Colexification Across the Globe

Teresa Davison

Table of contents

- Motivation
- Dataset exploration
- Data cleaning
- Feature extraction
- Models + Analysis
- Findings & future questions

Motivation

What is colexification?

When two or more distinct meaning share the same word form in a language.

- **Russian**: *нога* (noga) means both leg and foot.
- **English**: *to know* means knowing facts as well as being familiar with something.

Colexification patterns vary across languages

- English has only one form ‘to know’ while other languages like French, Spanish and German make a distinction.
- The concepts that are colexified can be close in meaning or more abstractly linked.
- *Дух* (spirit, breath) versus *Geist* (soul, mind)

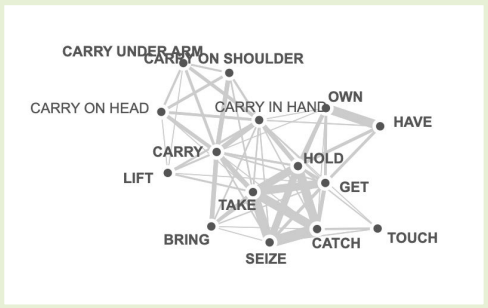
Are these patterns predictable?

Can the similarity between the types of words being colexified in each language be used to predict language family or macroarea?

Dataset - CLICS3 Database

What is CLICS3?

- The third installation of the Database of Cross-Linguistic Colexifications.
- Online interface but you can access underlying SQL database & networks.
- It provides:
  - List of concepts with links to lists of colexified concepts
  - Networks of related concepts (subgraphs)



Concepts

Showing 1 to 100 of 2,919 entries

← Previous 1 2 3 4 5 Next →

Details	Name	# varieties	# colexifications	Infomap	Subgraph
	Search	Search	Search		
more	GIVE	2441	41	GIVE	Subgraph GIVE
more	SAY	1561	37	SPEAK	Subgraph SAY
more	GET	753	28	SEIZE	Subgraph GET
more	EAT	2676	28	DRINK	Subgraph EAT
more	KILL	2125	28	KILL	Subgraph KILL

Dataset

- Extracted from SQL database
- Language, Form and Parameter DataFrames
  - Connected through IDs
  - 30 source datasets collated
- 3,248 **languages** represented
  - 6 Macroareas
  - 202 language families
  - Latitude and longitudinal data
- Over 2900 **concepts**
  - Semantic field, ontological category
- Almost 1.5 million **forms**
  - Different datasets have different conventions: transliteration, IPA, original script...

Form df

	ID	Glottocode	Concepticon_ID	dataset_ID	ID	Local_ID	Language_ID	Parameter_ID	Value
0	Venetianstd	vene1258	1692	logos	Venetianstd-350_friday-1	None	Venetianstd	350_friday	divendres
1	Hindistd	hind1269	1692	logos	Hindistd-350_friday-1	None	Hindistd	350_friday	शुक्रवार
2	Romagnolstd	roma1328	1692	logos	Romagnolstd-350_friday-1	None	Romagnolstd	350_friday	vèner
3	Latinstd	lati1261	1692	logos	Latinstd-350_friday-1	None	Latinstd	350_friday	Venerisdies
4	Galicianstd	gali1258	1692	logos	Galicianstd-350_friday-1	None	Galicianstd	350_friday	venres

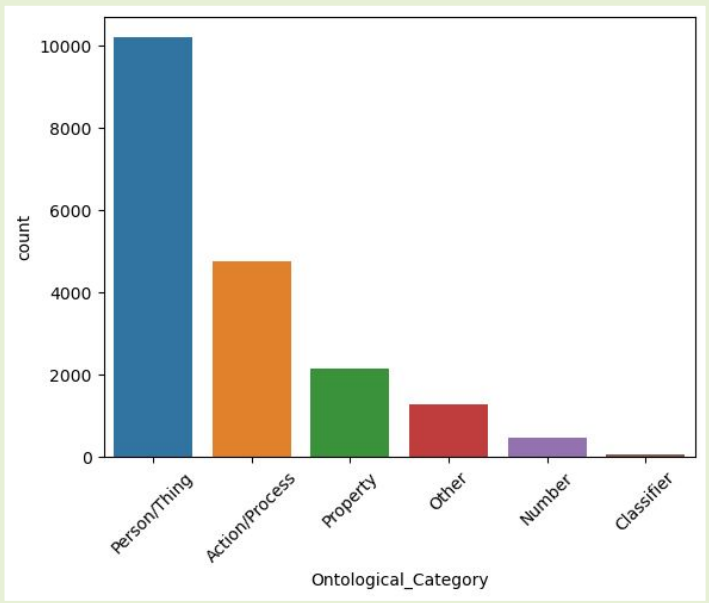
+

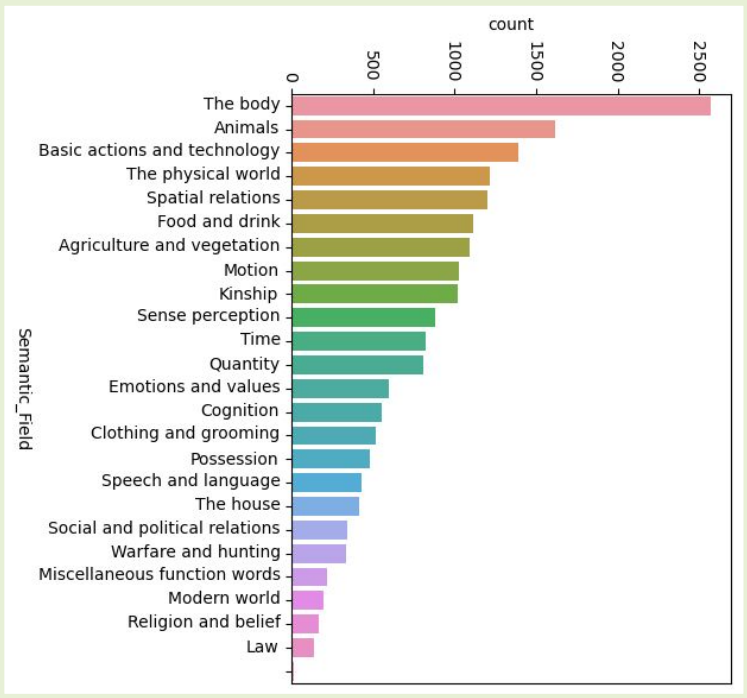
Colex_Key
vene1258_divendres
hind1269_शुक्रवार
roma1328_vèner
lati1261_venerisdies
gali1258_venres

Parameter df

	ID	Name	Concepticon_ID	Concepticon_Gloss	dataset_ID	Ontological_Category	Semantic_Field
0	3_earthgroundsoil	earth=ground, soil	1228	EARTH (SOIL)	logos	Person/Thing	The physical world
1	4_dust	dust	2	DUST	logos	Person/Thing	The physical world
2	5_mud	mud	640	MUD	logos	Person/Thing	The physical world
3	7_mountainhill	mountain, hill	2118	MOUNTAIN OR HILL	logos	Person/Thing	The physical world
4	8_cliffprecipice	cliff, precipice	618	PRECIPICE	logos	Person/Thing	The physical world

Distribution of concept types





Preparing data

	ID	Name	Glottocode	Glottolog_Name	ISO639P3code	Macroarea	Latitude	Longitude	Family	dataset_ID	Form_count
0	Venetianstd	Venetian-std	vene1258	Venetian	vec	Eurasia	45.503581	12.214167	Indo-European	logos	625.0
1	Hindistd	Hindi-std	hind1269	Hindi	hin	Eurasia	25.000000	77.000000	Indo-European	logos	3831.0
2	Romagnolstd	Romagnol-std	roma1328	Romagnol	rgn	Eurasia	44.234900	11.718900	Indo-European	logos	627.0
3	Latinstd	Latin-std	lati1261	Latin	lat	Eurasia	41.902600	12.450200	Indo-European	logos	4304.0
4	Galicianstd	Galician-std	gali1258	Galician	glg	Eurasia	42.244600	-7.534300	Indo-European	logos	620.0

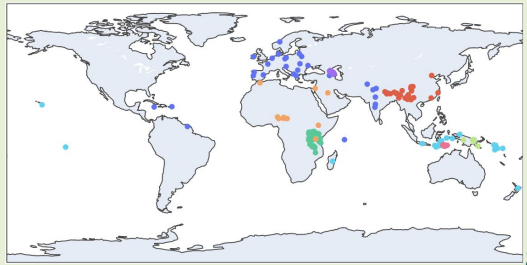
```
['Austronesian': 395,
'Nuclear Trans New Guinea': 307,
'Sino-Tibetan': 183,
'Indo-European': 173,
'Pama-Nyungan': 172,
'Atlantic-Congo': 133,
'Afro-Asiatic': 64,
'Nakh-Daghestanian': 55,
'Timor-Alor-Pantar': 43}
```

Before sampling: language families with more than 30 languages

Choose languages from these more populous families that:

- Lat/longitude, macroarea, and family info not null
- Have more than 600 forms
- If there are more than 30 for a family, choose random sample of size 30

Pama-Nyungan only had two languages that met these requirements so I dropped it.



Feature Extraction

def build_lang_df(glottocode):

Pick subset of Form df based on target Glottocode → Make list of concepts for every form in language → Choose forms w/ 2 different concepts (colexifications!) → Add semantic field and ontological category pairs

Repeat for each language in sample

Build feature df

Given all possible pair combos of sem field or ont cat → Create dictionary of counts for each pair in language → Insert pair counts to columns for that language's row

Repeat for each language in sample

Result of using build_lang_df once:

Actual form of word

Used to make sure multiple concepts share a form

	Colex_Key	dataset_ID	Form	Colex_IDs	Num_concepts	Concept_names	Semantic_field	Ontological_category
0	russ1263_batat	ids	batat	{410, 159}	2	[YAM, SWEET POTATO]	[Agriculture and vegetation, Agriculture and v...	[Person/Thing, Person/Thing]
1	russ1263_bedro	ids	bedro	{800, 1745}	2	[THIGH, HIP]	[The body, The body]	[Person/Thing, Person/Thing]
2	russ1263_čto	ids	čto	{1236, 1157}	2	[WHAT, BECAUSE]	[Cognition, Cognition]	[Other, Other]
3	russ1263_den	ids	den	{1807, 1225}	2	[AFTERNOON, DAY (NOT NIGHT)]	[Time, Time]	[Person/Thing, Person/Thing]
4	russ1263_derevo	ids	derevo	{906, 1803}	2	[TREE, WOOD]	[Agriculture and vegetation, The physical world]	[Person/Thing, Person/Thing]

Gloss for readability

Tuples we will use to build dictionaries of counts for pairs

Feature_df

Possible categories for classification

Pairs are sorted alphabetically for counts so no duplicates

	Glottocode	Macroarea	Family	Agriculture and vegetation:Agriculture and vegetation	Agriculture and vegetation:Animals	Agriculture and vegetation:Basic actions and technology	Agriculture and vegetation:Clothing and grooming
0	hind1269	Eurasia	Indo-European	5	1	0	0
1	gali1258	Eurasia	Indo-European	0	0	0	0
2	jian1239	Eurasia	Sino-Tibetan	1	2	2	1
3	nyam1276	Africa	Atlantic-Congo	0	0	1	0
4	asut1235	Africa	Atlantic-Congo	0	0	0	0

...

First 300 derived feature columns correspond to semantic field and next 21 to ontological category. These will be predictors

Count for particular pair of semantic fields among language's colexifications

Fitting models

1

X: Semantic field
y: Macroarea

Naive Bayes: 0.783
RandomForest: 0.639
SVC: 0.762

2

X: Ont category
y: Macroarea

Naive Bayes: 0.610

3

X: Semantic field
y: Language family

Naive Bayes: 0.735

4

X: Ont category
y: Language family

Naive Bayes: 0.505

X: Semantic field, y: Macroarea

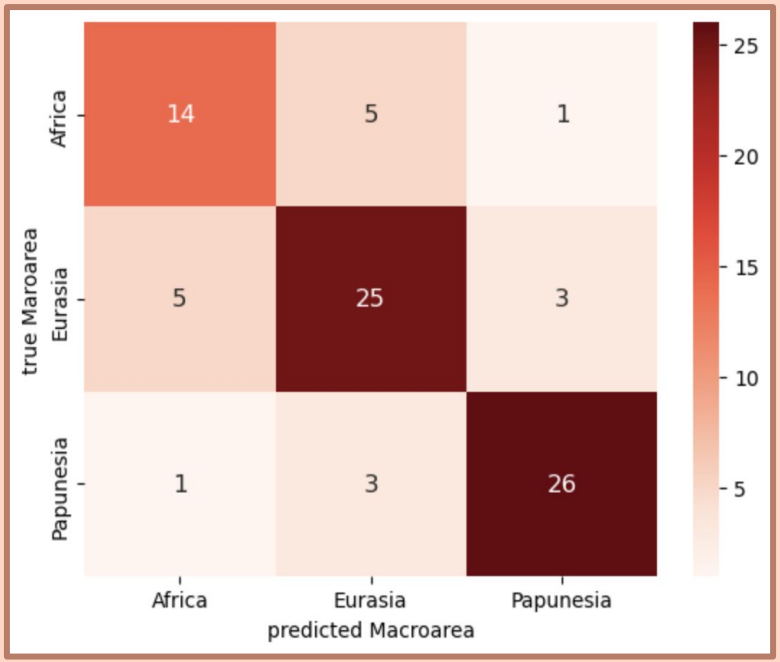
Naive Bayes

Top 5 features

Africa	**Eurasia:**	**Papunesia:**
Kinship:Kinship	Kinship:Kinship	The_body:The_body
The_body:The_body	The_body:The_body	Kinship:Kinship
Basic_actions_and_technology x2	Animals:Animals	The_physical_world x2
The_physical_world x2	The_physical_world x2	Basic_actions_and_technology x2
Animals:Animals	Spatial_relations x2	Motion:Motion

X: Semantic field, y: Macroarea

Naive Bayes

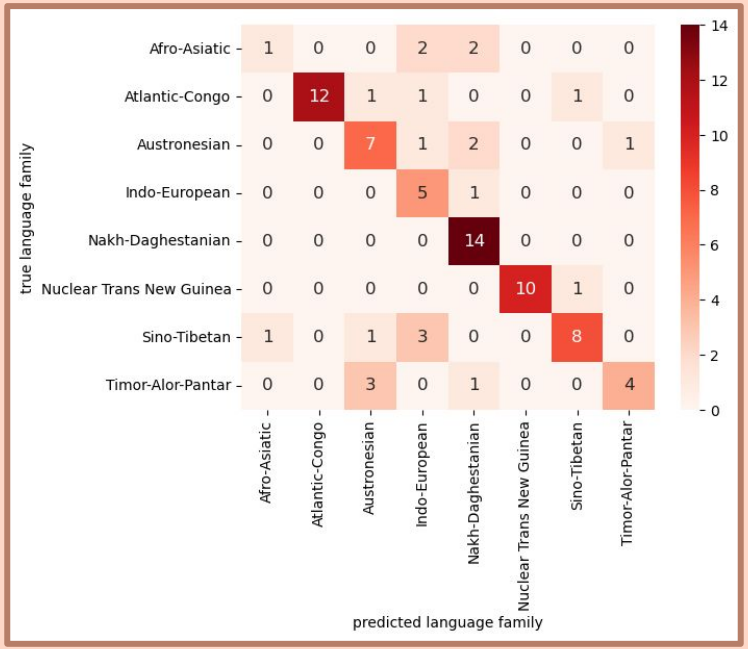


X: Semantic field, y: Language family
Naive Bayes

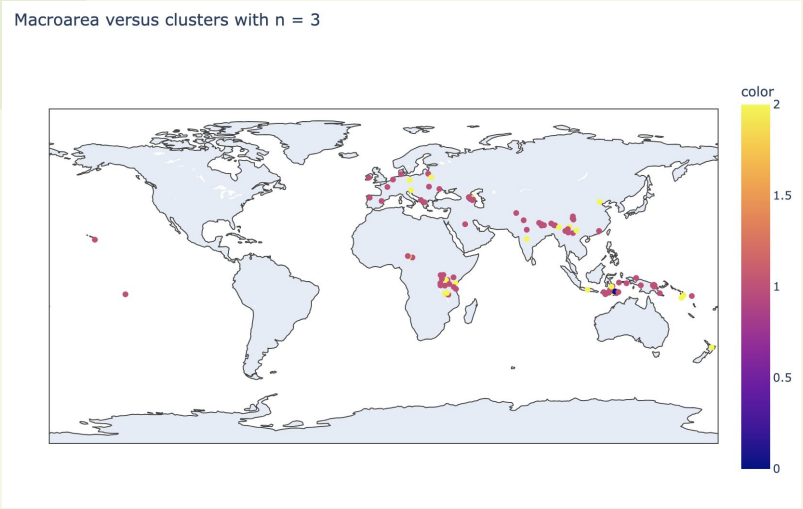
- **Afro-Asiatic**: Food_and_drink:The_physical_world, Speech_and_language:The_physical_world, Social_and_political_relations:The_physical_world, The_house:The_physical_world, Emotions_and_values:The_physical_world
- **Atlantic-Congo**: Animals:The_physical_world, Speech_and_language:The_physical_world, The_house:The_physical_world, Food_and_drink:The_physical_world, Agriculture_and_vegetation:The_physical_world
- **Austronesian**: Speech_and_language:The_physical_world, Food_and_drink:The_physical_world, The_house:The_physical_world, Animals:The_physical_world, Social_and_political_relations:Warfare_and_hunting
- **Indo-European**: Food_and_drink:The_physical_world, Speech_and_language:The_physical_world, Agriculture_and_vegetation:The_physical_world, The_house:The_physical_world, Social_and_political_relations:The_physical_world
- **Nakh-Daghestanian**: Food_and_drink:The_physical_world, Speech_and_language:The_physical_world, Agriculture_and_vegetation:The_physical_world, Emotions_and_values:The_physical_world, The_house:The_physical_world
- **Nuclear Trans New Guinea**: Modern_world:The_physical_world, Animals:The_physical_world, Food_and_drink:The_physical_world, Speech_and_language:The_physical_world, Motion:The_physical_world
- **Sino-Tibetan**: Food_and_drink:The_physical_world, Social_and_political_relations:The_physical_world, The_house:The_physical_world, Animals:The_physical_world, Basic_actions_and_technology:Social_and_political_relations
- **Timor-Alor-Pantar**: Speech_and_language:The_physical_world, Food_and_drink:The_physical_world, Agriculture_and_vegetation:The_physical_world, The_house:The_physical_world, Religion_and_belief:The_physical_world

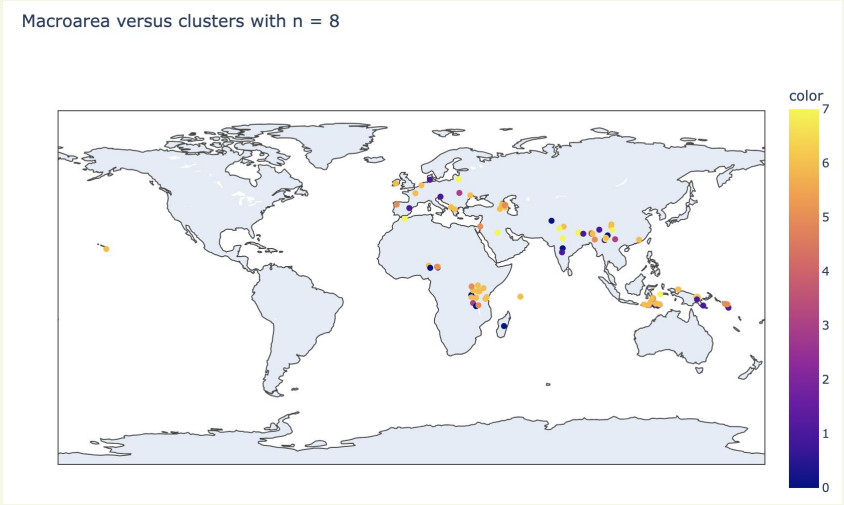
X: Semantic field, y: Language family

Naive Bayes



K-Means clustering using semantic field





Takeaways

- There is some relation between the type of colexifications in a language and its geographical region & language family
- Clustering did not work as well as I hoped
- Who decides what are separate concepts?
  - Limitations of having glosses in English
  - Need comparison to other languages? Or is this info in the lexicon?
- How representative is the data of the actual number of colexifications in a language?
- Connections between colexification types and typology of culture language developed in
  - E.g. hunter-gatherer vs agriculturally based?
- Subgraphs in network could help with word embeddings