# Unknown Word Handling
# Olso Summit

Francis **Bond**
and many, many more

Computational Linguistics Lab (凡土研)
Linguistics and Multilingual Studies,
Nanyang Technological University

<bond@ieee.org>

DELPH-IN 2017-08-07

# Overview

- Unknown words are handled differently in different tools
- Some of the ingredients are
  - grammar external POS (and probs)
  - some reg-exp in chart-mapping
  - lemmatization (with inflectional-rules)
  - creation of nonce- lex-entries for generation
  - generic lexical entries
- Sometimes projects handle this off-line
  - Batch parse, find all unknowns, create default lexicons
    - ★ allows parsing and generation
    - ★ fixes names for post-processing
    - ★ easy to then shift/correct into main lexicon
    - X requires at least two parsing runs

# Desiderata for unknown word handling

- uniform across tools
- can generate from the parse output
- predicate names follow MRS convention
  `_lemma_x_sense`
  `_tokenizations/NNS_u_unknown_rel` →
  `_tokenization_n_unk-NN:+PL`
  `_sapped/VBD_u_unknown_rel` →
  `_sa[p|pp|ppe]_v_unk-VBD-sapped`
- exploits existing inflectional machinery
- fits in stochastic model
- documented for end-users
- documented for grammarians
  with some coordination when tools change
- documented for developers

# Standard Process

1. one tool does something new with the ERG
   and sometime announces it
2. this filters across to other grammars slowly (after some breakage)
3. other tools implement it slightly differently call it something else and we fight about the name
4. we emerge with a better system with new capabilities
5. return to 1