



NUI Galway
OÉ Gaillimh



Insight

Linking languages across space and time

John P. McCrae

National University of Ireland Galway
Insight Centre for Data Analytics



Biography



Postdoctoral
Researcher 2009-2015

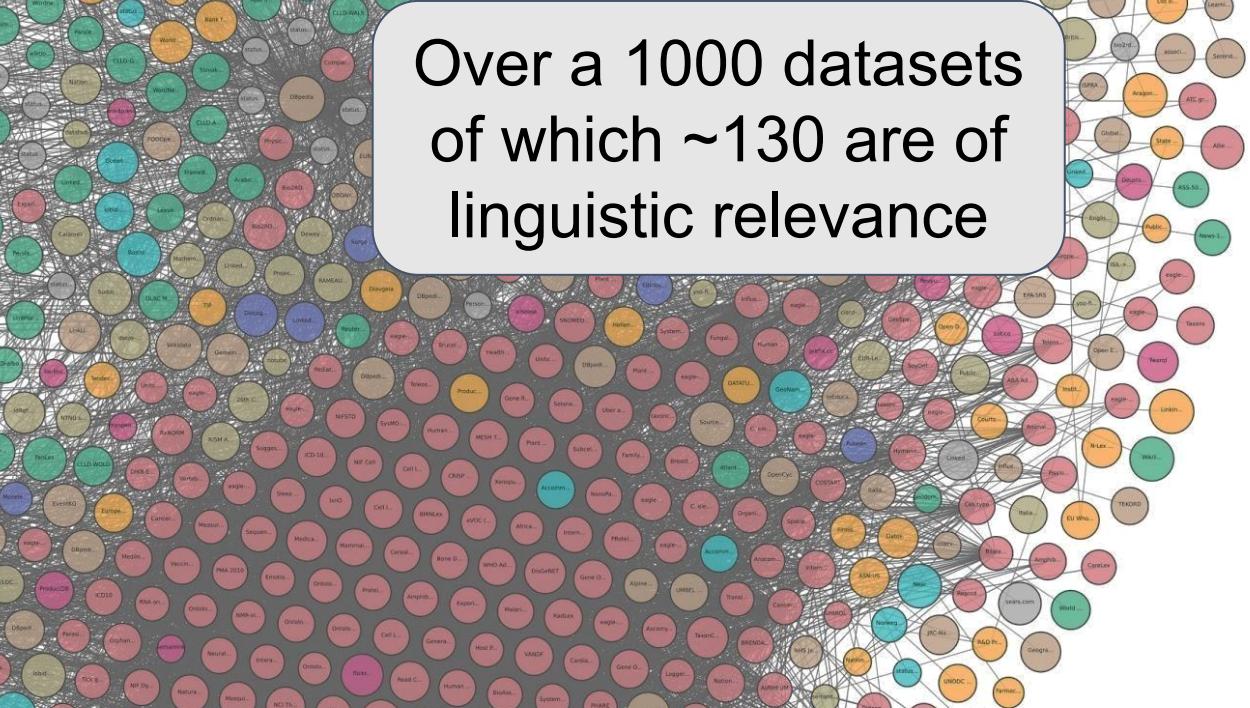
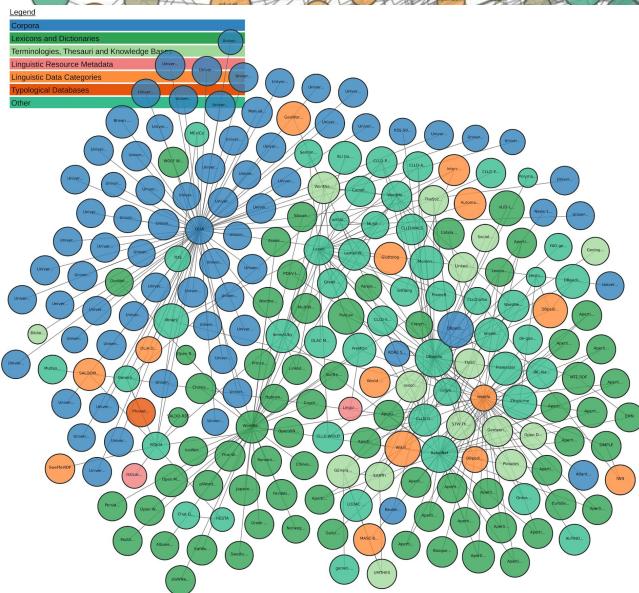
PhD 2006-2009

Universität Bielefeld



Lecturer
2015-

Linguistic Linked Open Data



(Linguistic) Linked Open Data

1. Use URLs to identify things
2. Use URLs that tell users how to find your data
3. When someone looks up an URL, provide useful information
4. Include links to other datasets.

linghub.org

Central aggregator of metadata
about language resources
across repositories

Expanding to 1,000,000
resources

Prêt-à-LLOD

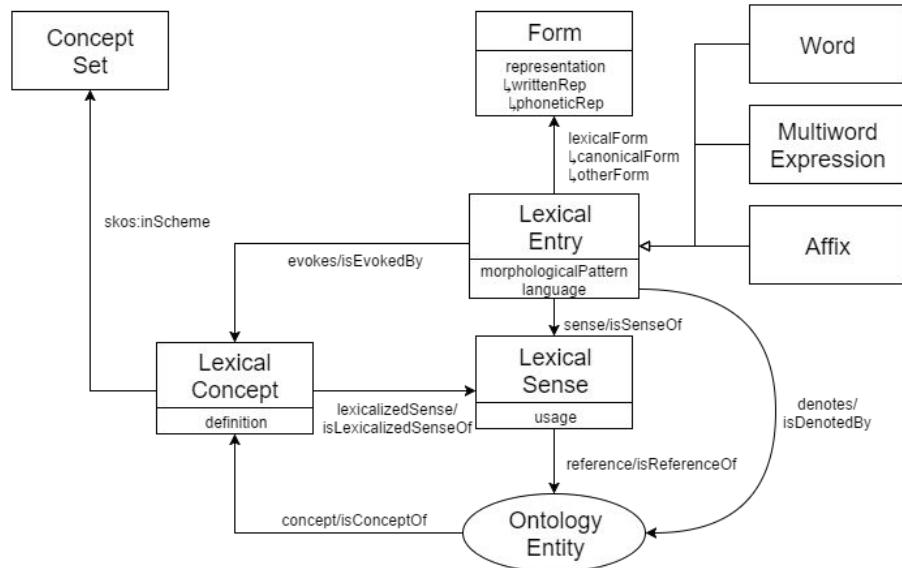
OntoLex

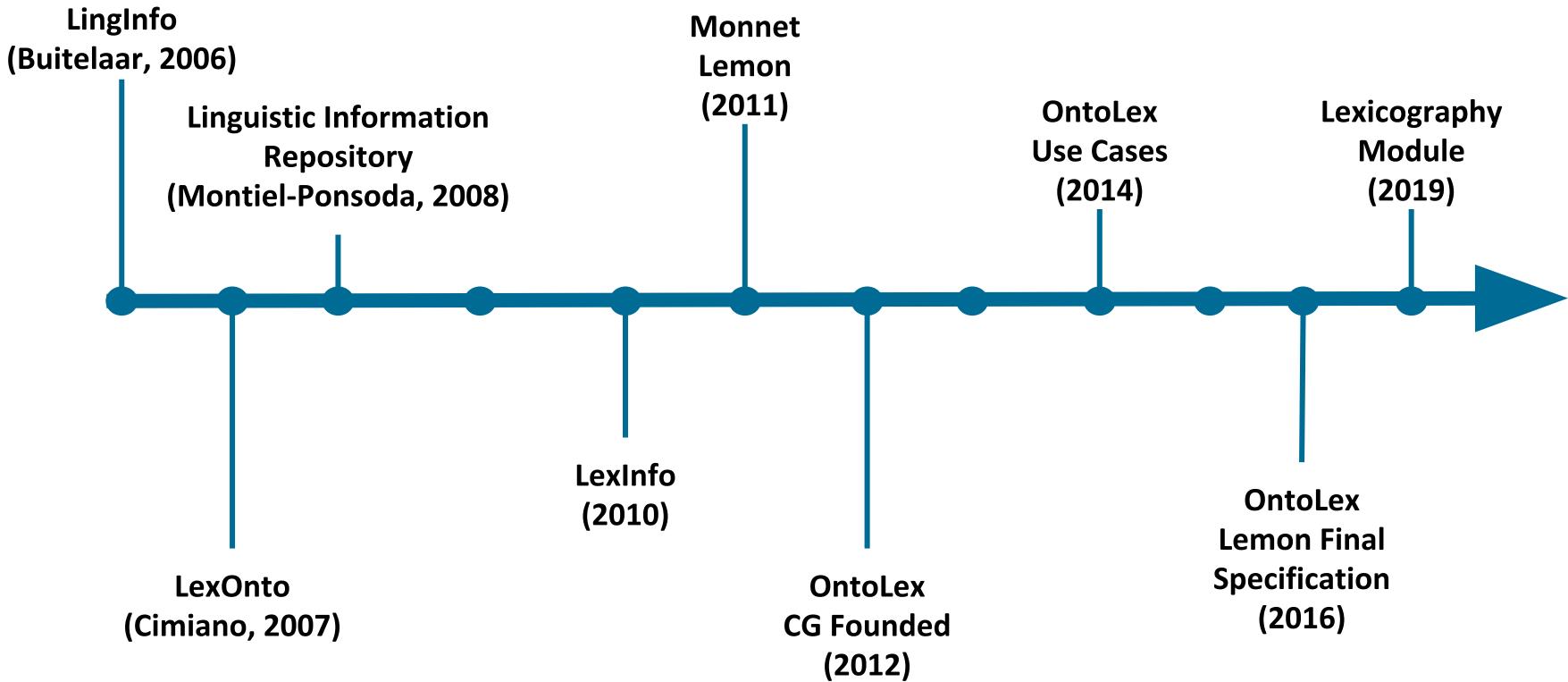
W3C Vocabulary for representing dictionaries as linked data.

<https://www.w3.org/2016/05/ontolex/>

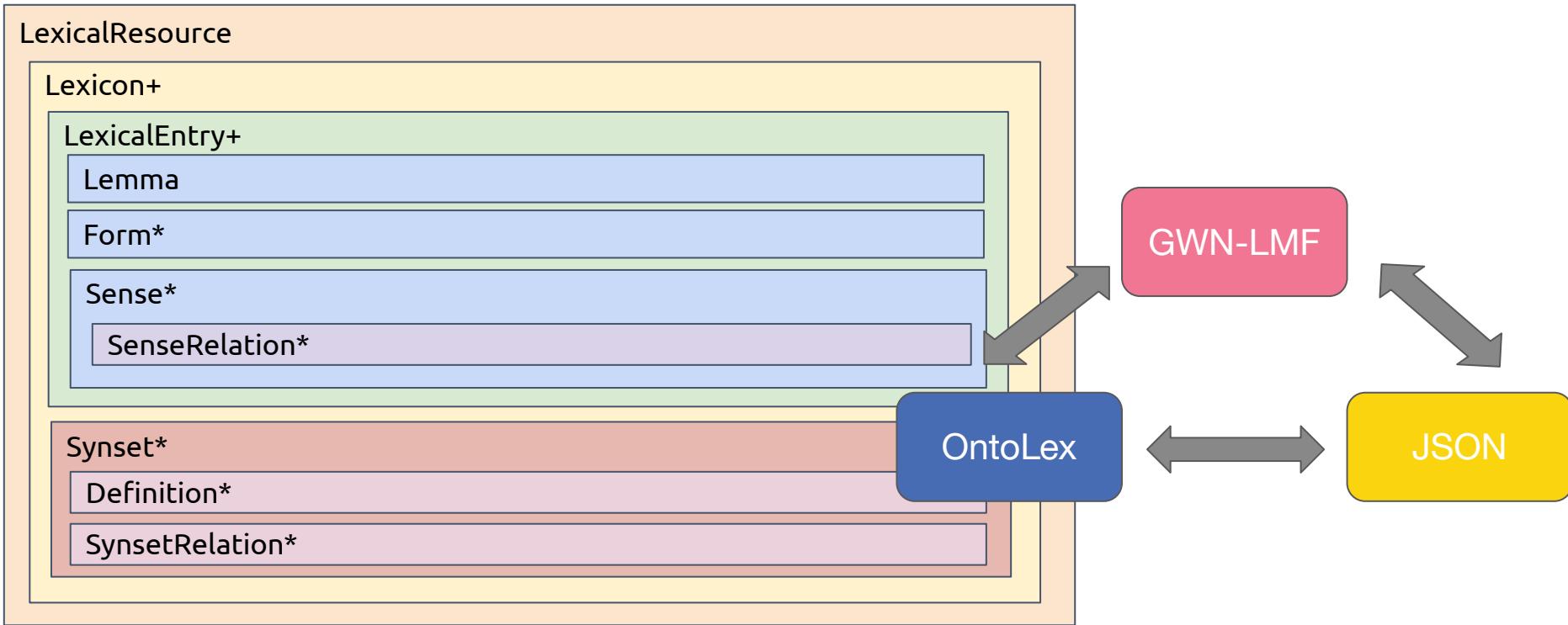
Community group is still working, please join!

<http://www.w3.org/community/ontolex>





Global WordNet Association Formats



OntoLex

GWN-LMF

JSON

Linking Dictionaries



1 Activity involving mental or physical effort done in order to achieve a purpose or result.

'he was tired after a day's work in the fields'

+ More example sentences

+ Synonyms

1.1 Mental or physical activity as a means of earning income; employment.

'I'm still looking for work'

+ More example sentences

+ Synonyms

1.2 The place where one is employed.

'I was returning home from work on a packed subway'

+ More example sentences

1.3 The period of time one spends in paid employment.

'he was going to the theatre after work'

+ More example sentences

1.4 *West Indian* [count noun] A job.

'I decided to get a work'

+ More example sentences

+ Synonyms

2 A task or tasks to be undertaken.

'they made sure the work was progressing smoothly'

+ More example sentences

+ Synonyms

2.1 The materials for a task.

'she frequently took work home with her'

+ More example sentences

2.2 (in combination or with modifier **works**) *British* Activity involving construction or repair.

'extra costs caused by additional building works'

1. (*heading, uncountable*) Employment.

1. Labour, occupation, job. [quotations ▾]

My work involves a lot of travel.

2. The place where one is employed.

He hasn't come home yet, he's still at work.

3. One's employer

"I want to go to the R.E.M. reunion concert but I'm not sure if my work will let me."

2. (*heading, uncountable*) Effort.

1. Effort expended on a particular task.

Holding a brick over your head is hard work. It takes a lot of work to write a difficult book.

2. Sustained human effort to overcome obstacles and achieve a result.

We know what we must do. Let's go to work.

3. Something on which effort is expended.

There's lots of work waiting for me at the office.

4. (*physics*) A measure of energy expended in moving an object; most commonly, force.

Work is done against friction to drag a bag along the ground.

5. (*physics, more generally*) A measure of energy that is usefully extracted from a process.

3. Sustained effort to achieve a goal or result, especially overcoming obstacles. [quotations ▾]

We don't have much time. Let's get to work piling up those sandbags.

4. (*heading*) Product; the result of effort.

1. (*uncountable, often in combination*) The result of a particular manner of production.

There's a lot of guesswork involved.

2. (*uncountable, often in combination*) Something produced using the specified material.

We've got some paperwork to do before we can get started. The piece was delivered yesterday.

3. (*countable*) A literary, artistic, or intellectual production. [quotations ▾]

It is a work of art.

the poetic works of Alexander Pope

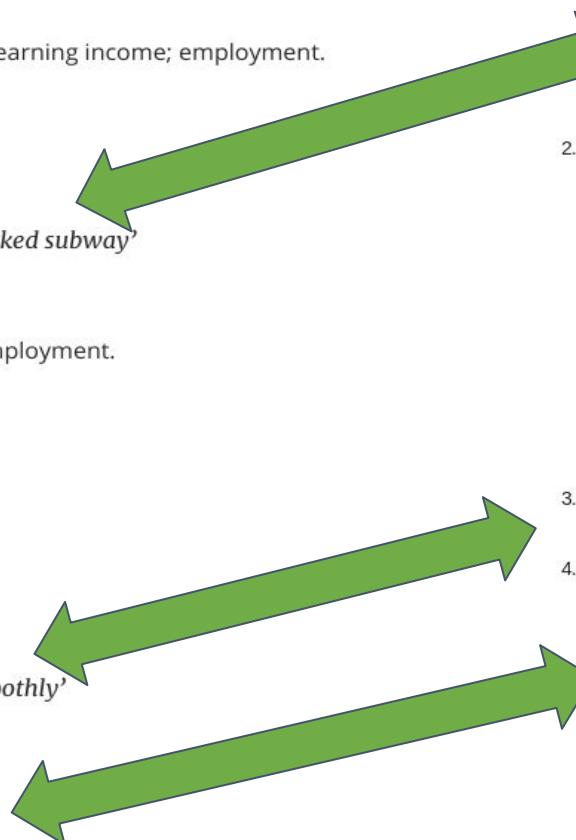
4. (*countable*) A fortification.

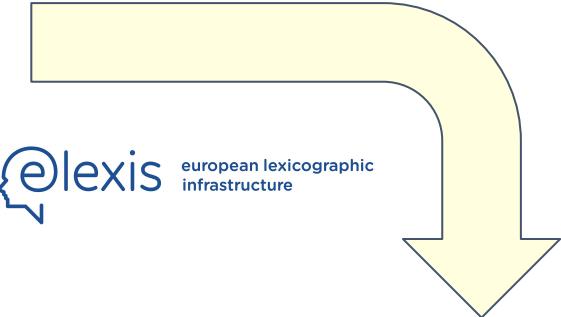
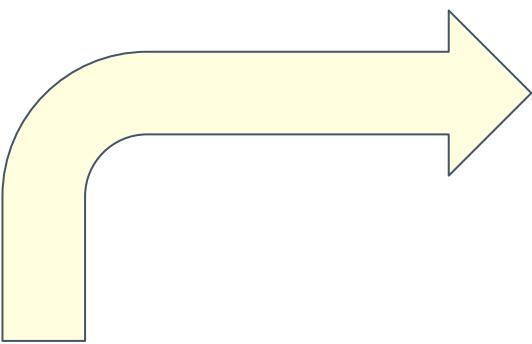
William the Conqueror fortified many castles, throwing up new ramparts, bastions, and towers.

5. (*uncountable, slang, professional wrestling*) The staging of events to appear as real.

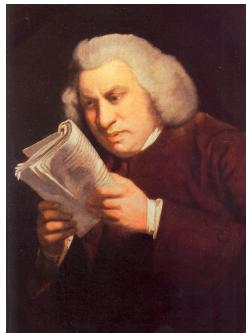
6. (*mining*) Ore before it is dressed.

(Can we [find](#) and [add](#) a quotation of Raymond to this entry?)





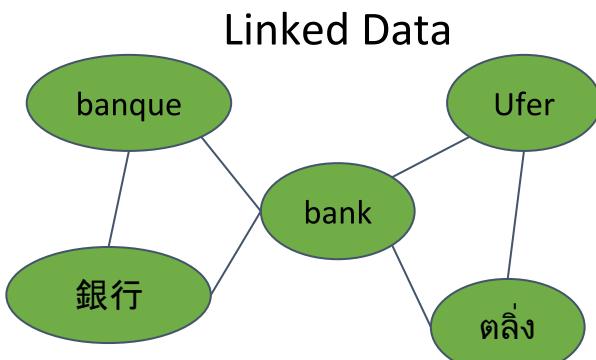
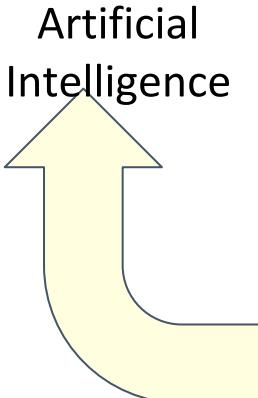
Dictionaries



The Virtuous Cycle of eLexicography

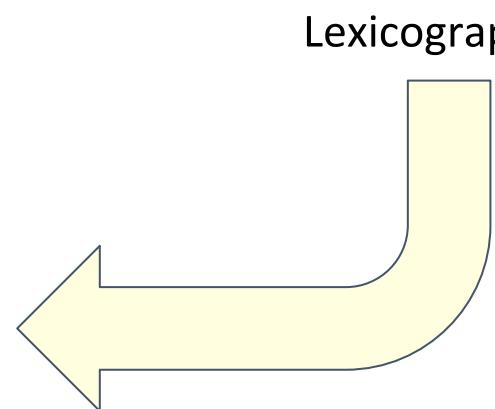


Artificial
Intelligence



Linked Data

Lexicographers



Lexicography

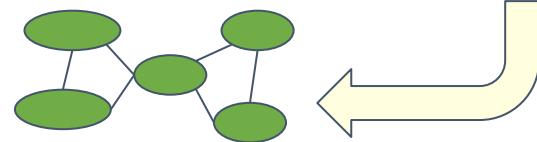


european lexicographic
infrastructure

Objectives:

- Support needs of lexicographic community
- Robust guidelines and best practices for workflows in lexicography
- New methods for retro-digitisation of lexicographic content

Interoperability and Linked (Open) Data



Objectives:

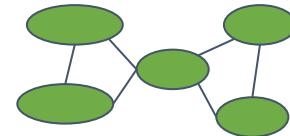
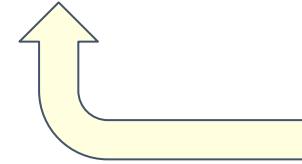
- Interoperability of existing lexical resources
- Establish cross-lingual links between dictionaries
- Validate the quality of dictionaries



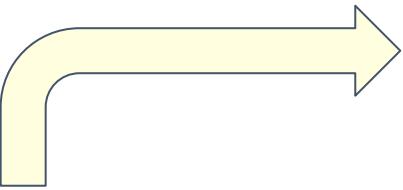
Lexicography for NLP

Objectives:

- Show impact of quality lexicographic data on NLP
 - Word Sense Disambiguation
 - Entity Linking
 - Semantic Parsing
- Address “knowledge acquisition bottleneck” in AI



NLP for Lexicography



Objectives:

- Automatic acquisition of lexicographic data
- Crowdsourcing and gamification
- Enrich dictionaries with multi-modal data

How to link your dictionary

Naisc



Step 1: Get two dictionaries and figure out which senses link

bank

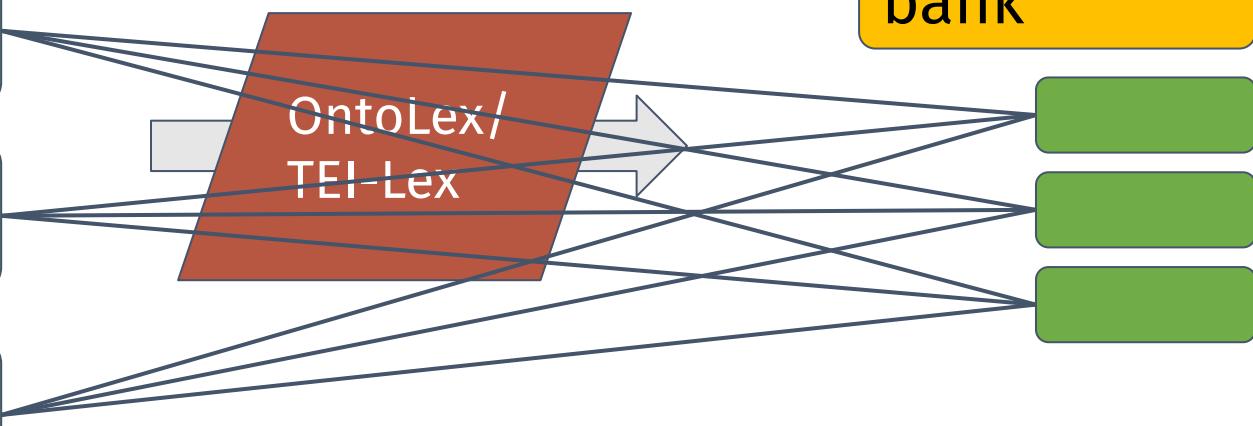
place for
keeping money

side of a river

turn as made by
a plane

bank

OntoLex/
TEI-Lex



Step 2: Extract facts from the entries

bank

Definition: place for keeping money

Example: merchant bank

Translation: 銀行

Step 3: Textual Similarity

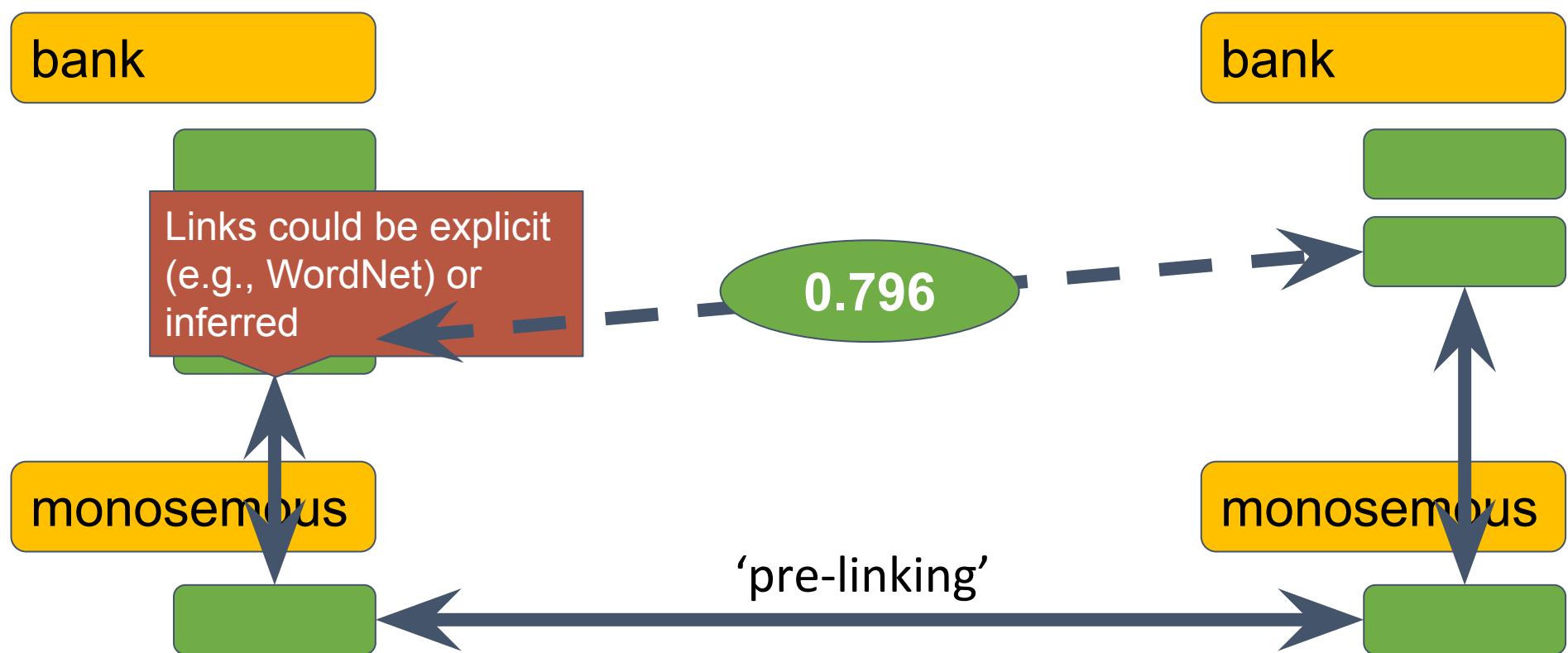
place for keeping money

An institution where one can place
and borrow money and take care of
financial affairs

Text Analysis/
Semantic Analysis/
Deep Learning

0.835

Step 3: Graph (contextual) Similarity



Step 4: Combine Features



0.796

0.853

0.223

0.156

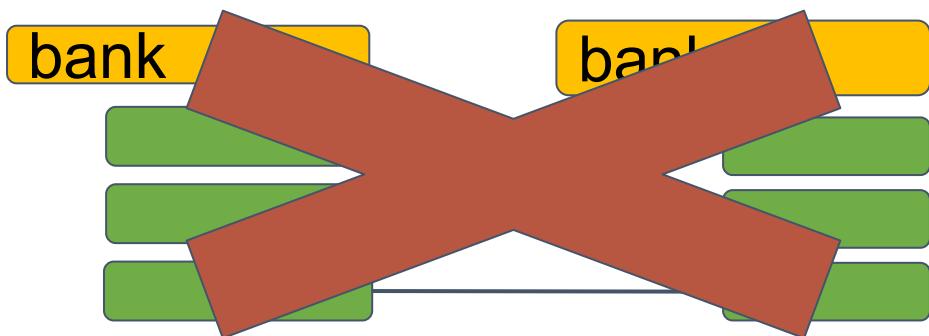
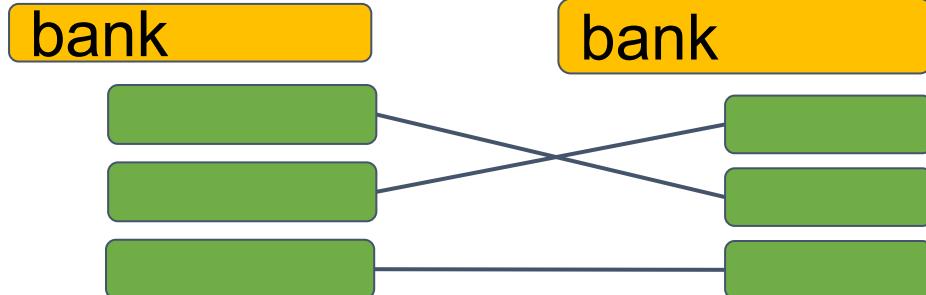
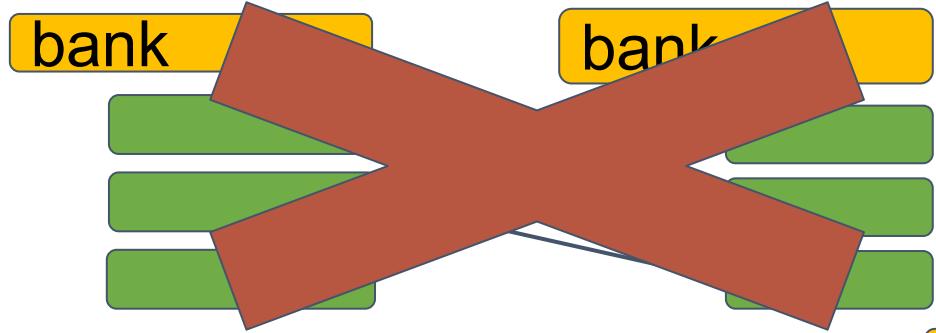
0.999

0.0



0.653 or
Top 20%

Step 5: Constrained optimization of matching



All approaches improve F-Measure from 20% to about 60~75% (on English)

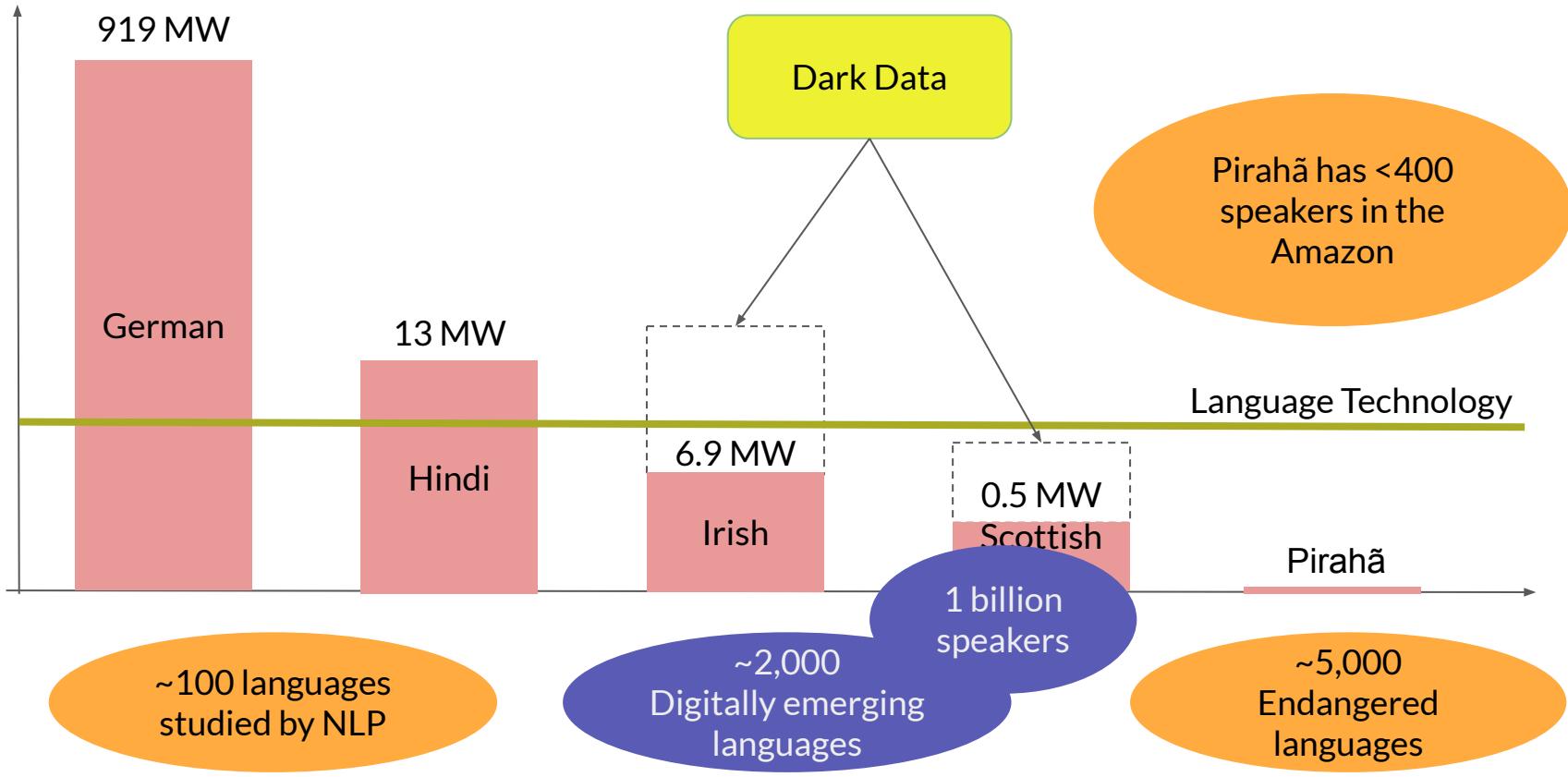
Resource-intensive and requires a lot of training

New benchmark and shared task on English, Dutch, Danish, Spanish, Italian, German, Slovene, Estonian, Bulgarian, Hungarian and Irish.

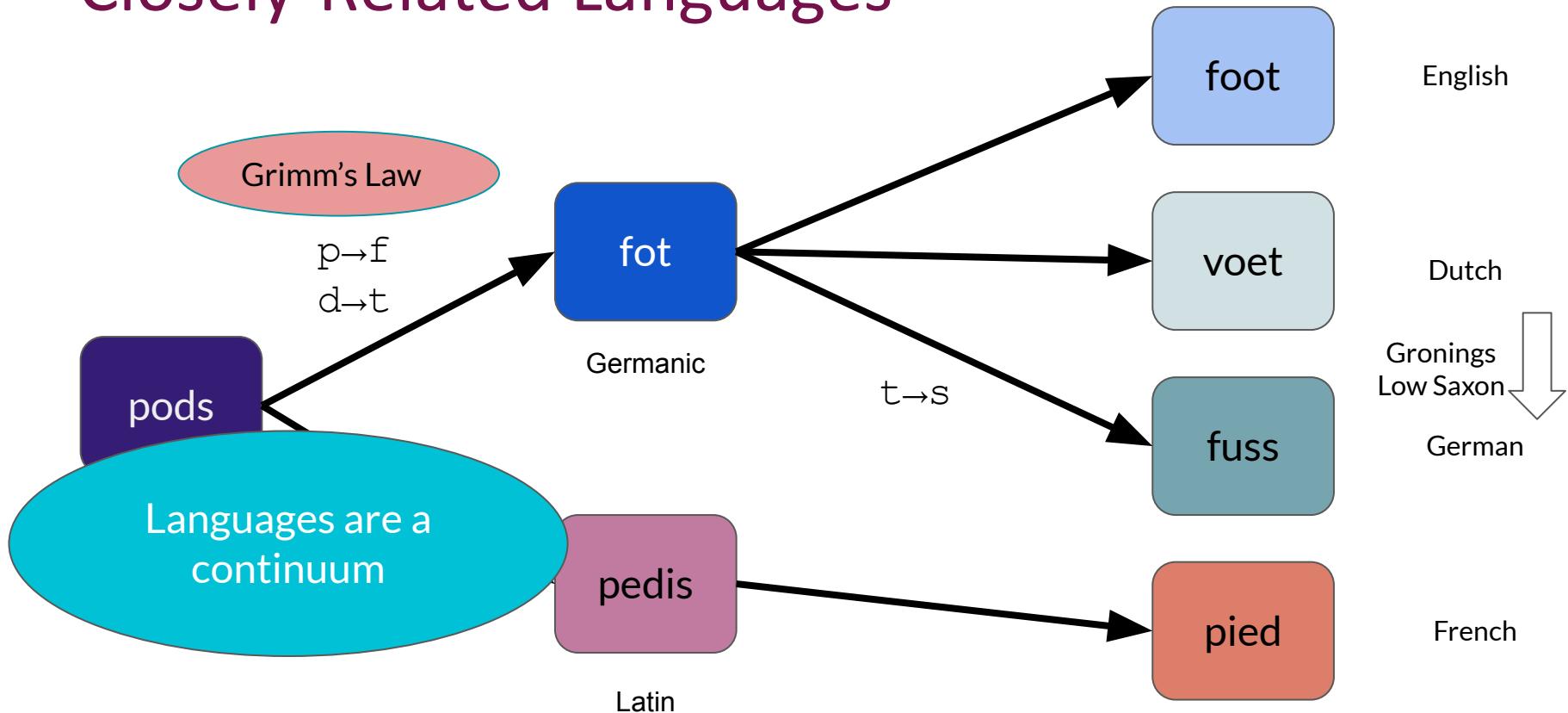
& your language?
If you have **2 open monolingual** dictionaries and about 4-5 hours.

Minoritized and Historical Languages

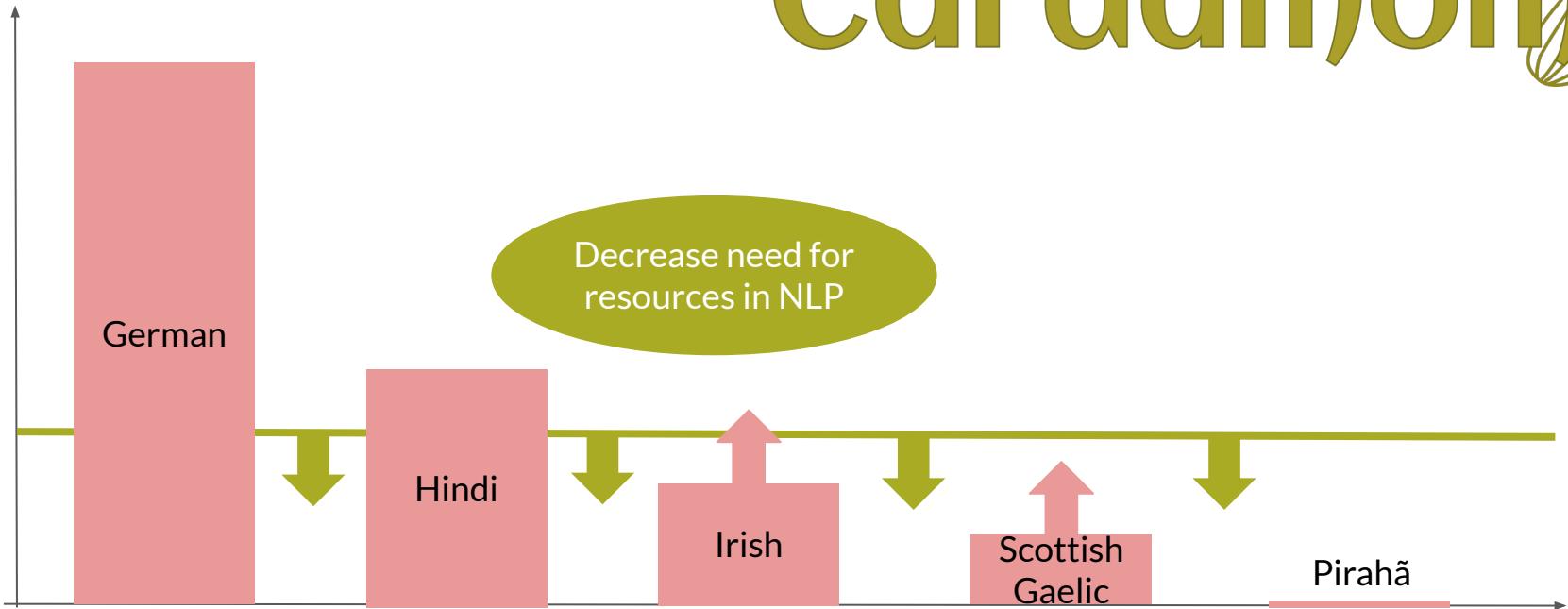




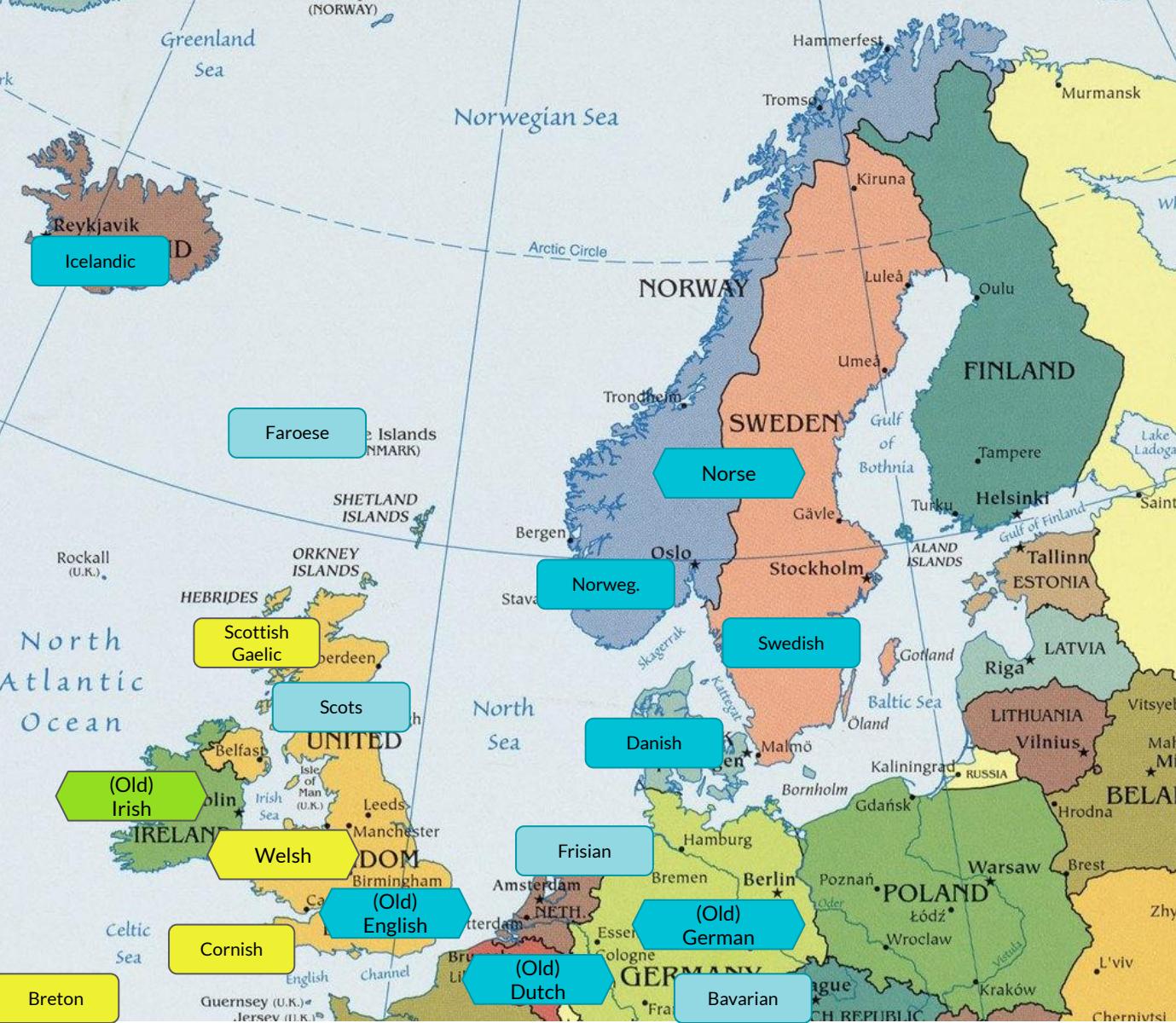
Closely-Related Languages

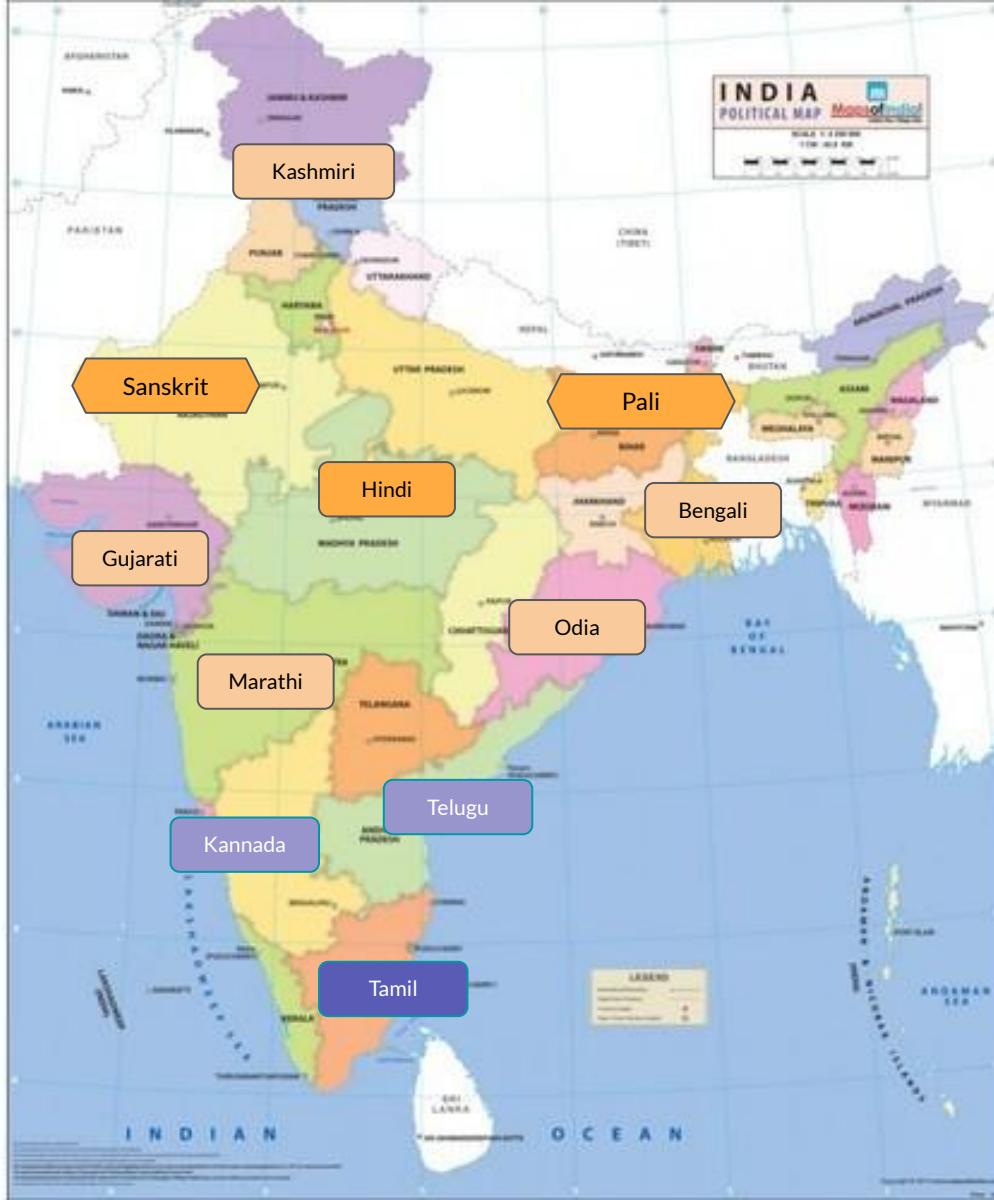


Cardamom



Icelandic	Germanic
Irish	Celtic
Hindi	Indo-Aryan
Tamil	Dravidian
(Old) English	Historical
Swedish.	National
Frisian	Minority





Icelandic	Germanic
Irish	Celtic
Hindi	Indo-Aryan
Tamil	Dravidian
(Old) English	Historical
Swedish.	National
Frisian	Minority

Towards a dictionary of all languages



Polylingual
WordNet

Automatically
constructed dictionary
of 26 EU languages

Expanding wordnets to new languages with
multilingual sense disambiguation. Arcan,
McCrae and Buitelaar, (2016).

Expand to more languages
through dynamic,
morphology-aware word
embeddings

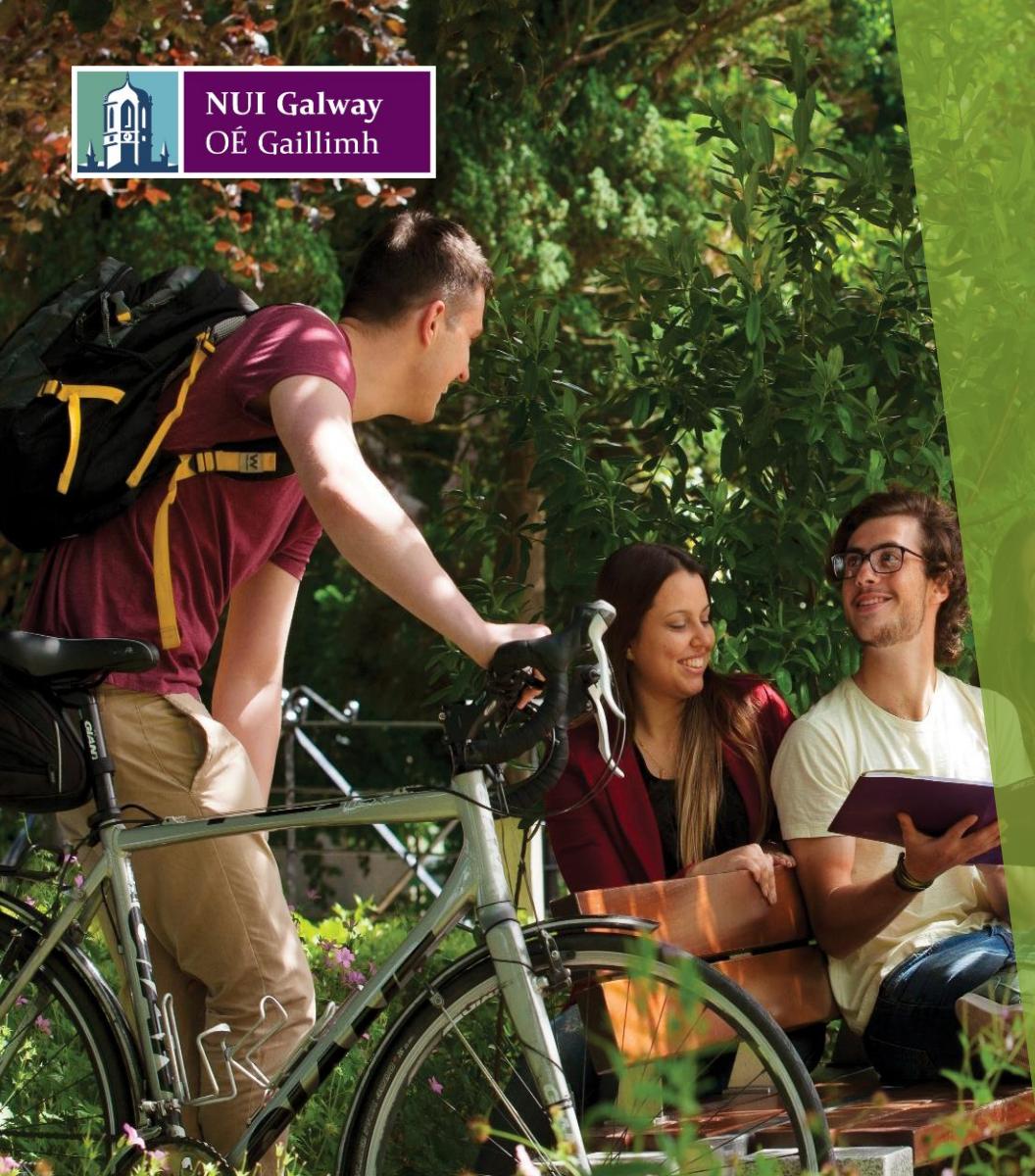
Crowd-sourcing through
computer-aided language
learning

Summary

- Dictionaries are a vital research tool for humanities and NLP
- Linking and integrating is challenging and important
- Link under-resourced languages as closely related languages
 - Historical → Modern
 - Minority→ National



NUI Galway
OÉ Gaillimh



Thanks,
questions?