Statistisch bewijs voor univesalia

Milan Lopuhaä[1]

[1]Institute of Mathematics, Astrophysics and Particle Physics
Radboud Universiteit Nijmegen

21 maart 2017

Universals

- Universal: eigenschap die alle natuurlijke talen hebben

Universals

- Universal: eigenschap die alle natuurlijke talen hebben
- Voorbeeld: ‘alle talen hebben klinkerfonemen’

Universals

- Universal: eigenschap die alle natuurlijke talen hebben
- Voorbeeld: ‘alle talen hebben klinkerfonemen’
- Probleem: er bestaan nauwelijks tot geen onbetwiste voorbeelden (‘Kabardisch heeft geen klinkerfonemen’)

Hoe vind je universals?

Methode 1 (voor generativisten):

- Bedenk een theorie over de werking van menselijke taal;

Hoe vind je universals?

Methode 1 (voor generativisten):

- Bedenk een theorie over de werking van menselijke taal;
- Leid daaruit universals af;

Hoe vind je universals?

Methode 1 (voor generativisten):

- Bedenk een theorie over de werking van menselijke taal;
- Leid daaruit universals af;
- Overtuig alle taalkundigen van je theorie.

Hoe vind je universals?

Methode 1 (voor generativisten):

- Bedenk een theorie over de werking van menselijke taal;
- Leid daaruit universals af;
- Overtuig alle taalkundigen van je theorie.

Voorbeeld: *Phase Impenetrability Condition: If X is dominated by a complement of a phase YP, X cannot move out of YP.*

Hoe vind je universals?

Methode 1 (voor generativisten):

- Bedenk een theorie over de werking van menselijke taal;
- Leid daaruit universals af;
- Overtuig alle taalkundigen van je theorie.

Voorbeeld: *Phase Impenetrability Condition: If X is dominated by a complement of a phase YP, X cannot move out of YP.*
Probleem: er is nauwelijks taalkundige theorie die door alle taalkundigen wordt gevolgd.

Hoe vind je universals?

Methode 2 (voor typologen):

- Bedenk een eigenschap die wel eens universeel zou kunen zijn;

Hoe vind je universals?

Methode 2 (voor typologen):

- Bedenk een eigenschap die wel eens universeel zou kunen zijn;
- Controleer het in een hele hoop talen.

Hoe vind je universals?

Methode 2 (voor typologen):

- Bedenk een eigenschap die wel eens universeel zou kunen zijn;
- Controleer het in een hele hoop talen.

Probleem: hoeveel talen zijn voldoende?

Hoeveel talen zijn voldoende?

- Bij geen enkele hoeveelheid heb je absolute zekerheid;

Hoeveel talen zijn voldoende?

- Bij geen enkele hoeveelheid heb je absolute zekerheid;
- Met statistiek kunnen we wel de onzekerheid kwantificeren.

Hoeveel talen zijn voldoende?

- Bij geen enkele hoeveelheid heb je absolute zekerheid;
- Met statistiek kunnen we wel de onzekerheid kwantificeren.
- Sowieso moet onze steekproef van talen bestaan uit talen uit zo veel mogelijk families en taalgebieden.

Hoeveel talen zijn voldoende?

- *World Atlas of Linguistic Structures*: 192 features van 2679 talen, lang niet allemaal compleet.

Hoeveel talen zijn voldoende?

- *World Atlas of Linguistic Structures*: 192 features van 2679 talen, lang niet allemaal compleet.
- Voorbeeld: *fixed stress location*: 502 talen onderzocht:

geen vaste klemtoon	220
eerste	92
tweede	16
derde	1
voorvoorlaatste	12
voorlaatste	110
laatste	51

Hoeveel talen zijn voldoende?

- één taal met vaste klemtoon op de derde lettergreep: Winnebago (Siouxtaal)

Hoeveel talen zijn voldoende?

- één taal met vaste klemtoon op de derde lettergreep: Winnebago (Siouxtaal)
- Als deze steekproef van talen Assiniboine (ook Siouxtaal, fonemische klemtoon) had gehad in plaats van Winnebago, hadden we *geen* talen met klemtoon op 3e lettergreep;

Hoeveel talen zijn voldoende?

- één taal met vaste klemtoon op de derde lettergreep: Winnebago (Siouxtaal)
- Als deze steekproef van talen Assiniboine (ook Siouxtaal, fonemische klemtoon) had gehad in plaats van Winnebago, hadden we *geen* talen met klemtoon op 3e lettergreep;
- in dat geval hadden we 'bewijs' gevonden van universal 'er is geen taal met vaste klemtoon op de 3e lettergreep'.

Maak je eigen universal

- Implicationele universal: ‘elke taal met X heeft Y.’

Maak je eigen universal

- Implicationele universal: ‘elke taal met X heeft Y.’
- Enige tegenvoorbeeld: wel X, niet Y.

Maak je eigen universal

- Implicationele universal: ‘elke taal met X heeft Y.’
- Enige tegenvoorbeeld: wel X, niet Y.
- Kies een zeldzame X en een vaak voorkomende $Y \Rightarrow$ universal gevonden!

Maak je eigen universal

- Implicationele universal: ‘elke taal met X heeft Y.’
- Enige tegenvoorbeeld: wel X, niet Y.
- Kies een zeldzame X en een vaak voorkomende $Y \Rightarrow$ universal gevonden!
- Voorbeelden:
  - ‘Elke taal met een vaste klemtoon op de 3e lettergreep heeft geen /ŋ/’ (174 talen)
  - ‘Elke taal met VOS heeft bilabiale fonemen’ (381 talen)
  - ‘Elke taal waarin meervoud met toon gemarkeerd wordt heeft geen syllabefinale stops’ (297 talen)

Statistische toets

- Probleem: hoe onderscheiden we heel zeldzame eigenschappen van nooit voorkomende eigenschappen?

Statistische toets

- Probleem: hoe onderscheiden we heel zeldzame eigenschappen van nooit voorkomende eigenschappen?
- Statistiek: maak een *toets* om van een eigenschap te kijken of deze zeldzaam, of nooit voorkomend is.

Statistische toets

- Probleem: hoe onderscheiden we heel zeldzame eigenschappen van nooit voorkomende eigenschappen?
- Statistiek: maak een *toets* om van een eigenschap te kijken of deze zeldzaam, of nooit voorkomend is.
- We nemen aan dat de eigenschap zeldzaam is (nulhypothese), en pas bij voldoende tegenbewijs nemen we aan dat de eigenschap nooit voorkomt (alternatieve hypothese).

Statistische toets

		Werkelijkheid	
		zeldzaam	nooit
Toets wijst uit	zeldzaam	*blijdschap!*	Type II
	nooit	Type I	*blijdschap!*

Statistische toets

		Werkelijkheid	
		zeldzaam	nooit
Toets wijst uit	zeldzaam	*blijdschap!*	Type II
	nooit	Type I	*blijdschap!*

- Idealiter: minimaliseer de kans op fouten Type I & II.

Statistische toets

		Werkelijkheid	
		zeldzaam	nooit
Toets wijst uit	zeldzaam	*blijdschap!*	Type II
	nooit	Type I	*blijdschap!*

- Idealiter: minimaliseer de kans op fouten Type I & II.
- Probleem: deze kansen zijn niet te bepalen! (hangt af van hoeveel eigenschappen 'echt' universals zijn, en dat weten we niet)

Statistische toets

		Werkelijkheid	
		zeldzaam	nooit
Toets wijst uit	zeldzaam	*blijdschap!*	Type II
	nooit	Type I	*blijdschap!*

- Idealiter: minimaliseer de kans op fouten Type I & II.
- Probleem: deze kansen zijn niet te bepalen! (hangt af van hoeveel eigenschappen 'echt' universals zijn, en dat weten we niet)
- Oplossing: kijk naar kans op Type I *gegeven dat de eigenschap die we bekijken zeldzaam is*.

Statistische toets

		Werkelijkheid	
		zeldzaam	nooit
Toets wijst uit	zeldzaam	*blijdschap!*	Type II
	nooit	Type I	*blijdschap!*

- Idealiter: minimaliseer de kans op fouten Type I & II.
- Probleem: deze kansen zijn niet te bepalen! (hangt af van hoeveel eigenschappen 'echt' universals zijn, en dat weten we niet)
- Oplossing: kijk naar kans op Type I *gegeven dat de eigenschap die we bekijken zeldzaam is*.
- Standaard in sociale wetenschappen: maak test zo dat deze kans $\geq 5\%$.

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.
- Vraag: hoe groot moet N zijn zodat de toets adequaat is?

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.
- Vraag: hoe groot moet N zijn zodat de toets adequaat is?
- Per definitie is een Type II-fout onmogelijk

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.
- Vraag: hoe groot moet N zijn zodat de toets adequaat is?
- Per definitie is een Type II-fout onmogelijk
- Wat is de kans op Type I-fout gegeven zeldzame eigenschap?

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.
- Vraag: hoe groot moet N zijn zodat de toets adequaat is?
- Per definitie is een Type II-fout onmogelijk
- Wat is de kans op Type I-fout gegeven zeldzame eigenschap?
  - Hangt af van de *kansverdeling* van de verdeling van deze eigenschap over de talen van de wereld

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.
- Vraag: hoe groot moet N zijn zodat de toets adequaat is?
- Per definitie is een Type II-fout onmogelijk
- Wat is de kans op Type I-fout gegeven zeldzame eigenschap?
  - Hangt af van de *kansverdeling* van de verdeling van deze eigenschap over de talen van de wereld
  - Simpelste model: elke taal heeft een vaste kans θ om de zeldzame eigenschap te hebben, en deze kansen zijn onafhankelijk

Statistische toets

- Onze toets: N talen, verspreid over zo veel mogelijk taalfamilies/gebieden.
  - Onderzoek deze talen op eigenschap X
  - 0 talen met X: 'niet X' is universal.
  - ≥ 1 talen met X: 'niet X' is niet universal.
- Vraag: hoe groot moet N zijn zodat de toets adequaat is?
- Per definitie is een Type II-fout onmogelijk
- Wat is de kans op Type I-fout gegeven zeldzame eigenschap?
  - Hangt af van de *kansverdeling* van de verdeling van deze eigenschap over de talen van de wereld
  - Simpelste model: elke taal heeft een vaste kans θ om de zeldzame eigenschap te hebben, en deze kansen zijn onafhankelijk
  - Onrealistisch: geen rekening gehouden met taalcontact, genetische verwantschap, etc...

Statistische toets

- Model kansverdeling: elke taal heeft een vaste kans θ om de zeldzame eigenschap te hebben, en deze kansen zijn onafhankelijk.

Statistische toets

- Model kansverdeling: elke taal heeft een vaste kans θ om de zeldzame eigenschap te hebben, en deze kansen zijn onafhankelijk.
- Onder deze aanname: kans op fout Type I gegeven zeldzame eigenschap: $(1-\theta)^N$.

Statistische toets

- Model kansverdeling: elke taal heeft een vaste kans θ om de zeldzame eigenschap te hebben, en deze kansen zijn onafhankelijk.
- Onder deze aanname: kans op fout Type I gegeven zeldzame eigenschap: $(1-\theta)^N$.
- Conclusie: N moet zo groot zijn dat $(1-\theta)^N \leq 0,05$.

Statistische toets

- Model kansverdeling: elke taal heeft een vaste kans θ om de zeldzame eigenschap te hebben, en deze kansen zijn onafhankelijk.
- Onder deze aanname: kans op fout Type I gegeven zeldzame eigenschap: $(1-\theta)^N$.
- Conclusie: N moet zo groot zijn dat $(1-\theta)^N \leq 0,05$.
- Voorbeeld:
  - $\theta = 0,1$: $N = 29$.
  - $\theta = 0,01$: $N = 299$.

Bepaling van θ

- Hoe bepalen we θ?
- Door alle zeldzame eigenschappen in de WALS te bestuderen, kunnen we θ benaderen.

Bepaling van θ

- Hoe bepalen we θ?
- Door alle zeldzame eigenschappen in de WALS te bestuderen, kunnen we θ benaderen.
- Methode:
  - Neem alle eigenschappen die $\leq 1\%$ voorkomen.
  - Gooi degenen eruit die een significant ander voorkomstpercentage hebben dan de rest (hier zijn statistische methodes voor).
  - Gooi de data samen en verkrijg θ.

Bepaling van θ

- Hoe bepalen we θ?
- Door alle zeldzame eigenschappen in de WALS te bestuderen, kunnen we θ benaderen.
- Methode:
  - Neem alle eigenschappen die $\leq 1\%$ voorkomen.
  - Gooi degenen eruit die een significant ander voorkomstpercentage hebben dan de rest (hier zijn statistische methodes voor).
  - Gooi de data samen en verkrijg θ.
- Resultaat: 28 features die $\leq 1\%$ voorkomen, 4 ervan uitgesloten vanwege te hoog voorkomstpercentage, totaal: $\theta = 0,0031$.

Bepaling van θ

- Hoe bepalen we θ?
- Door alle zeldzame eigenschappen in de WALS te bestuderen, kunnen we θ benaderen.
- Methode:
  - Neem alle eigenschappen die $\leq 1\%$ voorkomen.
  - Gooi degenen eruit die een significant ander voorkomstpercentage hebben dan de rest (hier zijn statistische methodes voor).
  - Gooi de data samen en verkrijg θ.
- Resultaat: 28 features die $\leq 1\%$ voorkomen, 4 ervan uitgesloten vanwege te hoog voorkomstpercentage, totaal: $\theta = 0,0031$.
- In dit geval: $N = 964$.

Kanttekeningen

- Als het acceptabel is dat bij een zeldzame eigenschap de toets het in hoogstens 5% van de gevallen bij het foute eind heeft, dan is het bekijken van 964 talen voldoende.

Kanttekeningen

- Als het acceptabel is dat bij een zeldzame eigenschap de toets het in hoogstens 5% van de gevallen bij het foute eind heeft, dan is het bekijken van 964 talen voldoende.
- Dit is maar bij 19 van de 192 onderzoeken in de WALS gedaan.

Kanttekeningen

- Als het acceptabel is dat bij een zeldzame eigenschap de toets het in hoogstens 5% van de gevallen bij het foute eind heeft, dan is het bekijken van 964 talen voldoende.
- Dit is maar bij 19 van de 192 onderzoeken in de WALS gedaan.
- Maar: we willen niet een toets om te zien of één specifieke eigenschap universeel is, we willen weten of er überhaupt universele eigenschappen bestaan.

Kanttekeningen

- Als het acceptabel is dat bij een zeldzame eigenschap de toets het in hoogstens 5% van de gevallen bij het foute eind heeft, dan is het bekijken van 964 talen voldoende.
- Dit is maar bij 19 van de 192 onderzoeken in de WALS gedaan.
- Maar: we willen niet een toets om te zien of één specifieke eigenschap universeel is, we willen weten of er überhaupt universele eigenschappen bestaan.
- Er zijn 2029 voorstellen voor universals in *The Universals Archive*; zelfs als het hier gaat om bijna altijd voorkomende eigenschappen ipv universals, dan zou onze test alsnog ± 101 universals aanwijzen.

Kanttekeningen

- Als het acceptabel is dat bij een zeldzame eigenschap de toets het in hoogstens 5% van de gevallen bij het foute eind heeft, dan is het bekijken van 964 talen voldoende.
- Dit is maar bij 19 van de 192 onderzoeken in de WALS gedaan.
- Maar: we willen niet een toets om te zien of één specifieke eigenschap universeel is, we willen weten of er überhaupt universele eigenschappen bestaan.
- Er zijn 2029 voorstellen voor universals in *The Universals Archive*; zelfs als het hier gaat om bijna altijd voorkomende eigenschappen ipv universals, dan zou onze test alsnog ± 101 universals aanwijzen.
- Om dit te voorkomen, moeten we de grens veel lager leggen dan 5%, bijvoorbeeld $0,1\%$.

Kanttekeningen

- Als het acceptabel is dat bij een zeldzame eigenschap de toets het in hoogstens 5% van de gevallen bij het foute eind heeft, dan is het bekijken van 964 talen voldoende.
- Dit is maar bij 19 van de 192 onderzoeken in de WALS gedaan.
- Maar: we willen niet een toets om te zien of één specifieke eigenschap universeel is, we willen weten of er überhaupt universele eigenschappen bestaan.
- Er zijn 2029 voorstellen voor universals in *The Universals Archive*; zelfs als het hier gaat om bijna altijd voorkomende eigenschappen ipv universals, dan zou onze test alsnog ± 101 universals aanwijzen.
- Om dit te voorkomen, moeten we de grens veel lager leggen dan 5%, bijvoorbeeld $0,1\%$.
- In dit geval zijn minstens 2225 talen nodig.

Kanttekeningen

- Wat als we wel rekening houden met taalcontact/genetische verwantschap?

Kanttekeningen

- Wat als we wel rekening houden met taalcontact/genetische verwantschap?
- Nog meer talen nodig (in je model lijken talen meer op elkaar, dus bekijk je effectief minder verschillende talen)

Kanttekeningen

- Wat als we wel rekening houden met taalcontact/genetische verwantschap?
- Nog meer talen nodig (in je model lijken talen meer op elkaar, dus bekijk je effectief minder verschillende talen)
- Stuk ingewikkelder om uit te rekenen!

Conclusie

- Steekproeven zijn ontoereikend om het bestaan van universals aan te tonen.

Conclusie

- Steekproeven zijn ontoereikend om het bestaan van universals aan te tonen.
- Wantrouw elke uitspraak van de vorm 'alle talen hebben eigenschap X'

Conclusie

- Steekproeven zijn ontoereikend om het bestaan van universals aan te tonen.
- Wantrouw elke uitspraak van de vorm 'alle talen hebben eigenschap X' .
- Met statistiek kunnen we onzekerheid kwantificeren en bepalen wat afdoende bewijs is in de taalwetenschap.