Linguistic Universals

Group 8

Greenberg Rules

In this project, we are testing Greenberg's (1963) word-order universals using the World Atlas of Language Structures (WALS) data.

We chose 16 universals and matched each one with the right WALS features that describe things like word order, adpositions, and noun phrase structure.

WALS code	Feature name	Why it matters
49A	number of cases	shows whether a language uses case marking at all (needed for Universal 41)
50A	asymmetrical case marking	helps refine whether subject and object cases differ; supports testing of Universal 41

⋮

Greenberg #	Universal	WALS feature IDs
3	Languages with dominant VSO order are always prepositional.	81A, 85A
6	All languages with dominant VSO order have the adjective after the noun.	81A, 87A

⋮

Contingency Tables and WALS Data

- To test each Greenberg universal on the WALS data, we translated them into contingency tables.
- For Rule 4 *“With overwhelmingly more than chance frequency, languages with normal SOV order are postpositional.”*

adp_type	postposition	preposition
word_order_group		
SOV	381	17
non-SOV	172	487

- In this case the contingency table supports the rule.

Contingency Tables and WALS Data - Results

- Overall we found that Greenberg rules fall into two categories with respect to their empirical support in the WALS data.
- Strongly supported
  - Rules linking word order and **adposition type**
  - About 90 % of VO languages use prepositions
  - About 95 % of OV languages use postpositions.
- Weakly supported
  - Rules linking word order and **adjective position**
  - A high number of counterexamples occur in both VO and OV languages.
  - For example Rule 24 "*If in a language the verb follows the object, the adjective likewise follows the noun.*" is only supported by 52 % of languages.

Greenberg Rule 20

If a language puts words like adjectives, numbers, or “this/that” before a noun, then it usually also puts the possessor (genitive) before the noun.

English puts modifiers before the noun:

- that house
- big house
- five houses

And English also puts the possessor before the noun:

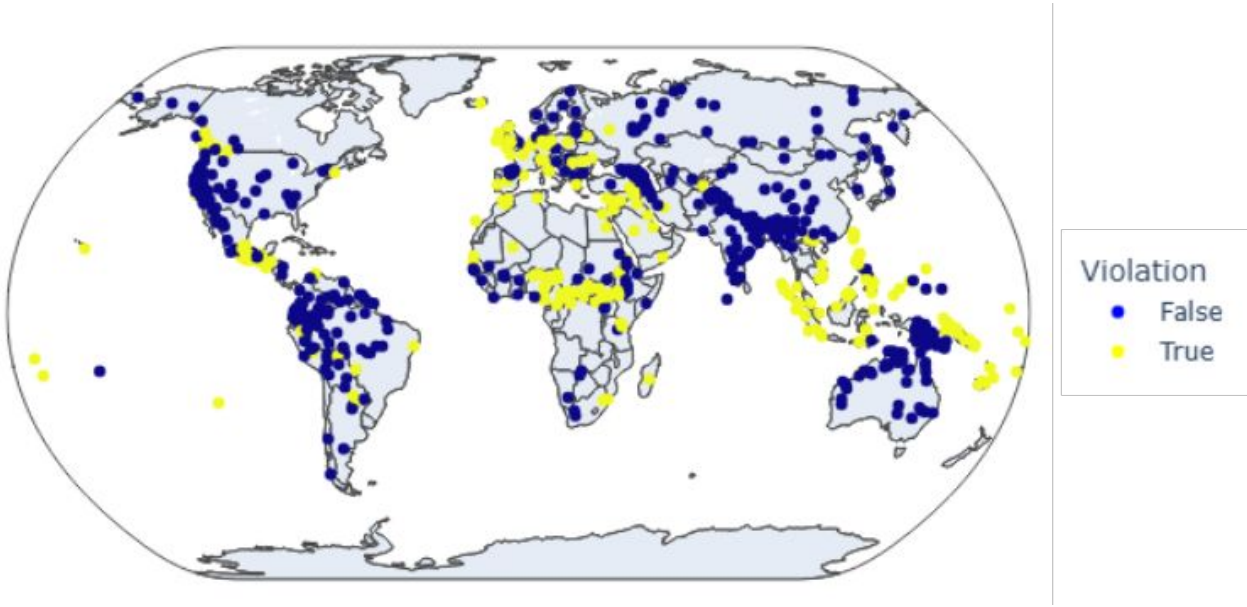
- Anna’s house
- the girl’s book

Greenberg Rule 20

Clusters can be seen in:

- Central and West Africa
- Southeast Asia
- Papuanesia
- Europe
- Mesoamerica

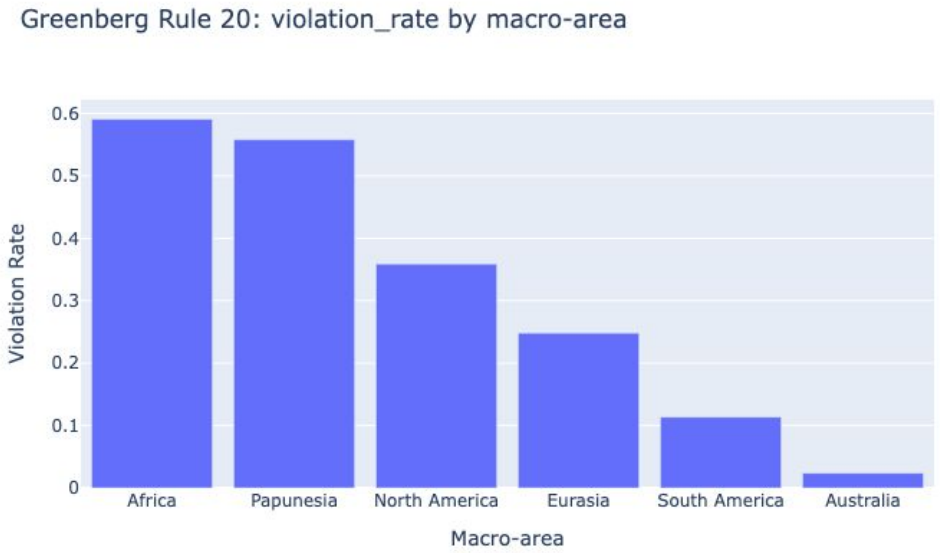
Australia is very consistent



Greenberg Rule 20

Rule 20 is not universal. It is regional dependent

- Africa and Papunesia have the highest violations with more than half the language breaking this rule
- In North America and Europe, the rule works in some languages but fails in many others
- In South America and especially Australia, there are almost no violations



Greenberg Rule 20

Here are some examples of exceptions showing the placement of the **modifier**, **possessor**, and **noun**:

African violation
- **Yoruba:** ilé ńlá yìí ti John (**house big this of John**)
- modifiers before noun, possessor after noun

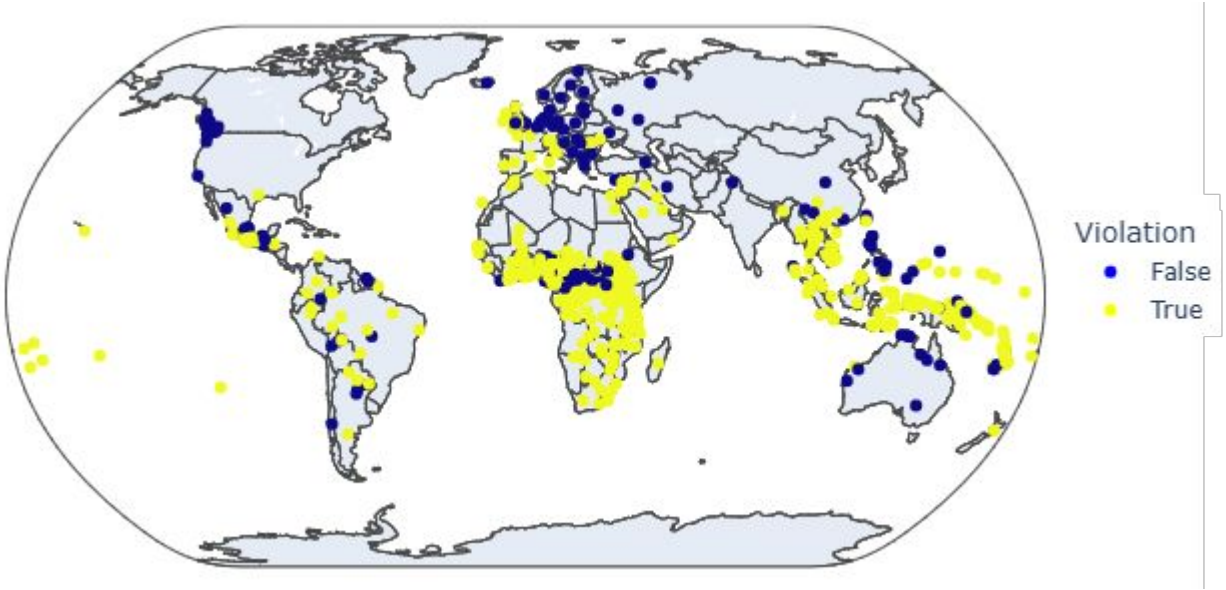
Southeast Asian violation
- **Indonesian: rumah besar itu milik John** (**house big** that **owned by John**)
- Possessor after noun

Pacific Northwest violation
- **Kwakw'ala: gan'əm'a baƛ^{w}as =əs John** (**big house John**)
- adjective position flexible, possessor consistently follows noun

European violation
- **German** allows two possessor constructions:
  - before noun: **Johanns großes Haus**
  - after noun: **das große Haus von Johann**
- So even though modifiers come before the noun, possessors do not have to, which breaks Rule 20.

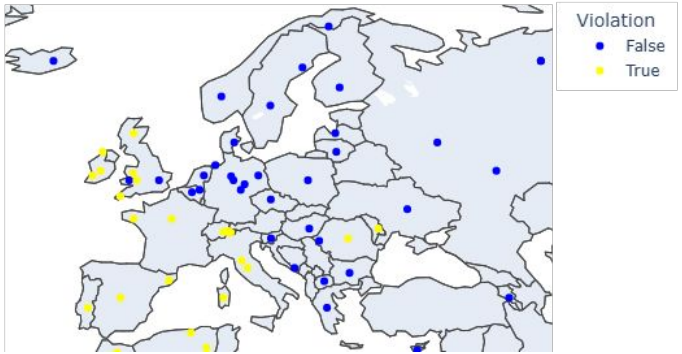
Greenberg Rule 23

If in a language the verb precedes the object, the adjective likewise precedes the noun.



Greenberg Rule 23

- Germanic languages
  - English “I bought a **red** car.”
  - German “Ich habe ein **rotes** Auto gekauft.”
  - Swedish “Jag köpte en **röd** bil.”
- Romance languages
  - Spanish “Compré un coche **rojo**.”
  - Italian “Ho comprato una macchina **rossa**.”
- Celtic languages
  - Welsh “Prynais gar **coch**.”
  - Irish “Cheannaigh mé carr **dearg**.”
- Slavic languages
  - Polish “Kupiłem **czerwony** samochód.”
  - Russian “Ya kupíl **krásnuyu** mashínu.”



Creating new rules through ML

- Can ML-methods on WALS-data be used to predict some attribute from other attributes?

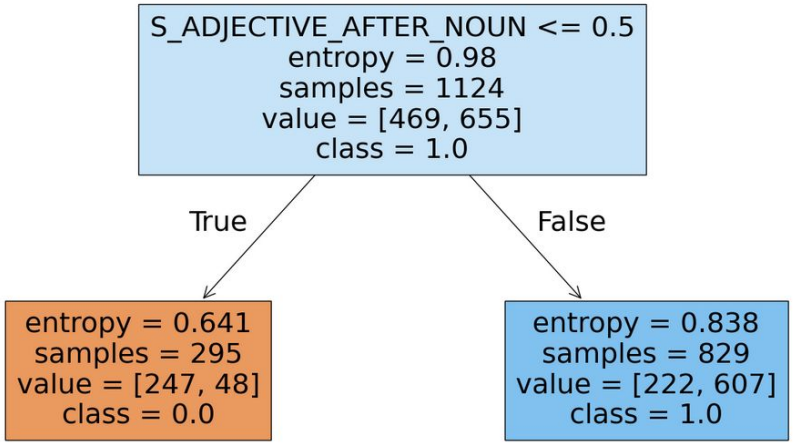
Creating new rules

- The original rules deduced by Greenberg in the 60's was only based on a set of 30 languages
- WALS data contains linguistic features for roughly 2,500 languages
- The goal was to see if we can use this large database to create new rules, by applying machine learning algorithms to predict a specific feature from the other features
- A white-box method was necessary since we wanted to be able to translate the results to the style of the original rules,
- e.g. “All languages with dominant VSO order have the adjective after the noun.”

Decision Tree for creating new rules

- We worked with a decision tree to get be able to deduce these rules

- First iteration with a very shallow tree, as to not create too complex rules
- The rules deduced from the decision tree was in many cases similar to the original rules



- **If the word order is SOV, then adpositions follow the noun.**
  Accuracy: 83.7 %, Correct/Incorrect: 525 / 102

  Related Greenberg rules:

  - Rule 4: “With overwhelmingly more than chance frequency, languages with normal SOV order are postpositional.”
  - Rule 17: “With overwhelmingly more than chance frequency, languages with dominant order SOV are postpositional.”
  - Rule 26: “If a language has dominant order SOV, it generally has postpositions.”

- **If demonstratives follow the noun, then adjectives follow the noun.** Accuracy: 91.3 %, Correct/Incorrect: 489 / 46

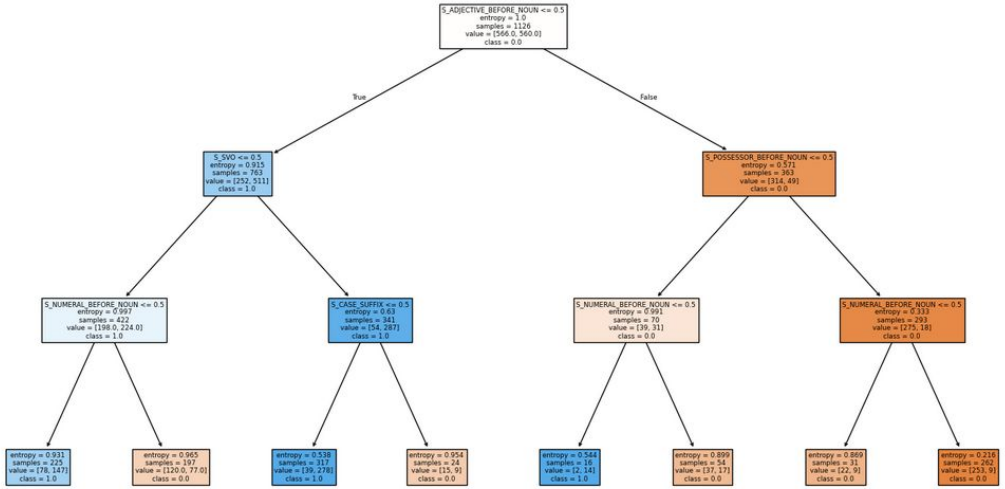
  Explanation: Demonstratives are used to indicate which entities are being referred to

  Related Greenberg rule:

  - Rule 19: “When the descriptive adjective follows the noun, the demonstrative and the numeral likewise follow.”

Deeper decision tree

- We extended the decision tree to be able to create more accurate rules, which are more complex
- Some rules where very accurate:



If demonstratives and possessors precedes the noun, and relatives follows the noun, then adjectives follows the noun
Accuracy: 97.5%, Correct/Incorrect: 199/5

If the word order is not SOV and demonstrative words follow the noun, then the language does not have a case system
Accuracy: 85.7%, Correct/Incorrect: 419/70

If adpositions and demonstratives follows the noun and relatives precedes the noun, then possessors precedes the noun Accuracy: 100 %, Correct/Incorrect: 206/0

Testing Greenberg’s Universals with German Corpus Data

Why Corpus Data?

- WALS: One general description per language
- UD: Real sentences with syntatic annotation

Data and method:

- German HDT treebank (Universal Dependencies)
- CoNLL-U format
- Dependency trees with morphology and word order
- Automatic extraction with Python

What do Universal Dependency Data look like?

...

	id	form	lemma	upos	xpos	feats	head	deprel	deps	misc
0	1	Konkursgerüchte	Konkursgerücht	NOUN	NN	{'Case': 'Nom', 'Gender': 'Neut', 'Number': 'P...	2	nsubj	None	None
1	2	drücken	drücken	VERB	VVFIN	{'Mood': 'Ind', 'Number': 'Plur', 'Person': '3...	0	root	None	None
2	3	Kurs	Kurs	NOUN	NN	{'Case': 'Acc', 'Gender': 'Masc', 'Number': 'S...	2	obj	None	None
3	4	der	der	DET	ART	{'Case': 'Gen', 'Definite': 'Def', 'Gender': '...	5	det:poss	None	None
4	5	Amazon-Aktie	Amazon-Aktie	NOUN	NN	{'Case': 'Gen', 'Gender': 'Fem', 'Number': 'Si...	3	nmod:poss	None	None

Sentence: Konkursgerüchte drücken Kurs der Amazon-Aktie.

- One row = one word
- Columns describe grammatical structure
- This structure allows automatic analysis

Greenberg Rule 20

Reminder: If modifiers precede the noun, the genitive possessor almost always precedes the noun.

Method:

- Modifier before noun
- Genitive possessor present
- Check genitive position

Results:

- 845 test cases
- 108 hold (13%)
- 736 exceptions (87%)

Examples:

- die Nutzung des Internet
- das letzte Quartal dieses Jahres

Greenberg Rule 23

Reminder: If the verb precedes the object, the adjective precedes the noun.

Method:

- Identify VO sentences
- Check adjective position

Results:

- 4,887 test cases
- 4,887 hold
- 0 exceptions
- Support rate: 1.0

Examples:

- der große Hund
- das alte Haus

Greenberg Rule 24

Reminder: If the verb follows the object, the adjective follows the noun.

Method:

- Identify OV sentences
- Check adjective position

Results:

- 7,571 test cases
- 0 hold
- 7,571 exceptions
- Support rate: 0.0

Conclusion Testing Greenberg Rules with UD Data

- WALS gives typological tendencies
- UD shows real usage
- Corpus testing is crucial for evaluating universals

Summary

- Greenbergs linguistic universals holds true in a lot of cases but there are also many exceptions
- We have tried to study these exceptions on a geographical level which gave some insights into trends of the exceptions, but also no clear patterns
- Using the WALS dataset, we were able to deduce some new rules which in some cases were similar to the original universals, but also create new once with a high accuracy
- Corpus testing is crucial for evaluating universals, and WALS data are simplifications, showing generalized properties of languages