Stability of Syntactic Dialect Classification Over Space and Time

Jonathan Dunn, Sidney Wong

Hiu Lam Choy
Heidelberg University - Diachronic LMs (6LP)

Content

1. Introduction
2. Approach
3. Results
4. Discussion

Introduction

What is a dialect?

- The variety of a language spoken within a country
- Usually mutually intelligible
- Difference in vocab, grammar and pronunciation
- Dialects change constantly
    - Make dialect models more robust

Dialect Classification

- Goal: Predicts the origin for the author of a sample
- Includes the entire population of the world (not only western/standard)
- Decay rate of classification performance
  - identify regions undergoing syntactic change
- Distribution of classification accuracy
  - homogeneity of the grammar

Aim

- To evaluate the robustness (spatial & temporal stability) of dialect models
- To understand the geographic variation in syntax
- Through a syntactic approach

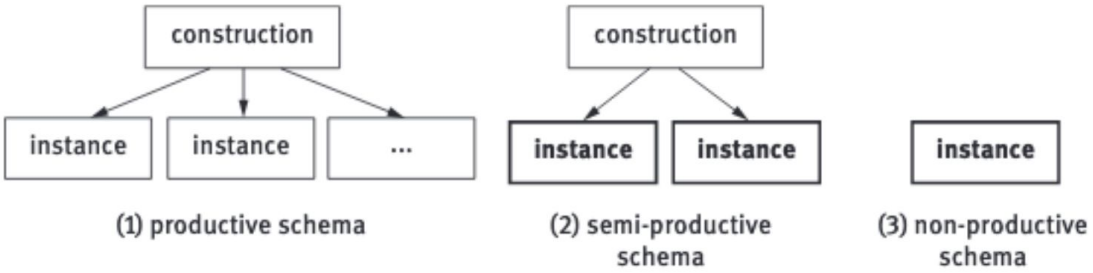
Approach

Construction grammar (CxG)

- Formulates syntactic representations through grammar induction
- Uses grammar induction to learn a grammar
- Dialects can be represented equally
- Ensures the model does not rely on extraneous information (toponymy, local topics…)
- Provides a syntactic feature space for modelling geographical variations

CxG - What is it?

- Usage-based
- Constructions are form-meaning pairings
- They have different levels of productivity
- Different levels of abstraction are captured using different types of slot-constraints



CxG - Core idea 1

1. Continuum between lexicon and syntax
- Verb valency can be fluid

(a) Peter *laughed*.
(b) The audience *laughed* Peter off the stage.
(c) His marriage *laughed* Peter into rehab.
(d) Peter *laughed* all the way to the bank.

CxG - Core idea 2

2. Syntactic structure varies in its level of abstraction
- Structures can be item-specific or idiomatic
- Productivity of the constructions can change syntax

(a) Peter *laughed*.
(b) The audience *laughed* Peter off the stage.
(c) His marriage *laughed* Peter into rehab.
(d) Peter *laughed* all the way to the bank.

CxG - Core idea 3

3. Constructions are constraint-based representations
- Slot constraints can be lexical, syntactic, and semantic

(a) Peter *laughed*.
(b) The audience *laughed* Peter off the stage.
(c) His marriage *laughed* Peter into rehab.
(d) Peter *laughed* all the way to the bank.

(e) [SYN:NP – SYN:VP]
(f) [SYN:NP – SYN:VP – SEM:*object* – SEM:*loc*]
(g) [SYN:NP – SYN:VP – LEX:*all the way to the bank*]

Dataset

- Corpus of Global Language Use (CGLU)
  - Geo-referenced tweets
  - Aggregates tweets from same place and time until sample reaches 500 words
- Separate groups for Inner- and Outer-Circle varieties
- Area of a city with 50 km radius

Circle	Region	Country	N. Cities	N. Words
Inner-Circle	Oceania	Australia	98	3.9 mil
Inner-Circle	Oceania	New Zealand	99	2.0 mil
Inner-Circle	North American	Canada	95	4.9 mil
Inner-Circle	North American	United States	86	4.5 mil
Inner-Circle	European	Ireland	100	3.6 mil
Inner-Circle	European	United Kingdom	89	5.5 mil
Total Inner-Circle	**3**	**6**	**567**	**24.4 mil**
Outer-Circle	African	Ghana	69	1.1 mil
Outer-Circle	African	Kenya	98	1.8 mil
Outer-Circle	South Asian	India	96	2.5 mil
Outer-Circle	South Asian	Pakistan	100	1.0 mil
Outer-Circle	Southeast Asian	Malaysia	99	0.8 mil
Outer-Circle	Southeast Asian	Philippines	91	1.1 mil
Total Outer-Circle	**3**	**6**	**553**	**8.37 mil**

Table 1: Inventory of Regions, Countries, and Cities for Data Collection (One Month)

Grammar

- Lexicon - most frequent 100k words across all varieties
- Syntactic - from the Universal Part-of-Speech tagset
- Semantic - fastText embeddings (cluster domain number of the words)

➔ One single grammar with 6119 constructions

Dialect models

- Model spatio-temporal variation
- Implemented as linear SVM
- Matrix of spatial weights
    - Row: constructions (features) in the grammar
    - Columns: dialects (country-level label)

1. Syntactic model
2. Function word model
3. Unigram lexical model

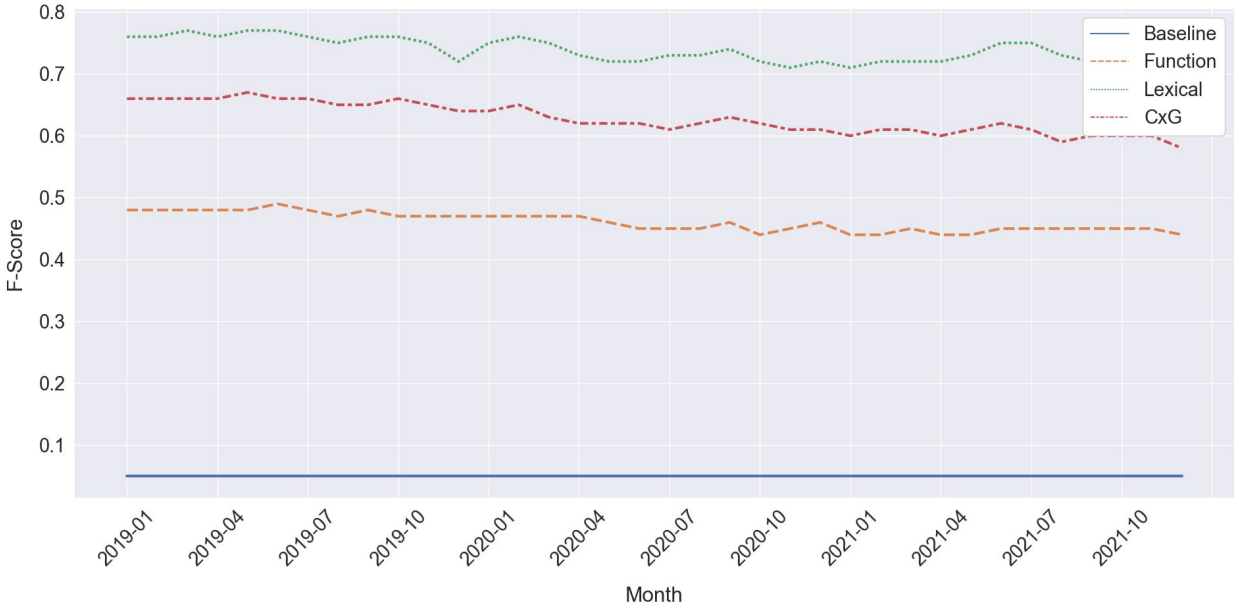
- INNER, OUTER, ALL

Experiment

- Training: July-Dec 2018
- Testing: monthly intervals from 2019-2021
- Same geographical distribution in testing and training

Results

Dialect models



AU	CA	IE	NZ
australia	canada	ireland	nz
australian	canadian	irish	zealand
mate	ontario	dublin	auckland
melbourne	trump	cork	jacinda
sydney	toronto	limerick	te
abc	vancouver	galway	kiwi
brisbane	trudeau	lads	liked
labor	km	hurling	lincoln
nsw	kpa	county	hamilton
turnbull	alberta	final	kph

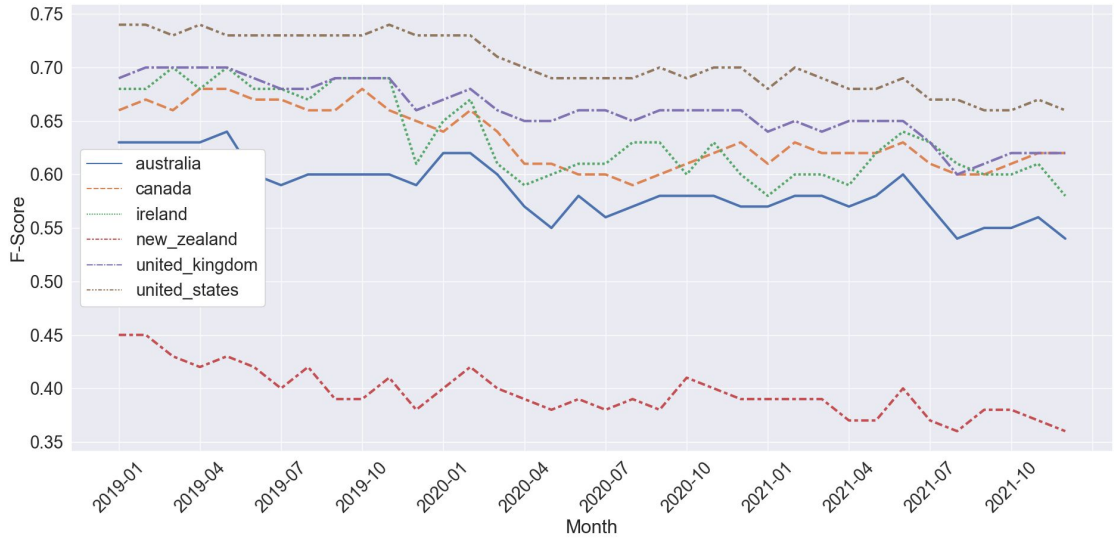
Table 2: Top Lexical Features By Country

Temporal evaluation

	Function	Grammar
Inner-Only, 2019-01	0.44	0.66
Inner-Only, 2021-12	0.40	0.59
Inner-Only Decline	*0.04*	*0.07*
Outer-Only, 2019-01	0.75	0.83
Outer-Only, 2021-12	0.66	0.75
Outer-Only Decline	*0.09*	*0.08*
All Dialects, 2019-01	0.48	0.66
All Dialects, 2021-12	0.44	0.58
All Dialects Decline	*0.04*	*0.08*

Table 3: Change in Performance Over Time by Model

Regression analysis



Vector Error Correction Model (VECM)

- Checks for relationship between different time periods
- Examine the relative frequency of false positives
- Identify any long-term trends in the error distribution

	AU	CA	IE	NZ	UK	US
AU	0	0	0	0	0	0
CA	0	0	0	0	0	0
IE	0	0	0	0	0	-.04
NZ	.16	0	0	0	.28	0
UK	0	0	0	0	0	0
US	0	0	0	0	0	0

	GH	IN	KE	MY	PK	PH
GH	0	-1.01	0	0	-.28	-.40
IN	0	0	0	-.91	0	0
KE	0	0	0	0	0	0
MY	0	0	0	0	0	0
PK	-.20	0	0	0	0	.13
PH	0	0	0	0	0	0

Spatial evaluation

	Moran's *I*	Mean Acc.	Min	Max
AU	0.30	61%	18%	83%
CA	0.54	65%	07%	100%
IE	0.17	58%	35%	89%
NZ	0.20	36%	08%	62%
UK	0.22	73%	41%	82%
US	0.18	79%	53%	97%

	Moran's *I*	Mean Acc.	Min	Max
GH	0.30	86%	42%	94%
IN	0.38	84%	27%	95%
KE	0.24	89%	62%	97%
MY	0.70	79%	50%	95%
PK	0.42	70%	15%	87%
PH	0.20	77%	37%	88%

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Covariance term

Inverse of variance ($1/s^2$)

Normalizing term

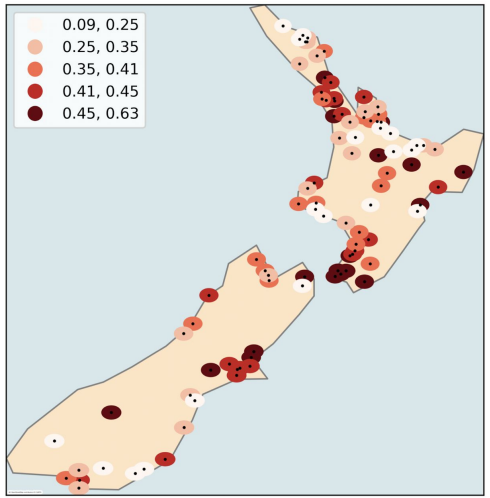


Figure 3: Map of Average City-Level Accuracy, NZ

Discussion

Limitations

- Samples might not represent the regional variation
- There is no clear boundary for dialects
- Limited resources
- Languages are constantly evolving

Questions

- What could be some possible NLP applications of dialect models?
- Are there other methods to capture the syntactic information other than POS tags?

Sources

Jonathan Dunn and Sidney Wong. 2022. Stability of Syntactic Dialect Classification over Space and Time. In Proceedings of the 29th International Conference on Computational Linguistics, pages 26–36, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hopper, P. (1987). Emergent Grammar. Annual Meeting of the Berkeley Linguistics Society, 13, 139. https://doi.org/10.3765/bls.v13i0.1834

Hilpert, M. (2021) What Is Construction Grammar? (2021). Ten Lectures on Diachronic Construction Grammar, 1–35. https://doi.org/10.1163/9789004446793_002

Boas, H. C. (2021). Construction Grammar and Frame Semantics. https://sites.la.utexas.edu/hcb/files/2021/07/Boas-CxG-and-FS-2021-DRAFT.pdf

https://amitness.com/2020/06/fasttext-embeddings/

THANK YOU FOR YOUR ATTENTION :)