Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

Mapping WordNet Instances to Wikipedia

John P. McCrae

Insight Centre for Data Analytics, National University of Ireland Galway

A World Leading SFI Research Centre

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN Science Foundation Ireland SFI For what's next

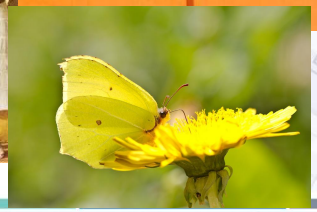
Lexical vs. Encyclopedic

Yellow (in a dictionary)

- Is a verb, noun and adjective
- Secondary synonyms: cowardly, warning (especially in soccer)

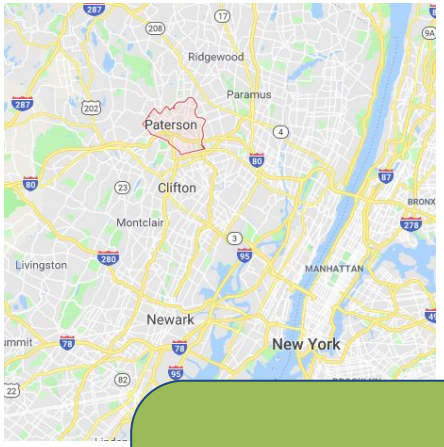
Yellow (in an encyclopedia)

- A colour
- 2 books
- 8 Films or TV shows
- 4 songs
- A butterfly

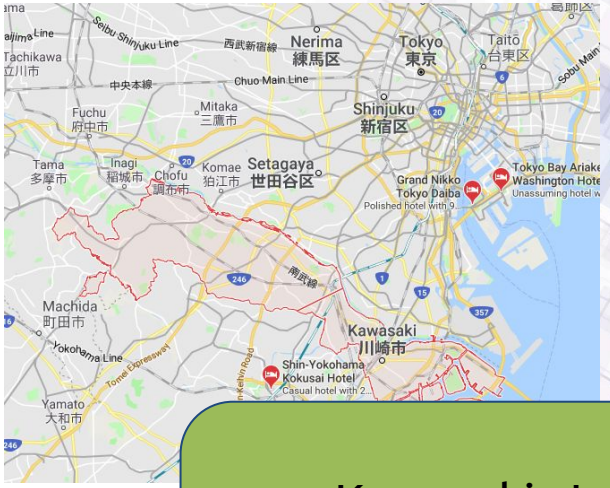


Princeton WordNet

- Is a lexical resource with some encyclopedic information
- This information is quite biased to Anglo-Saxon, American and even North Eastern US context.



Paterson, NJ, USA
Pop: 147,000
In Princeton WordNet



Kawasaki, Japan
Pop: 1,500,000
Not In Princeton WordNet

Wikipedia

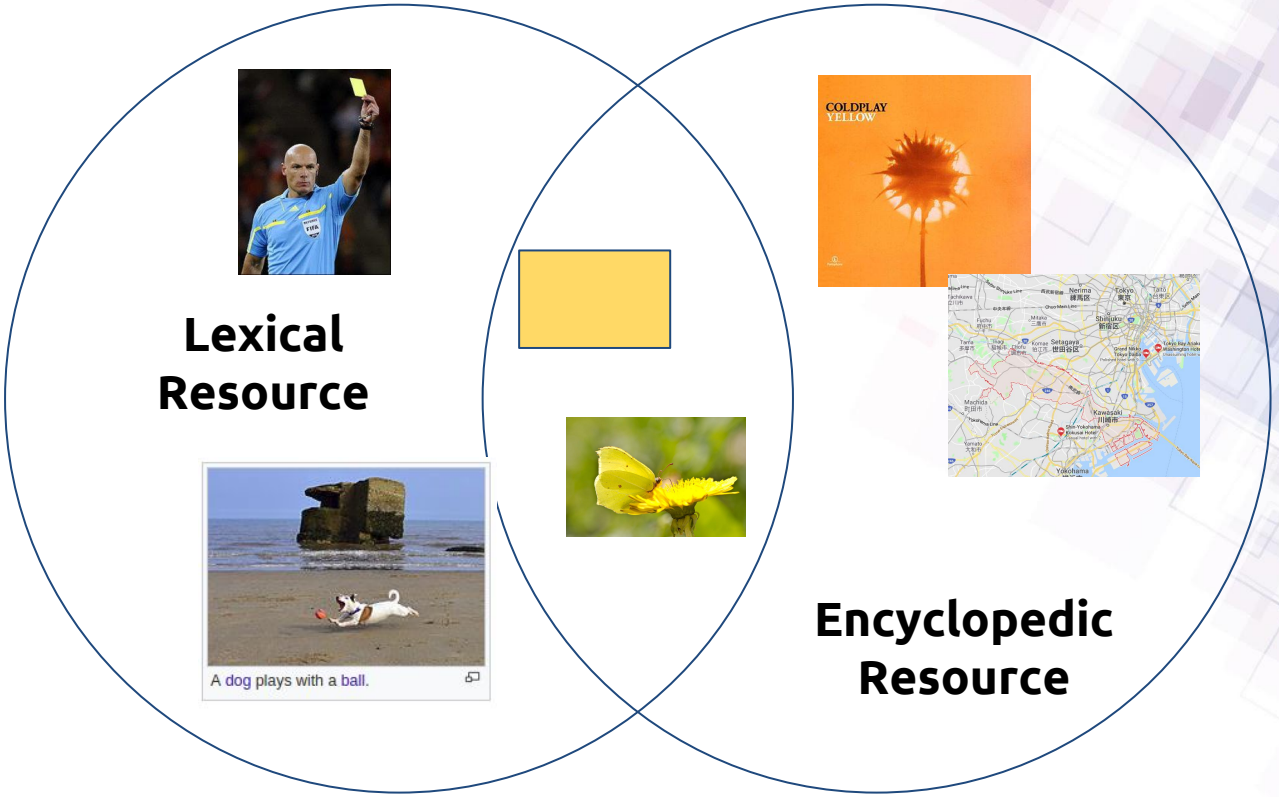
- Wikipedia is open and most-widely used encyclopedia
- Many lexical concepts are included, e.g.,
  - Play (activity)
    - https://en.wikipedia.org/wiki/Play_(activity)

A dog plays with a ball.

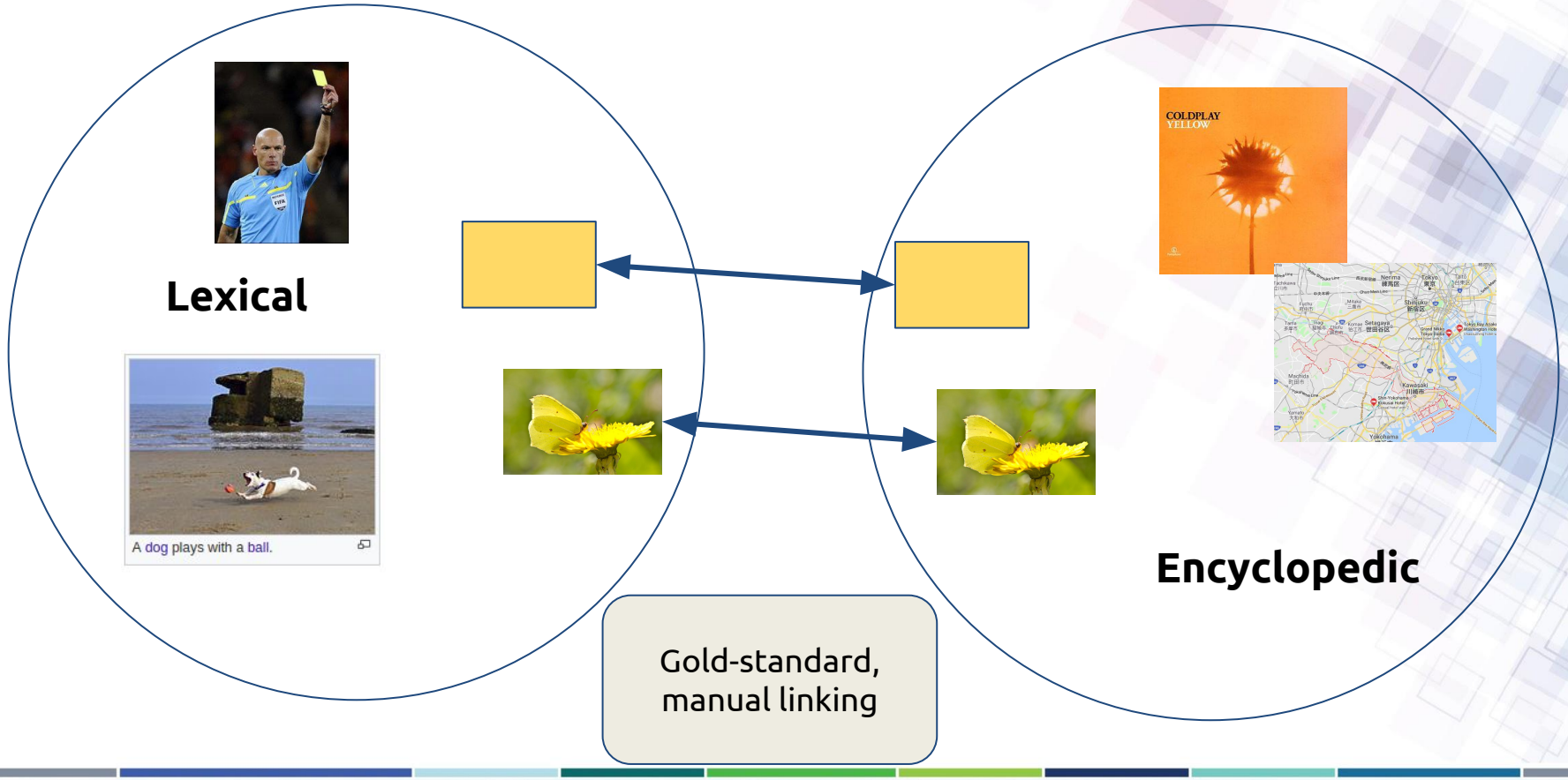
Cocker spaniel playing with a monkey doll

People having fun

Overlap between lexical and encyclopedic resources



Linking between resource types



Proper Nouns in WordNet

Identifying Proper Nouns in PWN

- Proper nouns are not a POS value in PWN
- But, proper nouns are capitalized in the PWN data
  - However, synsets often contain a mixture of capitalized and non-capitalized
  - (n) cat, Caterpillar - a large tracked vehicle that is propelled by two endless metal belts; frequently used for moving earth in construction and farm work
  - (n) domestic cat, house cat, Felis domesticus, Felis catus - any domesticated member of the genus Felis

Instance Hyponyms

- Instance hyponyms is one of the basic properties in PWN
- Represents instances - unique, singular objects
- Instance hyponyms should all be proper nouns!
  - 7,742 noun synsets of which 7,726 (99.8%) have a capitalized word.
  - 7 have non-standard orthography
    - “al-Muhajiroun”
  - 6 should be capitalized
    - “pampas”
  - 2 shouldn’t be instance hyponyms
    - “sierra”
  - 1 is puzzling
    - “church mouse” (a fictional mouse created by Lewis Carroll)

Instances

Major categories

- Named people (i35562)
- Named places (i35580)
- Body of water (i85104)
- Geographical features (i85439)
- Land (i85674)
- Gods (i86570)
- Wars (i35586)
  - and other events
- Terrorist organizations (i79103)
  - and other social groups
- Books (i69848)

Not included

- Linnaean Taxonomy
- Systems of belief (e.g., Buddhism)
- Languages and nationalities (e.g., German)

Related Work

Automated Approaches

- [Navigli and Ponzetto, 2012] Released as BabelNet with 82.7% F-Measure
- [Niemann and Gurevych, 2011] Released as UBY.
- [Fernando and Stevenson, 2012] F-Measure of 84.1%
- [Suchanek et al., 2008] Combination of categories (not instances) with high accuracy (97.7%) released as YAGO.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193:217–250.

Elisabeth Niemann and Iryna Gurevych. 2011. The people's web meets linguisti rdnet. In Proceedings of the Ninth International Conference on Computational Sema

Samuel Fernando and Mark Stevenson. 2012. Mapping WordNet synsets to W Conference on Language Resources and Evaluation, pages 590–596.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A ce, Services and Agents on the World Wide Web, 6(3):203–217.

Automatic approaches are still not able to get above 90% accuracy. Real-world results are often more disappointing

Linking from PWN to other resources

- [Mihalcea and Moldovan, 2000] SemCor to FrameNet and VerbNet
- [Palmer, 2009; Bonial et al. 2013] SemLink to PropBank, FrameNet and VerbNet
- [McCrae et al., 2012] To Wiktionary

Rada Mihalcea and Dan Moldovan. 2000. Semantic indexing using WordNet senses. In Proceedings of the ACL-2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval, pages 35–45. Association for Computational Linguistics.

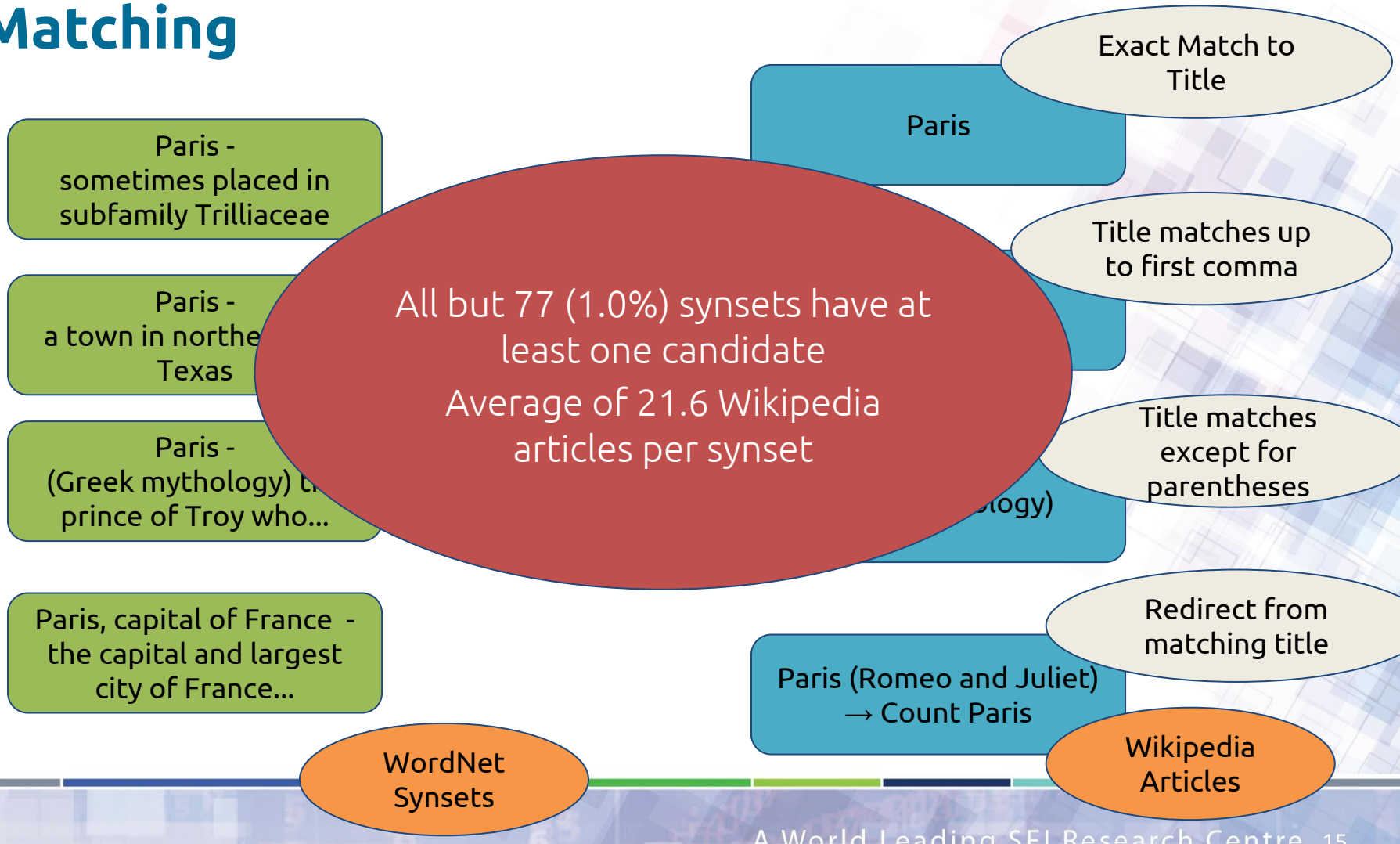
Martha Palmer. 2009. SemLink: Linking Propbank, VerbNet and FrameNet. In Proceedings of the generative lexicon conference, pages 9–15. Pisa Italy.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising SemLink. In The GenLex Workshop on Linked Data in Linguistics.

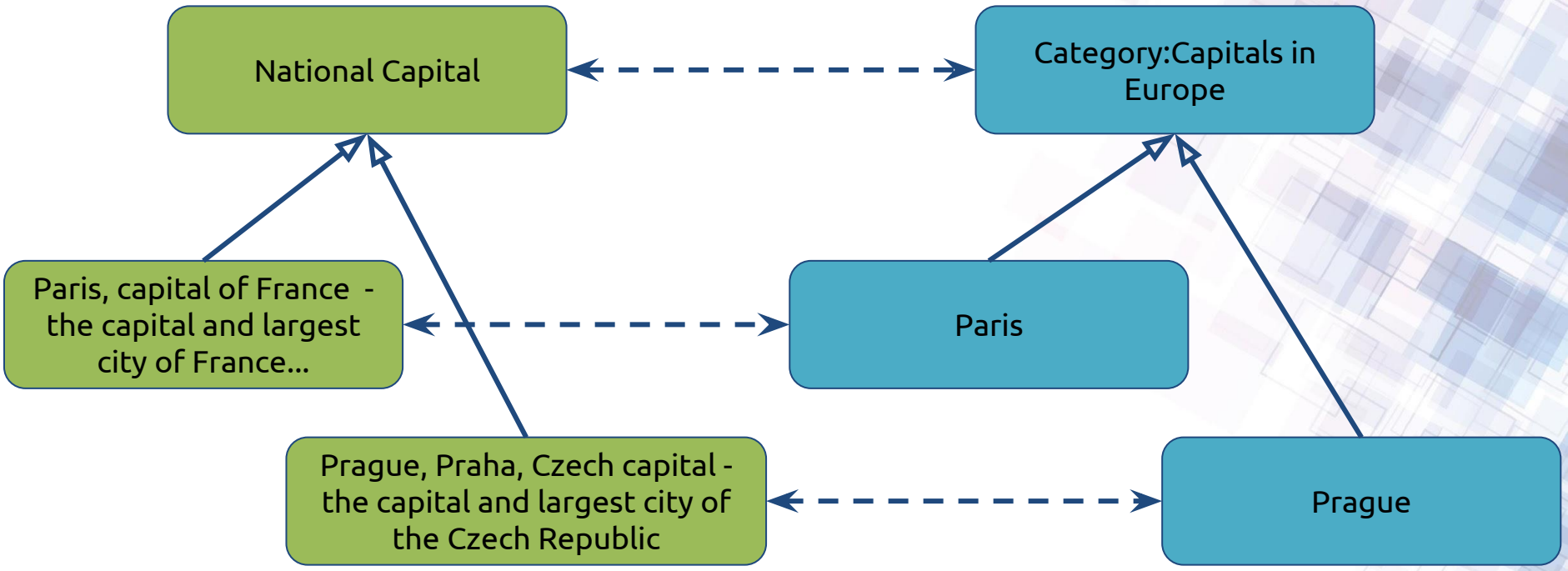
John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012. Integrating WordNet and Wiktionary with lemon. Linked Data in Linguistics, pages 25–34.

Methodology

Matching


Exact Match to Title
Paris
Paris - sometimes placed in subfamily Trilliaceae
Title matches up to first comma
Paris - a town in northe Texas
All but 77 (1.0%) synsets have at least one candidate
Average of 21.6 Wikipedia articles per synset
Title matches except for parentheses
Paris - (Greek mythology) t prince of Troy who...
ology)
Redirect from matching title
Paris, capital of France - the capital and largest city of France...
Paris (Romeo and Juliet) → Count Paris
WordNet Synsets
Wikipedia Articles

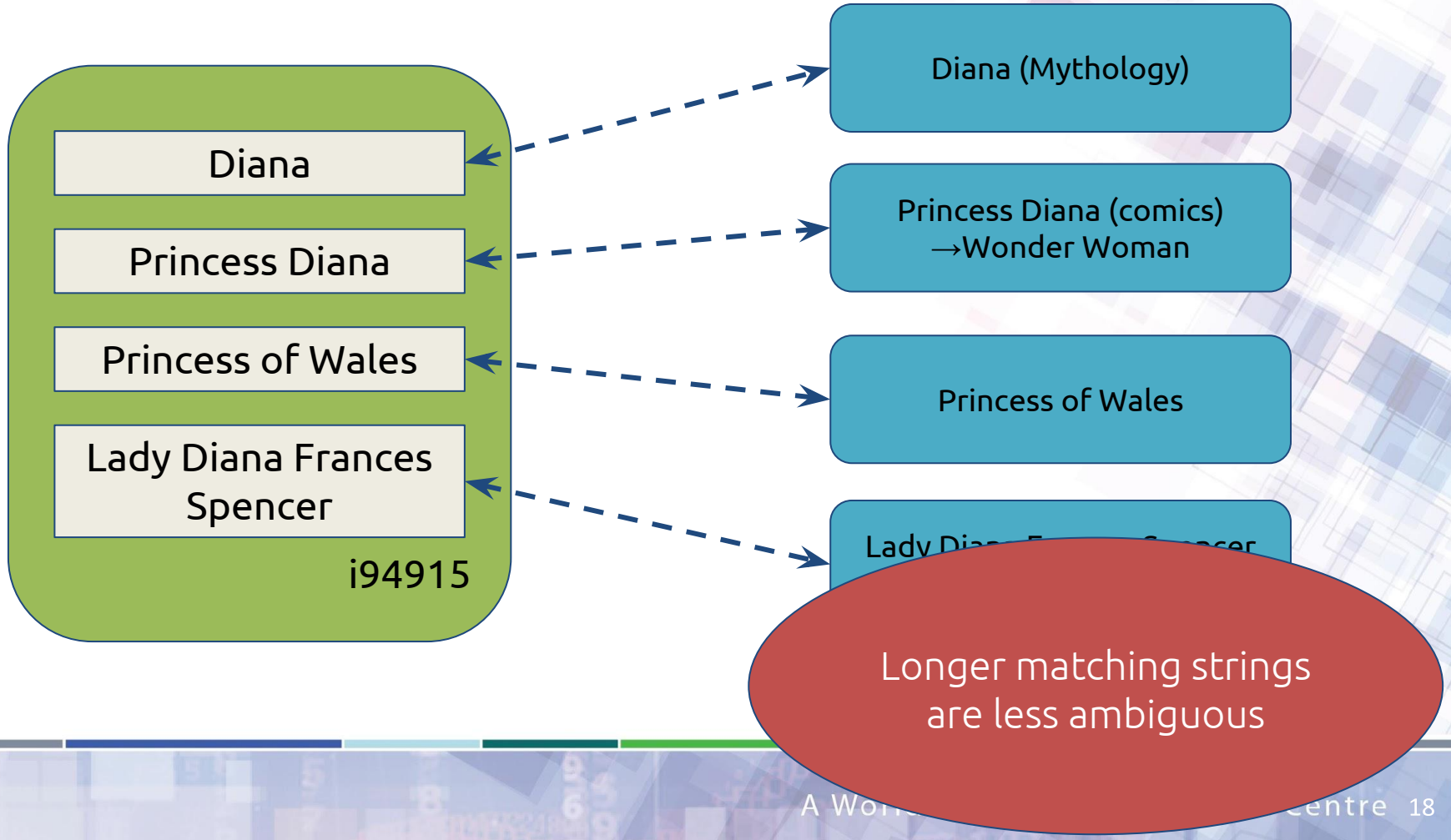
Category Matches (based on Suchanek's YAGO)



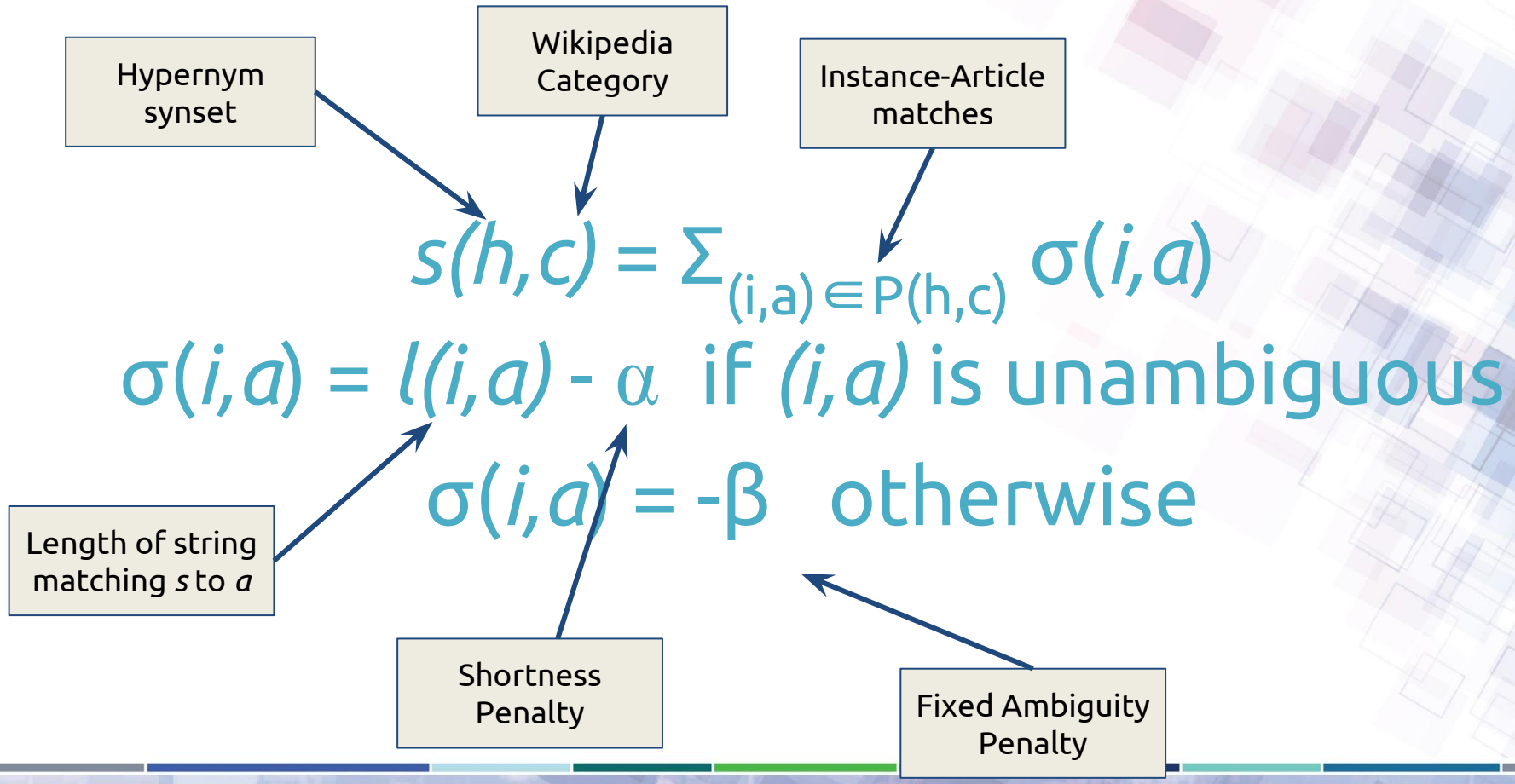
Category matches

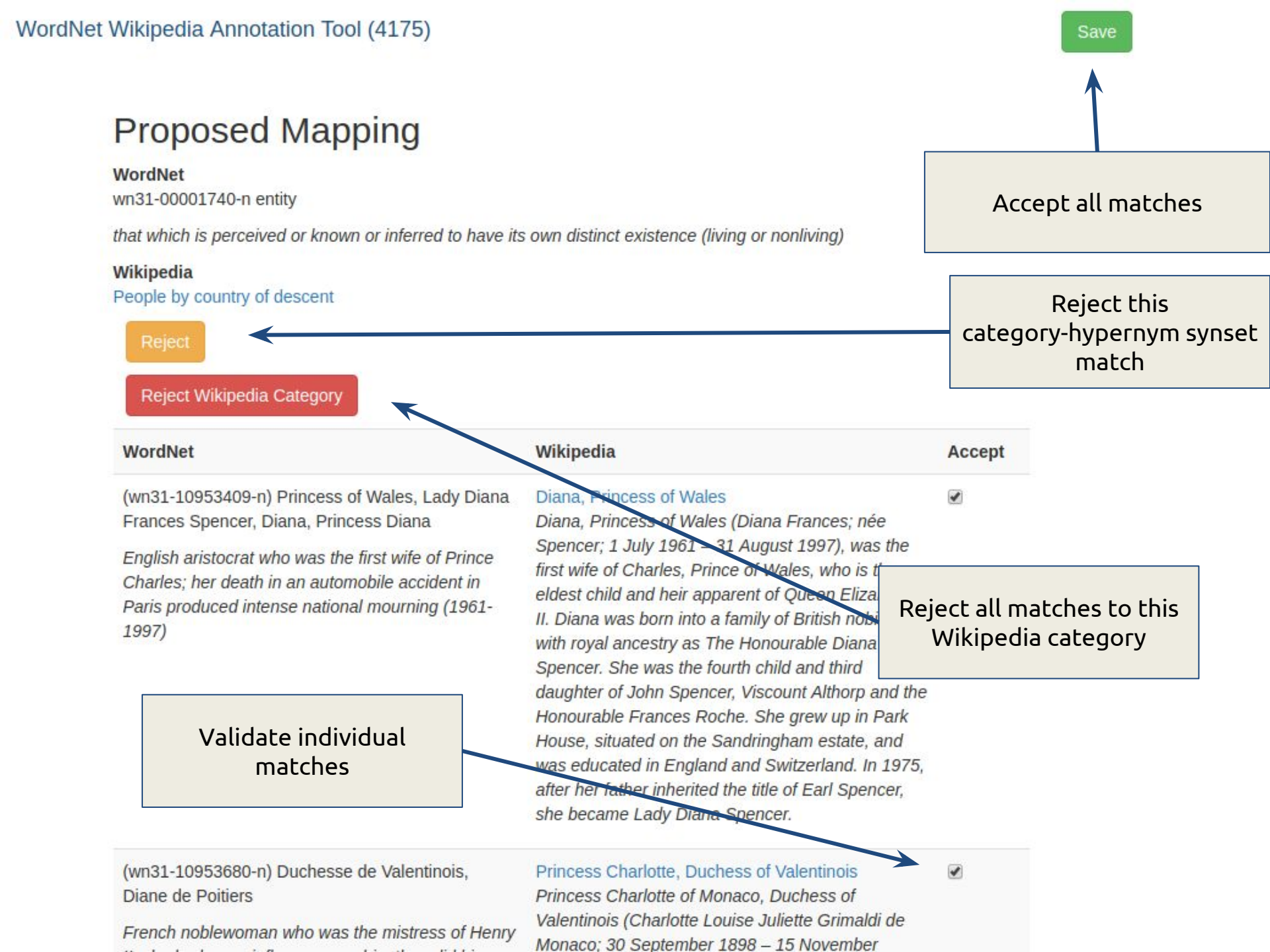
- A category match is mapping between a Wikipedia category and (non-instance) PWN synset.
- Produces a set of article/instance synsets matches that match on both category/hypernym synset and lemma/article title.
- We call a category match unambiguous if there does not exist
  - 2 hyponym synsets with matches to the same article
  - 2 articles with matches to the same hyponym synset
- Unambiguous matching was not sufficient.

Length based matches



Ranking Category matches




WordNet Wikipedia Annotation Tool (4175)
Save
Proposed Mapping
WordNet
wn31-00001740-n entity
that which is perceived or known or inferred to have its own distinct existence (living or nonliving)
Wikipedia
People by country of descent
Reject
Reject Wikipedia Category
WordNet
Wikipedia
Accept
(wn31-10953409-n) Princess of Wales, Lady Diana Frances Spencer, Diana, Princess Diana
English aristocrat who was the first wife of Prince Charles; her death in an automobile accident in Paris produced intense national mourning (1961-1997)
Diana, Princess of Wales
Diana, Princess of Wales (Diana Frances; née Spencer; 1 July 1961 – 31 August 1997), was the first wife of Charles, Prince of Wales, who is t
eldest child and heir apparent of Queen Eliza
II. Diana was born into a family of British nobi
with royal ancestry as The Honourable Diana
Spencer. She was the fourth child and third daughter of John Spencer, Viscount Althorp and the Honourable Frances Roche. She grew up in Park House, situated on the Sandringham estate, and was educated in England and Switzerland. In 1975, after her father inherited the title of Earl Spencer, she became Lady Diana Spencer.
(wn31-10953680-n) Duchesse de Valentinois, Diane de Poitiers
French noblewoman who was the mistress of Henry
Princess Charlotte, Duchess of Valentinois
Princess Charlotte of Monaco, Duchess of Valentinois (Charlotte Louise Juliette Grimaldi de Monaco; 30 September 1898 – 15 November
Accept all matches
Reject this category-hypernym synset match
Reject all matches to this Wikipedia category
Validate individual matches

Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

Resource and Evaluation

A World Leading SFI Research Centre

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN Science Foundation Ireland SFI For what's next

Resource

- Annotation was performed by a single annotator using interface and methodology
  - Approximately 30-40 hours
- 7,582 mappings were performed this way (239 unmapped)
- Second pass then performed
  - Broad/narrow/related mappings
  - Errors in mapping
  - Only mappings that did not satisfy:
    - Exact article title/lemma match
    - Article title of the form “X (Y)”, “X, Y” and X is a lemma of the synset and Y occurs in the definition
  - 1,733 mappings in second round

Second pass annotation

- **Exact**: Mapping is exact
- **Broad:** The Wikipedia article is broader, e.g., “Singapore, Singapore Island” (i83963), “Singapore, Republic of Singapore” (i83964) and “Singapore, capital of Singapore” (i83965) to “Singapore”
- **Narrower:** “Rameses - any of 12 kings of ancient Egypt between 1315 and 1090 BC” to “Rameses I”, “Rameses II”, etc.
- **Related: “**Hoover, William Hoover, William Henry Hoover” (i95579) to “The Hoover Company”
- **Unmapped**: No mapping is available
- **Error:** Annotation Error
- **Improved**: Annotation was improved (broad/narrow to exact)

Results

Exact	7,582
Broad	54
Narrow	21
Related	30
Unmapped	59
Errors	56
Improved	11

Princeton WordNet Improvement Suggestions

- Two synsets were duplicates
- One synset should be split (Netherlands != Kingdom of the Netherlands)
- 17 typos detected (Sir Richrd Steele)
- 2 incorrect synset relations
- 4 synsets described non-existent concepts (El Libertador, Changtzu, Ehadhamen, church mouse)
- 41 definitions with factual inaccuracies
- 1,062 new lemmas to introduce

Resource release

- Linked to ILI and PWN 3.1
- Now available as part of WordNet RDF
  - http://wordnet-rdf.princeton.edu
- Contributed to DBpedia project
  - https://github.com/dbpedia/links
- Will be available through the ILI
- Direct download link
  - https://jmccrae.github.io/wn-wiki-instances/ili-map-dbpedia.ttl
  - https://jmccrae.github.io/wn-wiki-instances/dbpedia-ili-map.ttl

Conclusion

Conclusion

- First large-scale, gold-standard mapping of instances from PWN to Wikipedia
  - Useful for NLP applications
  - Methodology can be used in future versions of PWN
- Mapping to Wikipedia can be (trivially) extended to:
  - DBpedia
  - WikiData
  - GeoNames
  - OpenStreetMap
  - Twitter Accounts
  - etc.