CILI: the Collaborative Interlingual Index

Francis Bond♠ Piek Vossen♢ John P. McCrae♣ Christiane Fellbaum♡

♠ Nanyang Technological University, Singapore
♢ VU University Amsterdam, The Netherlands
♣ Insight Centre for Data Analytics, NUI Galway, Galway, Ireland
♡ Princeton University, U.S.A.

bond@ieee.org,
piek.vossen@vu.nl,john@mccr.iae,fellbaum@princeton.edu

2016-01-27
Why the Collaborative ILI?

There are wordnets for many languages
  - Currently they link through PWN (3.0)

Many projects are adding new synsets
  - And not just synsets: lemmas, relations, POS, meta-data (domains, sentiment ...)

We want to be able to link them even if they are not in PWN

We want to minimize wasted effort
  - Adding the same thing in different projects
  - Fixing the same errors in different projects

We want to spread the burden of development
The solution: an InterLingual Index (ILI)

1. The Interlinear Index (ILI) should be a flat list of concepts (and instances).
2. The semantic and lexical relations should mean the same things for all languages.
3. Concepts should be constructed for salient and frequent lexicalized concepts in all languages.
4. Concepts linked to Multiword units (MWUs) in wordnets should be included.
5. A formal ontology could be linked to but separate from the wordnets.

Basic idea well known (Fellbaum and Vossen, 2008).
The license must allow redistribution of the index

ILI IDs should be persistent: we never delete, only deprecate or supercede; we should not change the meaning of the concept

Each new ILI concept should have a definition in English, as this is the only way we can coordinate across languages. The definition should be unique, which is not currently true, and preferably also parse and sense tag information should be included. Definition changes will be moderated.
Each new ILI concept should link to a synset in an existing project that is part of the GWG with one of a set of known relations (hypernymy, meronomy, antonymy, …)

This synset should link to another synset in an existing project that is part of the GWG and links to an ILI concept.

⇒ each concept is linked to another concept through at least one wordnet in the grid

Any project adding new synsets should first check that they do not already exist in the CILI

- New concepts are added through their existing in a wordnet
- If something fulfills the criteria is proposed
- If no objections after three months then it is added
License

- The CILI itself is licensed under the Creative Commons Attribution (CC BY 4.0)
- Wordnets in the grid should be compatible with CC BY SA
  - CC BY
  - CC BY SA
  - Wordnet
  - dis-preferred: ODC BY, CECILLE, Apache 2.0, MIT
- Ideally the license should be available as a LICENSE file and pointed to within the LMF
  - dc:rights "Copyright GWA; License: CC BY 4.0"
  - cc:license <http://creativecommons.org/licenses/by/4.0>
Persistence

- We want to make it possible for wordnets to interlink even if not all of them have the resources to follow changes.
- We will keep all old concepts (from PWN30 on).
- Concepts can be deprecated (no interface for this).
- Concepts can be superseded:
  - e.g. `<ili82222> schema:supercededBy <ili82221>;`  
  `<ili82221> owl:deprecated "true"^^xsd:boolean ;`
  - Effectively this merges the two entries for *never*.
- The open multilingual wordnet will try to deal with these intelligently.
Definitions

- Each new ILI concept should have a definition in English
  - necessary to coordinate across languages
  - **NOT** imposed on individual wordnets

- The definition should be unique
  - Not true for about 2,000 in PWN 3.0
  - We have fixed all but a handful
    - “a variety of aster ⟨Eurybia radula⟩”
    - “one of the Gorgons” → “the eldest Gorgon”
    - some are merge candidates

- Changes to PWN definitions will need to be coordinated with the gloss corpus

- Changes to definitions can change the meaning of the concept
  this should be avoided

More on these from Christiane
Fixing Definitions or Synsets

- **phylum Protozoa** “in some classifications considered a superphylum or a subkingdom; comprises flagellates; ciliates; sporozoans; amoebas; foraminifers”
  → “in some classifications considered a superphylum or a subkingdom; comprises flagellates, ciliates, sporozoans, amoebas, foraminifers”

- **unite, unify** “bring together for a common purpose or action or ideology or in a shared situation —the Democratic Party platform united several splinter groups”

- **rule in, rule out** “include or exclude by determining judicially or in agreement with rules”
Changing Definitions

04303258-n  staple – a short U-shaped wire nail for securing cables

10168  The chimney is wide, but is barred up by four large staples.

07844042-n  milk – a white nutritious liquid secreted by mammals and used as food by human beings

10403  Well, a cheetah is just a big cat, and yet a saucer of milk does not go very far in satisfying its wants, I daresay.

02930766-n  cab – a car vehicle driven by a person whose job is to take passengers where they want to go in exchange for money

11588  A cab had driven up whilst the American had been talking.

Add typically or characteristically; change cab to vehicle (it is horse-drawn here)?
All new concepts should be linked to an existing concept through one of the wordnets

Should we allow all kinds of links?
  ▶ hypernym, meronym, antonym\textsubscript{sense}
  ? see also, fuzzynym, near synonym, role, involved, manner, ...
  ?? derivation\textsubscript{sense}, pertainym\textsubscript{sense}...
Should we merge some derivational links?

- quick, quickly
- today$_n$, today$_v$
The details

- ILI as RDF — shared on github
- Each new wordnet release can suggest candidates
  - Mark as ili='ni' “new ILI”
  - Only projects can add new ILI entries
- Review period (1–3 months)? — accept if no comments
- ili.ttl hosted in github
  - URLs hosted at VU, proxied through GWA
  - OMW and merging hosted at NTU, proxied through GWA
    http://compling.hss.ntu.edu.sg/omw/
    show provenance and confidence
  - Individual wordnets can be uploaded
- ILI file

Bond, Vossen, McCrae and Fellbaum
Grandfathering

- Start off with Princeton Wordnet 3.0 Synsets even if they have
  - Non-unique definitions almost fixed
  - Ill-formed definitions (‘;’, ‘,’), too short; short definitions need to be revised very hard for e.g. quantifiers
  - Not linked to another synset (orphans) any volunteers?
Where do we go from here?

- Convert your wordnet to LMF
  - Validate it
    Possibly revise the LMF as needed
- Upload it to the Grid
  - We will validate it again
  - We add the wordnet to OMW: linking through ILI
  - Look for ILI='in' “ili new”
  - Check the definition is good (and not too close)
  - Check it is linked to something existing
- Finish New OMW interface
- Add an interface for changing definitions, deprecating and superseding
Where do we go from here?

- New POS: ’x’, ’q’, ’c’, ’p’, ...
- Orthographic variant explicitly modeled
- More languages, links, words
- Sense tagged corpora and live examples
- More sense groupings
- Decompositional semantics: un-clasp, today “this day”
- External DBS: species, geo-wordnet, images, ...
Not sili at all.

CILI
the hot new ILI that is also cool!