### Problems with WordNet as Lexicography

#### Inadequate Definitions
- English definitions are indicative, not definitive
  - One English meaning ascribed to multiple terms in multiple languages
  - Mixed quality; e.g. elevator car: where passengers ride up and down
  - Not accurate for all members, e.g. *eat, feed*: take in food; used of animals only
  - Tautologies, e.g. *visit*: pay a brief visit
  - Outright errors, e.g. *law practice*: the practice of law

#### Translation Term Problems
- Incorrect glosses, often matching the wrong sense of homographs
- "Kitchen sink" collections of somewhat related terms
- Semantic drift: correct glosses of English terms that are a bit off from synset definition
- Meanings mostly assumed from English definitions
- No information beyond lemmatic spelling: what is usually known is that one form of a term is a near equivalent of something in English

#### Erratic Coverage
- Missing senses, e.g. *light*: traffic signal
- Missing terms, e.g. *lightsaber*
- Spotty relationships, e.g. no pair between boat and *ship*, a tie for *japoly* and *bus"
- Random named entities, often figures of US or UK cultural significance from a by-gone era
- Cultural focus on US and UK concepts, e.g. *shortstop* (inherent to any English-based elicitation list)

#### Lumping and Splitting

#### OMW: One Big Concept, Separated per Language
- Terms from one language WordNet can be seen together with:
  - "Same" idea terms in that language
  - English synset definition
- Matching clusters from other languages
- Each term in one cluster is parallel to each term in all other clusters

#### OMW: One Big Concept, Separated per Language
- 11 Japanese terms = 20 Arabic terms:

#### Synsets: From Lumps to Lemmas
- All English members of a synset are ascribed the same meaning and usage examples
- Most other languages are ascribed the same meaning
- 200+ terms lumped together for rag-song

### Solutions: WordNet as Seed Data

#### Crowd Review of Existing Data
- Competition for best-written definitions per concept/spelling entity
  - If PWN definition is good, it will win
  - Validation/rejection of bilingual matches by bilingual speakers
- English <-> WordNet X
- Wordnet X <-> WordNet Y

#### Data in Own Languages
- Definitions of terms in their own language
- *dedo* in Portuguese is different from the English elicitation term
- Own-language definitions can be translated to English or other languages
- Usage examples from own-language sources, e.g. blogs and tweets
- Additional lexical data
  - Inflected forms
  - Pronunciations
  - Etc, etc...

#### More Data for English Terms
- Descriptions of differences between synset members, e.g. *snuggle vs. nestle*
- User-curated usage examples, video links, images
- Geo-tagged pronunciations
- Geo-tagged usage sightings
- Lexicalized etymologies for historical and comparative linguistics
- BabelNet and other linked data
- Etc... (bringing in data from wheels that have already been invented, working with partners on new ways to enhance English data)

#### More Terms, More Languages
- Compare WordNet to other sources to find omissions
- Terms from bilingual dictionaries can address cultural bias
- Candidates for WordNet inclusion could be selected based on popularity (search logs, number of languages that choose to translate)
- When Kamusi processes produce entries linked to WordNet in languages that do not already contain them, WordNets for those languages are created or expanded
- Data merged from existing sources is directly matched to synset senses

#### Synsets Are Not (necessarily) Synonyms
- Within a language, subtle differences exist for important reasons, e.g. nuance among (approximate, estimate, gauge, guess, judge)
- Larger English synsets inspire very large translation synsets
- Translation introduces semantic drift
- Especially notable in larger synsets
- 1/5 English terms[^2] odds that a term in one translation language will equate with a given term in another language

#### Synsets Are Topical Relationships
- Synset definition describes the semantic relationship
- Synset relations perform like horizontal ontologies
- Members share a certain property (top) but independent essences
- Members should generally have independent definitions and examples in addition to synset topical guides
- Imputed translations among languages should be seen as only topically indicative until human verified

#### Substitutes
- Degree of equivalence between terms:
  1. Parallel – basically the same idea
  2. Similar – substantial overlap, but noteworthy differences
  3. Explanatory – invented term in one language to fill lexical gap for a concept indigenous to another
- A term can be parallel to one synset/translation set member but similar or explanatory to another (programming complexity)
- Differences can be elaborated in definition-like field

#### Joints
- Degree of separation between terms:
  - WordNet data has (often faulty) presumption of 1st manual validity for:
    - Same-language synset members
    - Links to English
  - Relationships via intermediate languages are mapped transitively, i.e. A <-> B <--> C <--> D
  - A and D are 3rd generation links (3+1)
  - For OWN data, “B” is always English, most computed pairs are 2
  - Evaluating predicted bilingual joints is future work via crowd systems