Sharing Language Resources
and linking them in the Open Multilingual Wordnet

Francis Bond
and many, many more

Computational Linguistics Lab
Linguistics and Multilingual Studies,
Digital Humanities Research Cluster
Nanyang Technological University

<bond@ieee.org>

NTU Data Sharing 2019-03-06
Background

- I worked at NTT in Japan on Machine Translation
  - Language Resources were key to quality: lexicons, corpora, grammars
    - All the commercial systems were separately building the same resources

- I left to join the National Institute for Information and Communications Technology (NICT)
  - We built open language resources: wordnets, grammars, corpora
  - ALAGIN: Advanced LAnGuage INformation forum
    - Still hard to cooperate outside of NICT

- I came to NTU to work on Natural Language Understanding
  - Great collaboration within and without NTU
    - funding short term and unpredictable

- We developed informal inter-institute groups with open resources
  - Deep Linguistic Processing with HPSG Initiative (DELPH-IN)
  - Global Wordnet Association (GWA)
    - Data mainly hosted at institute sites, issues when people move

- Minor contributor to MULE (Multilingual Emacs)
Why Open? — Not just the warm inner glow I

- Research should be open in principle:
  
  ...the principle of openness in research - the principle of freedom of access by all interested persons to the underlying data, to the processes, and to the final results of research - is one of overriding importance.

- Making resources Open Source removes difficulties in distributing work or in continuing work at another institution.

- Computational Linguistics needs serious resources
  
  ▶ They cannot be built and maintained by a single group
  ▶ Open Source is the only practical way of achieving flexible multi-group collaboration

- CL needs standards and historically the successful ones have been created bottom-up.
Why Open? — Not just the warm inner glow II

- Seeing one’s work used by other groups is very rewarding.
- People are generally enthusiastic about contributing to widely used work.
- Researchers are evaluated by the impact that their work has. Open Source work generally has more impact.
- Licensing attempts are extremely time-consuming.
- In my personal opinion, Restricting access to information about human language, the great achievement of our species, is immoral.
- People use language resources for many things beyond research: teaching, learning, playing, …
Most researchers very forward about sharing their data
  - Will share on request
  - Can often download (even for ELRA licensed Wordnets)!
  - Data occasionally redistributed under different license
  - Lack of metadata causes problems

Some reasons given for opening data
  - Desire to make language resources publicly available
  - To encourage use and avoid license issuing overhead

Some reasons given for not opening data
  - Financial — get money to support further construction
  - Quality Control — want to stop low quality forks
Effects of different licenses

<table>
<thead>
<tr>
<th>Size</th>
<th>Date</th>
<th>Open</th>
<th>Free</th>
<th>Non free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>2009</td>
<td>Danish/Thai</td>
<td>8/4</td>
<td>Korean</td>
</tr>
<tr>
<td>Large</td>
<td>2008</td>
<td>Japanese</td>
<td>Dutch</td>
<td>24</td>
</tr>
<tr>
<td>Small</td>
<td>2008</td>
<td>French</td>
<td>Slovenian</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bulgarian</td>
<td>13</td>
</tr>
</tbody>
</table>

Citations for wordnets release under different licenses (Bond and Paik, 2012)

*Uptake of a resource partially depends on how usable (legally accessible) the resource is. This has been shown across many fields*
What is a free license?

www.opensource.org/docs/definition.php

The Open Source Definition

1. Free Redistribution
2. Source Code
3. Derived Works
4. Integrity of The Author’s Source Code
5. No Discrimination Against Persons or Groups
6. No Discrimination Against Fields of Endeavor
7. Distribution of License
8. License Must Not Be Specific to a Product
9. License Must Not Restrict Other Software
10. License Must Be Technology-Neutral
WordNet

- WordNet is an open-source electronic lexical database of English, developed at Princeton University
  
  http://wordnet.princeton.edu/

- Made up of four separate semantic nets, for each of nouns, verbs, adjectives and adverbs

- WordNets exist for many languages, at LMS we work on:
  - Japanese
  - Bahasa Malay/Indonesian
  - Chinese
  - The shared **Open Multi-lingual Wordnet** (150+ languages)
    
    http://compling.hss.ntu.edu.sg/omw/
Wordnet Structure

- Lexical items are categorised into ~115K (and counting) glossed synsets (= synonym sets)
- Lexical relations at either the synset level or sense (= combination of lexical item and synset) level
- Strongly lexicalist (originally):
  - synsets only where words exist
  - but many multiword expressions (≈ 50%)
The synset for *Harimau* “tiger”

- **kucing besar 'big cat'**
  - hyponym

- **large feline of forests in most of Asia having a tawny coat with black stripes; endangered**
  - 02129604-n

- **eng: tiger, panthera tigris**
- **zsm: harimau, pak belang**
- **ind: harimau, macan**

- **harimau benggala 'bengal tiger'**
  - hyponym

- **anak harimau 'tiger cub'**
  - hyponym

- **NOLEX 'tigress'**
  - hyponym

- **Panthera 'genus Panthera'**
  - member

- **Feline**
  - SUMO
Wordnets in the world 2008-06

Many wordnets, but few free: we decided to build on open wordnet for Japanese

Green is free; Blue is research only; Brown costs money
Wordnets in the world 2011-06

Free wordnets for French, Catalan, Polish and Japanese

Green is free; Blue is research only; Brown costs money
Wordnets in the world 2012-06

Spanish freed (and Galician and Basque), Farsi, Norwegian, Swedish, Bahasa (Malay and Indonesian)

Green is free; Blue is research only; Brown costs money
Chinese Open Wordnet

Green is free; Blue is research only; Brown costs money
Wordnets in the world 2014-06

Swedish

Green is free; Blue is research only; Brown costs money
Wordnets in the world 2016-01

Dutch, Icelandic, Lithuanian, Romanian, ...

Green is free; Blue is research only; Brown costs money
Wordnets in the world 2017-12

German, Xhosa, Isizulu, Myanmar...

Green is free; Blue is research only; Brown costs money
Wordnets in the world 2019-02

Turkish, Urdu …

Green is free; Blue is research only; Brown costs money
• Every new wordnet makes the network much richer $n(n - 1)$ lexicons!
• Multilingual disambiguation makes it easier to add new languages
• New languages add new phenomena
• How did we do it?
  ▶ Leading by example (we made open wordnets for Japanese, Bahasa, Chinese)
  ▶ Appeal to self-interest: open resources are cited more (Bond & Paik 2016)
  ▶ Simple format for sharing (tsv) — I wrote many converters
    Many checks for ill-formed lexicons
  ▶ Public praise for freed resources
  ▶ Private persuasion for non-open resources
  ▶ Open website
    ★ online interface
    ★ download in multiple formats
    ★ linking to other resources
  ▶ Used by other projects: Google Translate; Natural Language Toolkit (NLTK)
```python
>> import nltk

>>> from nltk.corpus import wordnet as wn

>>> wn.synsets('dog')
[Synset('dog.n.01'), Synset('frump.n.01'),
 Synset('dog.n.03'), Synset('cad.n.01'),
 Synset('frank.n.02'), Synset('pawl.n.01'),
 Synset('andiron.n.01'), Synset('chase.v.01')]

>>> wn.synsets('dog')[0].definition()
'a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds'
```
>>> wn.synsets('dog')[0].lemmas(lang='jpn')
    [Lemma('dog.n.01.イヌ'), Lemma('dog.n.01.ドッグ'),
     Lemma('dog.n.01.洋犬'), Lemma('dog.n.01.犬')]

>>> wn.synsets('dog')[0].lemmas(lang='zsm')
    [Lemma('dog.n.01.anjing')]

>>> wn.synsets('dog')[0].hypernym_paths()[0]
    [Synset('entity.n.01'), Synset('physical_entity.n.01'),
     Synset('object.n.01'), Synset('whole.n.02'),
     Synset('living_thing.n.01'), Synset('organism.n.01'),
     Synset('animal.n.01'), Synset('chordate.n.01'),
     Synset('vertebrate.n.01'), Synset('mammal.n.01'),
     Synset('placental.n.01'), Synset('carnivore.n.01'),
     Synset('canine.n.02'), Synset('dog.n.01')]
Extending Wordnets

- Original wordnet only has Nouns, Verbs, Adjectives, Adverbs
  - Added pronouns (*me*), classifiers (個), quantifiers (*some*), greetings/exclamations (*Hello, Thank you*)
  - Adding modals (*must*), bound morphemes (*re-, un-*)

- Tagging Corpora
  how is this word used in practice

- Many new languages
- Richer representation of idioms
- Internet slang
Where do we go from here

- The interlingual index
  - 御飯 gohan “cooked rice”, 米 kome “raw rice”

- Better disambiguation

- Sub-word meaning
  - Everybody

- More languages
Why we can’t include non-commercial licensed wordnets

- The NC licenses may not permit some uses of your work that you would like others to make. For example, not all educational uses are necessarily NonCommercial uses, so your use of an NC license may preclude use of your work in some educational contexts.
- The NC licenses are not compatible for remixing with many works. You may not remix BY-SA content (such as Wikipedia content or many of the existing wordnets) with BY-NC content.
- NC licenses do not qualify as open licenses under the Open Definition, and works licensed under an NC license are not considered Free Cultural Works. This may be important if you want others to further distribute your work on Wikipedia, Wikimedia Commons, or other platforms requiring a license that meets the Open Definition or the Definition of Free Cultural Works.

Adapted from Creative Commons
How do we share data?

- Lab or individual web sites
  - University sites like DR-NTU
- LDC, ELRA (proprietary organizations)
- Conferences often include data with proceedings (CDROMs, USB)
- CLARIN, ALAGIN, META-NET (government sponsored organizations)
  - GWA, DELPH-IN (informal organizations)
- sourceforge, github, wikimedia (open-source infrastructure)
  - university hosted version control systems (LOGON, Oslo)
- increasing emphasis on documentation, metadata and LOD (interoperability)
- data is dynamic, we often improve our understanding and re-annotate
  - the Redwoods treebank is now in its 9th major release