Om bruken av semantisk kart i språkforskning

Dag Haug

22. april 2024

Da og når

- Harald Eias teori

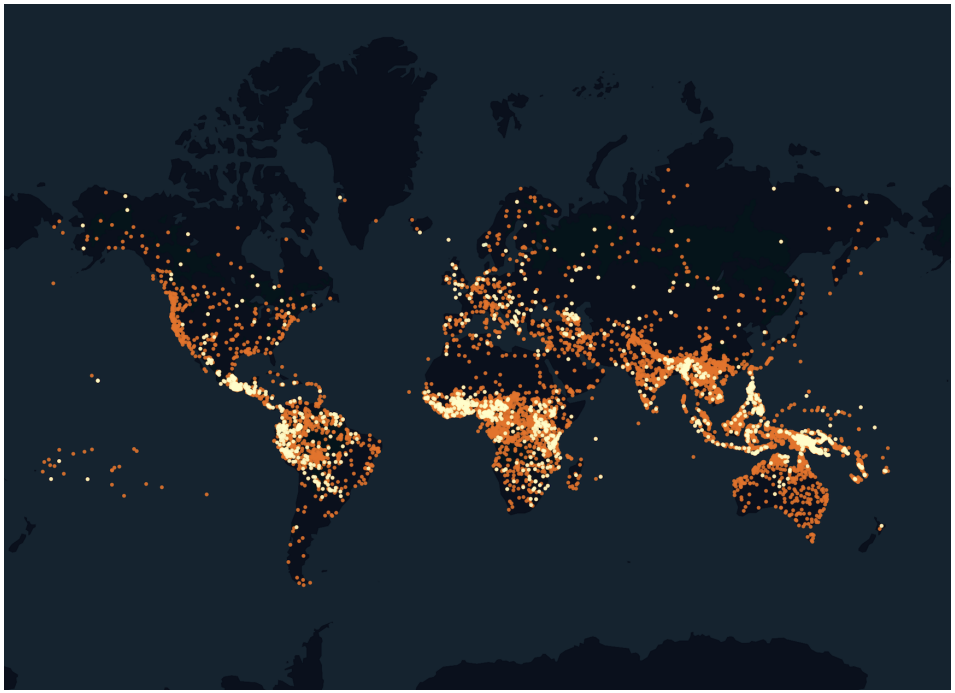
Da og når

- Harald Eias teori
- Er det sant at det ikke er noen forskjell?
- Kan vi gjenfinne forskjellen utafor de germanske språkene?
- Og hvordan?
- Dette var utgangspunktet for Haug & Pedrazzini (2023)

Probabilistiske semantiske kart

- Vi kan få nyttige data ved å studere **oversettelsesekvivalenter** til f.eks. *when* på tvers av mange språk
- Vi kan bruke de kontekstene *when* opptrer i som representanter for ulike deler av det funksjonelle domenet til *when*
- Systematiske likheter og forskjeller i hvordan ulike språk deler opp dette domenet kan representeres i probabilistiske semantiske kart (Wälchli & Cysow, 2012)

Data



Figur 1: Arealdistribusjon (etter Glottolog) i vårt datasett (gul) vs. verdens språk (oransje)

Stort parallellkorpus:
NT-oversettelser til >1400 *languoider* (137 familier/isolatspråk)
(Mayer & Cysouw, 2014)

Parallellstiling og distansematrise

- 1-til-1 ordparallellstilling (automatisk med SyMGIZA++)

Parallellstiling og distansematrise

- 1-til-1 ordparallellstilling (automatisk med SyMGIZA++)
- Trekke ut *when* og parallellene:

	eng	mri	por	fin	...	kaz	kor
1	*when*	no	quando	kun		қашан	때에
2	*when*	ka	quando	jolloin		кейін	때에
n	...	...	...	...	...	...	...

Figur 2: Eksempeldata

Parallellstiling og distansematrise

- 1-til-1 ordparallellstilling (automatisk med SyMGIZA++)
- Trekke ut *when* og parallellene:

	eng	mri	por	fin	...	kaz	kor
1	*when*	no	quando	kun		қашан	때에
2	*when*	ka	quando	jolloin		кейін	때에
n	...	...	...	...	...	...	...

Figur 2: Eksempeldata

- **Hamming-distanse**:
  - Hver rad i Figur 2 $\approx$ et bruksområde for WHEN

Parallellstiling og distansematrise

- 1-til-1 ordparallellstilling (automatisk med SyMGIZA++)
- Trekke ut *when* og parallellene:

	eng	mri	por	fin	...	kaz	kor
1	*when*	no	quando	kun		қашан	때에
2	*when*	ka	quando	jolloin		кейін	때에
n	...	...	...	...	...	...	...

Figur 2: Eksempeldata

- **Hamming-distanse**:
  - Hver rad i Figur 2 $\approx$ et bruksområde for WHEN
  - "Avstand" mellom to bruksområder $\approx$ *hvor mange språk som bruker forskjellige ord* for dem

Parallellstiling og distansematrise

- 1-til-1 ordparallellstilling (automatisk med SyMGIZA++)
- Trekke ut *when* og parallellene:

	eng	mri	por	fin	...	kaz	kor
1	*when*	no	quando	kun		қашан	때에
2	*when*	ka	quando	jolloin		кейін	때에
n	...	...	...	...	...	...	...

Figur 2: Eksempeldata

- **Hamming-distanse**:
  - Hver rad i Figur 2 $\approx$ et bruksområde for WHEN
  - "Avstand" mellom to bruksområder $\approx$ *hvor mange språk som bruker forskjellige ord* for dem
  - Vi måler alle avstandene og setter dem forskjellen mellom alle radene og representerer i en avstandsmatrise

Multidimensjonal skalering

- Fra en avstandmatrise kan vi konstruere en mengde punkter slik at avstandene mellom dem svarer til matrisen.

Multidimensjonal skalering

- Fra en avstandmatrise kan vi konstruere en mengde punkter slik at avstandene mellom dem svarer til matrisen.
- En nøyaktig representasjon vil kreve svært mange dimensjoner, men vi fokuserer på de to dimensjonene som inneholder de største avstandene

Multidimensjonal skalering

- Fra en avstandmatrise kan vi konstruere en mengde punkter slik at avstandene mellom dem svarer til matrisen.
- En nøyaktig representasjon vil kreve svært mange dimensjoner, men vi fokuserer på de to dimensjonene som inneholder de største avstandene
- Metoden er grundig forklart på https://daghaug.github.io/mds/

Multidimensjonal skalering

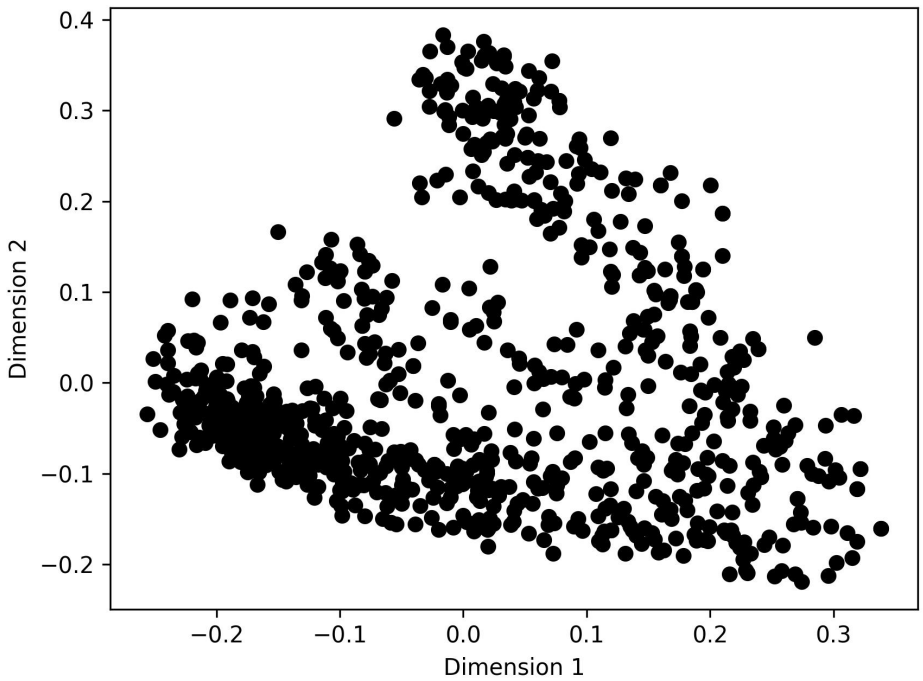
- Fra en avstandmatrise kan vi konstruere en mengde punkter slik at avstandene mellom dem svarer til matrisen.
- En nøyaktig representasjon vil kreve svært mange dimensjoner, men vi fokuserer på de to dimensjonene som inneholder de største avstandene
- Metoden er grundig forklart på https://daghaug.github.io/mds/
- Da får vi punkter som vi kan plotte som et kart.

	0	1	2	...	473	474	475
0	0.000000	0.717813	0.826331	...	0.697500	0.774194	0.839394
1	0.717813	0.000000	0.831435	...	0.681905	0.733471	0.786543
2	0.826331	0.831435	0.000000	...	0.797583	0.828767	0.889262
3	0.651341	0.459270	0.770732	...	0.598077	0.676275	0.791045
4	0.802083	0.762994	0.575499	...	0.731507	0.783951	0.839344
...	...	...	...	...	...	...	...
471	0.812709	0.793103	0.843373	...	0.715254	0.747368	0.810277
472	0.786372	0.739071	0.855172	...	0.654851	0.740000	0.832962
473	0.697500	0.681905	0.797583	...	0.000000	0.657382	0.792023
474	0.774194	0.733471	0.828767	...	0.657382	0.000000	0.823151
475	0.839394	0.786543	0.889262	...	0.792023	0.823151	0.000000

	x	y
0	0.153564	-0.077936
1	-0.180358	-0.079010
2	0.294392	-0.123286
3	-0.208843	-0.074415
4	0.052290	0.275351
...	...	...
471	-0.116570	-0.113176
472	-0.093750	-0.159694
473	0.177257	-0.140021
474	-0.170266	0.051007
475	-0.090192	0.092315

Figur 3: **Venstre**: Matrise med Hamming-avstander . **Høyre**: 2-dimensjonal matrise etter MDS med dimensjonsreduksjon

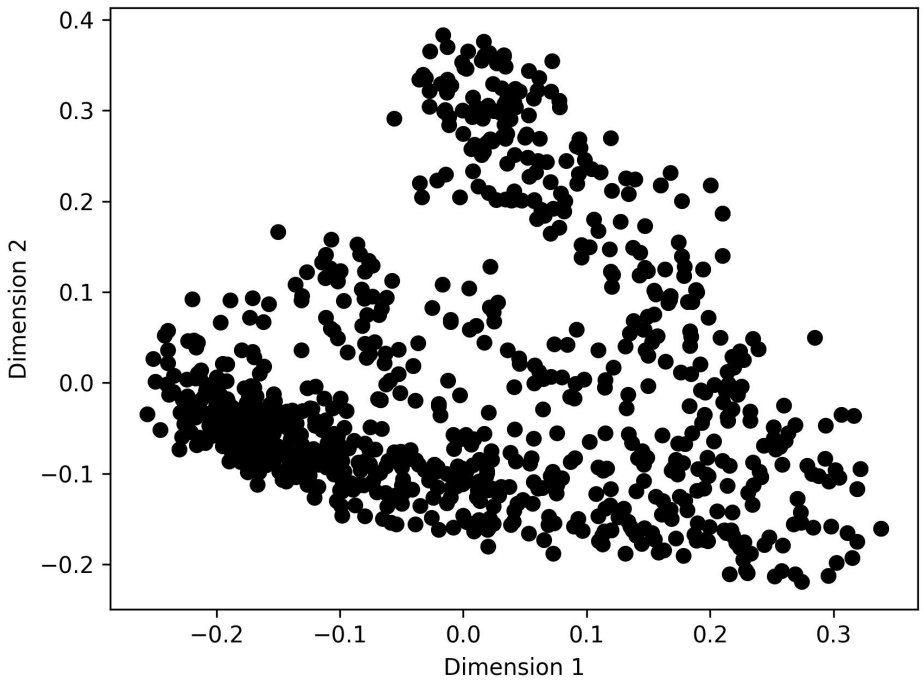
Semantisk kart over WHEN



Figur 4: WHEN-kart fra todimensjonal MDS

- Det "rå" MDS-kartet antyder at det er to-tre "bånd" med observasjoner

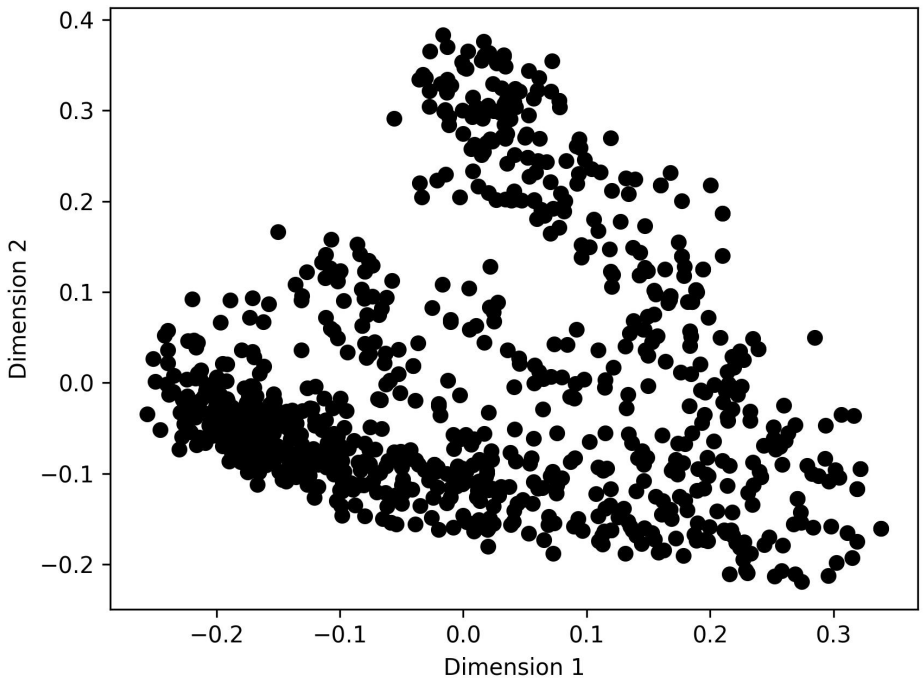
Semantisk kart over WHEN



Figur 4: WHEN-kart fra todimensjonal MDS

- Det "rå" MDS-kartet antyder at det er to-tre "bånd" med observasjoner
- Men hva betyr de? Vi har ingen forhåndsdefinerte merkelapper som definerer individuelle konstruksjoner

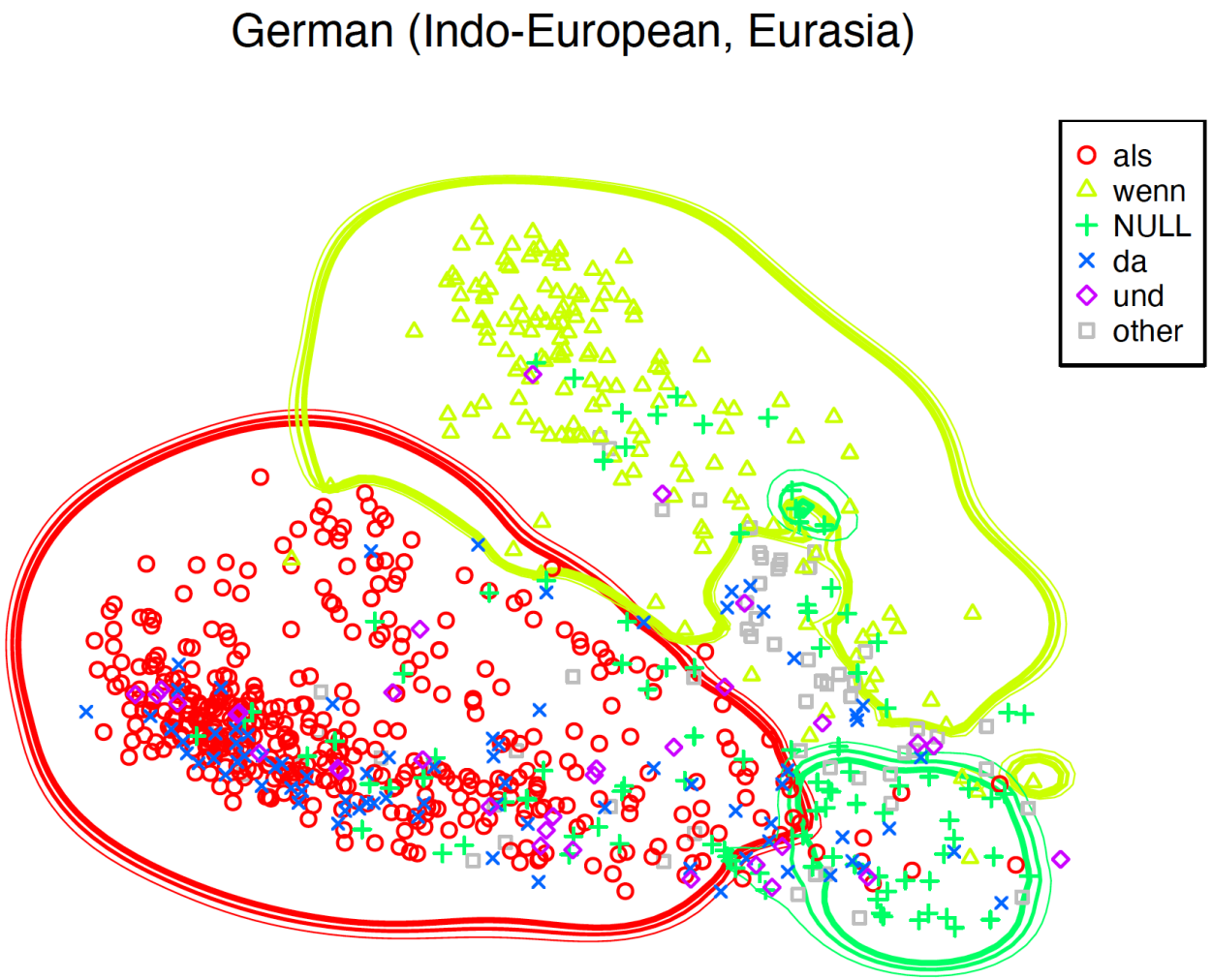
Semantisk kart over WHEN



Figur 4: WHEN-kart fra todimensjonal MDS

- Det "rå" MDS-kartet antyder at det er to-tre "bånd" med observasjoner
- Men hva betyr de? Vi har ingen forhåndsdefinerte merkelapper som definerer individuelle konstruksjoner
- Men for hvert språk, kan vi markere punktene med hvilken konstruksjon som brukes

Tysk

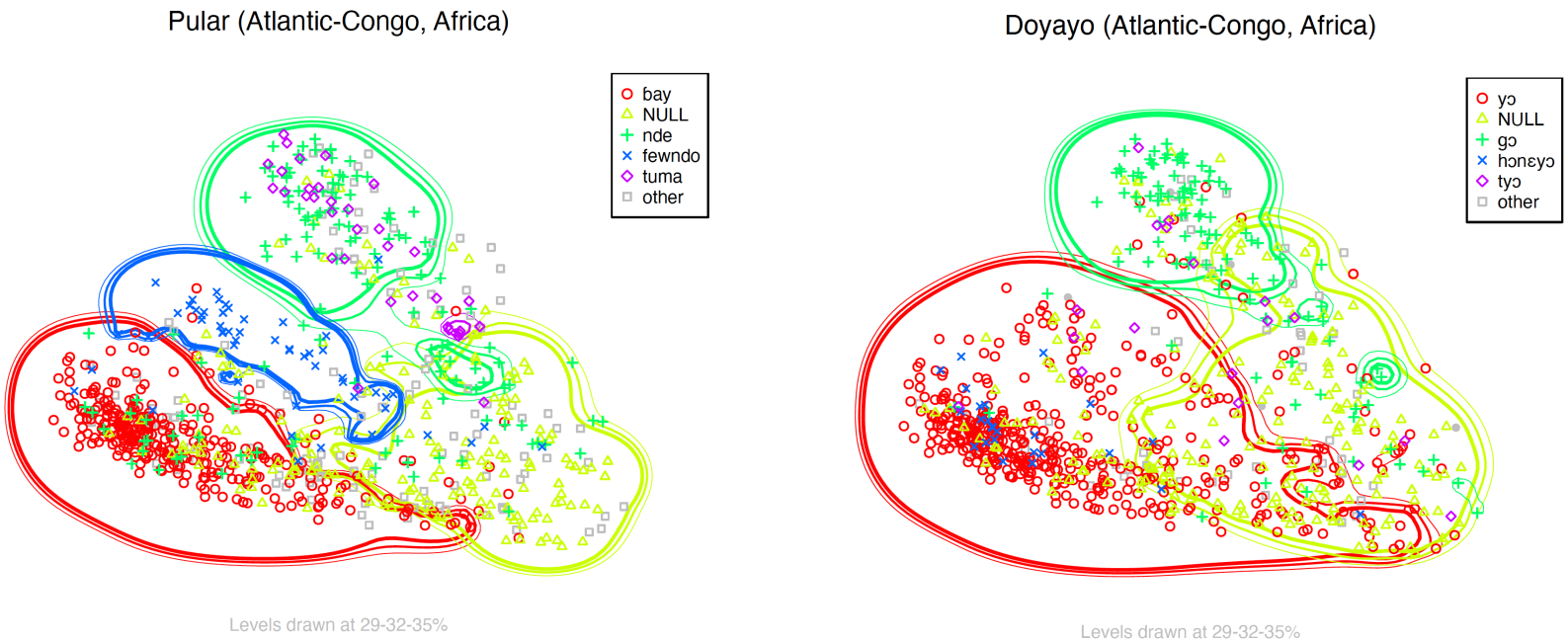

German (Indo-European, Eurasia)
als
wenn
NULL
da
und
other
Levels drawn at 29-32-35%

Romlig interpolasjon gjennom kriging

- Her hadde vi samlet punktene i områder gjennom *interpolasjon*
- Hartmann et al. (2014) bruker *kriging* som interpolasjonsmetode
- Kriging-områder kan overlappe og dermed vise *konkurranse* mellom forskjellige konstruksjoner:

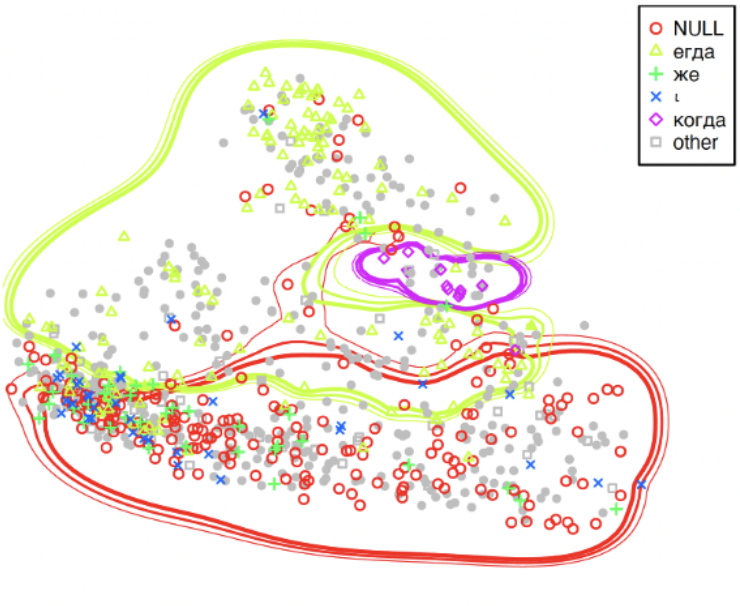
Romlig interpolasjon gjennom kriging

- Her hadde vi samlet punktene i områder gjennom *interpolasjon*
- Hartmann et al. (2014) bruker *kriging* som interpolasjonsmetode
- Kriging-områder kan overlappe og dermed vise *konkurranse* mellom forskjellige konstruksjoner:



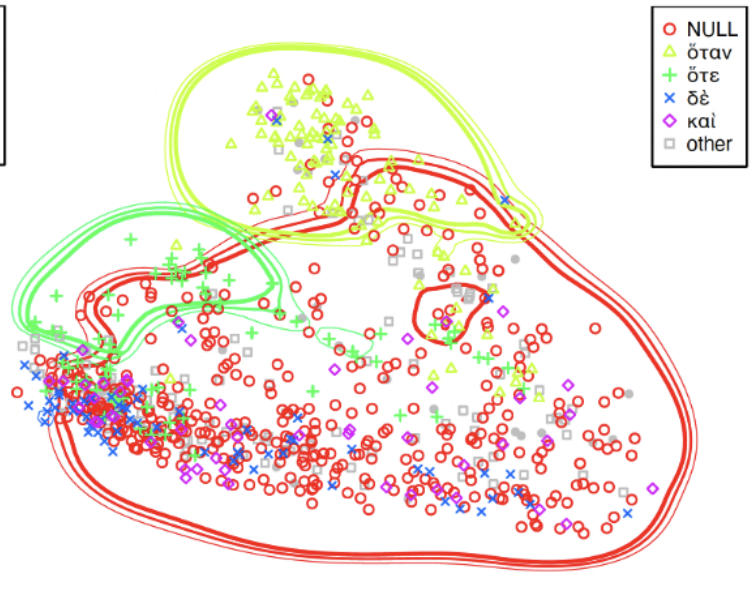
Figur 5

Old Church Slavonic (Indo-European, Eurasia)



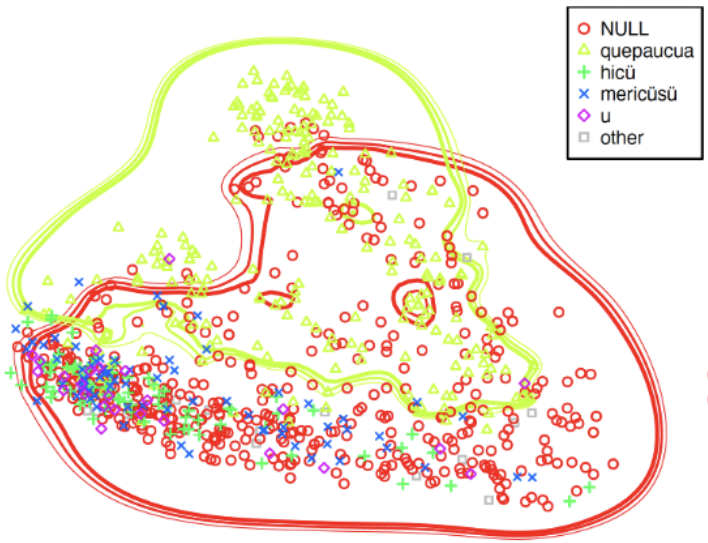
Levels drawn at 29-32-35%

Ancient Greek (Indo-European, Eurasia)



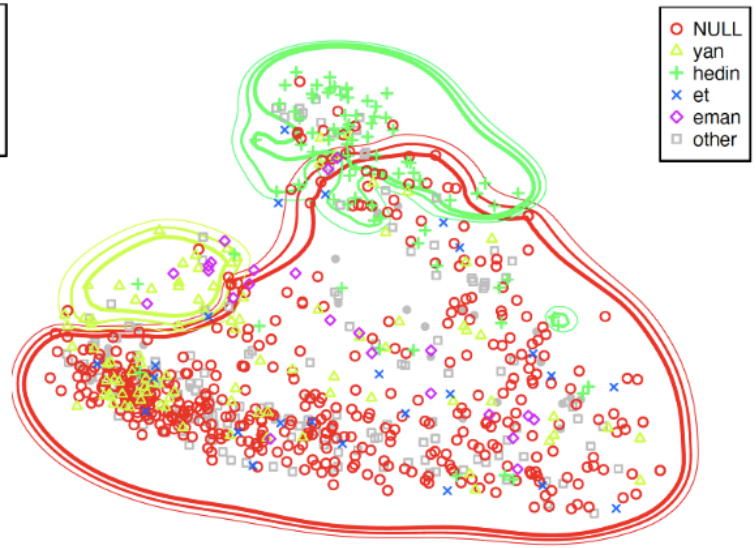
Levels drawn at 29-32-35%

Huichol (Uto-Aztecan, North America)



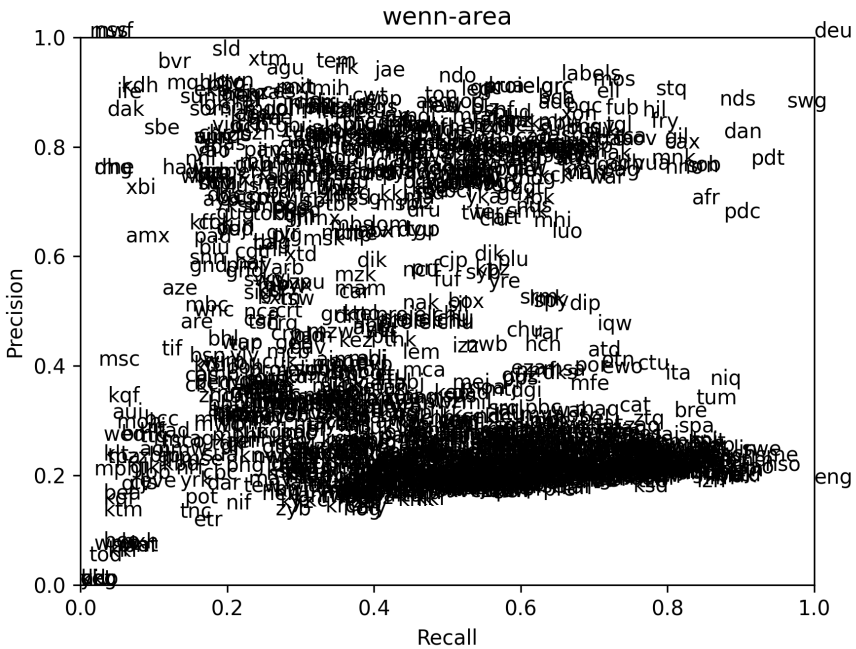
Levels drawn at 29-32-35%

Keley-i Kallahan (Austronesian, Papunesia)



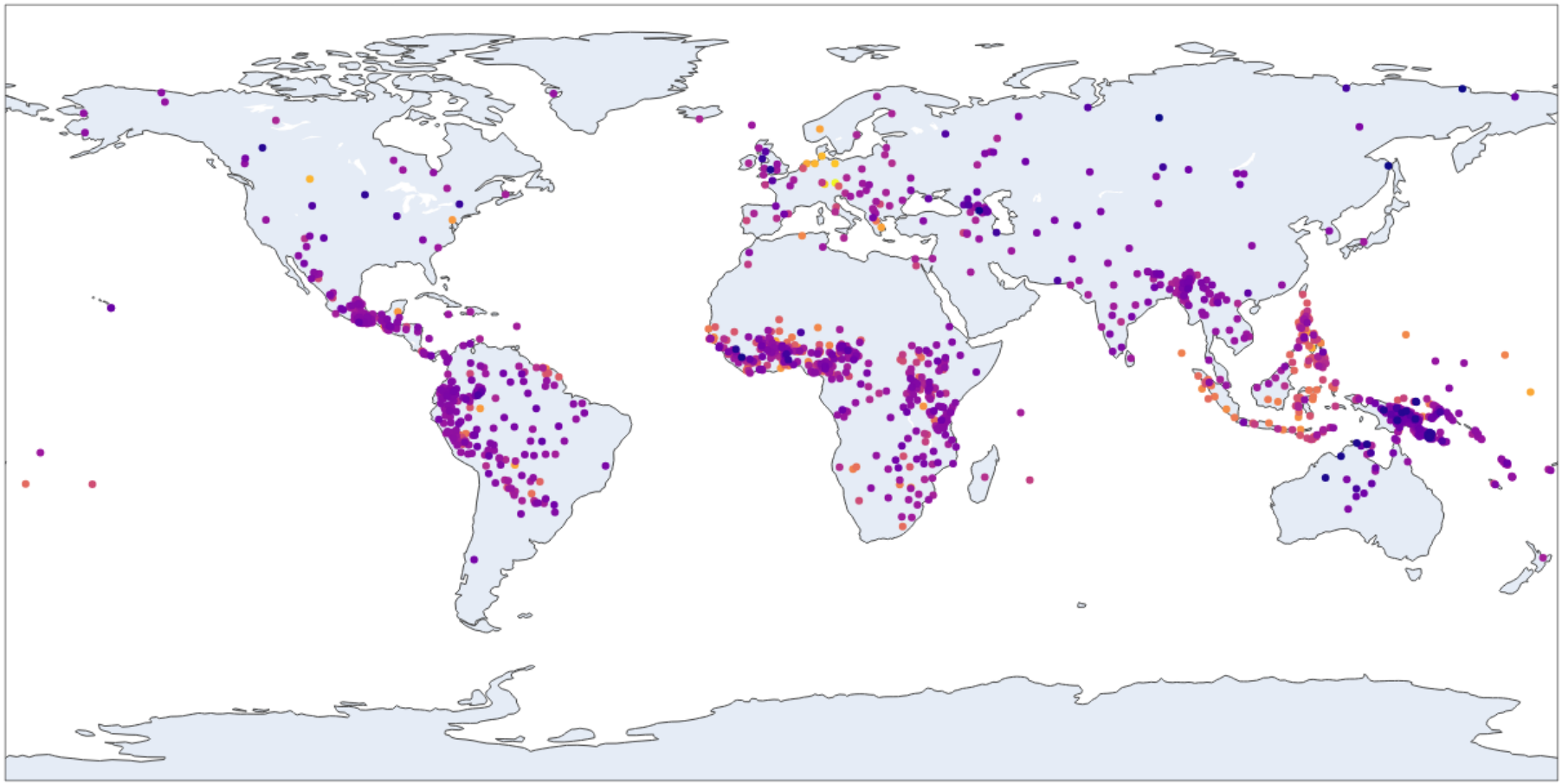
Levels drawn at 29-32-35%

Sammelikning av konstruksjoner



- Vi kan sammenlikne to konstruksjoner fra forskjellige språk
- Da måler vi hvor presist et ord svarer til et annet og hvor dekkende (*recall*) det er
- Vi tar (det harmoniske) gjennomsnittet av presisjon og dekningsgrad
- Og gjentar for alle språk

Universal WHEN



References I

Hartmann, Iren, Martin Haspelmath & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38(3). 463–484.

Haug, Dag & Nilo Pedrazzini. 2023. The semantic map of when and its typological parallels. *Frontiers in Communication* 8. doi:10.3389/fcomm.2023.1163431. https://www.frontiersin.org/articles/10.3389/fcomm.2023.1163431.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3158–3163. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf.

Wälchli, Bernhard & Michael Cysow. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50. 671–710. https://doi.org/10.1515/ling-2012-0021.