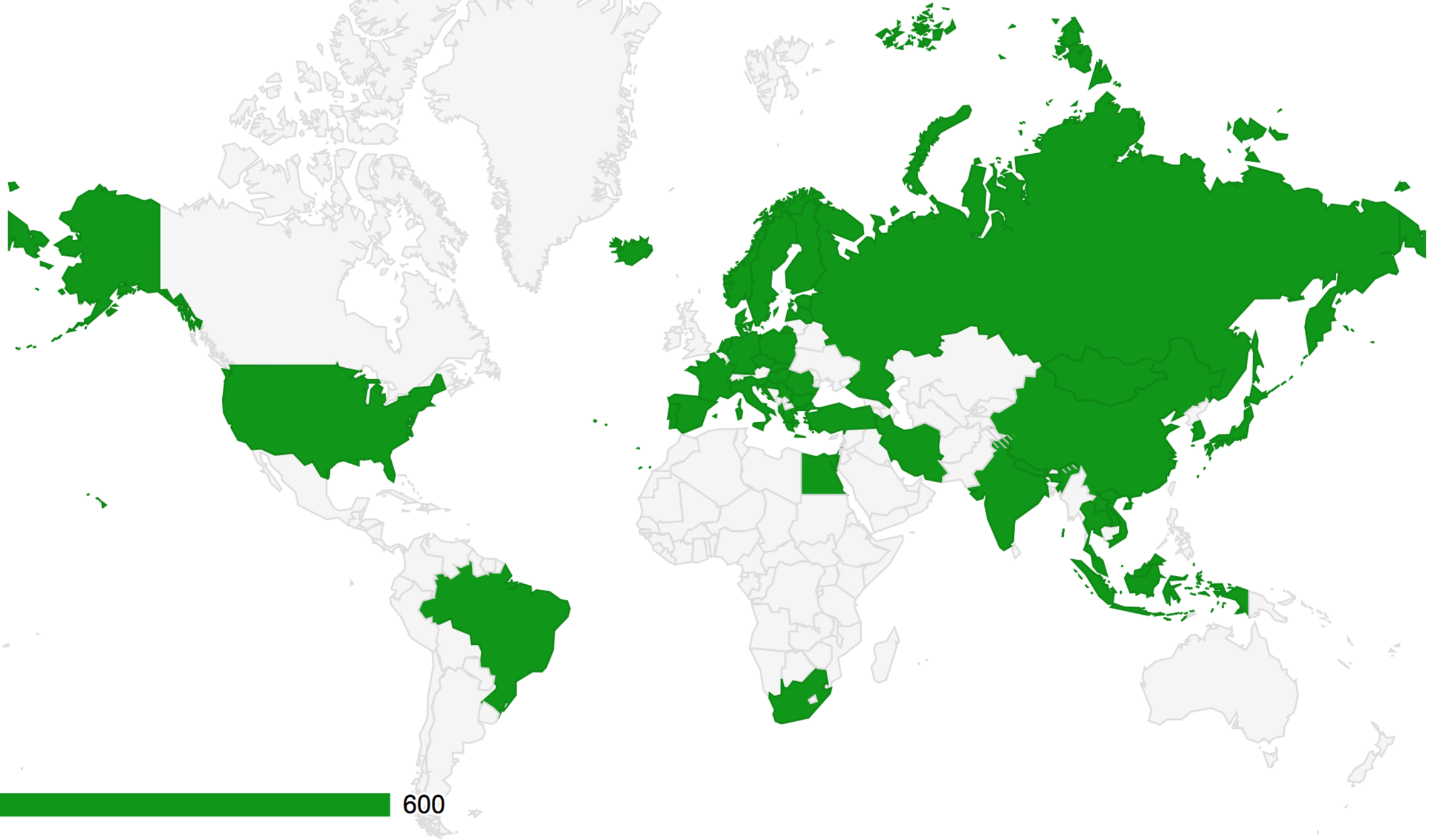
A Collaborative Interlingual Index for harmonizing word nets

Piek Vossen
VU University Amsterdam

25 years of wordnets

- Princeton WordNet since 1990 (1.5, 1.6, 2.0, 3.0)
- EuroWordNet
- Multiwordnet, BalkaNet
- Indowordnet, Asian wordnet, African wordnet
- Open Multilingual Wordnet
- BabelNet, Wiktionary

Global Wordnet Map



Bond & Paik (2012)

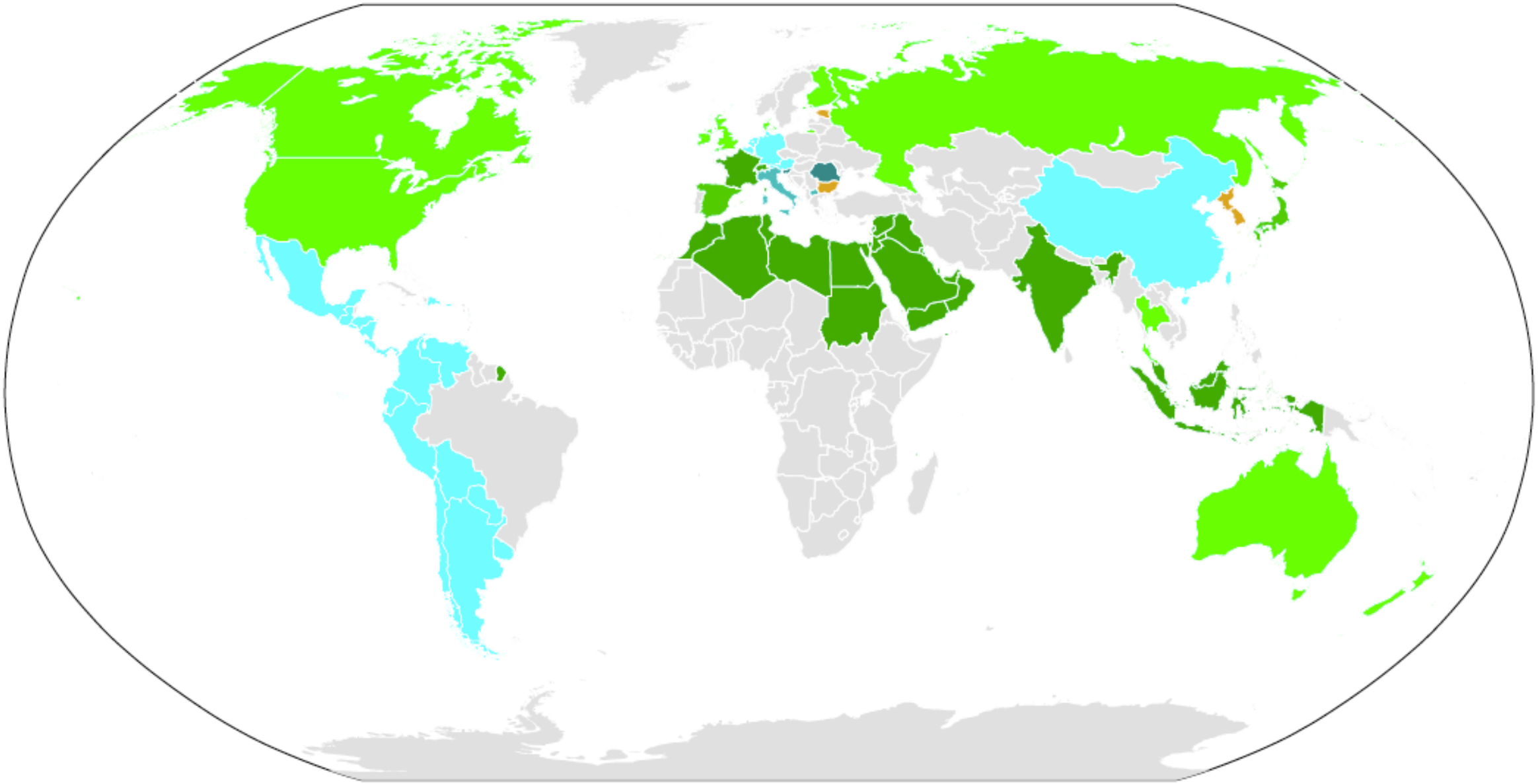


Figure 1: Map of Countries showing WordNet availability

Countries with **open source wordnets** are in green; **free for research** wordnets are in blue; **non free** wordnets are in brown. The lighter the color, the more synsets.
Citation counts from Google Scholar (accessed on 2011-09-23)

EuroWordNet design

- Each wordnet is structured according to the Princeton model: synsets & semantic relations between them
- Synsets have equivalence relations to the Inter-Lingual-Index (ILI) = fund of concepts provided by Princeton (IndoWordnet uses Hindi as ILI)
- Many different equivalence relations are allowed that mimic the wordnet-relations
- No semantic relations imposed on the ILI, relations are expressed through each wordnet -> obtain any structuring as desired

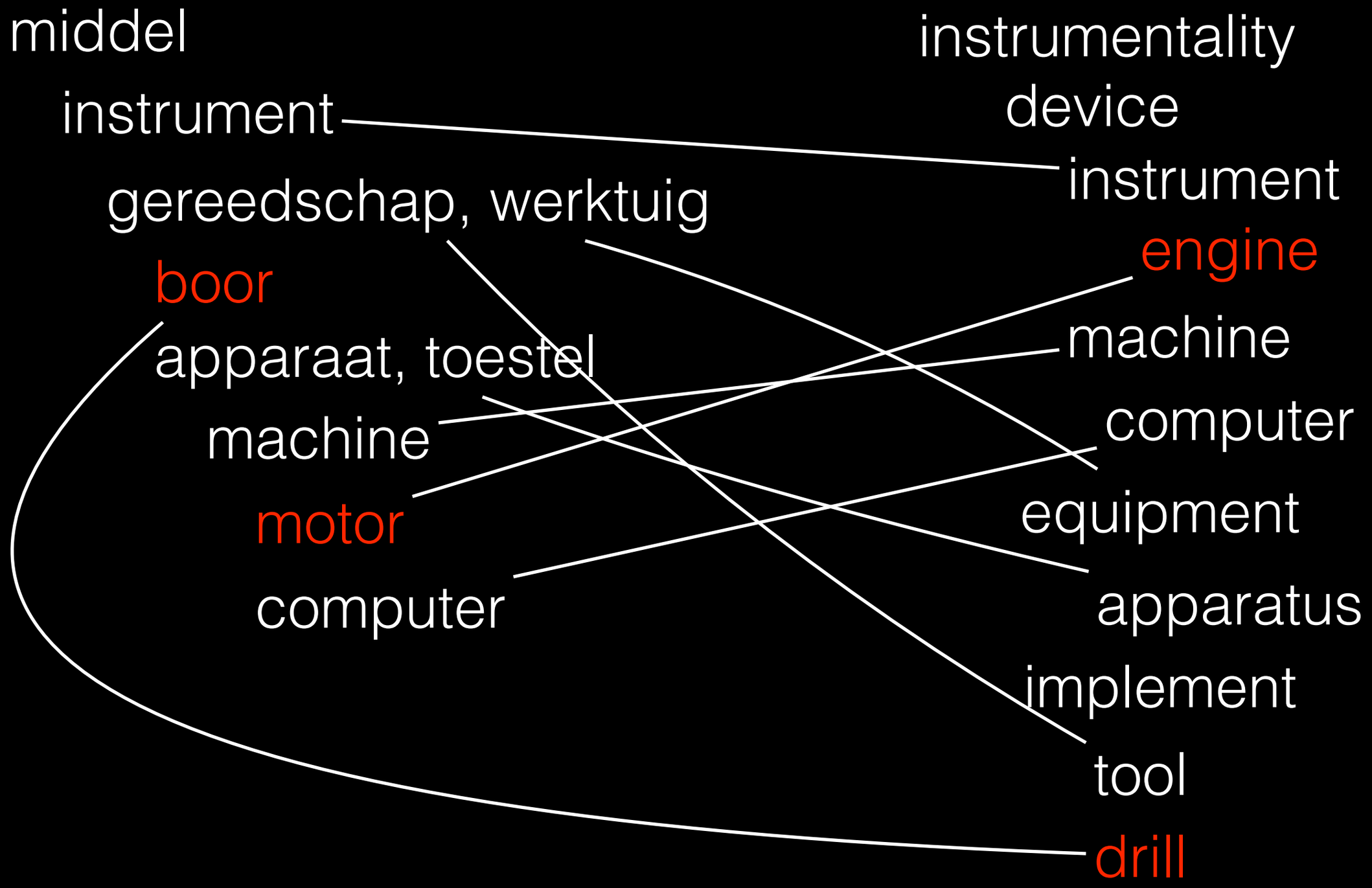
Francis Bond and Kyonghee Paik (2012) A survey of wordnets and their licenses. GWC 2012. Matsue. 64–71

MERGE

Name	Language	# Synsets	Release	License	Citation	Count
Open Source						
Princeton WN*€	English	155,000	1991	WordNet	Fellbaum (1998)	6,821
FinnWordNet	Finnish	117,700	2010	WordNet	Lindén and Carlson. (2010)	0
Russian WN	Russian	117,000	2004	Wordnet	Balkova et al. (2008)	15
Thai Wordnet	Thai	73,593	2007	WordNet	Thoongsup et al. (2009)	4
DanNet*	Danish	65,000	2008	WordNet	Pedersen et al. (2009)	8
Japanese WN*	Japanese	57,000	2009	WordNet	Isahara et al. (2008)	24
Catalan WN*	Catalan	42,000	1999	GPL	Benítez et al. (1998)	17
LSG	Irish Gaelic	32,742	?	GNU FDL	`http://borel.slu.edu/lsg/`	—
Hindi WN	Hindi	28,687	?	GNU FDL	Jha et al. (2001)	10
WOLF	French	22,000	2009	Cecill-C†	Sagot and Fišer (2008)	22
Wordnet Bahasa*	Malay, Indonesian	20,000	2011	MIT	Nurril Hirfana et al. (2011)	—
Spanish WN*⊙€	Spanish	15,556	2006	LGPL	Farreres et al. (1998)	65
Catalan WN*⊙€	Catalan	15,556	2006	LGPL	Benítez et al. (1998)	17
Arabic WN*	Arabic	11,269	2008	CC BY SA	Black et al. (2006)	28
Hebrew WN*	Hebrew	5000	2006	GPL	Ordan and Wintner (2007)	0
Free for Research						
Chinese WN*	Chinese	115,424	2008	res/com	Xu et al. (2008)	0
KorLex*⊙	Korean	90,000	2007	res/com	Yoon et al. (2009) (nouns)	—
Spanish WN*€	Spanish	62,000	1999	res/com	Farreres et al. (1998)	65
Cornetto*€	Dutch	70,371	2009	res/com	Vossen et al. (2008)	19
GermaNet*€	German	69,594	2011	res/com	Kunze and Lemnitzer (2002)	52
MultiWN*€	Italian	38,877	2008	res/com	Pianta et al. (2002)	143
MWN*	Macedonian	33,276	2010	CC BY NC	Saveski and Trajkovski (2010)	0
Ro-WordNet*	Romanian	30,000	soon	no-deriv.	Tufiş et al. (2008)	9
Czech WN *€	Czech	29,000	1999	res/com	Pala and Smrž (2004)	34
SloWnet*	Slovene	20,000	2010	CC BY NC SA	Fišer and Sagot (2008)	13
Non Free (Available for Research)						
KorLex*	Korean	130,878	2007	res/com	Yoon et al. (2009)	5
Estonian*€	Estonian	47,000	—	ELRA	Kerner et al. (2010)	0
EuroWordNet					Vossen (1998)	728
Dutch	Dutch	44015	1999	ELRA	ELRA-M0016	
Spanish	Spanish	23370	1999	ELRA	ELRA-M0017	
Italian	Italian	48529	1999	ELRA	ELRA-M0018	
German	German	15,132	1999	ELRA	ELRA-M0019	
French	French	22,745	1999	ELRA	ELRA-M0020	
Czech	Czech	22,745	1999	ELRA	ELRA-M0021	
Estonian	Estonian	9,317	1999	ELRA	ELRA-M0022	
ItalWordNet	Italian	49,360	1999	ELRA	ELRA-M0018	
BasqWN	Basque	30,281	?	ELRA	Pociello et al. (2011)	0
BulNet*⊙	Bulgarian	23,715	2004	ELRA	ELRA-M0041 (Koeva, 2008)	3

Table 1: Catalog of WordNets

Merging wordnets



Status of the global net

- Wordnets built using different methods: merge or expand, manual or (semi-)automatic
- Different sets of relations were used
- Different interpretations of relations (with the same name)
- Different ways of defining synonyms (strict, loose)
- Different degrees of polysemy
- Differences in coverage

Status of the global net

- Linked to different versions of the English WordNet
- Released in different formats
- Using different license schemes
- Fixed Anglo-Saxon ILI: changes through English WordNet
- No central hosting of the ILI

Global Wordnet Grid

- All wordnets linked to a single fund of concepts.
- Merge of concepts in all languages, not depending on English WordNet.
- Available as LOD, one license: CC-BY-SA3.0, CC-BY-SA4.0.
- Adaptable by the wordnet-language community.
- As many wordnets as possible linked to the Grid and available as LOD through the same license.

Motivation for GWG

- From natural language text to concepts: populating the LOD from textual data.
- Understanding any language by machines in a similar way.
- Platform for achieving conceptual interoperability.
- NewsReader case: read text in 4 languages to achieve a single unified RDF representation of changes reported in the news.
  - Thursday, session O23, Hackaton London June 10th!

Why adapt the ILI?

- Better mapping across languages following a merge approach.
- Bypass English gaps to map languages.
- Study universals and idiosyncrasies across languages.
- Share resources across languages: ontologies, domains, sense-tagged corpora.
- Harmonize the semantics of wordnets across languages: -> definition of synonymy and relations

Not lexicalized in English

- Cultural specific concepts
  - $\text{klunen}_{\text{Dutch}}$ = walk on skates over land from one frozen water to another
  - $\text{Udhiyah}_{\text{Arabic}}$ = slaughtering of a lamb during the period of Eid-Aladha
- Pragmatic concepts
  - Gender variants:
    - $\text{Lehrer}_{\text{German}}$ and $\text{Lehrerin}_{\text{German}}$ = teacher

Not lexicalized in English

- Aspect variants in Slavic languages:
  - vypít$_{Czech}$ = to drink up,
  - pít$_{Czech}$ = to be drinking
- Lexical inclusion:
  - hilamos$_{Tagalog}$ = to wash one's face
  - alevín$_{Spanish}$ = small fish
  - bemahlen$_{German}$ = paint something (obligatory obj)
- Compounds:
  - kindermeel$_{Dutch}$ = flour for children
  - tarwemeel$_{Dutch}$ = flour made of oats

Should there be a limit?

- NO! we could even include adjective-noun or verb-object pairs
- What about productivity?
- What about duplicate concepts?
- Productivity and compositionally need to be observed.
- Cross-lingual lexicalization determines value:
  - what is linked works, what is not linked disappears

Bulk import of ILI records

- Open Dutch Wordnet:
  - 95,674 PWN synsets
  - 14,523 new synsets (linked to a PWN parent)
    - taartschep = shovel for tart or pie
    - vorstperiode = period of frost
    - deltavliegen = to fly with a delta wing

```
<cdb_synset c_sy_id="odwn-10-101348407-n" posSpecific="NOUN" comment="eq-parent-match">
    <synonyms>
        <synonym c_lu_id="r_n-36788" c_lu_id-previewtext="taartschep:1" status="cdb2.2_None"/>
    </synonyms>
    <wn_internal_relations>
            <relation relation_name="HAS_HYPERONYM" target="eng-30-04208210-n" source="pwn"/>
            <relation relation_name="CO_INSTRUMENT_PATIENT" target="eng-30-07623933-n" source="own"/>
    </wn_internal_relations>
</cdb_synset>
<cdb_synset c_sy_id="eng-30-04208210-n" posSpecific="NOUN" comment="EQ_SYNONYM">
    <synonyms>
        <synonym c_lu_id="r_n-32893" c_lu_id-previewtext="schep:1" status="cdb2.2_Manual"/> <!-- shovel -->
        <synonym c_lu_id="o_n-102248675" c_lu_id-previewtext="schepje:1" status="cdb2.2_Manual"/>
    </synonyms>
    <definition><![CDATA[a hand tool for lifting loose material; consists of a curved container or scoop and a handle]]></definition>
    <wn_internal_relations>
            <relation relation_name="HAS_HYPERONYM" target="eng-30-03489162-n" source="pwn"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-01578821-v" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-01310660-v" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-01311103-v" source="own"/>
            <relation relation_name="HAS_HYPERONYM" target="eng-30-04451818-n" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-01310660-v" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-01311103-v" source="own"/>
            <relation relation_name="HAS_HYPERONYM" target="eng-30-04451818-n" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-01584701-v" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="odwn-10-103015499-v" source="own"/>
            <relation relation_name="ROLE_INSTRUMENT" target="eng-30-02090990-v" source="own"/>
    </wn_internal_relations>
</cdb_synset>
<cdb_synset c_sy_id="eng-30-07623933-n" posSpecific="NOUN" comment="EQ_SYNONYM">
    <synonyms>
        <synonym c_lu_id="r_n-36785" c_lu_id-previewtext="taart:1" status="cdb2.2_Manual"/> <!-- tart -->
    </synonyms>
    <definition><![CDATA[a small open pie with a fruit filling]]></definition>
    <wn_internal_relations>
            <relation relation_name="HAS_HYPERONYM" target="eng-30-07625493-n" source="pwn"/>
            <relation relation_name="CO_PATIENT_INSTRUMENT" target="odwn-10-101348407-n" source="own"/>
            <relation relation_name="CO_PATIENT_INSTRUMENT" target="odwn-10-101403873-n" source="own"/>
    </wn_internal_relations>
```

Bulk import of concepts

- BabelNet 2.5
  - Merge of WordNet, Open Multilingual WordNet, Wikipedia, OmegaWiki, Wikidata, Wiktionary
- 9.3M synsets, 21,7M definitions
- 7.7M images linked to synsets
- creativecommons.org/licenses/by-nc-sa/3.0/

Why not use an ontology?

- Lack of consensus on ontologies.
- Axiomizing concepts in an ontology is more complex.
- Coverage of ontologies is too low.
- Ontologies can be linked to the ILI as well: we can do both!

Axiomizing concepts
SUMO

```
(=>
  (instance ?UR UdhiyahRitual)
  (exists (?S ?EA ?P)
    (and
      (instance ?EA EidAladha)
      (during ?UR ?EA)
      (attribute ?S Udhiyah)
      (agent ?UR ?P)
      (attribute ?P Muslim)
      (patient ?UR ?S))))
```

```
(=>
  (attribute ?S Udhiyah)
  (exists (?UR)
    (and
      (instance ?S Lamb)
      (instance ?UR UdhiyahRitual
      (patient ?UR ?S)))))
```

What defines a concept?

- Keep it as simple as possible:
  - A unique IRI
  - English gloss
  - **LINKED** to a Synset in at least one language through a sameAs relation
  - **LINKED** to a hypernym synset or being a well-defined top: -> no orphans

Design properties

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:gwa="http://www.globalwordnet.org/" .etc...>

<rdf:Description rdf:about="http://www.globalwordnet.org/ili/version1/ILI-01321770">
    <gwa:gloss gwa:lang="eng" rdf:datatype="http://www.w3.org/2001/XMLSchema#string">a young animal
                                without a mother </gwa:gloss>
    <gwa:glossili gwa:lang="eng" rdf:datatype="http://www.w3.org/2001/XMLSchema#string"> a ILI-00345645 ILI-87687
                                without a ILI-63342258 </gwa:glossili>
     <gwa:gloss gwa:lang="nld" rdf:datatype="http://www.w3.org/2001/XMLSchema#string">een jong dier zonder
                                een moeder</gwa:gloss>
    <gwa:provenance source="WordNet" version="3.0"  lang="eng"/>
    <gwa:status gwa:onionlayer="3" gwa:moderation="required"/>
    <gwa:creationTime/>
    <gwa:lastModifiedTime/>
    <gwa:editHistory/> <!-- provides all details on the provenance of the concept: who made what changes -->
    <gwa:wordnetLinks> <!-- list of wordnet synsets that link to this with the type of link eq_syn or eq_near_syn
        <gwa:hasHypernym rdf:resource="http://www.w3.org/2006/03/wn/wn20/instances/wordsense-animal-noun-1"/>
        <gwa:sameAs rdf:resource="http://www.w3.org/2006/03/wn/wn20/instances/wordsense-orphan-noun-1"/>
    </gwa:wordnetLinks>
    <gwa:ontoLinks/> <!-- list of ontology classes that link to this with the type of link similarAs or subclassO
    <gwa:communityVotes/>
    <rdfs:comment/>
</rdf:Description>
</rdf:RDF>
```

Design properties

- Glosses in other languages optional, English gloss required!
- Special gloss: content words replaced by ILI concepts
- Concept identifiers are unique and never deleted or modified.
- No further semantics is imposed on the ILI.
    - A concept can be linked to any other concept in a wordnet or ontology.
    - The linked wordnet or ontology provides a specific model imposed on the ILI.
- Concepts have NO PART OF SPEECH
- Linking: gwa:sameAs or gwa:similarTo, no other links are allowed

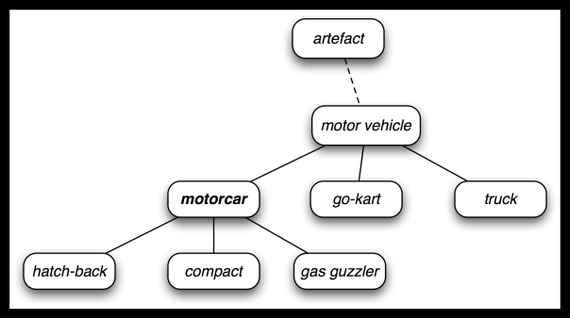
Protocol for changes

- New concepts can be proposed but need to be linked to at least one wordnet.
- The only way to change a concept is by changing its English gloss.
- Concepts can be voted for and commented on.
- Concepts never disappear but can be ignored
- Duplication check: plagiarism!

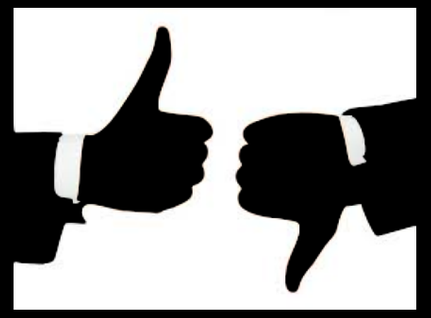
ONION MODEL

- Kernel of fund consists of concepts that:
  - shared by all associated wordnets
  - sufficiently voted for (defined sufficient: nr., global/cultural spread)
  - axiomized through an ontology
  - passed the consistency checking
- Outer layer contains the most recently proposed new concepts linked to a single language.
- In between layers link to more languages, are moderated

ONION MODEL



MODERATED



VALIDATED

VOTED

.....

ILI-IRI shared by all

.....

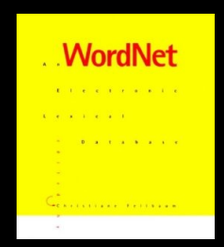
ONTOLOGIZED

New ILI-IRI linked to 1

ILI-IRI shared by 2

ILI-IRI shared by 2

New ILI-IRI linked to 1



ILI Community platform

- The community:
  - wordnet-members: wordnet builders and ontologizers
  - ili-moderator: moderator for the overall platform
    - wordnet-moderators: moderators for each language-wordnet community
- Every member belongs to a group associated with the wordnet of a language
- Add new members: ili moderator

ILI Community platform

- What can members do:
  - vote for concepts: the whole world?
  - comment on concepts: members
  - modify concepts: members
  - promote concepts to inner layers: language moderators

ILI Community platform

- Define your preference for alerts:
  - any modification
  - modification of concepts linked to your resource
  - modification of concepts related to concepts in your resource

Domain communities

- Specialists in domains should include their terminology for a languages and the corresponding concepts for the ILI
- Use the Collaborative environment for achieving cross-lingual interoperability in domains

Discrimination of concepts

- Gloss similarity can be used to find ILI concepts that are similar
- Semantic relations (of any linked wordnet) can be used to find glosses of siblings or co-hyponyms
- ILI groupings can be created for too fine-grained concepts

Technical implementation

- Specification of the data model to store concepts, editing history, provenance
- Github repository for hosting the ILI-concepts
- Social community software for voting and editing
- Export functions to WN-LMF, WN-RDF, LEMON,…
- Versioning
- Hosting

Consistency checking

- Check relations imposed on concepts from any linked wordnet or ontology
- How many hypernym matches across word nets?
- How consistent are antonymy relations?
- etc…..

Statistics

- Cross-wordnet sharing of concepts (gwa:sameAs & gwa:similarTo)
- Different parts-of-speech realizations
- Linkage in external wordnet:
  - subclass relation: top-leaf-middle, depth
  - other relations

Project plan

- Design the data structure
- Set up the repository with versioning and onion layers
- Bulk import of ILI records
- Bulk import of linked wordnets
- Set op social community platform
- Develop tools for checking and gloss comparison/ suggestion