UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

Know
dive

# Diversity-aware Multilingual Lexical Semantic Resources Management

Freihat Abed Alhakim

03 25, 2018

# TOC

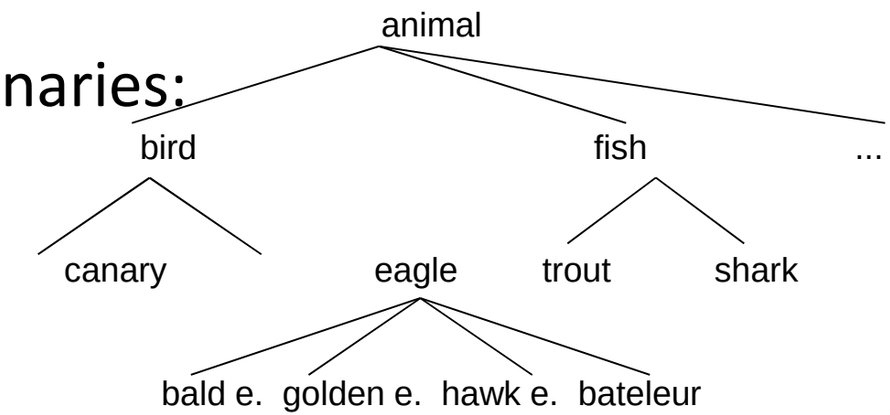
- What is Lexical semantics
- Lexical semantics resources
- Applications of Lexical Resources
- Managing multilinguality in lexical semantics resources
- The Universal Knowledge Core (UKC)
- Localization of the UKC
- Diversity management in the UKC

# What is Lexical semantics

- Subfield of linguistic Semantics
  - Classification of lexical items
    - Parrot is a bird
  - Relations between lexical Items
    - Lexical Relations : written is derivationally related to write
    - Semantic Relations: wheel is a part meronym of vehicle
  - How to map lexical items to Concepts
    - *big* and *large:* Do they denote the same concept (in some context )?
  - How to identify the domain of groups of concepts
    - *The cooking domain: boil*, *bake*, *fry*, and *roast*,…
  - How to map lexical items to events, states, properties…
    - The game started
    - The door is closed
    - The sky is blue

# Lexical semantics resources

- Machine readable lexical databases that organize lexical items based on lexical semantics theory
- In contrast to traditional alphabetic dictionaries:
  - They are conceptual dictionaries
  - Divided into POS-categories
    - Nouns, verbs, adjectives, adverbs
  - Each concept is denoted by synset
    - love, enjoy -- (get pleasure from; "I love cooking")
  - Monolingual vs. multilingual
  - Famous Lexical resources:
    - Princeton WordNet, EuroWordNet , MultiWordNet, …



# Applications of Lexical Resources

- Machine Translation
- Information retrieval
- Word Sense disambiguation
- Knowledge representation and reasoning
- Semantic Web
- Digital and smart societies
- Dictionaries
- …

# Managing multilinguality in lexical semantics resources

- Two or more  lexical resources linked together
  - Choose one of these lexical resources as refrence and link all other lexical resources to it
  - Example: Open Multilingual Wordnet
    - 34 Open Wordnets
    - Princeton WordNet as a refernce

# Managing multilinguality in lexical semantics resources

- Problems:
  - Inherit all problems of the reference lexical resource
  - What to do if the inference  is biased, contains errors, … ?
  - How to manage diversity if all lexical resources are linked to one reference (≈ one language, one culture)
  - How to link new items if they do not exist in the reference?
- Lexical gap:  A lexical item exists in some language and does not exist in other languages
  - Bike, cornfield, …
  - last straw, kick the bucket,  …
  - uncle , aunt, brother, sister, …

# The Universal Knowledge Core (UKC): Idea

- To solve the problems of using a lexical resource as a reference in multilingual lexical resource:
  - Organize the resource into different layers:
    - Knowledge layer:  language independent
    - Language layer: the language dependent representations of the knowledge layer
    - Other layers (entity layer, domain layer)
  - Use the Knowledge layer as a reference for all languages.

# The Universal Knowledge Core (UKC): Definition

- The Universal Knowledge Core (UKC) is multilingual, high quality, large scale, and diversity aware machine readable lexical resource.
- Organization:
  - The concept core (CC): The knowledge layer of UKC
  - The language core (LC): The language layer of the UKC
  - Classification of relations:
    - The relations are concepts
    - Semantic Relations: (language independent) relations
      - used in the concept core only
    - Lexical Relations: Language dependent
      - used in the language core only

# The Universal Knowledge Core (UKC): Concept core

- A  set of connected nodes forming a directed acyclic graph. (DAG).
- Each node in this DAG corresponds to a concept.
- Concept:  a language independent representation of some thing or a happening.
- The concepts are organized through semantic relations such as hypernymy (is-a), the meronymy (part-of) relations.

# The Universal Knowledge Core (UKC): Language core

- The lexicalization of the concepts in the concept core in one or more natural languages.
- Lexicalization is performed through synsets, and lexical gaps.
- Synset:
  - a group of lexical units (synonyms) that express a concept.
  - a natural language description of the concept (gloss), and
  - one or more (optional) examples that help in clarifying the usage of the concept
- Lexical gap: indicates the absence of the lexicalization of a concept if it is unknown in some language.

# Localization of the UKC: Current state

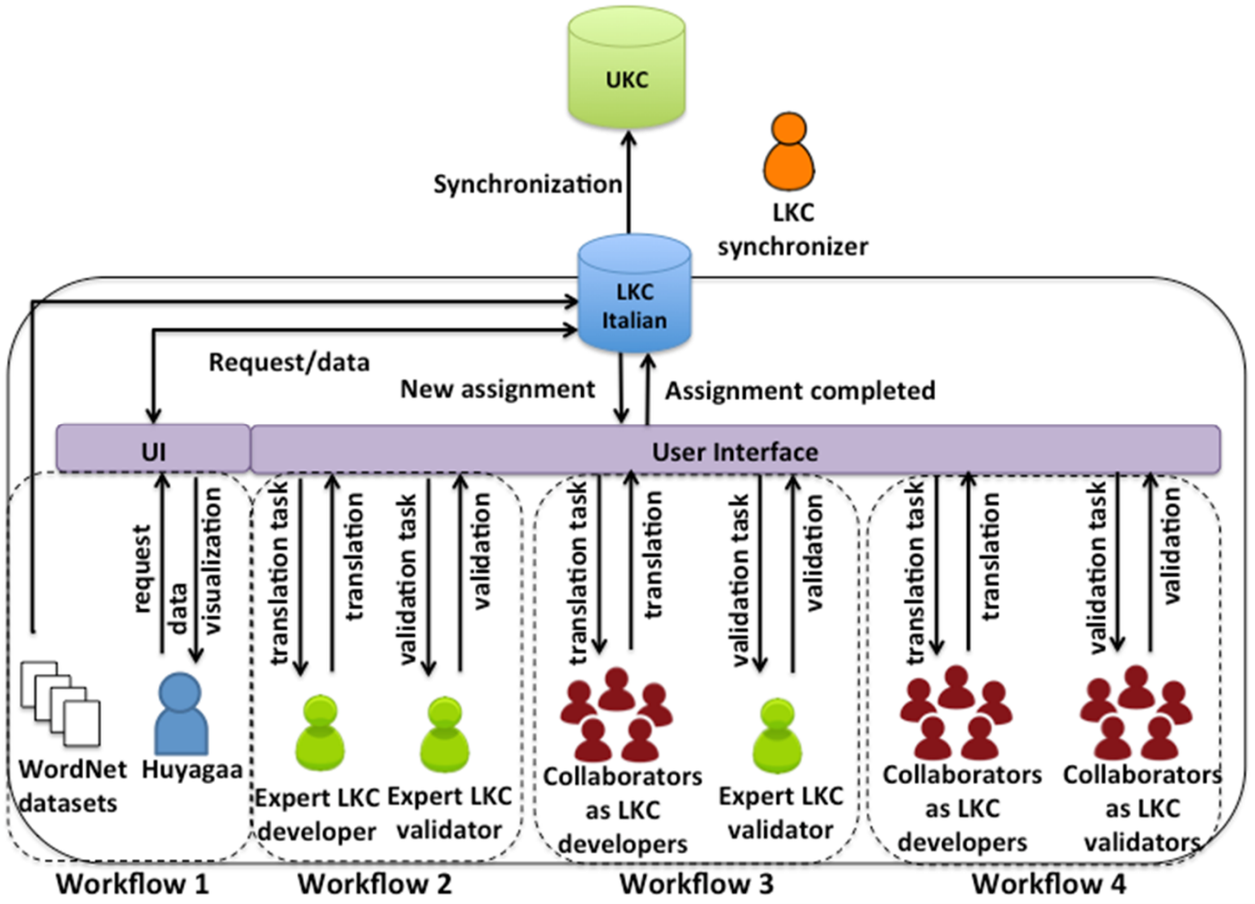
- The UKC currently:
  - The concept core contains more than 117.000 concepts
  - The language core contains 350 languages
  - These languages localize (partially) the concept core concepts
- UKC is evolving and continues to grow in terms of quality and quantity.
  - adding new languages,
  - expanding the coverage of the existing languages.
- Current active projects:
  - South African languages, Indian languages, Gaelic,  Romanian, Italian

# Localization of the UKC: Enviornmet

- A collaborative lexical resources development
  - Involve linguistic experts in the lexicalization process
    - provide and evaluate translations produced in their own language
  - Local Knowledge Core (LKC)
    - <Source language, Target language>
      - For the same language possibly different source languages

    Different LKCS
      - Example: Arabic
        - Source language for Arabs from North Africa : French
        - Source Language for Arabs frim Asia : English

# Localization of the UKC: Framework

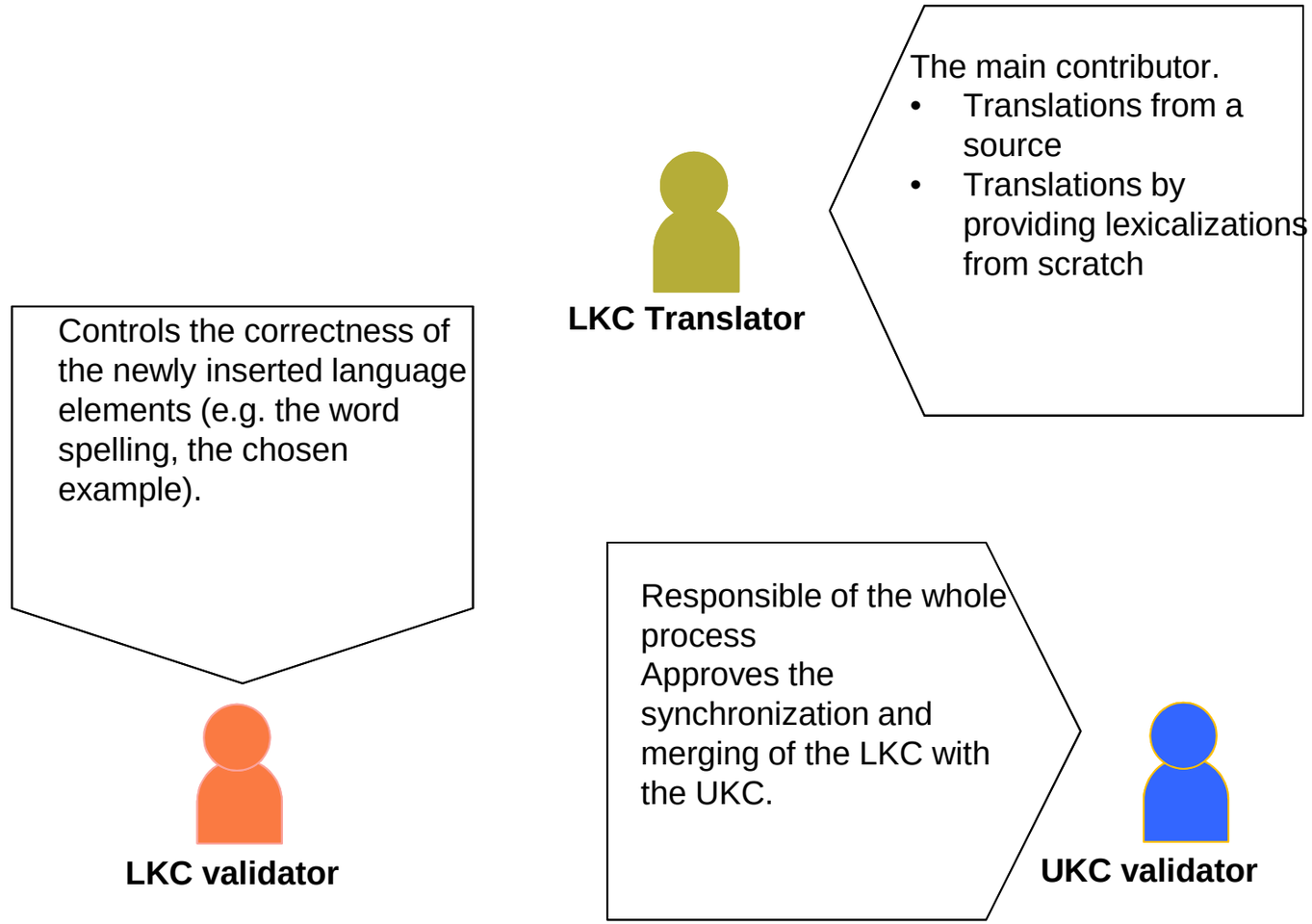


# Localization of the UKC: Experts

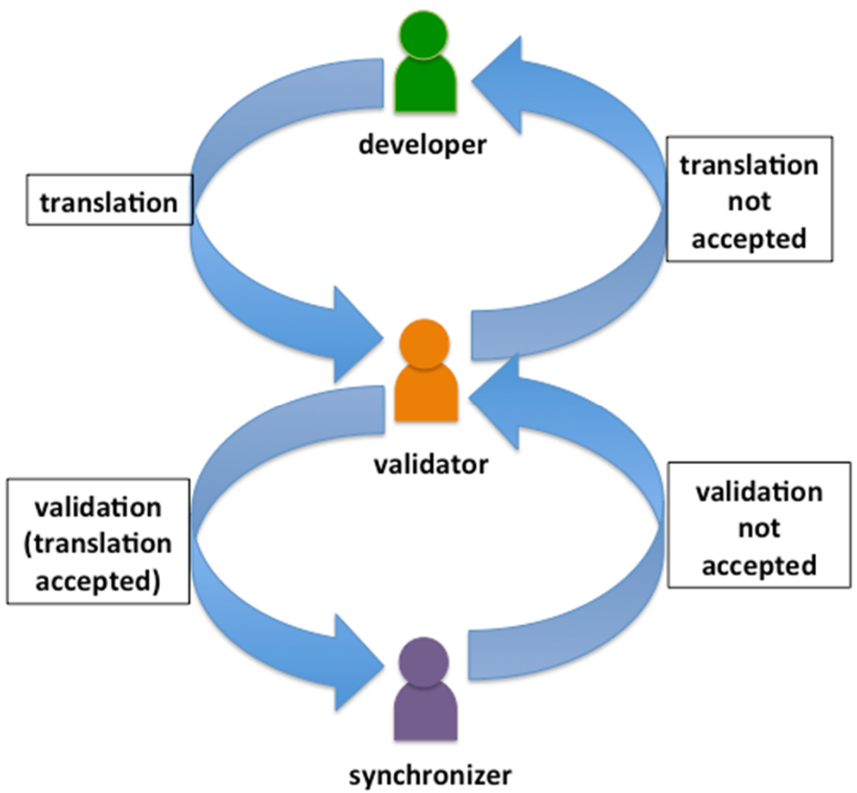
## Who are the experts?

- Native speakers of the target language
- Competent level of the source  language
- Extended knowledge of the two languages

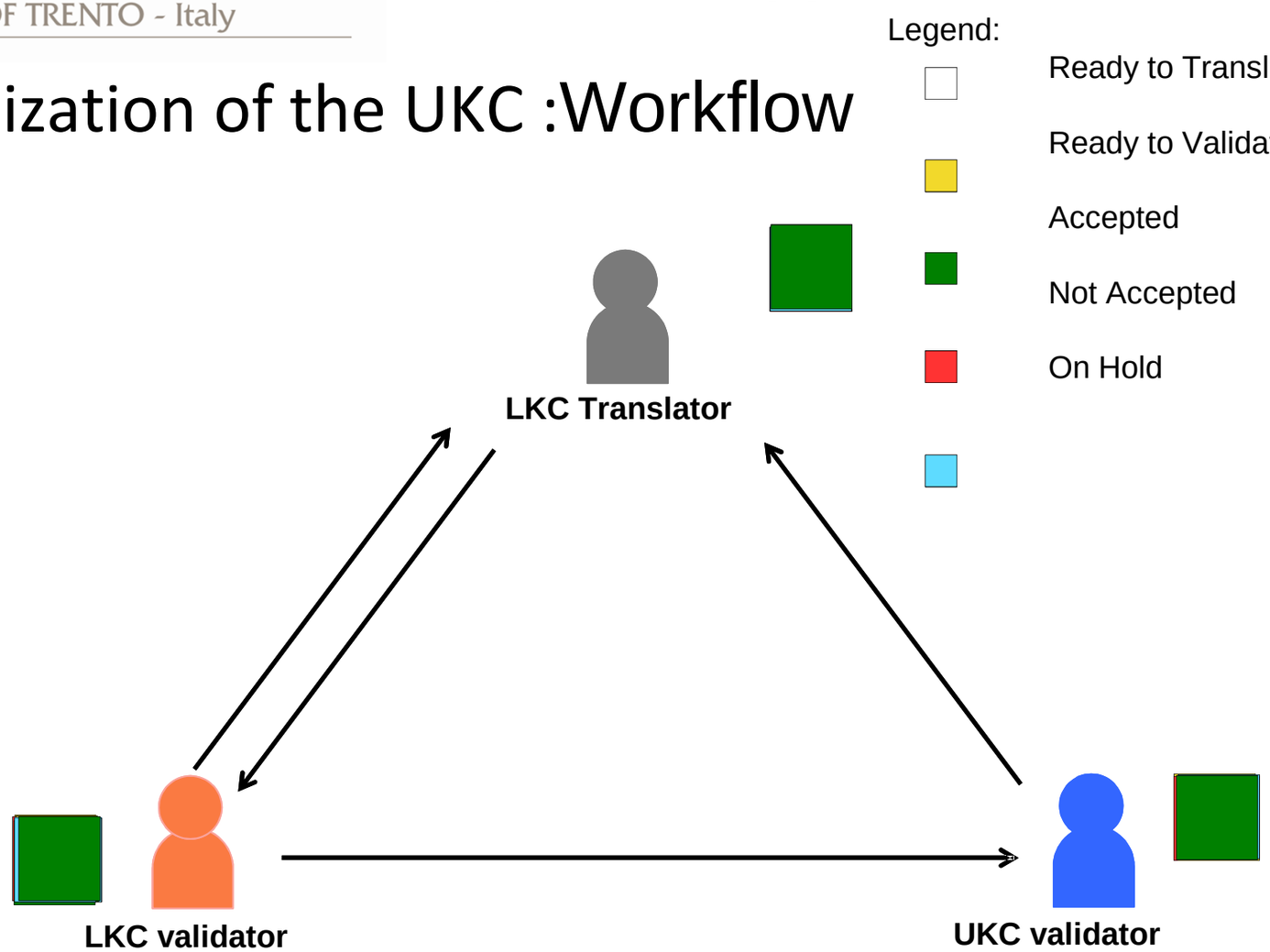
# Localization of the UKC: Roles



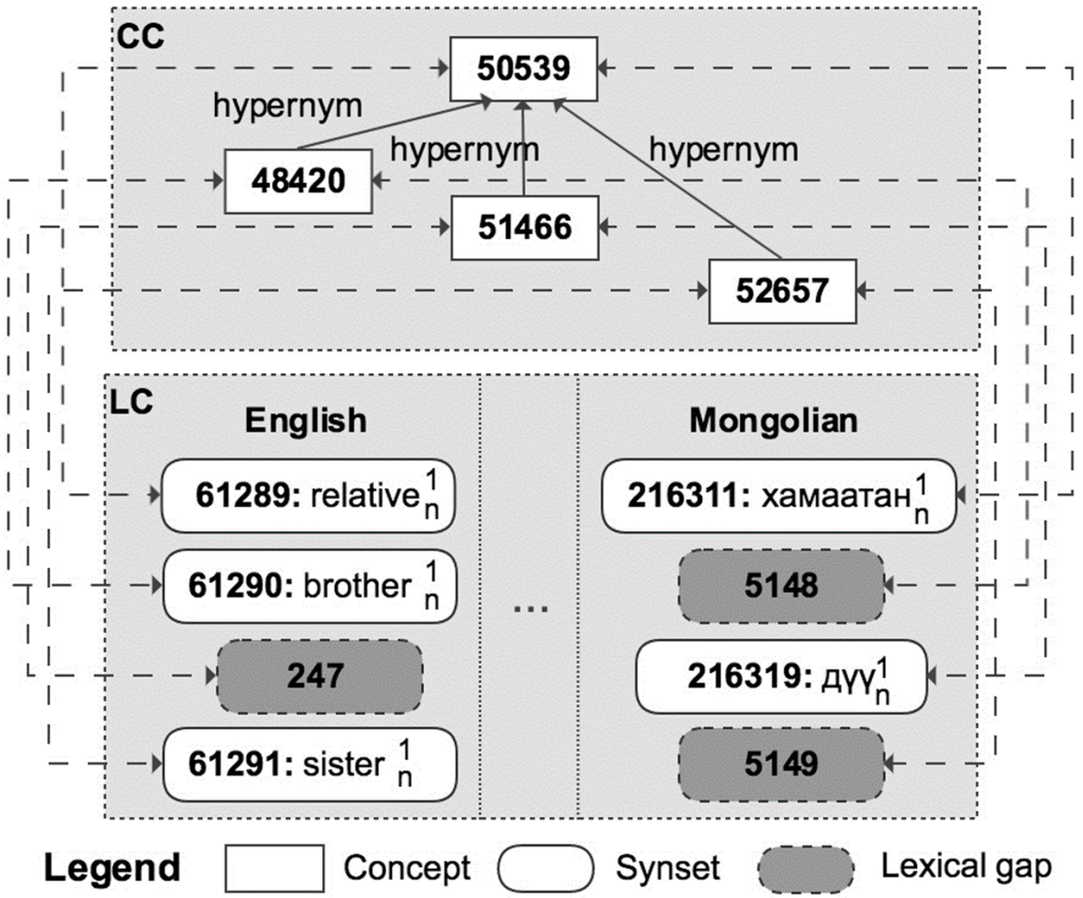
# Localization of the UKC: Collaboration



# Localization of the UKC :Workflow



# Diversity management in the UKC Example



# Diversity-aware Multilingual Lexical Semantic Resources Management

Questions?

Thank you