile, a do-not-use noun** title

p Delete the **dumpfile**. Delete the dump file. p

Term: 'dumpfile'
Status: Do not use
Part of speech: noun
Replace with : dump file
Edit Flag
Ignore Flag

termweb

Concept ID 2316

Term ID 2316-2

English
host name
Usage status Preferred
Part of speech Noun

Term ID 2316-1

English
hostname
Usage status Constrained
Part of speech Noun
Usage note Use only as a variable, otherwise, use "host name."

termweb

Concept ID 2317

Term ID 2317-1

English
dump file
Usage status Preferred
Part of speech Noun

Term ID 2317-2

English
dumpfile
Usage status Rejected
Part of speech Noun

Ter·mo·lo·gic

The screenshot displays the Termweb application interface. At the top left is the 'termweb' logo. Below it, the 'Dictionary' is set to 'MiniT Test' and the view is 'All sections, No additional dictionaries'. The 'Source' is 'English' and the 'Filter' is '<No filter>'. A search box contains 'catastrophic error' with a search button. Below the search box is a list of search results, with 'catastrophic error' selected. To the right of the search box are icons for 'New', 'Edit', 'Copy', 'Delete', and 'Admin'. Below the search box is a 'Target' dropdown set to 'Custom' and a 'View' button. The main content area shows the details for the selected term, 'catastrophic error', with its Term ID (2318-1), Usage status (Rejected), and Part of speech (Noun). Below this are three other terms: 'unrecoverable error' (Term ID 2318-2, Usage status Preferred, Part of speech Noun), 'irrecoverable error' (Term ID 2318-3, Usage status Allowed, Part of speech Noun), and 'fatal error' (Term ID 2318-4, Usage status Rejected, Part of speech Noun).

termweb

Dictionary: MiniT Test All sections, No additional dictionaries < Change

Source English

Filter <No filter>

catastrophic error Search

Rikstermbanken

Domain search

catastrophic error

category data

category predictor

category response variable

category scale

category variable

category

category data

category variable

Cauchy

Cauchy distribution

Cauchy distribution (also called Lorentz distribution or Breit-Wagner distribution)

cause

cause

cause and effect diagram

cause-and-effect diagram

cause-and-effect diagram (also called C&E diagram, C&E matrix, Ishikawa diagram, or fishbone diagram)

cause for concern

CCF

C chart

CCk

New Edit Copy Delete Admin

Target Custom Custom view... View

Concept ID 2318

Term ID 2318-1

English

catastrophic error

Usage status Rejected

Part of speech Noun

Term ID 2318-2

English

unrecoverable error

Usage status Preferred

Part of speech Noun

Term ID 2318-3

English

irrecoverable error

Usage status Allowed

Part of speech Noun

Term ID 2318-4

English

fatal error

Usage status Rejected

Part of speech Noun

Ter·mo·lo·gic

section
title Troubleshooting title
p If you encounter a catastrophic error, you need to apply the cumulative fix. p
section

title Troubleshooting title
p If you encounter a catastrophic error, you need to apply the cumulative fix. p
section
conbody
concept

Term: 'catastrophic error'
Status: Do not use
Part of speech: noun

Replace with :
unrecoverable error

Edit Flag
Ignore Flag

Step-through Mode
Previous Flag
Next Flag

you need to apply the cumulative fix. p

Term: 'cumulative fix'
Status: Do not use
Part of speech: noun

Replace with :
fix pack


Edit Flag
Ignore Flag

Step-through Mode
Previous Flag
Next Flag

Ter·mo·lo·gic

Search for:

Language: in English in

Did you know?  Who has been recognized as a Search Champion of the m Search? [Here they are.](#)

1 - 10 of about 1,834 for catastrophic error sorted by relevance

Did you mean: [fatal error](#) , [irrecoverable error](#) , [unrecoverable error](#) ?

Ter·mo·lo·gic

The screenshot shows the termweb interface. At the top left is the 'termweb' logo. Below it, the 'Dictionary' is set to 'MiniT Test' and 'All sections, No additional dictionaries' is selected. The 'Source' is 'English' and the 'Filter' is '<No filter>'. A search box contains 'modular arithmetic' and a search button is visible. Below the search box, there is a 'Rikstermbanken' section with a 'Domain search' checkbox. A list of search results is shown on the left, with 'modular arithmetic (also called clock arithmetic)' selected. On the right, the 'Target' is 'Custom' and the 'View' is 'condensed'. The main content area displays three terms related to 'modular arithmetic':

- Term ID 1140-3**
English
modular arithmetic (also called clock arithmetic)
Usage status: Constrained
Part of speech: Noun
Term type: Surface
Register: Technical
- Term ID 1140-2**
English
clock arithmetic
Usage status: Rejected
Part of speech: Noun
Register: Technical
- Term ID 1140-1**
English
modular arithmetic
Usage status: Preferred
Part of speech: Noun
Register: Technical

Ter·mo·lo·gic

```
<termEntry>
...
  <tig>
    <term>modular arithmetic</term>
    <termNote type="partOfSpeech">Noun</termNote>
    <termNote type="Register">Technical</termNote>
    <termNote type="Usage_Status">Preferred</termNote>
  </tig>
  <tig>
    <term>clock arithmetic</term>
    <termNote type="partOfSpeech">Noun</termNote>
    <termNote type="Register">Technical</termNote>
    <termNote type="Usage_Status">Rejected</termNote>
  </tig>
  <tig>
    <term>modular arithmetic (also called clock arithmetic)</term>
    <termNote type="partOfSpeech">Noun</termNote>
    <termNote type="Register">Technical</termNote>
    <termNote type="Usage_Status">Constrained</termNote>
    <termNote type="Form">Surface</termNote>
  </tig>
...
</termEntry>
```

Challenges using an existing termbase

- Few synsets
- Lack of required data categories (usage status, part of speech, conditions of restriction)
- Presence of terms not required for CA
- Problems introduced by automatic pos-assignment of the CA tool

partition

noun

1. Definition: A portion of a page set. Each partition corresponds to a single, independently extendable data set. Partitions can be extended to a maximum size of 1, 2, or 4 gigabytes, depending on the number of partitions in the partitioned page set. All partitions of a given page set have the same maximum size.
Source: *DB2 Glossary, DB2 UDB for z/OS, Translation Services Center*
2. Definition: On a personal computer hard disk, one of four possible storage areas of variable size; one may be accessed by DOS and each of the others may be assigned to another operating system.
Source: *Dictionary of Printing, iSeries, Translation Services Center*
3. Definition: A subset of the active cluster nodes that result from a network failure. Members of a partition maintain connectivity with each other.
Source: *iSeries, Translation Services Center*
4. Definition: A logical division of storage on a fixed disk.
Source: *AIX Glossary, Dictionary of Printing, iSeries, Translation Services Center, Tivoli Product Terminology, System p5, eServer p5 and i5 and OpenPower servers, x430 and NUMA-Q Glossary*
Comment: Partitions make it easier to organize information. Each partition can be formatted for a different file system. A partition must be completely contained on one physical disk, and the partition table in the Master Boot Record for a physical disk can contain up to four entries.
5. Definition: In VSE, a division of the virtual address area that is available for program execution.
Source: *CICS Glossary, Dictionary of Printing, Translation Services Center*
Comment: CICS/VSE runs in a VSE/ESA region, usually referred to as the CICS region.
6. Subject Area: Hardware \ Printers
Definition: In FD:OCA, a conceptual subdivision of a string of data fields. A partition can be further divided into subpartitions.
Source: *Dictionary of Printing, Translation Services Center*
Related Terms: [partitioning](#)

Ter·mo·lo·gic

Only one of multiple senses is required for CA.



Concept ID 2319	
Term ID 2319-1	
English	
partition	
Usage status	Constrained
Part of speech	Noun
Usage note	Use "swimlane" in the context of flow charts.
Term ID 2319-2	
English	
swimlane	
Usage status	Preferred
Part of speech	Noun

Ter·mo·lo·gic

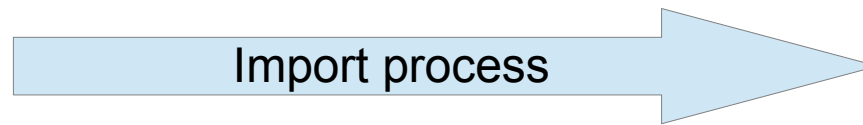


Concept ID 2320	
Term ID 2320-1	
English	
readme	
Usage status	Rejected
Part of speech	Noun
Term ID 2320-2	
English	
readme file	
Usage status	Preferred
Part of speech	Noun

The part-of-speech data category is important for disambiguating homographs.

Concept ID 2321	
Term ID 2321-1	
English	
readme	
Usage status	Rejected
Part of speech	Adjective
Usage note	Do not use "readme" as a modifier. Use only in conjunction with a head noun, such as "readme file."

Automatic pos assignment of the CA tool



Termbase	CA tool
1. (n) readme file (preferred) readme (prohibited)	1. (n) readme file (preferred) readme (prohibited)
2. (adj) readme (<i>as in readme file</i>)	2. (n) readme

conflict !

What should the CA tool do when it comes across “readme” in the noun position?

Spell checking

- Internal dictionaries
- Company-specific dictionaries
- Acronyms

Conclusion

- Extensive testing is required, using texts seeded with the problematic terms
 - in various syntactic positions
 - alone and in compounds
- Simply importing an existing termbase into a CA tool causes more damage than good
 - Only a fraction of the total existing entries are useful.