# Modeling Global Syntactic Variation

## (in English) (using Dialect Classification)


Jonathan Dunn

jonathan.dunn@canterbury.ac.nz

www.jdunn.name

UC
UNIVERSITY OF CANTERBURY
*Te Whare Wānanga o Waitaha*
CHRISTCHURCH NEW ZEALAND

# Goals

(i) Identify dialects with syntax features

(ii) Explore grammar adaptation for dialects

# Steps

**(1) Finding national dialects of English**

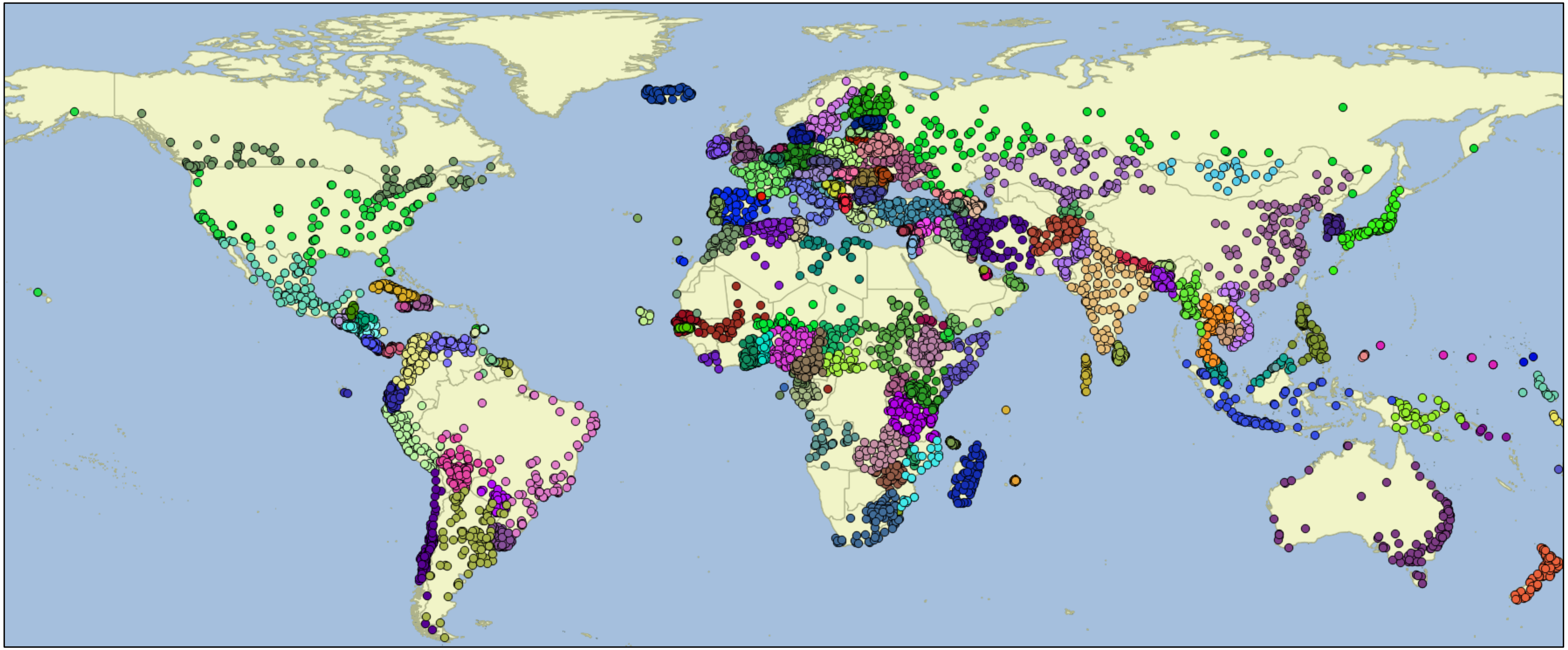
(2) Finding syntactic variants in English

(3) Modeling dialects using classification

# Finding national dialects

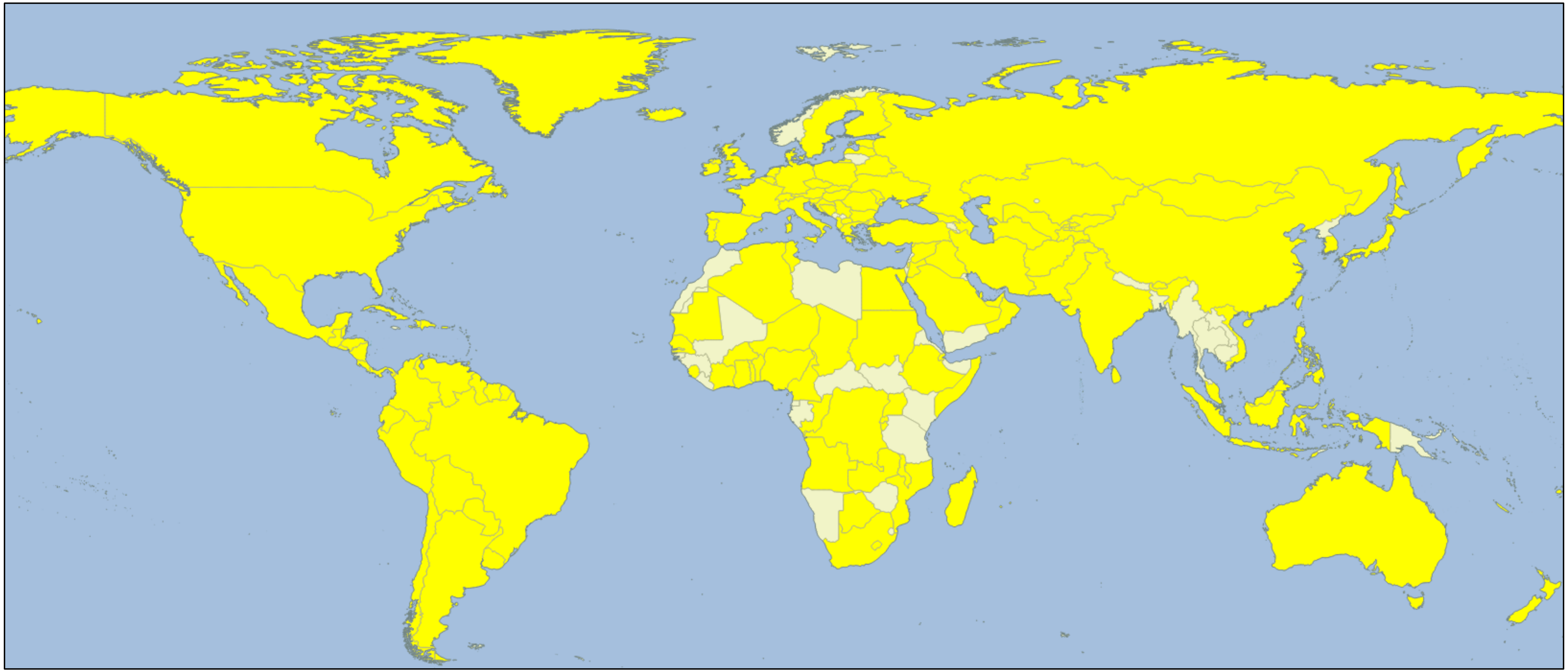
Countries in the World

# Finding national dialects



Twitter Collection by City

# Finding national dialects

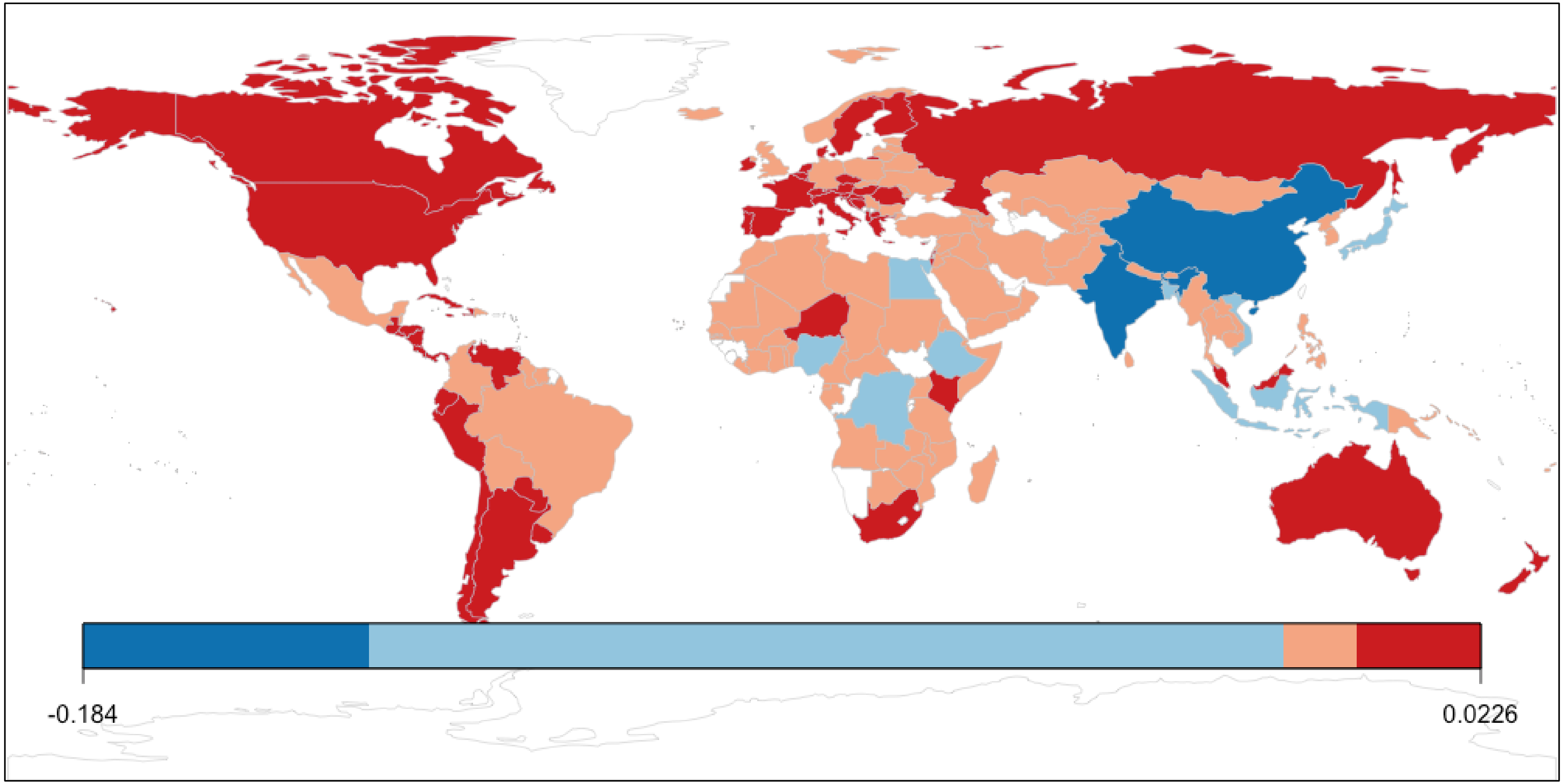


Web Collection from the Common Crawl

# Finding national dialects

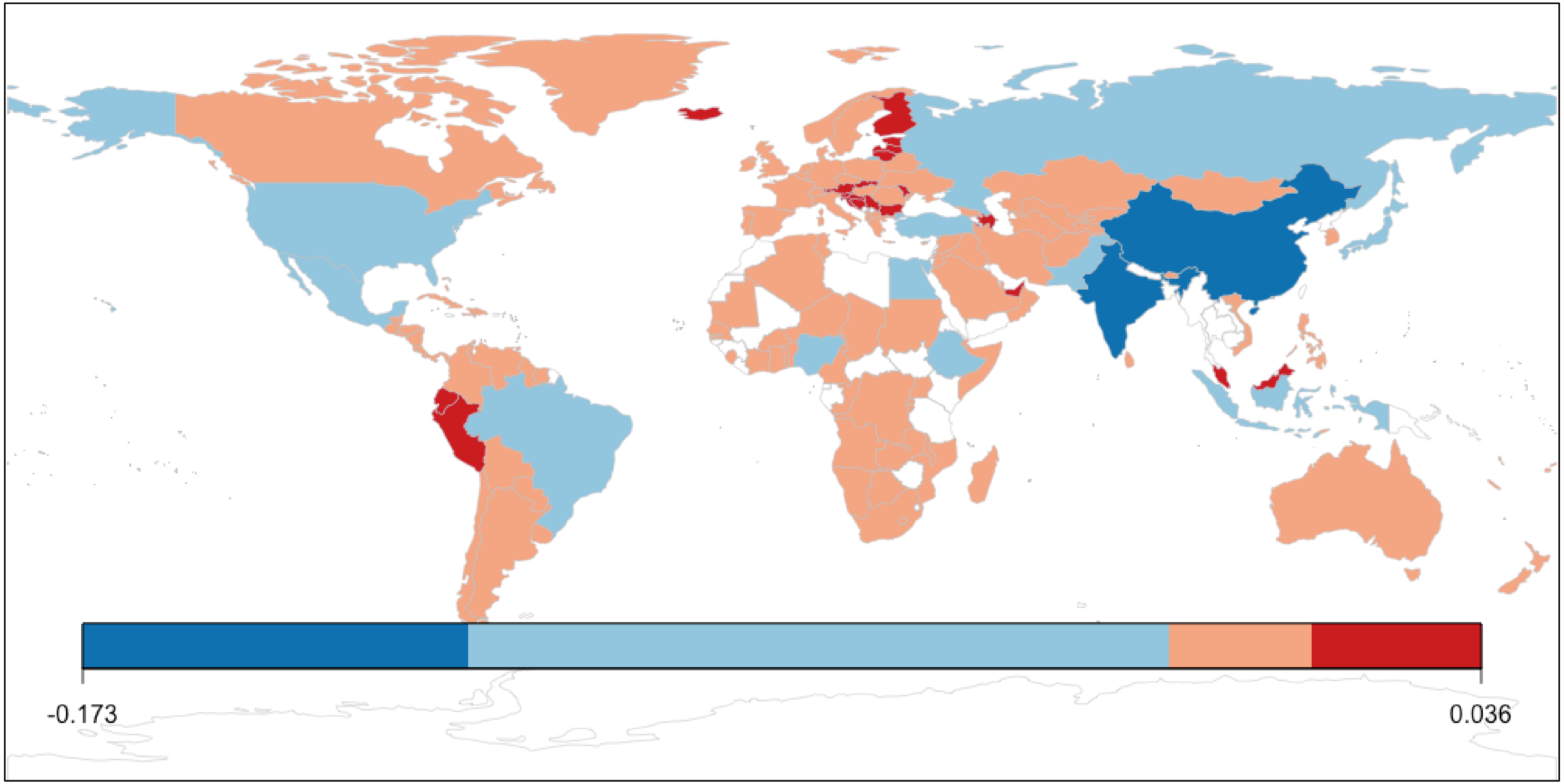
| Region | Population | Twitter | Common Crawl |
|---|---|---|---|
| Africa, North | 3% | 2% | 0.7% |
| Africa, Southern | 1% | 2% | 0.4% |
| Africa, Sub-Saharan | 10% | 6% | 2% |
| America, Brazil | 2% | 2% | 1% |
| America, Central | 2% | 9% | 5% |
| America, North | 4% | 8% | 1% |
| America, South | 2% | 9% | 7% |
| Asia, Central | 2% | 2% | 5% |
| Asia, East | 22% | 2% | 13% |
| Asia, South | 23% | 8% | 2% |
| Asia, Southeast | 8% | 5% | 12% |
| Europe, East | 2% | 7% | 27% |
| Europe, Russia | 2% | 2% | 0.6% |
| Europe, West | 5% | 19% | 14% |
| Middle East | 4% | 5% | 4% |
| Oceania | 1% | 5% | 1% |
| **TOTAL** | **7.35 billion (People)** | **4.14 billion (Words)** | **16.65 billion (Words)** |

# Finding national dialects




Population-to-Corpus Comparison, Twitter

# Finding national dialects



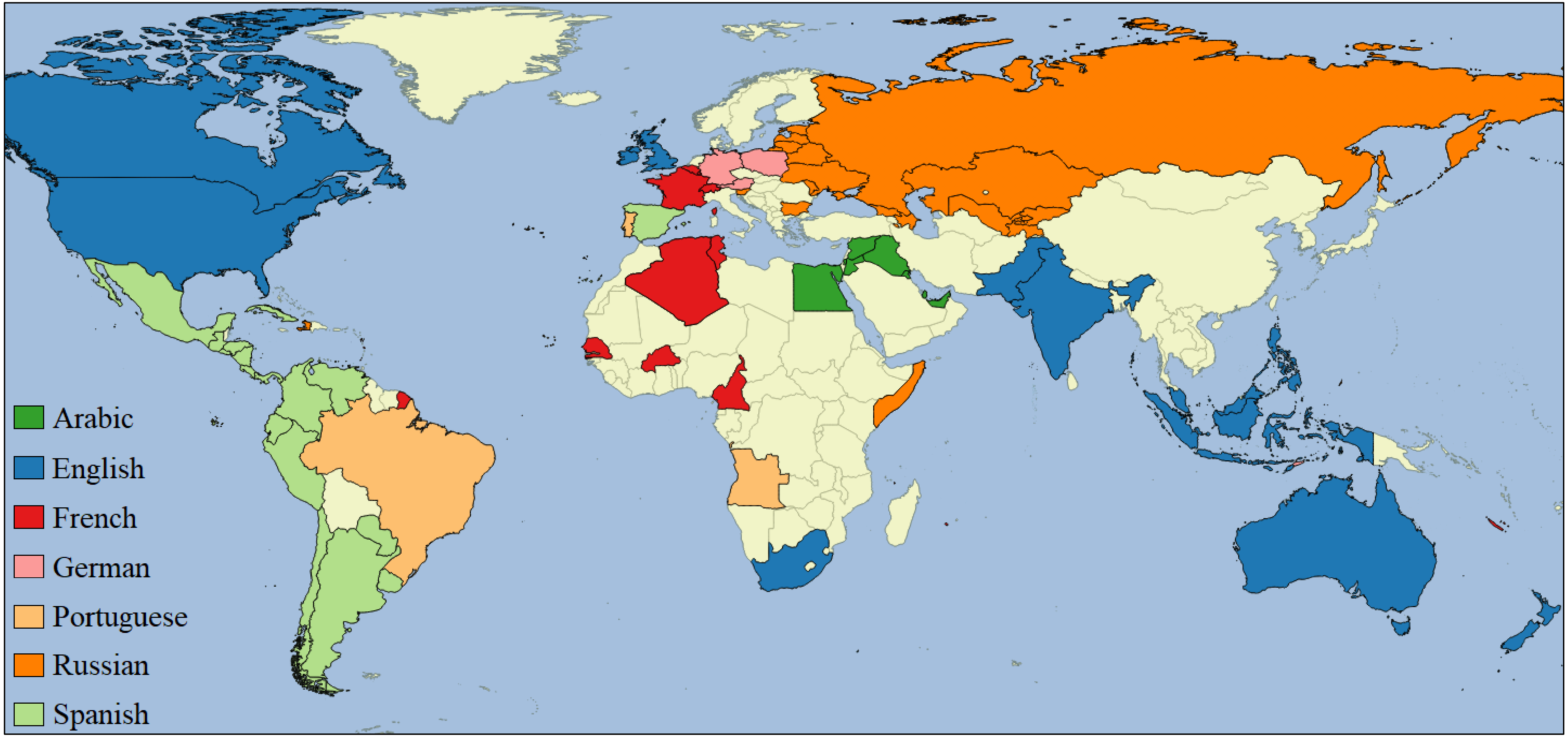

Population-to-Corpus Comparison, Common Crawl

# Finding national dialects

| Country | Twitter (Words) | Common Crawl (Words) | Circle |
|---|---|---|---|
| (au) Australia | 29.1 mil | 98.9 mil | Inner |
| (ca) Canada | 149.8 mil | 97.8 mil | Inner |
| (ie) Ireland | 43.9 mil | 46.0 mil | Inner |
| (nz) New Zealand | 87.9 mil | 37.4 mil | Inner |
| (uk) United Kingdom | 62.8 mil | 43.3 mil | Inner |
| (us) United States | 42.8 mil | 220.9 mil | Inner |
| (in) India | 71.2 mil | 80.0 mil | Outer |
| (my) Malaysia | 198.5 mil | 18.2 mil | Outer |
| (ni) Nigeria | 113.9 mil | 29.3 mil | Outer |
| (ph) Philippines | 209.4 mil | 19.7 mil | Outer |
| (pk) Pakistan | 140.1 mil | 34.0 mil | Outer |
| (za) South Africa | 53.4 mil | 57.0 mil | Outer |
| (ch) Switzerland | 15.4 mil | 17.7 mil | Expanding |
| (pt) Portugal | 20.9 mil | 23.3 mil | Expanding |
| **TOTAL** | **1.23 billion** | **0.82 billion** | |

English Data by Source

# Finding national dialects



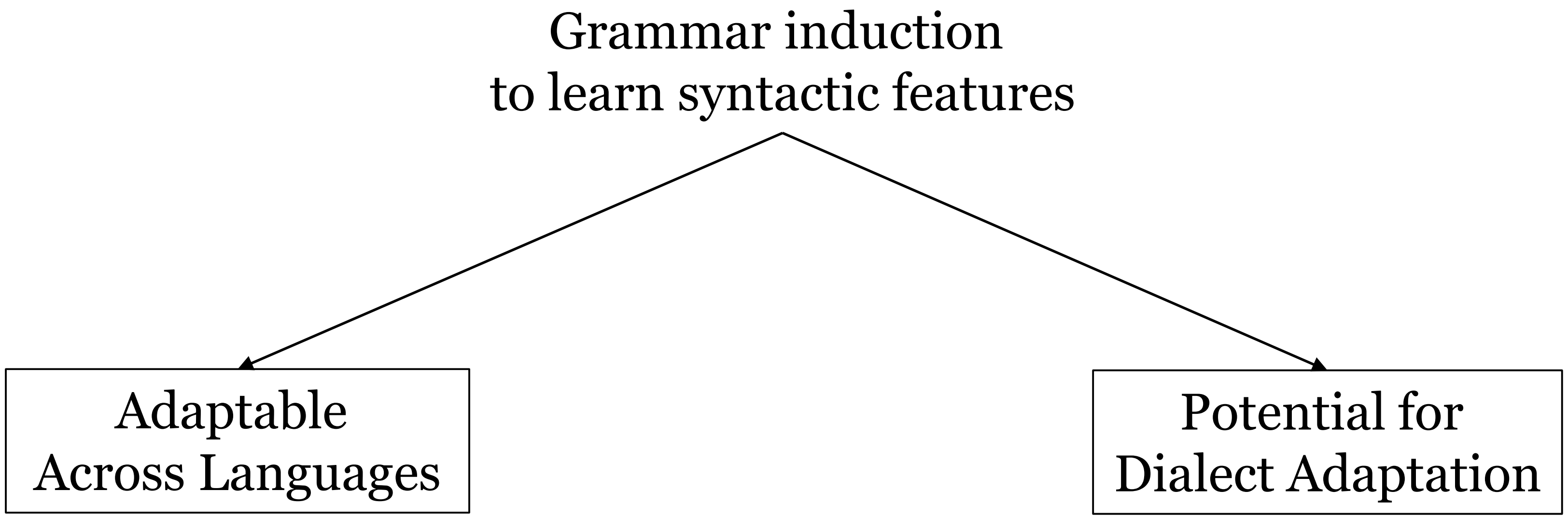
# Steps

(1) Finding national dialects of English

**(2) Finding syntactic variants in English**

(3) Modeling dialects using classification

# Finding syntactic variants



# Finding syntactic variants

Computational Construction Grammar

# Finding syntactic variants

Computational Construction Grammar

CxG represents grammar using constraint-based *constructions*

# Finding syntactic variants

Computational Construction Grammar

CxG represents grammar using constraint-based *constructions*

Each construction is made up of slots,

# Finding syntactic variants

Computational Construction Grammar

CxG represents grammar using constraint-based *constructions*

Each construction is made up of slots, each of which is defined by a *constraint*

# Finding syntactic variants

Computational Construction Grammar

CxG represents grammar using constraint-based *constructions*

Each construction is made up of slots, each of which is defined by a *constraint*

(1a) [SYN:NOUN — SEM-SYN:TRANSFER[V] — SEM-SYN:ANIMATE[N] — SYN:NOUN]
(1b) "He gave Bill coffee."
(1c) "He gave Bill trouble."
(1d) "Bill sent him letters."
(2a) [SYN:NOUN — LEX:"give" — SEM-SYN:ANIMATE[N] — LEX:"a hand"]
(2b) "Bill gave me a hand."

# Finding syntactic variants

Poster @ CMCL today, 2:30-3:30

CxG-1

Template-based Selection Algorithm
using Frequency measures

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| LEX | "he" | "mailed" | "George" | "a package" |
| SYN | **Noun** | Verb | **Noun** | Noun |
| SEM-SYN | ANIMATE[N] | **TRANSFER[V]** | PERSON[N] | **OBJECT[N]** |

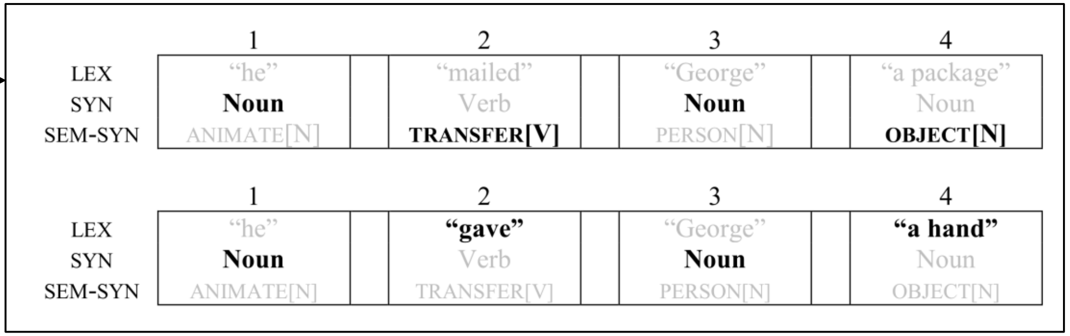
| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| LEX | "he" | **"gave"** | "George" | **"a hand"** |
| SYN | **Noun** | Verb | **Noun** | Noun |
| SEM-SYN | ANIMATE[N] | TRANSFER[V] | PERSON[N] | OBJECT[N] |

# Finding syntactic variants

Poster @ CMCL today, 2:30-3:30

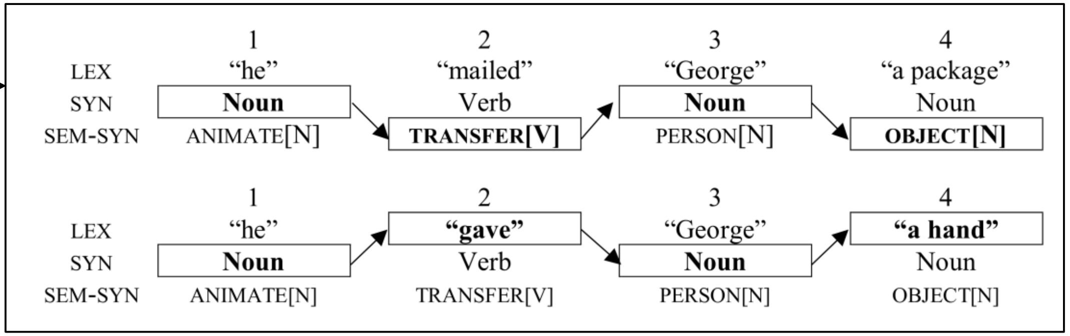
CxG-1

Template-based Selection Algorithm
using Frequency measures




CxG-2

Transition-based Selection Algorithm
using Association measures



# Finding syntactic variants

Grammars are learned using other web corpora (i.e., ukWac)

(Not learned using Twitter data)

# Finding syntactic variants

Adapting grammars to regional dialects

False Positives

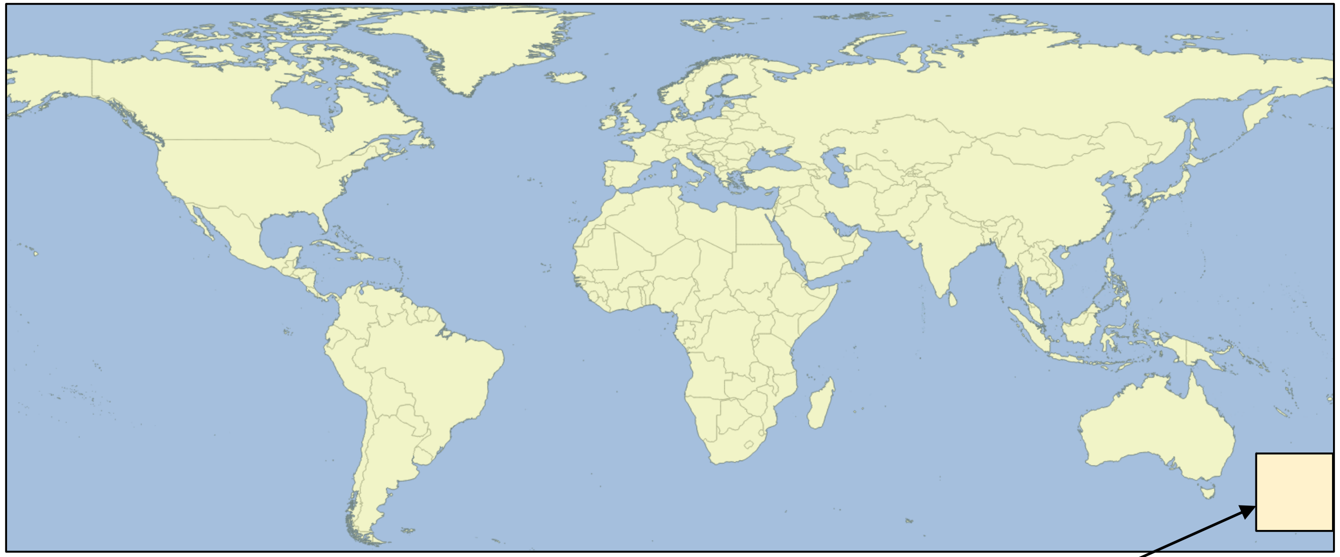
False Negatives

# Finding syntactic variants

| Country | Twitter | Common Crawl | Circle |
| --- | --- | --- | --- |
| (au) Australia | + 5.28% | + 8.15% | Inner |
| (ca) Canada | + 2.77% | + 5.17% | Inner |
| (ie) Ireland | + 8.56% | + 18.62% | Inner |
| (nz) New Zealand | + 5.32% | - 0.59% | Inner |
| (uk) United Kingdom | + 9.71% | + 13.98 % | Inner |
| (us) United States | - 0.18% | - 1.90 % | Inner |
| (in) India | - 9.39% | - 10.38% | Outer |
| (my) Malaysia | - 9.22% | - 11.51% | Outer |
| (ni) Nigeria | - 0.10% | - 0.78% | Outer |
| (ph) Philippines | - 4.96% | - 17.39% | Outer |
| (pk) Pakistan | - 11.24% | - 17.25% | Outer |
| (za) South Africa | + 3.78% | + 4.62% | Outer |
| (ch) Switzerland | + 4.82% | + 13.96% | Expanding |
| (pt) Portugal | - 5.34% | - 4.70% | Expanding |

Relative Average Feature Density (CxG-2)

# Why syntactic models?

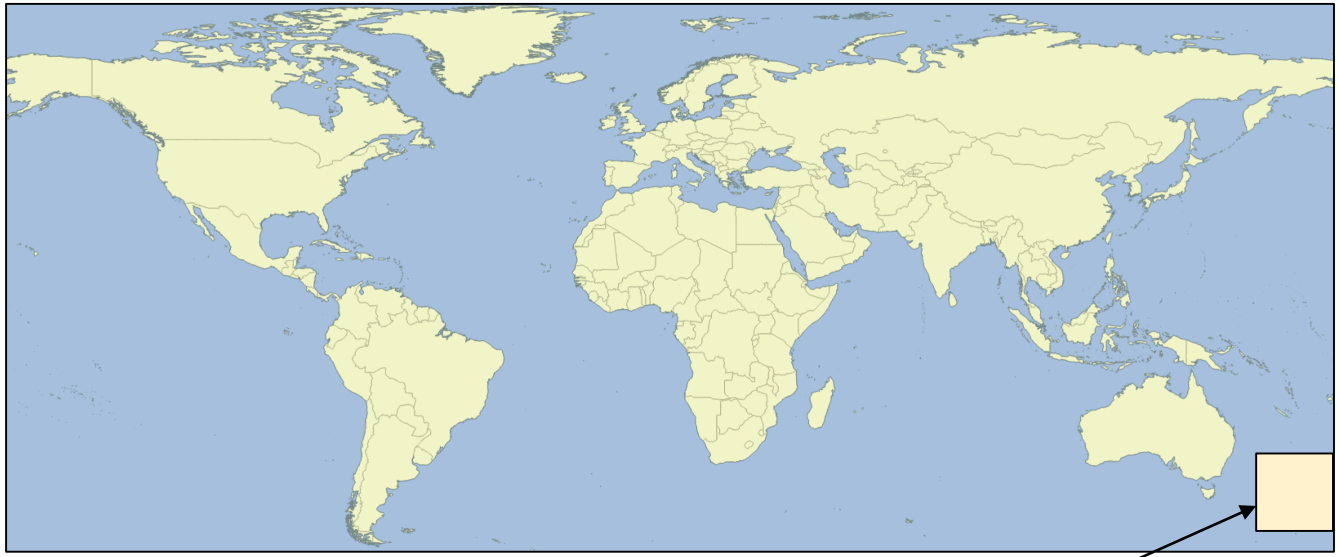


A Place

# Why syntactic models?


*Human Geography*: Place Names (Mt. Cook vs. Aoraki; Canterbury vs. Waitaha)

# Why syntactic models?



A Place

*Human Geography*: Place Names

*Human Geography*: Culture (Kapa haka, Cricket, Freedom Camping)

# Why syntactic models?

A Place

*Human Geography*: Place Names

*Human Geography*: Culture

*Human Geography*: Events (World Buskers Festival, Well-being budget)

# Why syntactic models?


*Human Geography*: Place Names

*Human Geography*: Culture

*Human Geography*: Events

*Linguistics*: Dialect (Dative vs. Ditransitive; Gerund vs. Infinitive)

# Steps

(1) Finding national dialects of English

(2) Finding syntactic variants in English

**(3) Modeling dialects using classification**

# Dialect Classification

(1) Fixed training / testing sets (327k/66k and 308k/64k)

# Dialect Classification

(1) Fixed training / testing sets

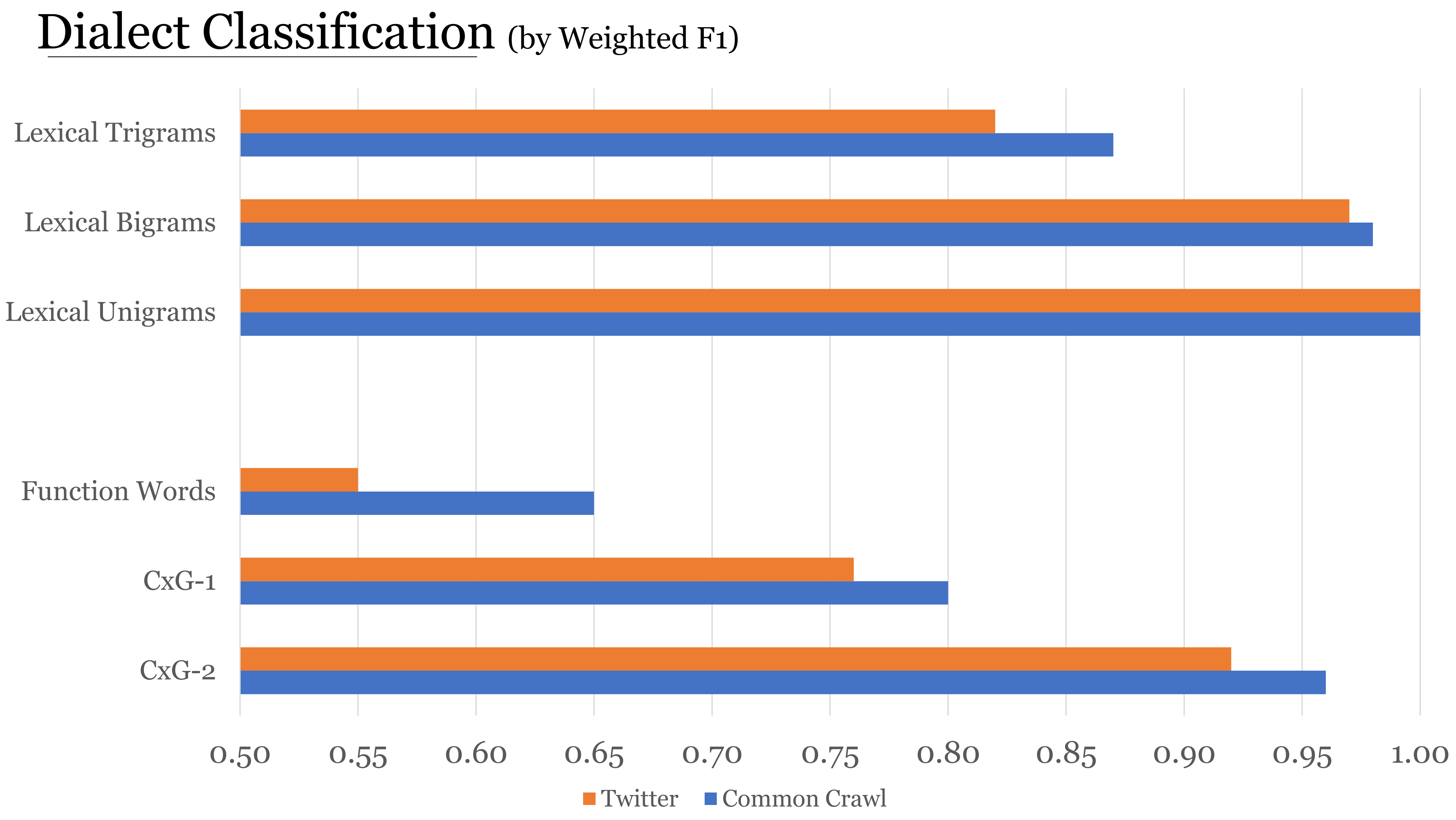
(2) Linear SVM (with unmasking)

# Dialect Classification

(1) Fixed training / testing sets

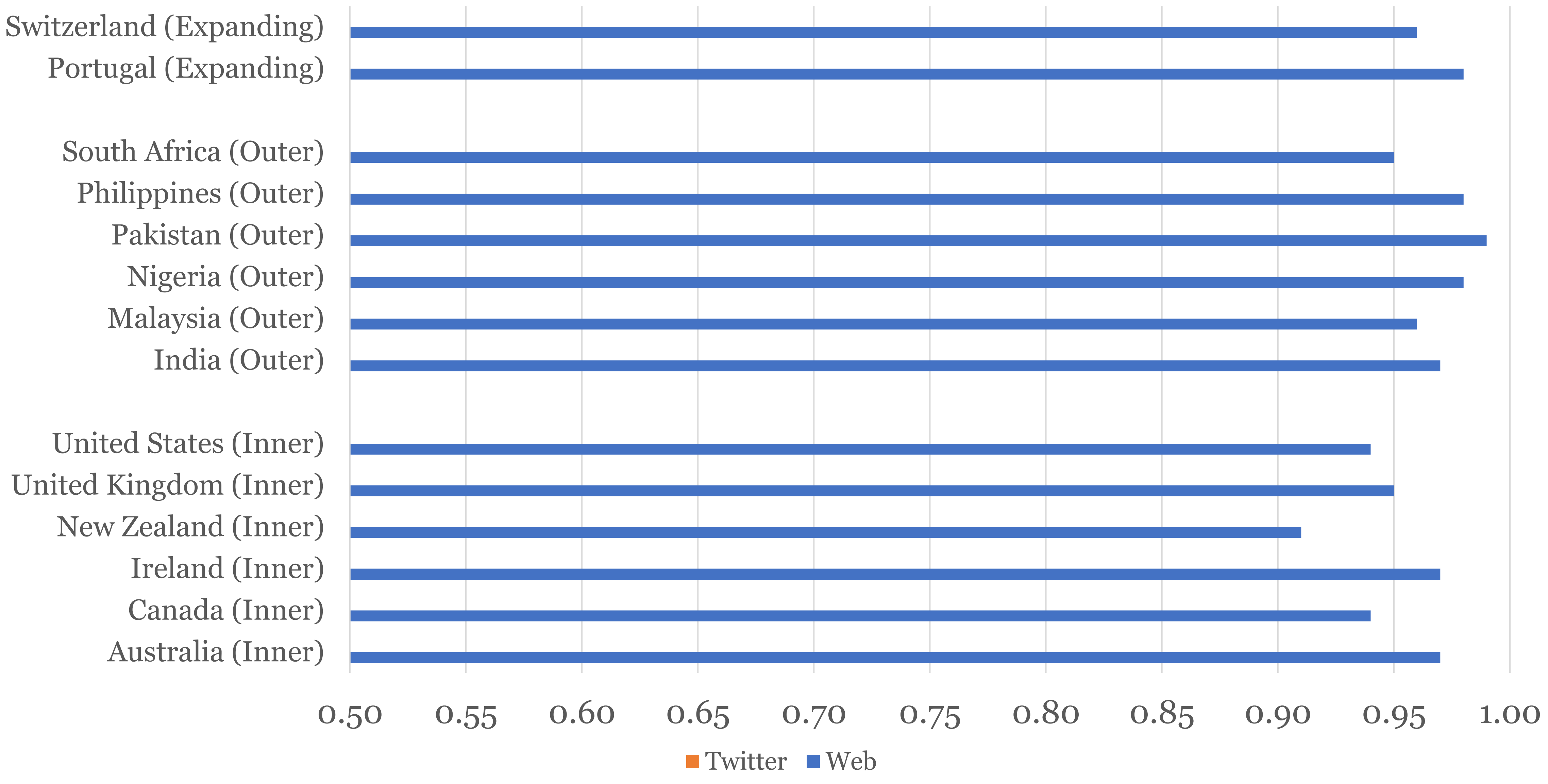
(2) Linear SVM

(3) Sample size: 1k words

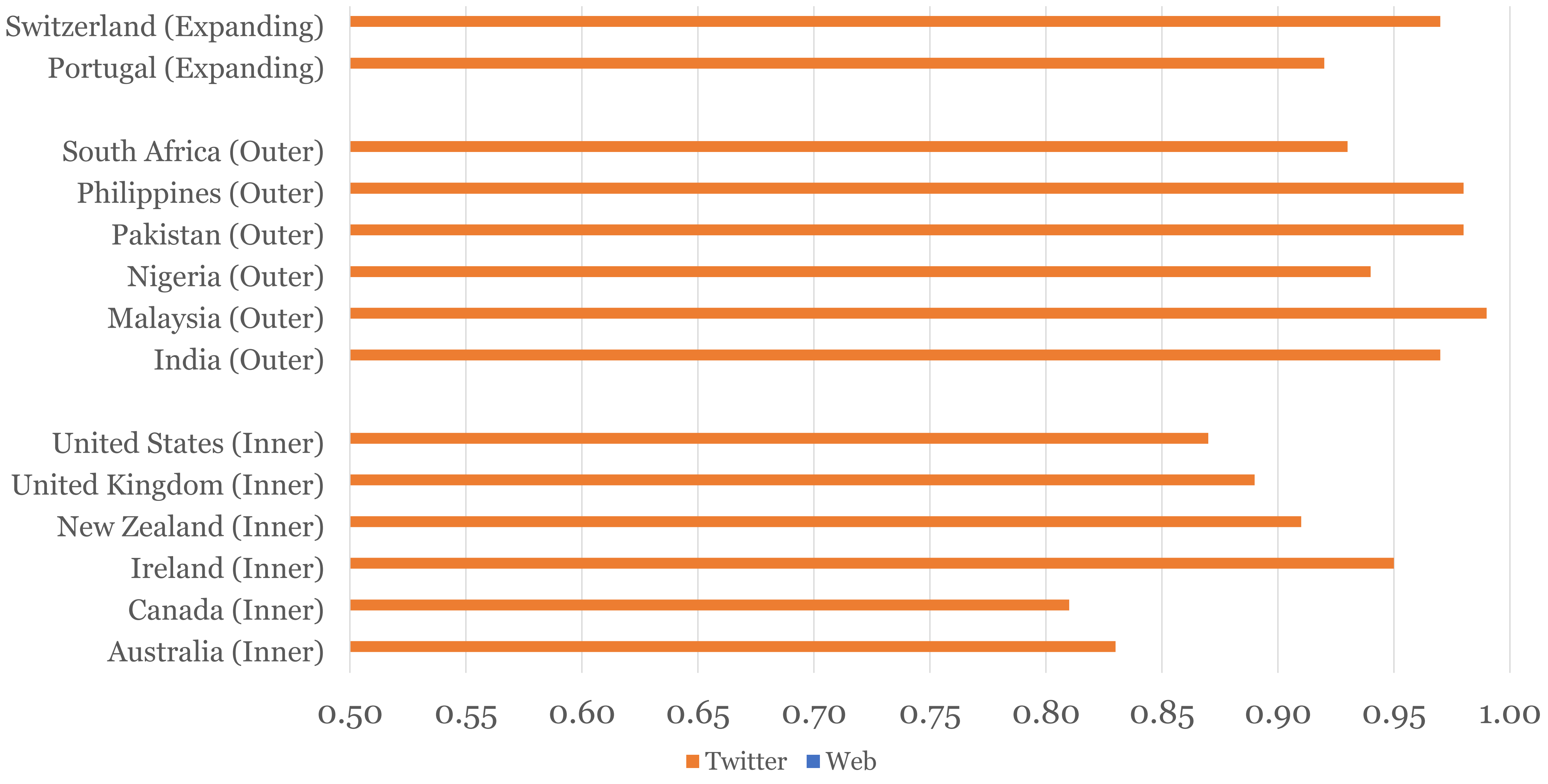


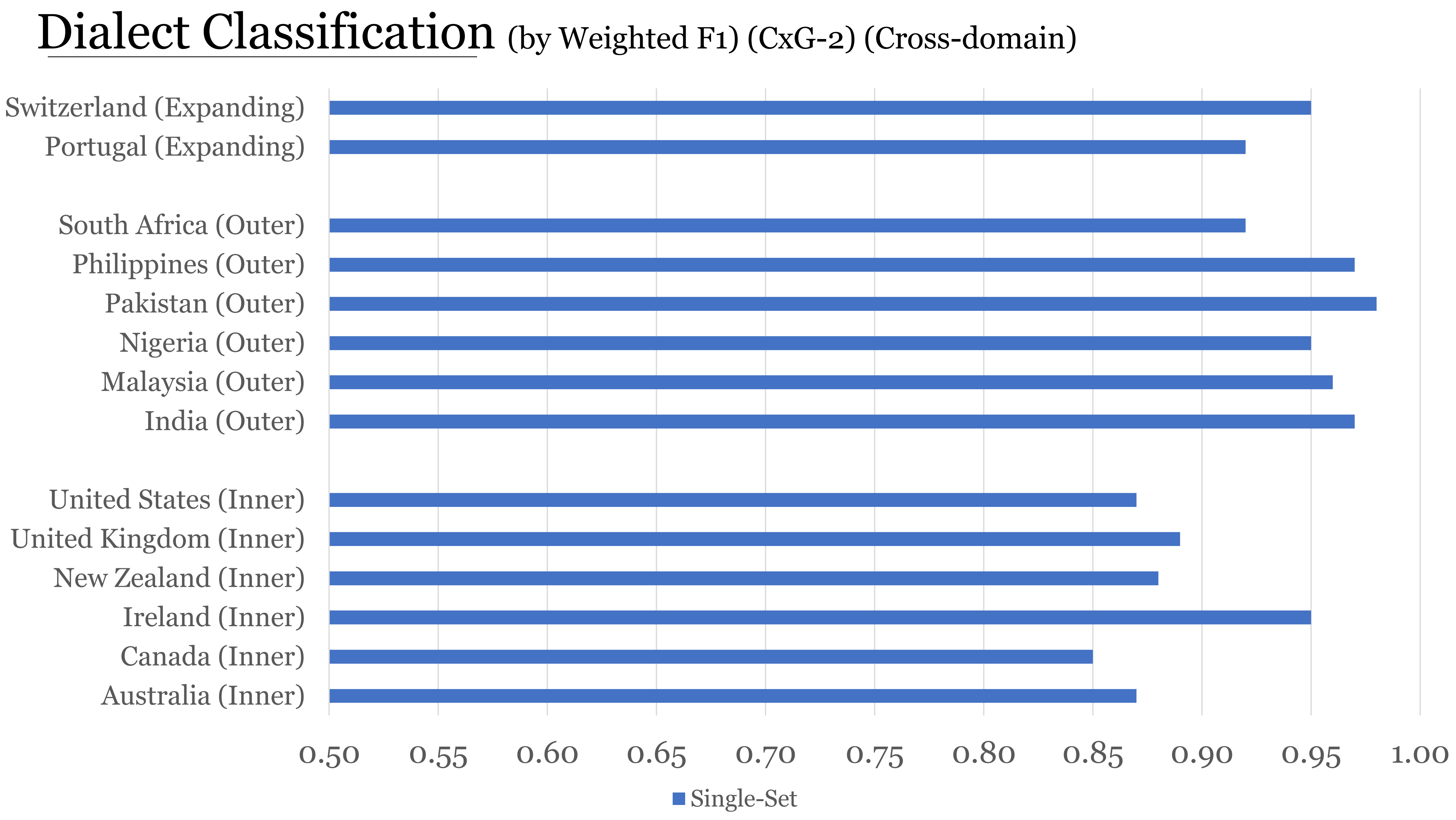
Dialect Classification (by Weighted F1)
Lexical Trigrams
Lexical Bigrams
Lexical Unigrams
Function Words
CxG-1
CxG-2
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Twitter
Common Crawl

# Dialect Classification (by Weighted F1) (CxG-2)

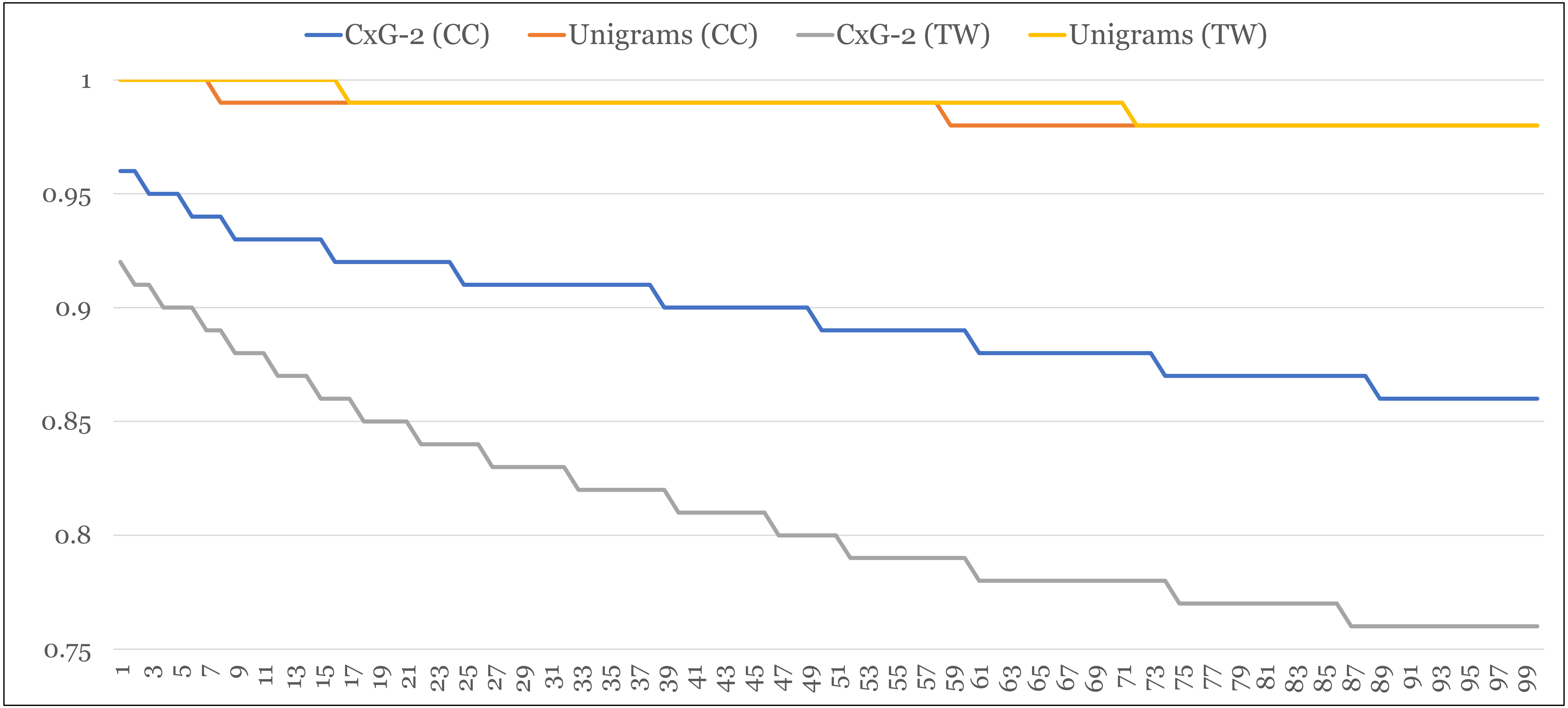


# Dialect Classification (by Weighted F1) (CxG-2)




Dialect Classification (by Weighted F1) (CxG-2) (Cross-domain)
Switzerland (Expanding)
Portugal (Expanding)
South Africa (Outer)
Philippines (Outer)
Pakistan (Outer)
Nigeria (Outer)
Malaysia (Outer)
India (Outer)
United States (Inner)
United Kingdom (Inner)
New Zealand (Inner)
Ireland (Inner)
Canada (Inner)
Australia (Inner)
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Single-Set

# Dialect Classification (by Weighted F1) (Across unmasking rounds)



# Conclusions

1. Mixing *inner*- and *outer*- and *expanding*- circle varieties seems to work fine

# Conclusions

1. Mixing *inner*- and *outer*- and *expanding*- circle varieties seems to work fine

2. Inner-circle varieties have the best fit with a generic grammar…

# Conclusions

1. Mixing *inner*- and *outer*- and *expanding*- circle varieties seems to work fine

2. Inner-circle varieties have the best fit with a generic grammar…

3. But *outer*- and *expanding*- circle varieties are more distinct (negative evidence?)

# Conclusions

1. Mixing *inner*- and *outer*- and *expanding*- circle varieties seems to work fine

2. Inner-circle varieties have the best fit with a generic grammar…

3. But *outer*- and *expanding*- circle varieties are more distinct (negative evidence?)

4. *Within-domain* models work well; *cross-domain* models are bad

# Conclusions

1. Mixing *inner*- and *outer*- and *expanding*- circle varieties seems to work fine

2. Inner-circle varieties have the best fit with a generic grammar…

3. But *outer*- and *expanding*- circle varieties are more distinct (negative evidence?)

4. *Within-domain* models work well; *cross-domain* models are bad

5. Lexical models (human geography?) have better accuracy and stability over features

# Conclusions

1. Mixing *inner*- and *outer*- and *expanding*- circle varieties seems to work fine

2. Inner-circle varieties have the best fit with a generic grammar…

3. But *outer*- and *expanding*- circle varieties are more distinct (negative evidence?)

4. *Within-domain* models work well; *cross-domain* models are bad

5. Lexical models (human geography?) have better accuracy and stability over features

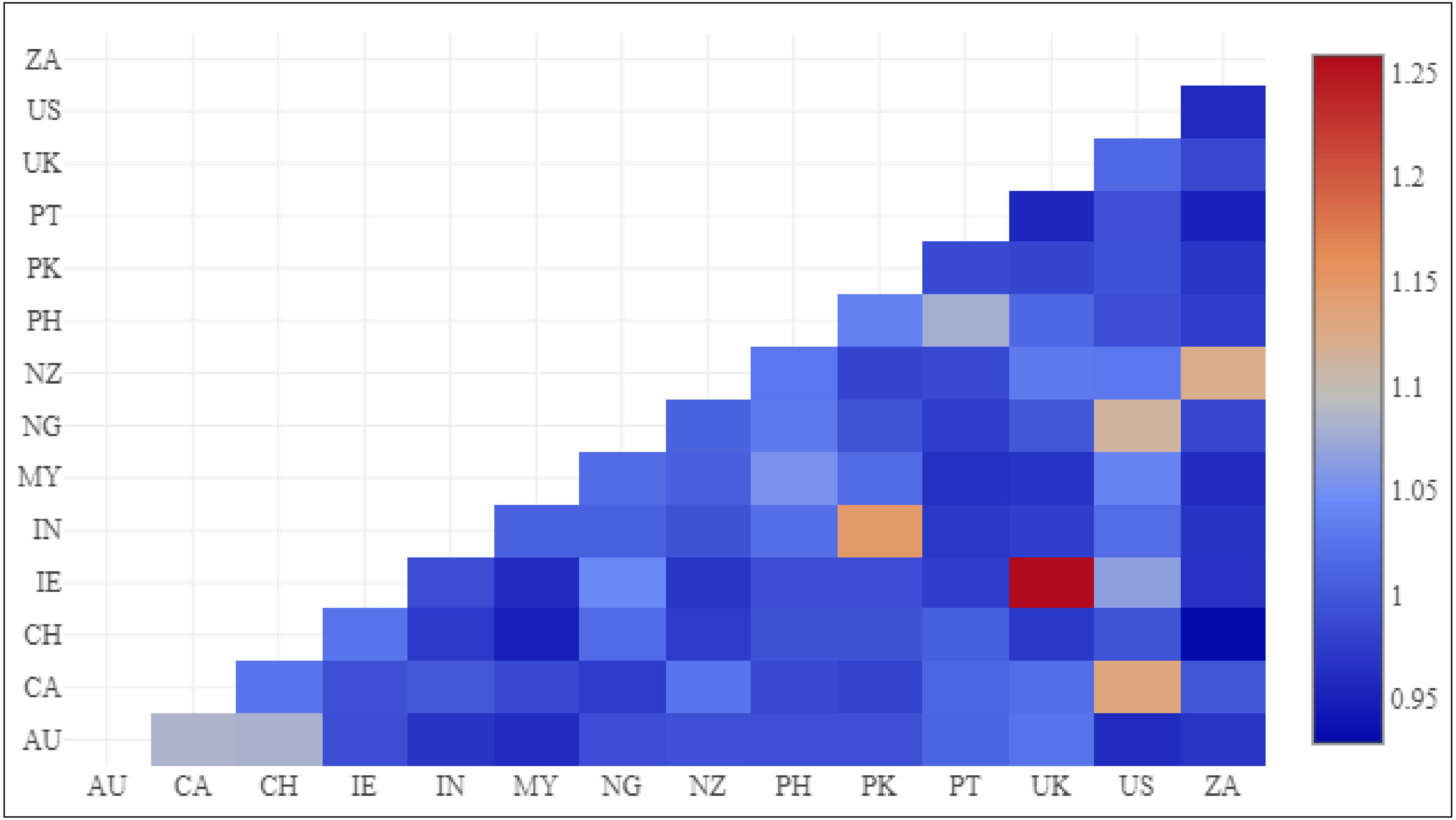
6. Pruning an umbrella-grammar to fit a dialect is fine… adding constructions is a challenge

# Thanks!

Jonathan Dunn
jonathan.dunn@canterbury.ac.nz
www.jdunn.name

UC
UNIVERSITY OF CANTERBURY
*Te Whare Wānanga o Waitaha*
CHRISTCHURCH NEW ZEALAND

# Dialect Classification (Similarity by cosine distance) (CxG-2 model, Common Crawl)



# Dialect Classification (grammar evaluation)

CxG-1 CxG-2

| | Frequency | Association | P |
|---|---|---|---|
| ara | 44.08% | **29.45%** | 0.0001 |
| deu | 52.49% | **18.69%** | 0.0001 |
| eng | 51.80% | **23.11%** | 0.0001 |
| fra | 43.28% | **40.52%** | 0.0037 |
| por | 45.13% | **38.91%** | 0.0137 |
| rus | 54.14% | **13.93%** | 0.0001 |
| spa | 60.34% | **26.36%** | 0.0001 |
| zho | 57.01% | **37.96%** | 0.0030 |

Compression = MDL Score / Baseline

(smaller is better)