Ter·mo·lo·gic

# 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

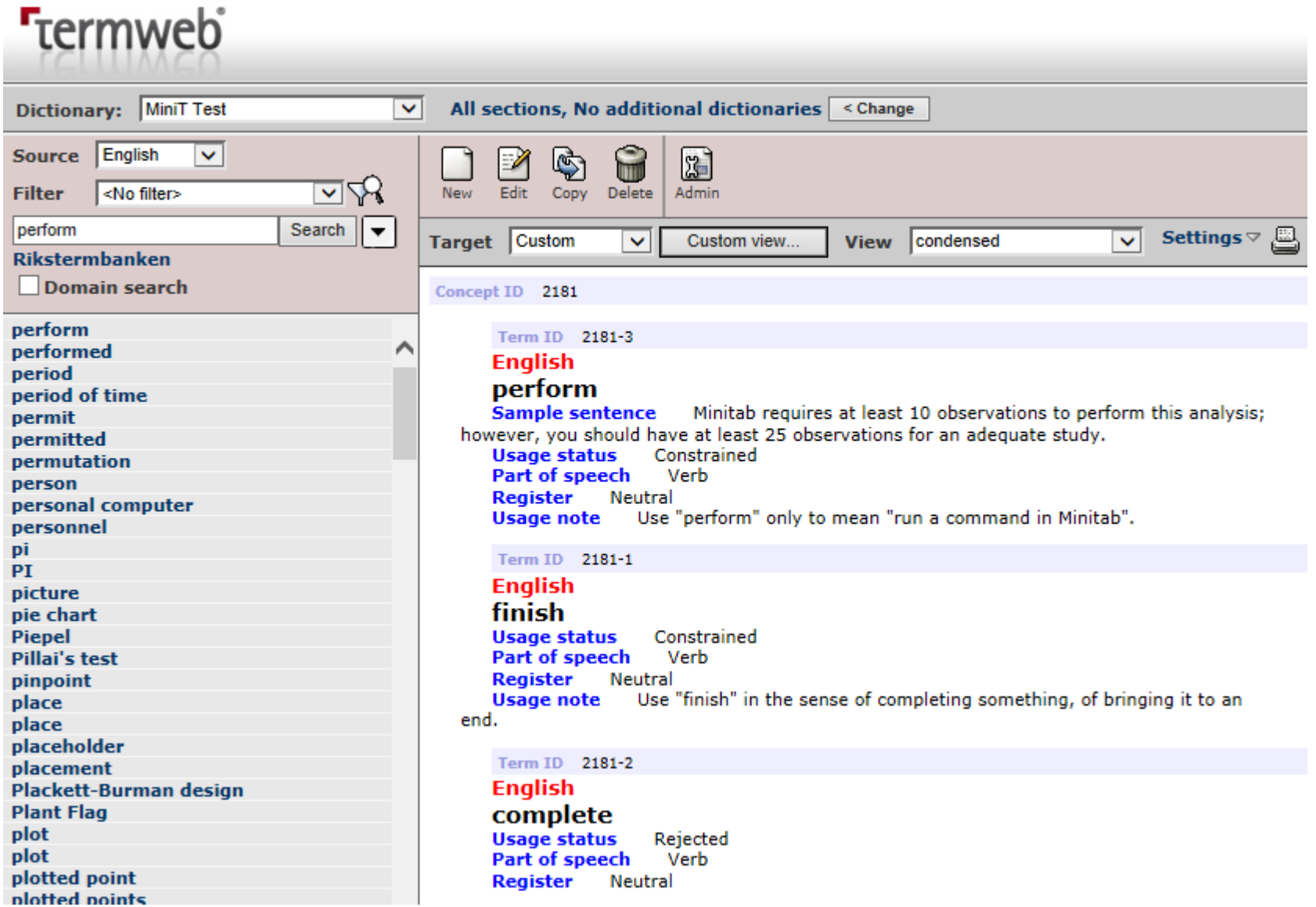
- Minimize synonyms and variants
- Minimize homonyms
- Manage abbreviations
- Protect non-translatable text

# Lexical resources for CA

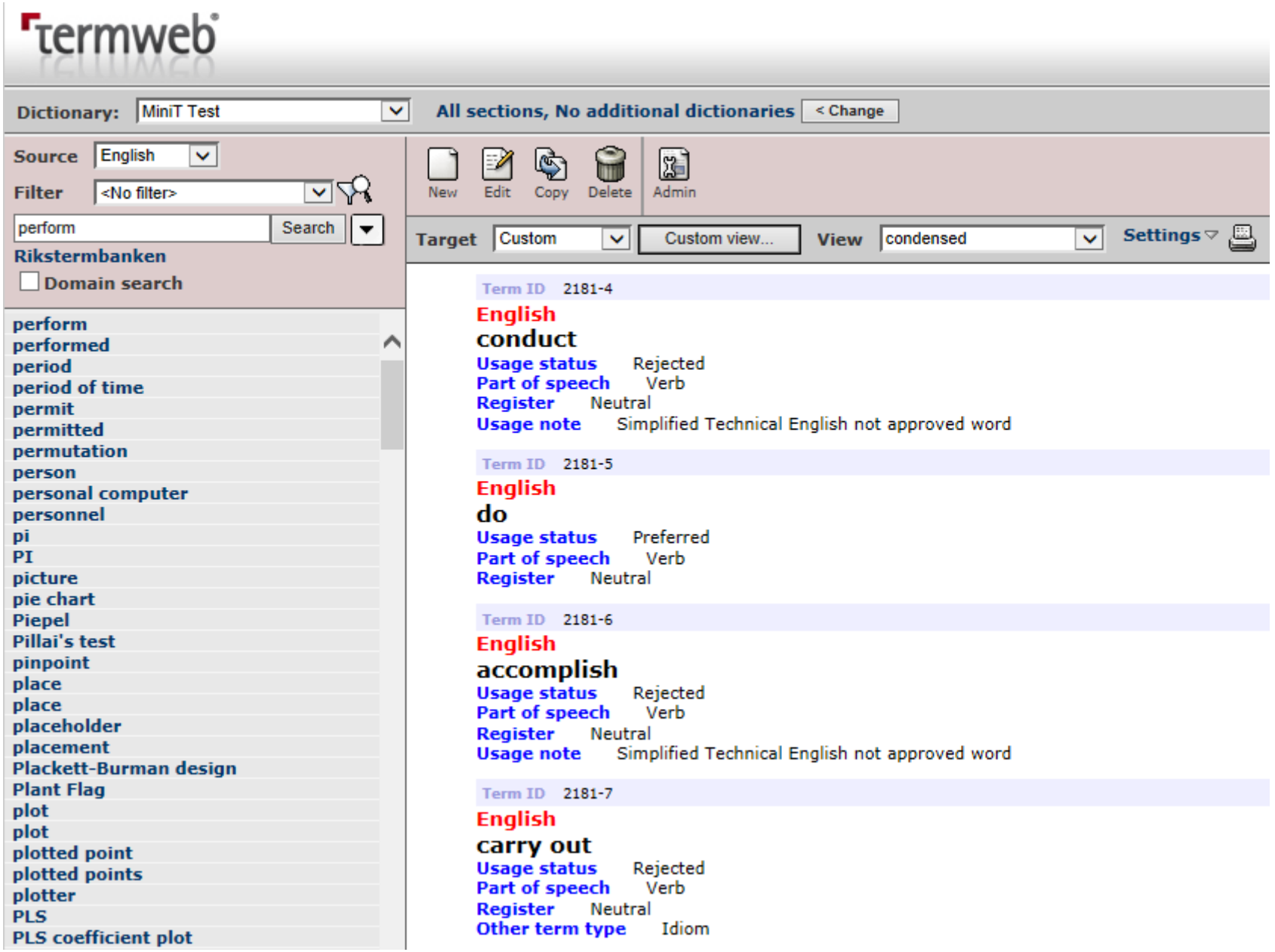
- Differ from “terminology”
  - include general lexicon expressions
  - more non-nominal units
- Are not “lexicographical”
  - Require the concept-oriented approach to manage synsets
  - ISO standards: 16642, 30042.

# 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


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
plotted points
plotter
PLS
PLS coefficient plot
New
Edit
Copy
Delete
Admin
Target Custom
Custom view...
View condensed
Settings
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."

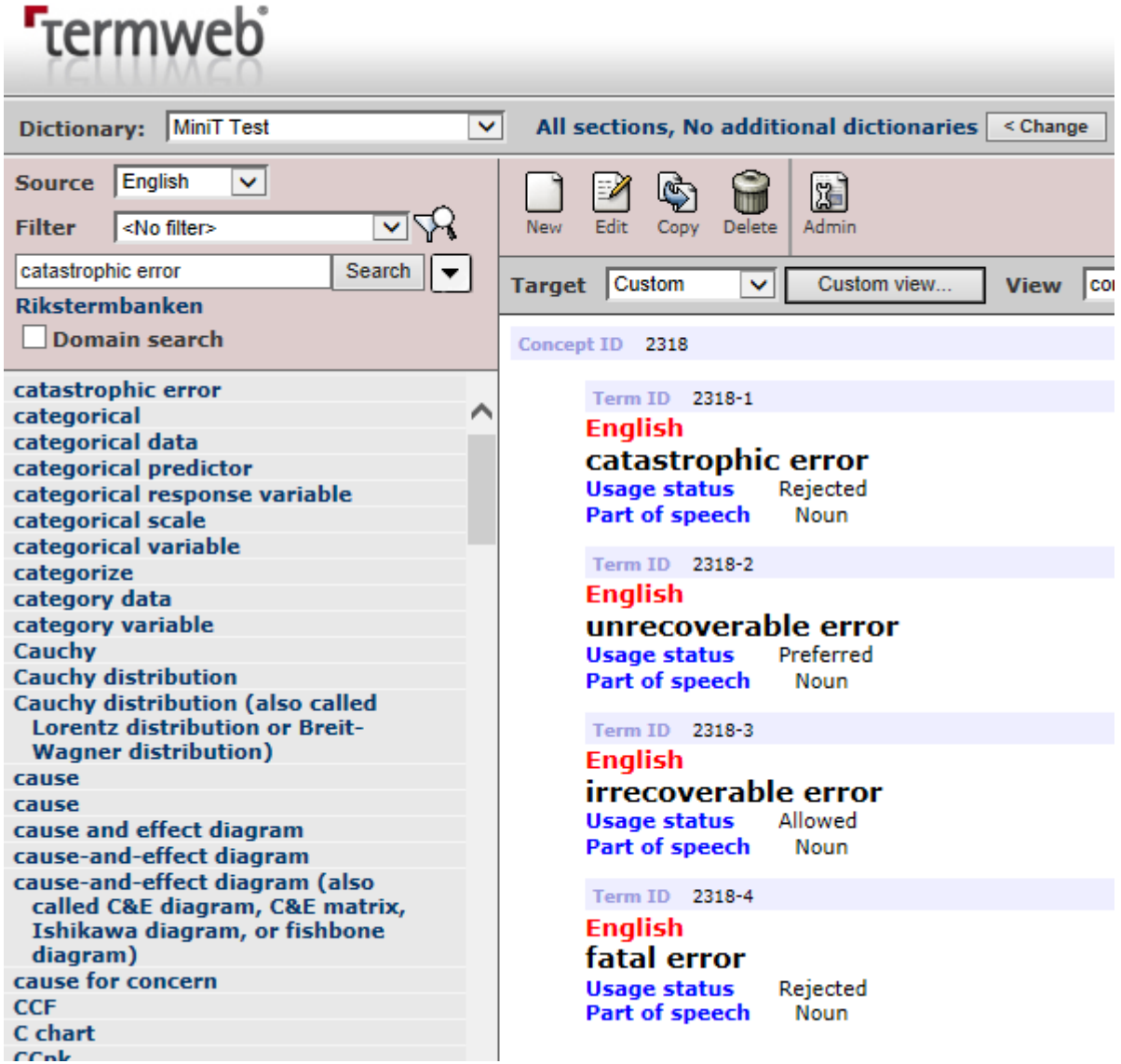
termweb

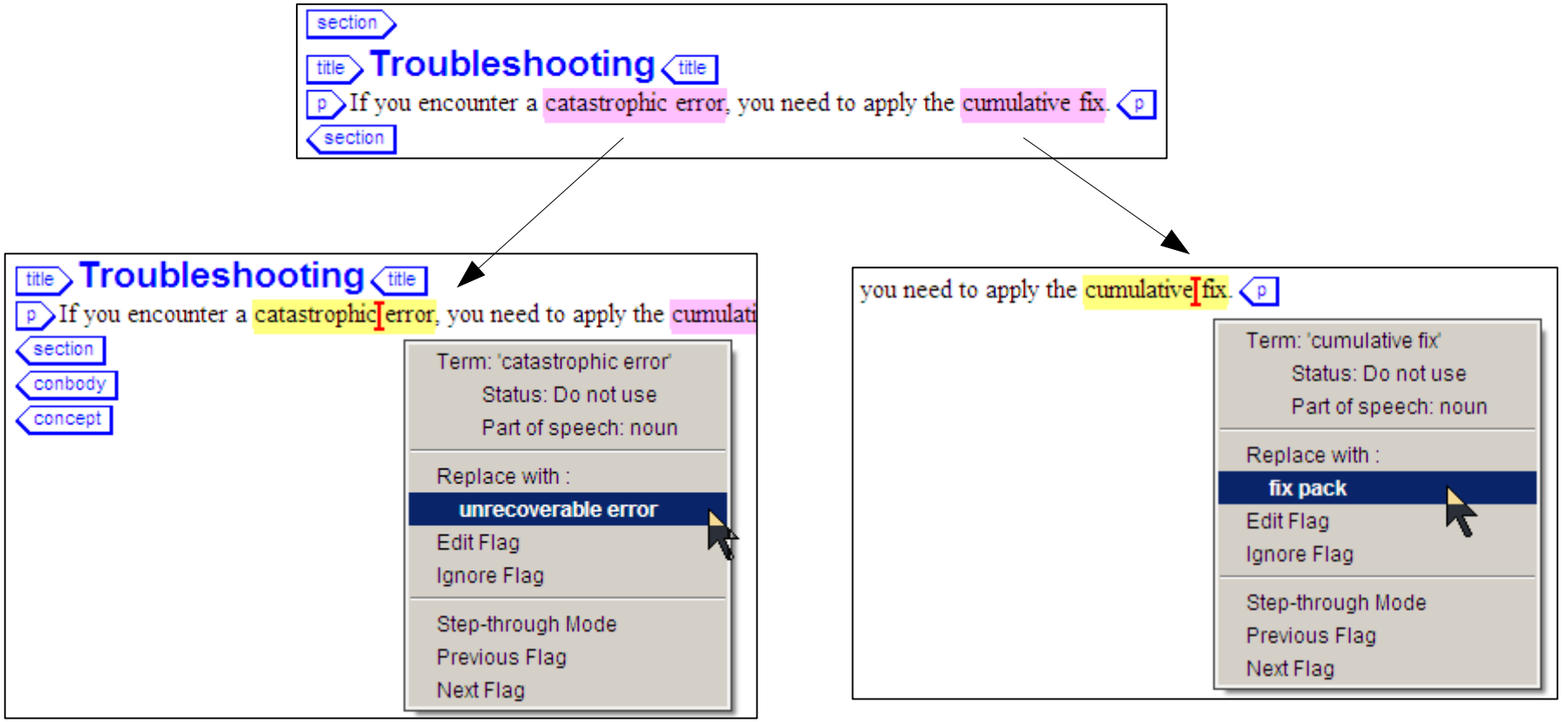
Concept ID 2317

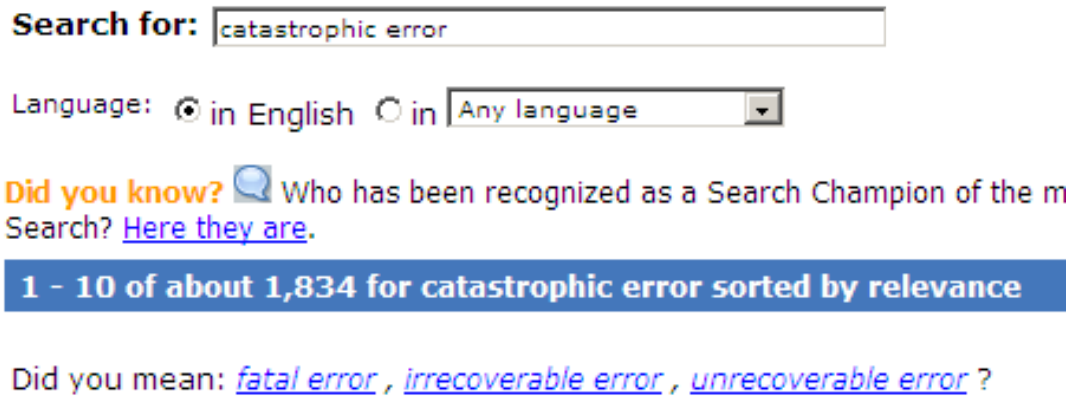
Term ID 2317-1
English
**dump file**
Usage status Preferred
Part of speech Noun

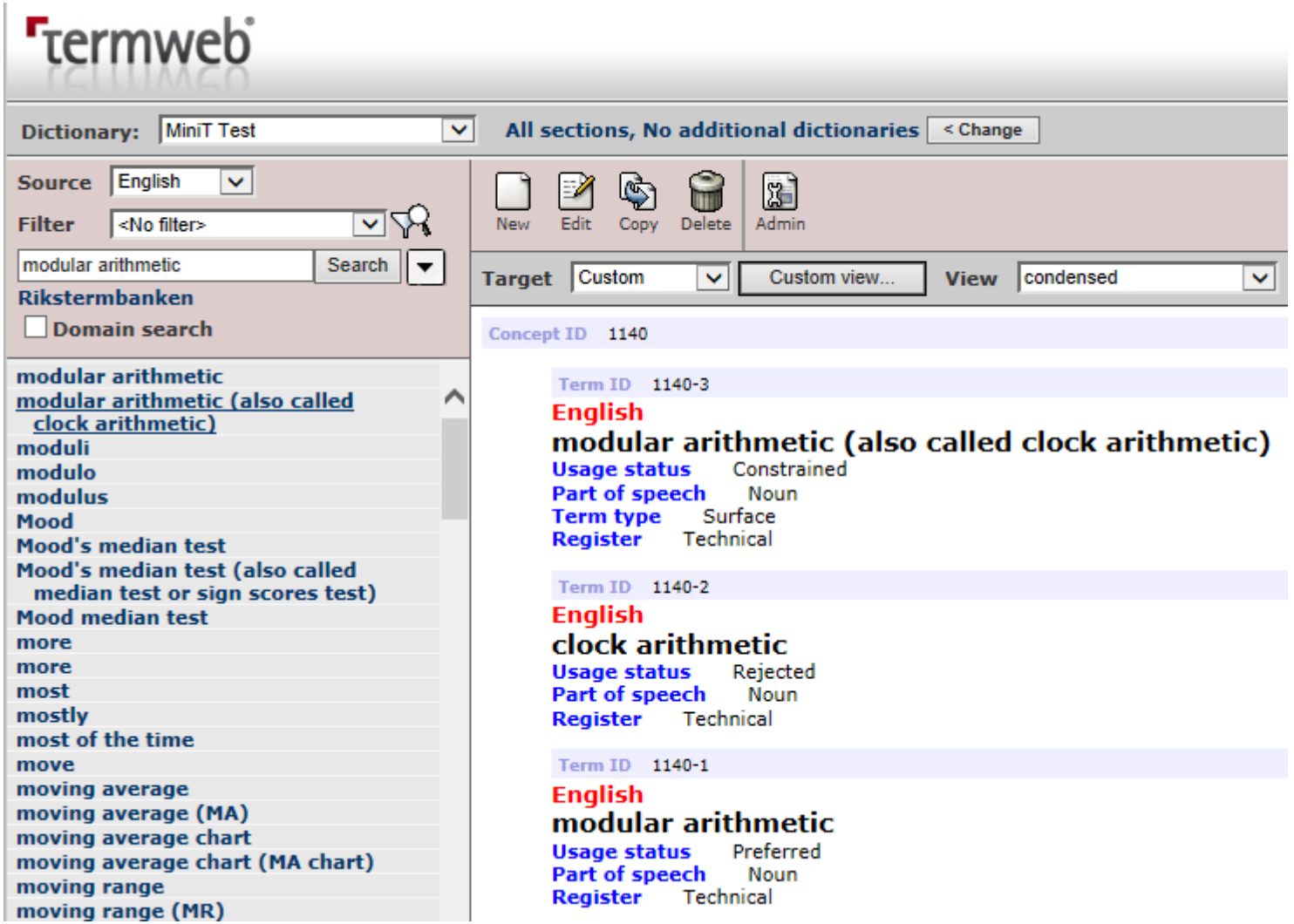
Term ID 2317-2
English
**dumpfile**
Usage status Rejected
Part of speech Noun

# Ter·mo·lo·gic









```
<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

Only one of multiple senses is required for CA.

termweb

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

termweb

Concept ID 2320

Term ID 2320-1
English
**readme**
Usage status Rejected
Part of speech Noun

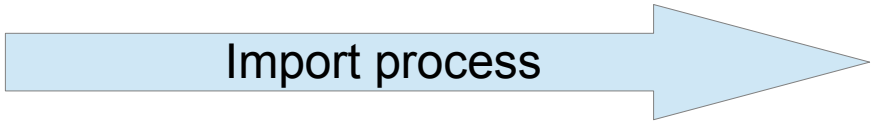
Term ID 2320-2
English
**readme file**
Usage status Preferred
Part of speech Noun

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)<br>readme file (preferred)<br>readme (prohibited) | 1. (n)<br>readme file (preferred)<br>readme (prohibited) |
| 2. (adj)<br>readme *(as in readme file)* | 2. **(n)**<br>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.