



# One Model, Many Languages?

An approach to developing global language content

Euralex, Ljubljana, July 2018

Judy Pearsall  
Oxford University Press



# Oxford Global Languages

---

## Key areas:

- 1. Technology**
  - Central platform
  - Internal and external use cases
- 2. Methodological**
  - Single lexicographical framework
  - Appropriate tools
- 3. Ways of working**
  - Iterative, Agile approach
  - Strong collaboration editorial/technology
- 4. Data representation**
  - Single data model
  - Syntactic/semantic interoperability



# Oxford Global Languages

---

## **The vision:**

To bring lexical content online for 100 of the world's languages and make it available to developers, consumers, licensees, and researchers for a wide variety of uses

## **The mission:**

To improve the quality and breadth of global linguistic knowledge and communication, giving voice to all people in a rapidly changing world



# Oxford Global Languages

The screenshot shows the top navigation bar of the isiZulu Oxford Living Dictionaries website. It features a blue header with the 'isiZulu Oxford Living Dictionaries' logo on the left, a search bar in the center with a dropdown menu set to 'ISIZULU - ENGLISH', and 'SIGN IN' and 'ULIMI LWESAYITHI' links on the right. Below the search bar, there are advertisements for 'next' and 'Discover Money Pools'.



Yiba yingxeny  
yesichazamazwi esiphilayo:  
faka igama



Translation strategies quiz:  
test yourself

**Northern Sotho**  
Oxford Living Dictionaries

NORTHERN SOTHO - ENGLISH

Type word or phrase

NORTHERN SOTHO - ENGLISH

ENGLISH - NORTHERN SOTHO

**THERE'S A  
TO CHIP IN WITH FRIENDS.**

Laura  
£70

Valentine  
£70

Christian  
£70

Discover Money Pools



O tseba kanegelo ya Mma Winnie Mandela  
gakaakang? Iteke go marara a rena!



E ba karolo ya pukuntšu ye  
phelago: tsenya lentšu



Tiragalo ya ntlo ye e  
buletšwego go aga  
pukuntšu: Janaware 2018



Bolela: etela lekgotla



**Dr Langa Khumalo at the launch of Oxford  
Global Languages on 2 September 2015**

# Oxford Global Languages

So far...

---

18 languages launched



# Oxford Global Languages

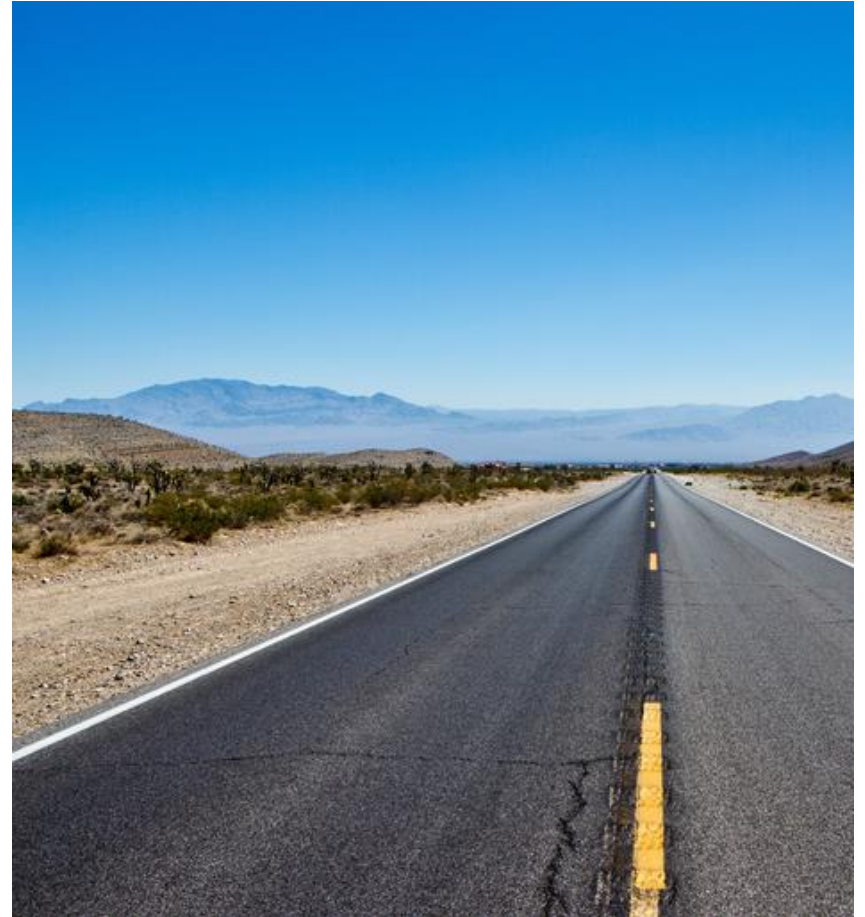
The screenshot shows a web browser window with the URL <https://developer.oxforddictionaries.com>. The page header includes the Oxford Dictionaries logo, navigation links for 'DOCUMENTATION' and 'SUPPORT', and a 'GET YOUR API KEY' button. The main content area is a grid of diverse human faces, with the text 'Oxford Dictionaries API' and 'Enhance your app with our world-renowned dictionary data.' prominently displayed. A blue button labeled 'GET YOUR API KEY' is positioned below the text. At the bottom of the page, a blue cookie consent banner reads: 'We use cookies to enhance your experience on our website. By clicking "continue" or continue using our website, you accept the use of cookies. You can change the setting of cookies at any time.' The banner includes 'CONTINUE' and 'FIND OUT MORE' buttons. The Windows taskbar at the bottom shows various application icons and the system clock indicating 13:56 on 30/06/2018.

<https://developer.oxforddictionaries.com>

# Scaling Oxford Global Languages

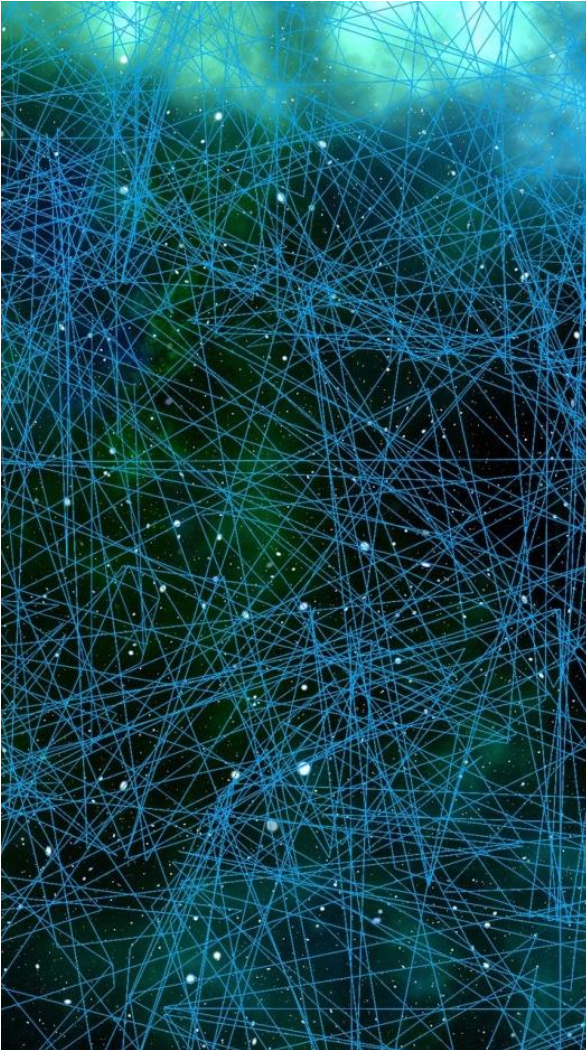
---

- 30 languages by March 2020
- Focus on Indian languages
- New architecture and Data Model
- Content creation model
- Experimental and Agile



# What do we mean by one model?

---



**A standardized representation of  
lexical data and language data**

**Allows terminology, tools, and  
outputs to be interoperable between  
languages**

**Scalable to new languages and data  
features**



# Why one model?

---



**Answers needs of our customers**



**Single system and set of tools**



**Cost efficiency**



**Multilingual linking**



**Efficient data management**

# Main challenges and objections

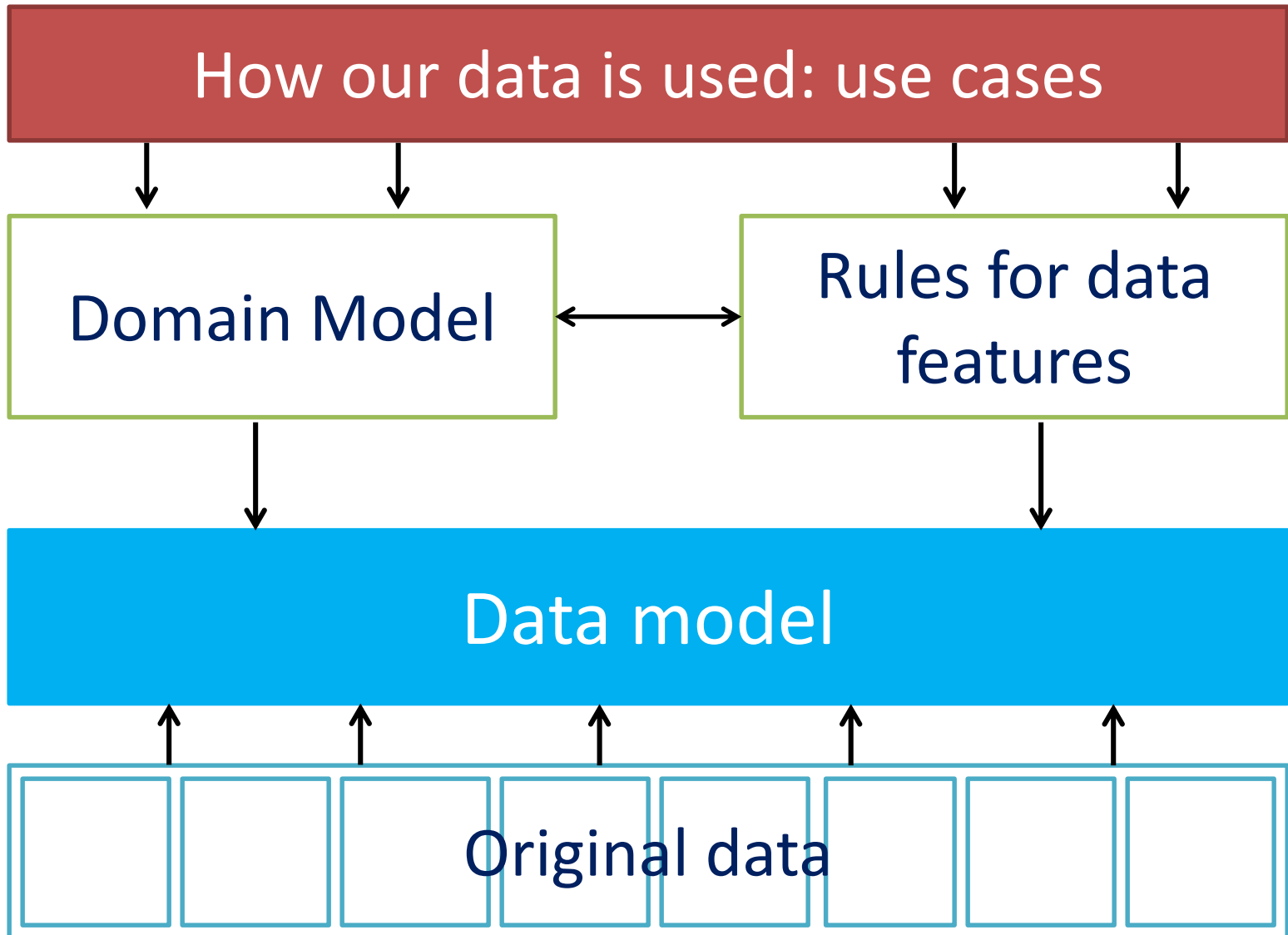
---

- **Size of task**
- **Complex legacy data from disparate sources**
- **Languages are not the same!**
  - **Alignment can flatten valid distinctions**
  - **Potential for false analogy**
- **Dealing with different lexical traditions**
- **Tendency for English/European languages to predominate**



# Dictionaries Domain Model

---



# The print business

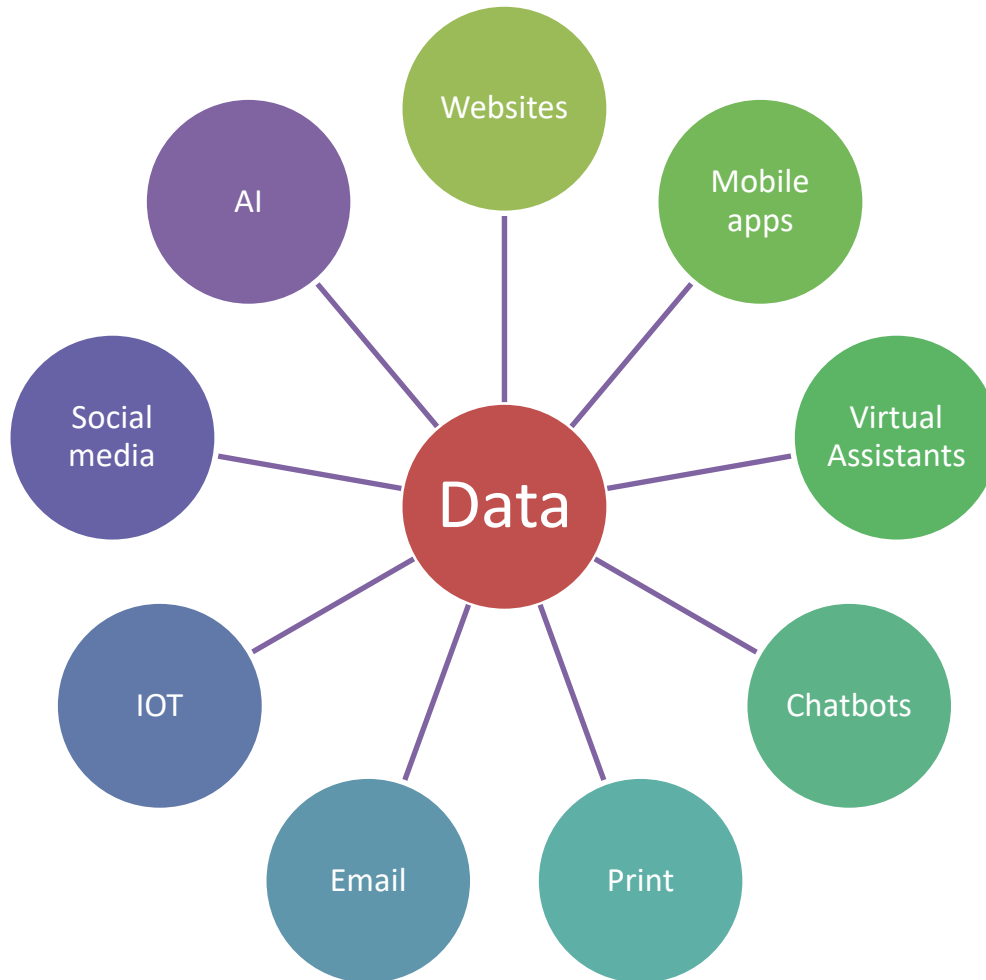
---

Each dictionary is a data silo



# Our business has changed

---

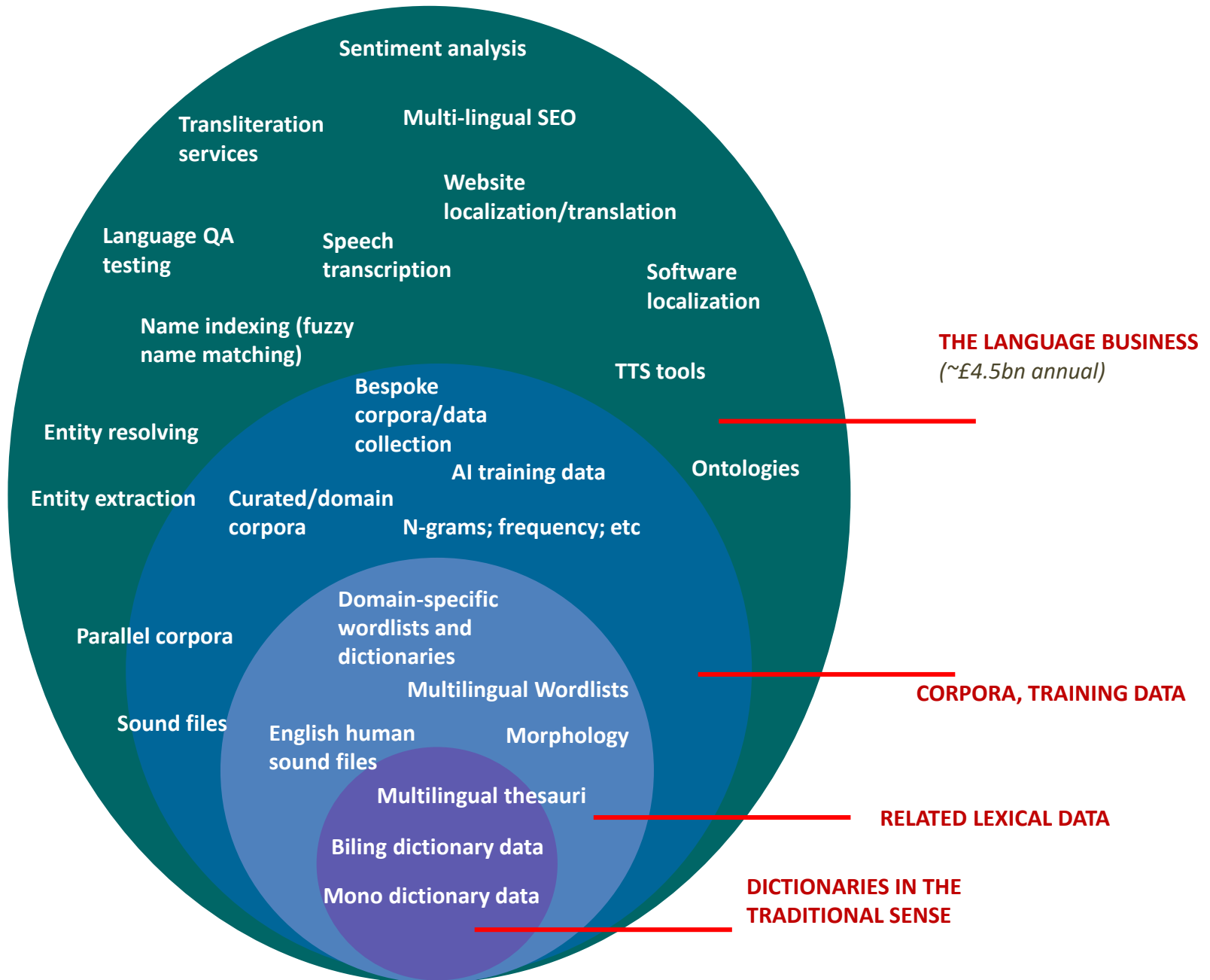


# The market shift

---



# Language Data and Services Sector (not to scale)



# Domain model

A map of the domain to help solve real-world business, data, software, and design problems

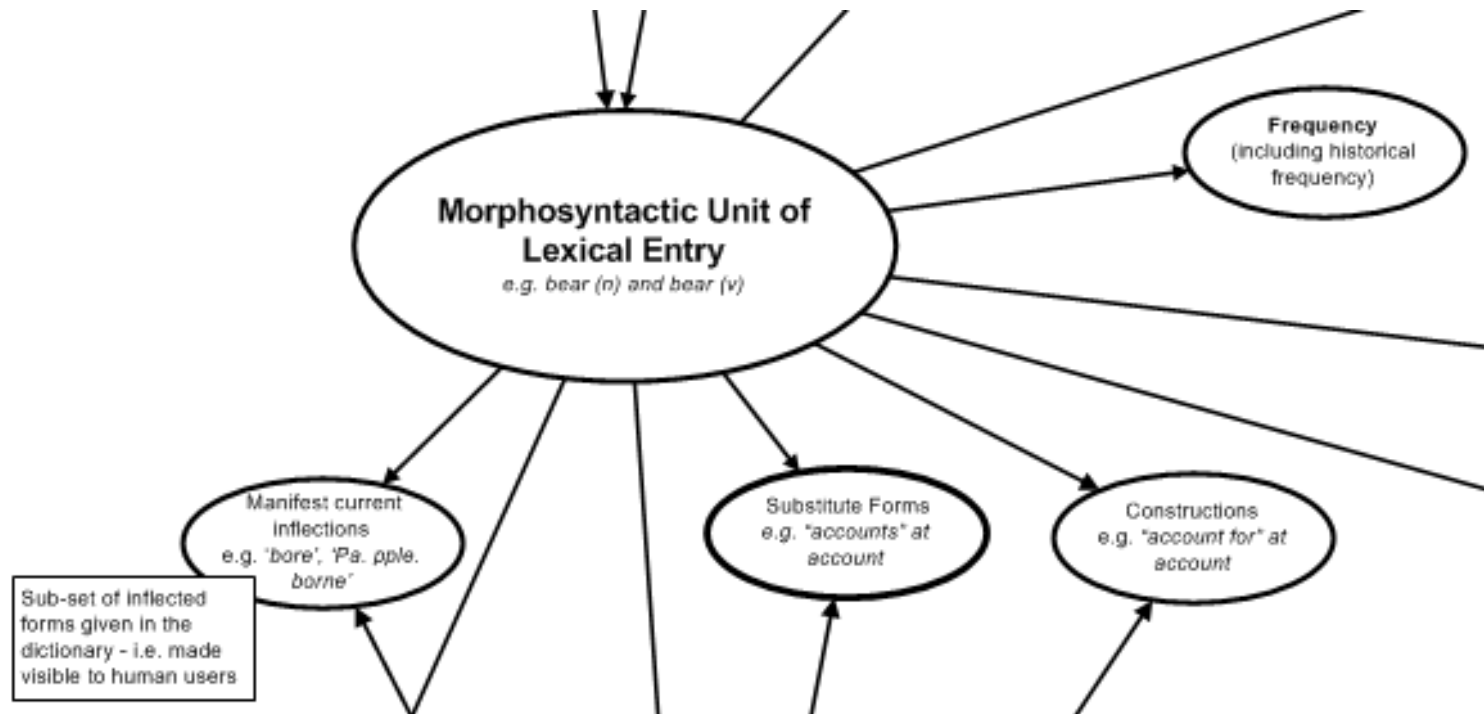
Data  
models

Technology  
decisions

Interface  
decisions

Business  
decisions

Product  
decisions

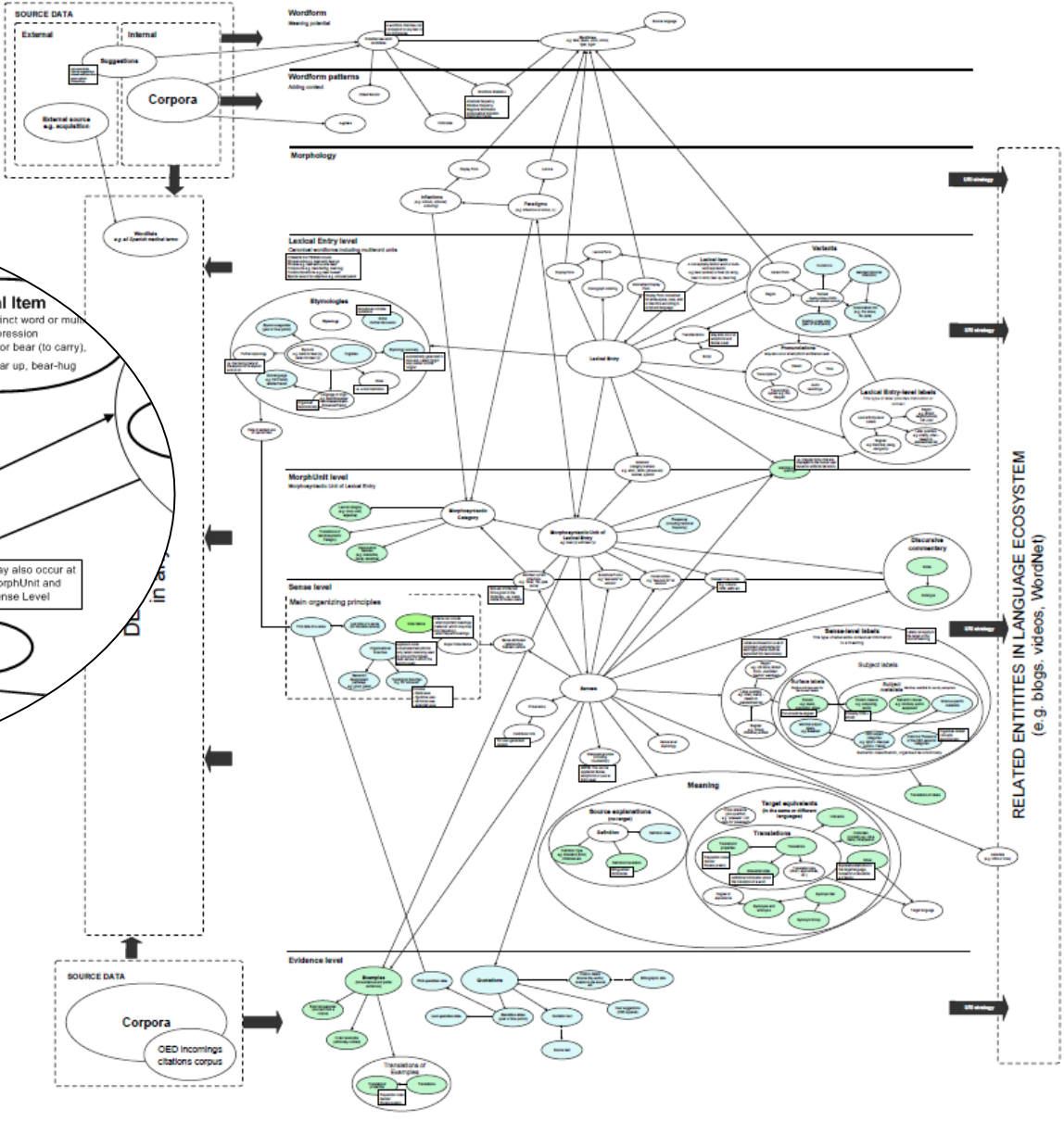
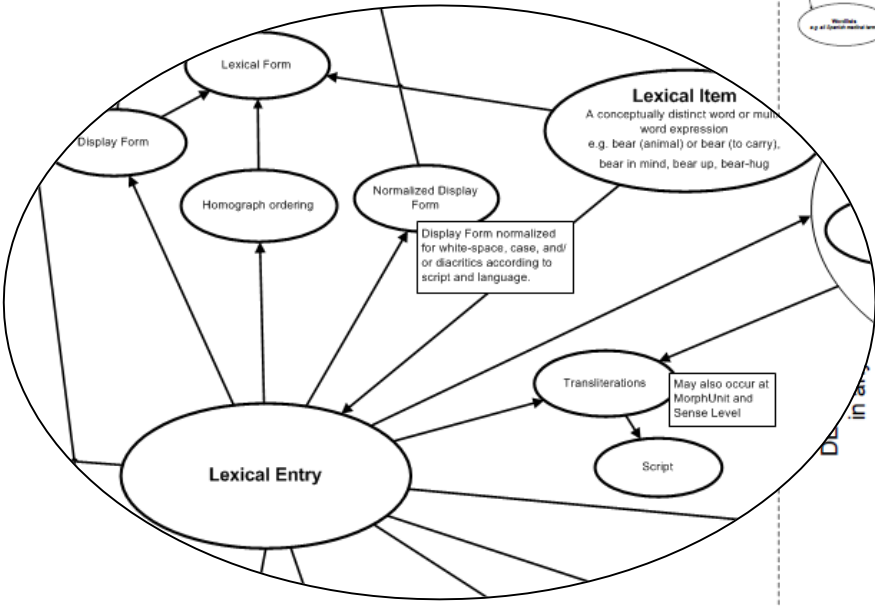




# Domain model

Dictionaries domain model v2.0

Legend External Internal Internal Internal

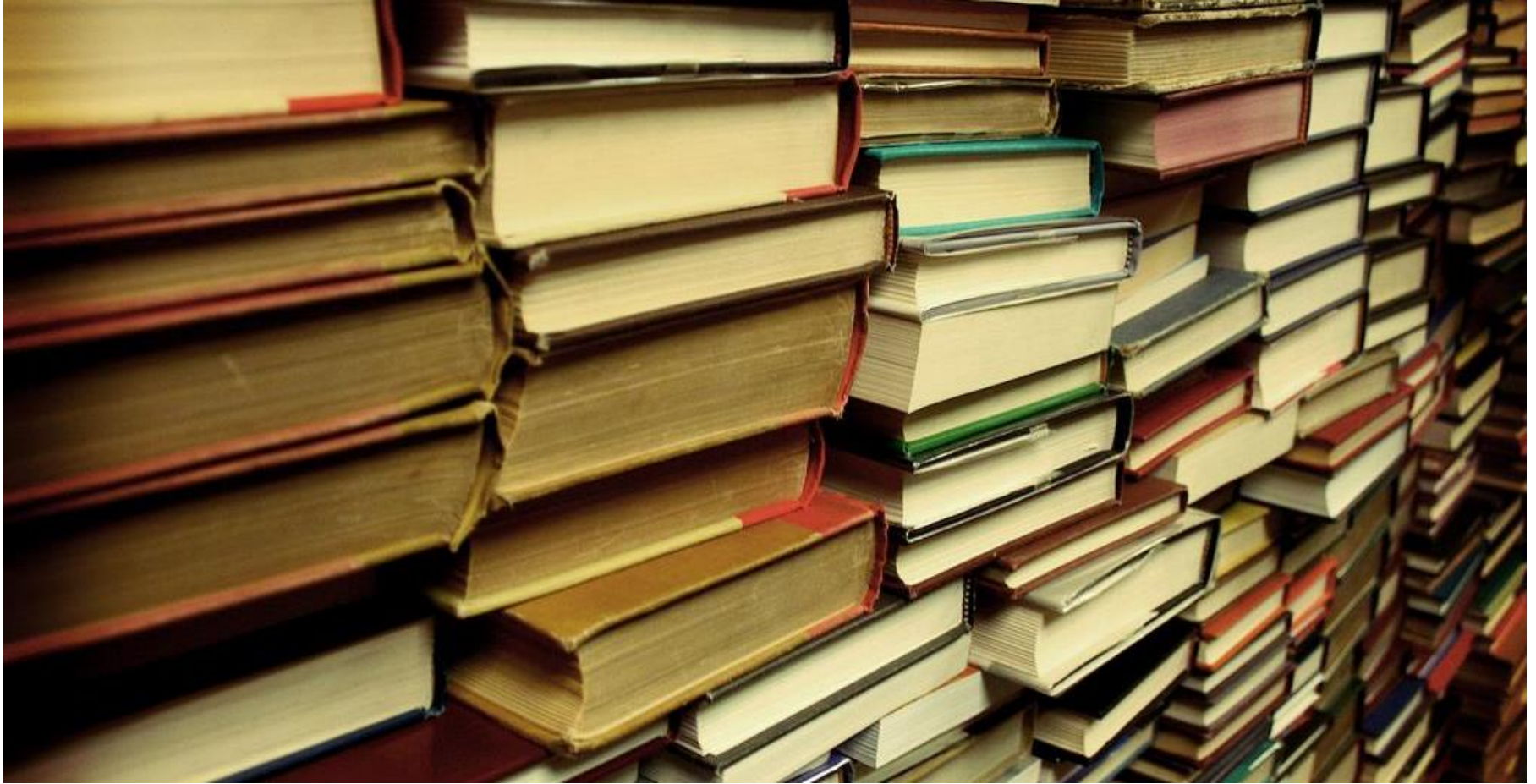


RELATED ENTITIES IN LANGUAGE ECOSYSTEM  
(e.g. blogs, videos, WordNet)

More entities in the ecosystem

# Different data sources

---



# Domain Model: single set of terminology

---

Lemma

Part of  
Speech

Headword

Entry

Label

Word



# Domain Model: print structures

---

## Space-saving treatments in print

Parts of speech combined

**elate** 

---

**VERB**

[WITH OBJECT]

(usually as adjective **elated**)

Make (someone) ecstatically happy.

Words embedded within another item

## **beta rhythm**

---

**NOUN**

[MASS NOUN] *Physiology*

The normal electrical activity of the brain when conscious and alert, consisting of oscillations ( **beta waves**) with a frequency of 18 to 25 hertz.

# Domain Model: print structures

---

## Space-saving treatments in print



Cross references to other words

**wildebeest**

**NOUN**

another term for [gnu](#)

**gnu**

Pronunciation [/\(g\)nju:/](#)  [/\(g\)nu:/](#) 

**NOUN**

A large dark antelope with a long head, a beard and mane, and a sloping back.

Genus *Connochaetes*, family Bovidae: two species, in particular the abundant brindled gnu or blue wildebeest (*C. taurinus*)

Also called [wildebeest](#)

# Domain Model: derivatives and compound forms

Definition of *asylum* in English:

**asylum** 

**NOUN**

- 1 (also **political asylum**) [*mass noun*] The protection granted by a state to someone who has left their home country as a political refugee.

**political asylum** *n.* asylum granted by a country to a political refugee from another country.

1852 *Times* 28 Feb. 4/4 These cantons, with their free press, their political asylum, and their creed, are intolerable to the jealous eye of neighbouring despotism.

Thesaurus »

**asilo**


PT. 

BR. PT. 

Translation of *asilo* in English:  
masculine noun

(proteção)

**asylum**

*asilo político* PT. 

BR. PT. 

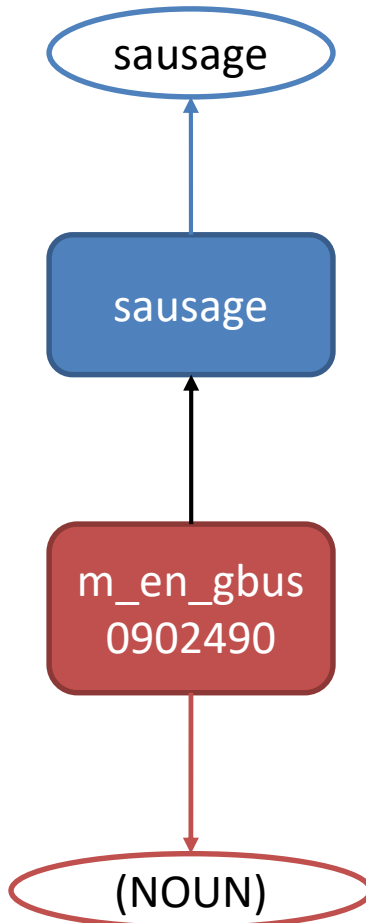
**political asylum**















**political asylum**

Translation of *political asylum* in Russian:  
noun

**политическое убежище** 

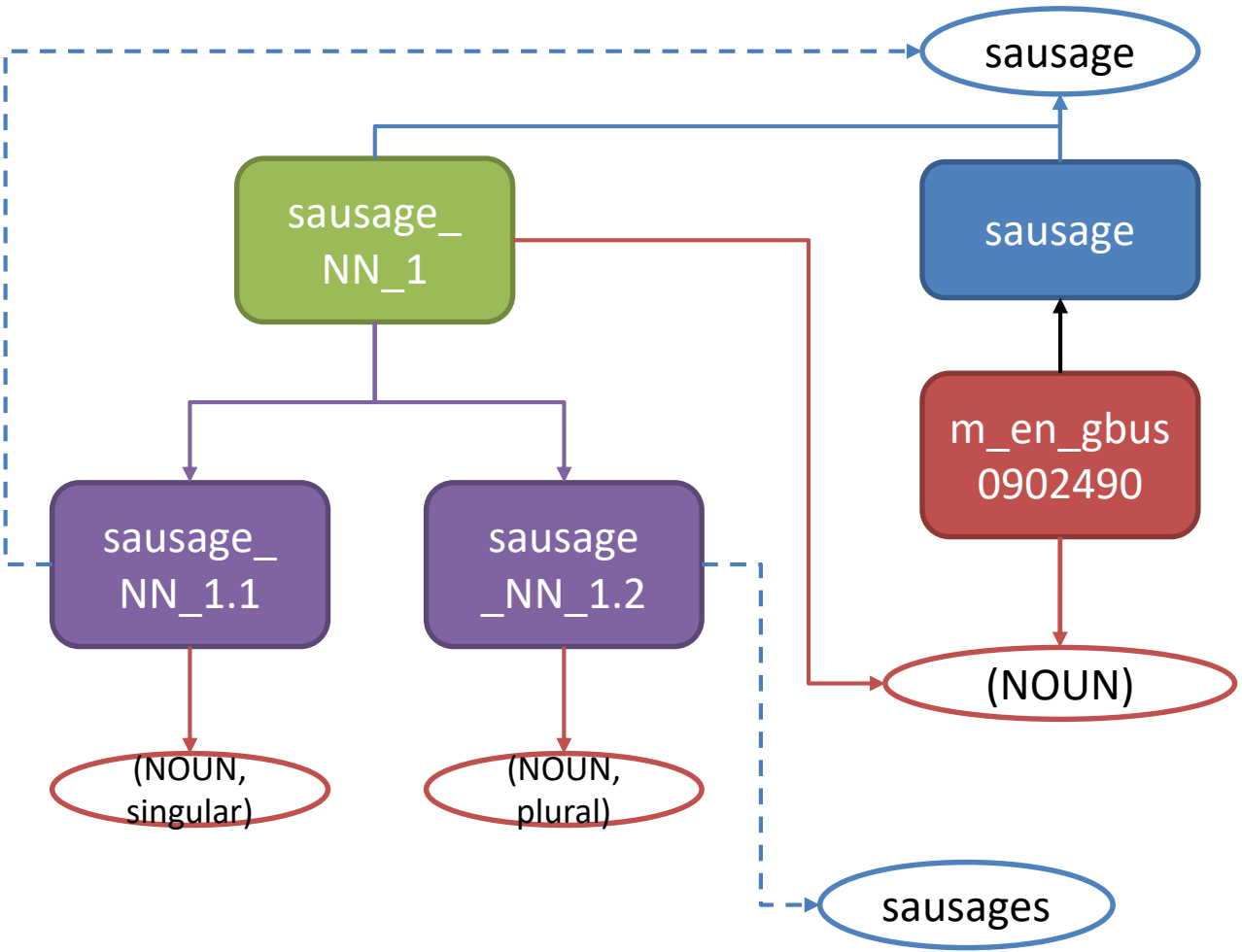
# Lexical entry, Lemma



-  Lexical Entry
-  MorphUnit
-  Paradigm
-  Inflection
-  Wordform
-  MSCat
-  Has Lexical Entry
-  Lemma/Display Form
-  y Form
-  Normalized Form
-  Form
-  Has MsCat
-  Has Inflection
-  Has Wordform

# Lexical entry, Lemma + Morphology

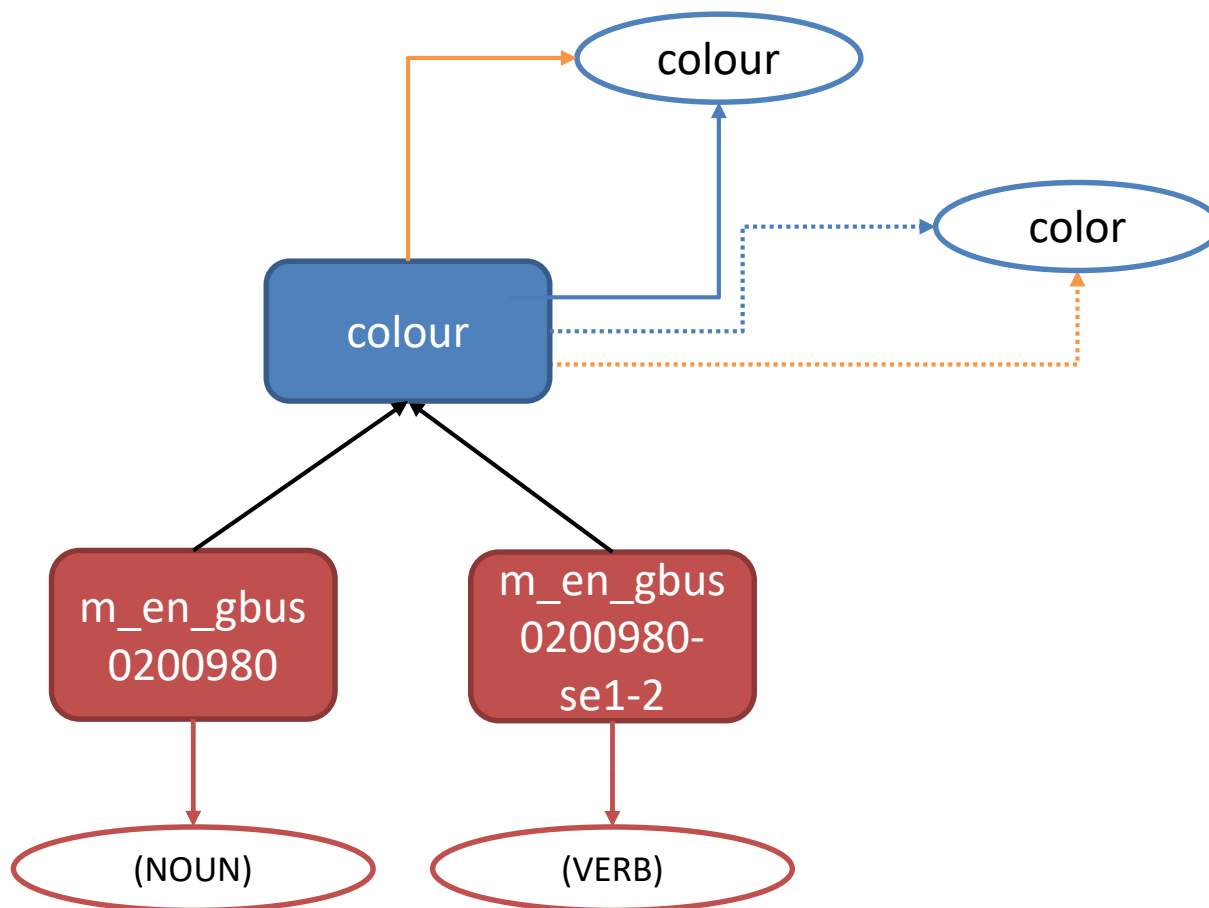
- Lexical Entry
- MorphUnit
- Paradigm
- Inflection
- Wordform
- MSCat
- Has Lexical Entry
- Lemma/Display Form
- y Form
- Normalized Form
- Form
- Has MsCat
- Has Inflection
- Has Wordform





# Lexical entry, lemma + Variant

- Lexical Entry
- MorphUnit
- Paradigm
- Inflection
- Wordform
- MSCat
- Has Lexical Entry
- Lemma/Display Form
- Normalized Form
- Has MsCat
- Has Inflection
- Has Wordform



# Domain Model Terminology

---

- **Lexical Item**
  - The abstract concept of a “word” or “discrete linguistic item” as distinct from the canonical spelling of it
- **Display Form**
  - the **wordform** typically used to write a **Lexical Item**
- **Lexical Form**
  - The combination of wordform and homograph number that uniquely identifies a **Lexical Item** (e.g. "march 1")

# Tonality

## Igbo:

a standardized code is used

Igbo											English		
Word							Sense				Word		
Word	Phon	Inflection	Inflection 2	Gra1	Gra2	Wfm	Indicator	Cat	Reg	Sty	Prep	Note	Word
✓ ᠔ᠠ gbapu	AE			2	0								᠔ᠠ rupture
													᠔ᠠ puncture

## Yoruba:

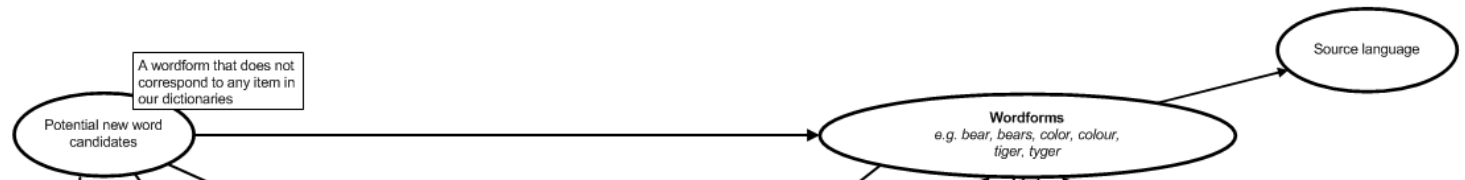
a system of accents is used as part of the translation

	A	B	C	I	N	V
	No.	ID	type	English	Indicator	LOTE
1						
2	1	11649	sense	convention	norm	ilànà
3	2	11650	sense	convention	assembly	ipèṣo
4	3	11651	sense	execution	carrying out	ìse
5	4	11652	sense	execution	killing	pípa
6	5	11653	sense	accessible		bí a ṣe lè débì kan
7	6	11654	sense	mate	companion, fel	egbé

# Domain Model: corpora

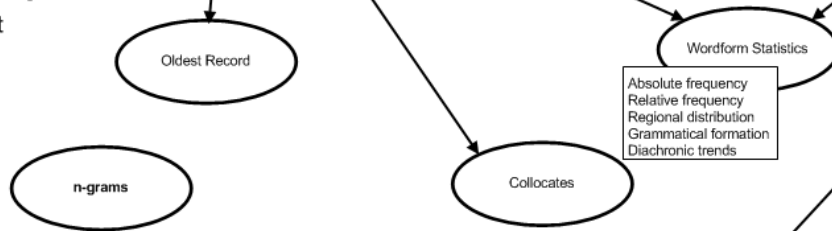
## Wordform

Meaning potential

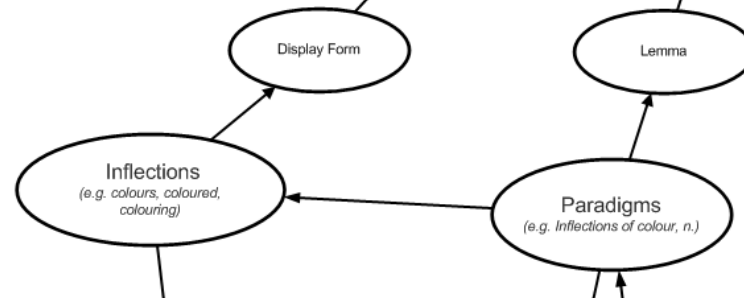


## Wordform patterns

Adding context



## Morphology



# Basic Rules for All Dictionaries

---

## 9. DISPLAY FORMS - BRAD

### GENERAL PRINCIPLES

1. Schemas should change rarely
2. Controlled vocabularies should be used and adhered to
3. The same tag names should be used for every dataset
4. Mixed content should be avoided
5. Every tag / object / item should be one thing and only one thing

### What is a display form?

A formal representation of the lexical item in a particular language, used as the header of a particular lexical entry. This may be the way in which it is typically written in a language, or it may differ (for example Arabic display forms will have vowels).

Each lexical entry has its own display form and has only one display form. For example, march<sup>1</sup> has the display form "march", and march<sup>2</sup> has the display form "march".

A lexical entry can only have one display form, but there may be variant forms (e.g. **color** as a variant of **colour**), which we might choose to display instead of the display form for particular use cases. We will discuss this further under "variants".

## 12. HOMOGRAPHS - BRAD

### GENERAL PRINCIPLES

1. Schemas should change rarely
2. Controlled vocabularies should be used and adhered to
3. The same tag names should be used for every dataset
4. Mixed content should be avoided
5. Every tag / object / item should be one thing and only one thing

### What is a Homograph?

- A homograph is lexical item that shares the display form with another lexical item, e.g. **bank** "money institution" versus **bank** "river bank".
- A near homograph is a lexical item that shares a normalised form with another lexical item, but differs in upper/lower case, diacritics, punctuation.  
E.g. **polish** vs. **Polish**, **rose** vs. **rosé**, Urdu **جَل** vs. **جُل** vs. **جُلّ**, **am** vs. **a.m.**, **resign** vs. **re-sign**
  - In abugidas (e.g. most Indian scripts), display forms with differences in diacritics are treated as entirely separate lexical items, not as homographs or near-homographs. E.g. Hindi **जन** and **जेन**.

# Rules in Action

*Ex. 1) If a variable display form is provided, the data must allow expanded versions of all implied lexical forms*

## swipe right (or left)

### PHRASE

#### *informal*

(on the online dating app Tinder) indicate that one finds someone attractive (or unattractive) by moving one's finger to the right (or left) across an image of them on a touchscreen.

*'I swiped right, but sadly for me, she swiped left'*

*figurative* 'are elephants more likely to 'swipe right' when they see mates with longer trunks?'

## bloody (or bloodied) but unbowed

### PHRASE

Proud of what one has achieved despite having suffered great difficulties or losses.

*Ex. 2) Separate words with multiple morphosyntactic categories*

## al fresco



### ADVERB & ADJECTIVE

In the open air.

*[as adjective]* 'an al fresco supper'

*[as adverb]* 'in the unlikely event of some sunshine you can even dine al fresco'

# Master Data Model

---

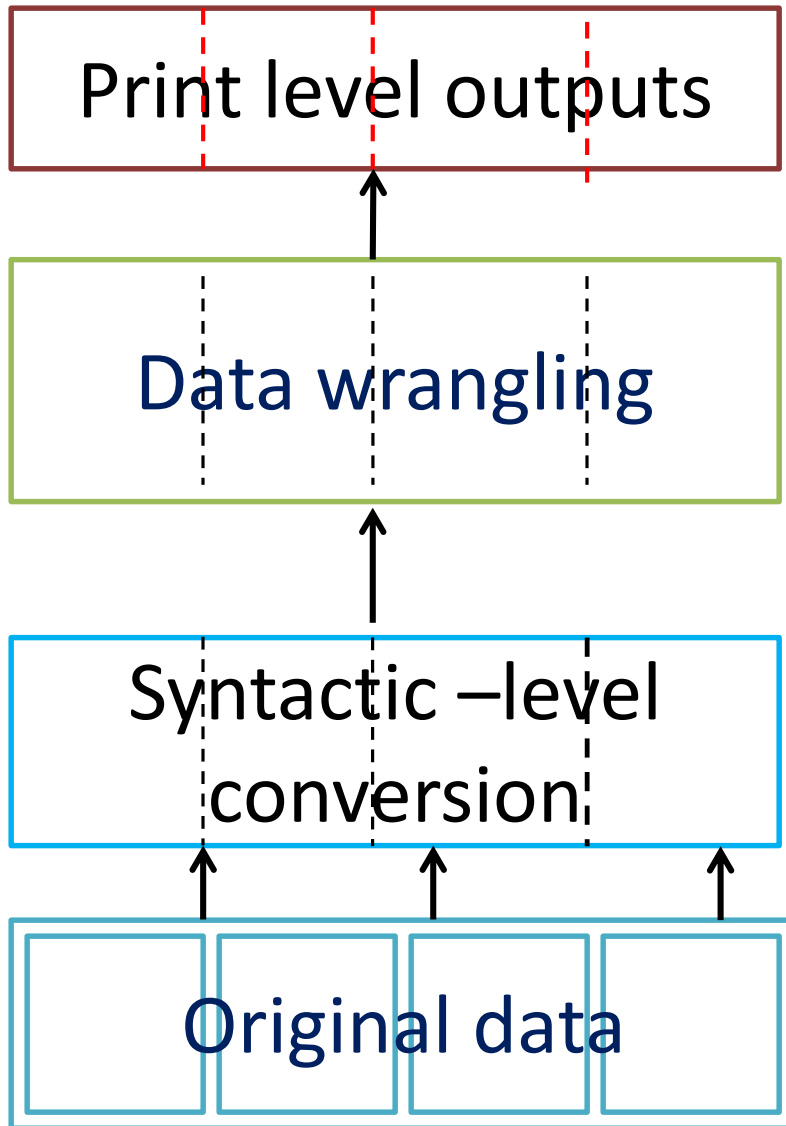
## What is it?

- A data model for JSON, applicable to all types of lexical data.
- Takes the Domain Model as its base, and follows the principles of the defined Rules.
- Covers monolinguals, bilinguals, thesauruses, morphology, wordlists

```
{
  "$schema": "../schema/sense.schema.json",
  "dataset": "en-gb_dict",
  "id": "m_en_gbus0290820.006",
  "idSource": "DPS",
  "morphUnitRef": "m_en_gbus0290820",
  "senseOrder": 1,
  "definitions": [
    {
      "type": "standard",
      "text": ["likely to have an unfortunate and inescapable outcome; ill-fated"]
    },
    {
      "type": "short",
      "text": ["likely to have unfortunate and inescapable outcome"]
    }
  ],
  "exampleRefs": [
    "01234567-1234-1234-1234-000000000004"
  ],
  "datasetCrossLinks": [
    {
      "targetType": "morphUnit",
      "targetDataset": "en_thes",
      "target": "t_en_gb0004383"
    }
  ]
}
```

# Data harmonization, part 1

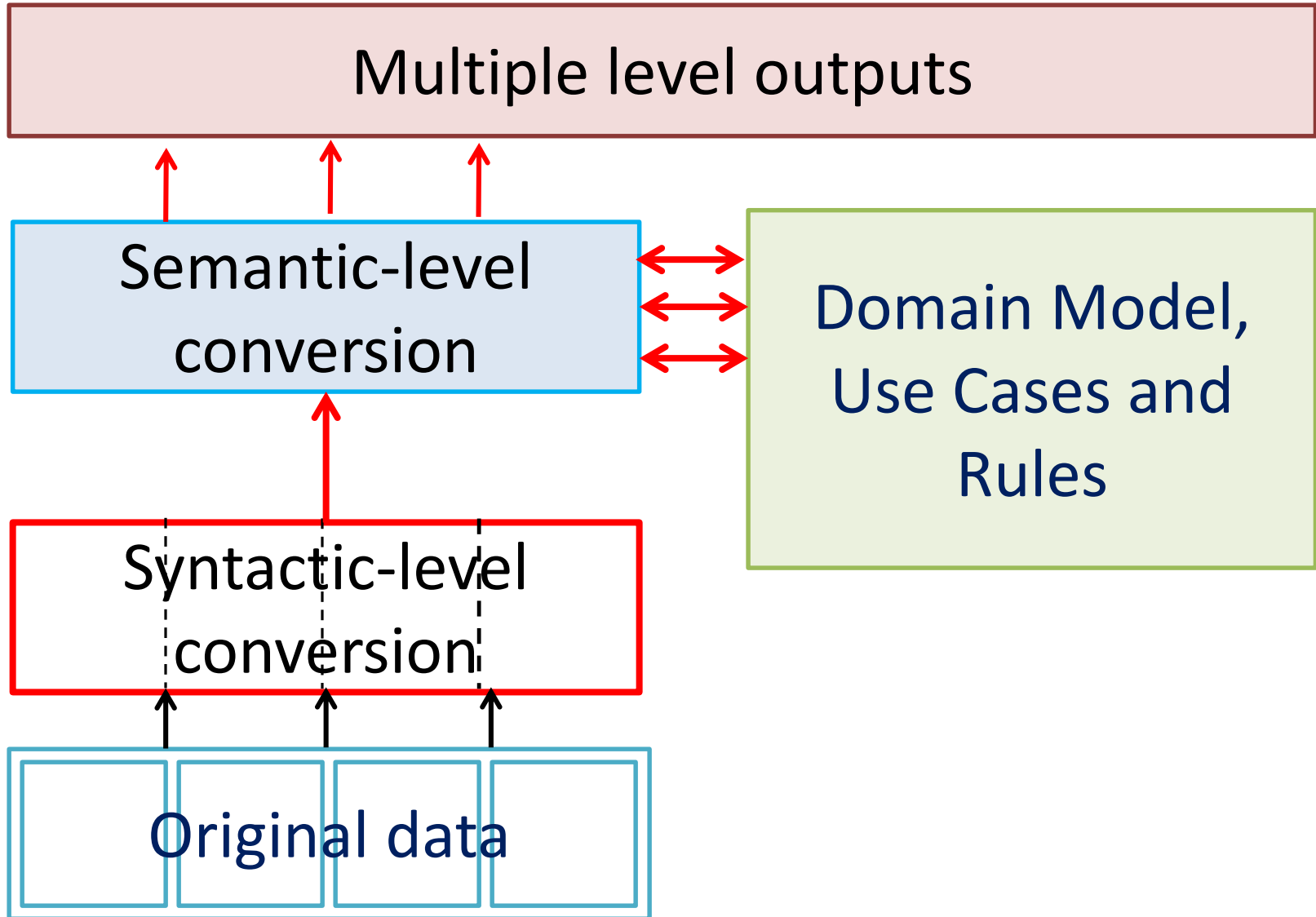
---





# Data harmonization, part 2

---



# New editing tools

The screenshot shows a web-based lexical editing interface. On the left is a dark blue sidebar with navigation icons. The main area is divided into a top header, a left sidebar, and a central editing pane. The top header shows 'DATASET: EN\_RO' and a search bar. The left sidebar contains a 'WORKLIST: 4 / 4' and a list of items: 'cat', 'help', 'helper', and 'water'. The central editing pane is for the word 'water' and includes sections for 'PRONUNCIATIONS', 'TRANSCRIPTIONS' (with IPA and TEXT tabs), and 'LEXICAL ENTRY MORPHUNIT'. The 'MSCAT' section is expanded to show 'LEXICAL CATEGORY' with a dropdown menu where 'noun' is selected. On the right, a light blue box contains a disclaimer, followed by the word 'water' and its IPA transcription. Below this are sections for 'ADJECTIVE', 'NOUN', and 'VERB', each with a list of related terms and their part-of-speech tags.

DATASET: EN\_RO

Quick search Tamara Bowler

WORKLIST: 4 / 4

cat help helper water

DISPLAY FORM water

PRONUNCIATIONS

TRANSCRIPTIONS

IPA TEXT 'wɔːtə(r)

LEXICAL ENTRY MORPHUNIT

MSCAT

LEXICAL CATEGORY

Select one (required)

- conjunction
- contraction
- determiner
- ideophone
- idiomatic
- interjection
- noun**
- summat

LEXICAL ENTRY MORPHUNIT

Please note that this is an early prototype, and only a limited number of features have been mapped. The data shown is almost certain to be incomplete at this stage.

**water**

'wɔːtə(r)

**ADJECTIVE**

1 apă

**NOUN**

1 medicine nautical  
(med.) lichid amniotic  
(naut.) apă (dintr-un râu, lac etc.)

**NOUN**

1 cu apă  
de apă  
de apă  
de prelucrare a apei

**VERB**

1 a uda  
a iriga  
a adăpa

**VERB**

1



# Thank you

With special thanks to Khalil Ahmed, Imogen Foxell, Marie-Claire Holmes, Ruth Madder, Roser Sauri, Richard Shapiro, Angus Stevenson and all the Oxford Global Languages team

Judy Pearsall  
Oxford University Press



# Appendix

---