Semantic Typology and Parallel Corpora

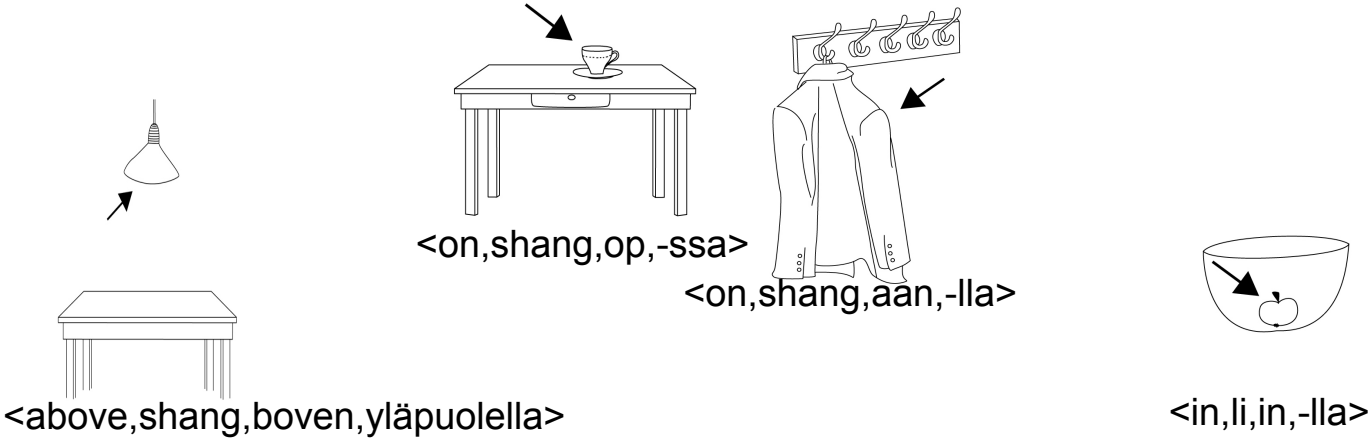
Something about Indefinite Pronouns

Barend Beekhuizen Julia Watson Suzanne Stevenson

Department of Computer Science
University of Toronto

CogSci 2017

- Modeling meaning requires representation space
- Typology: the more languages co-categorize two entities, the more conceptually similar they are (Gentner & Bowerman 2009; Beekhuizen et al. 2014).
- How to obtain such data?



		English	Mandarin	Dutch	Finnish	
Horiz., no contact	Lamp above table	< above	shang	boven	yläpuolella	>
Stable support	Cup on table	< on	shang	op	-ssa	>
Tenuous support	Coat on hook	< on	shang	aan	-lla	>
Containment	Apple in bowl	< in	li	in	-lla	>

Semantic typology: data acquisition

- Elicitation (Berlin & Kay 1969),
- Secondary sources (Haspelmath 1997),
- Primary text (Cysouw & Wälchli 2009)
  - translated parallel data (subtitles, bibles)
  - reflects actual usage patterns
  - can be used for more abstract domains

Semantic typology: data acquisition

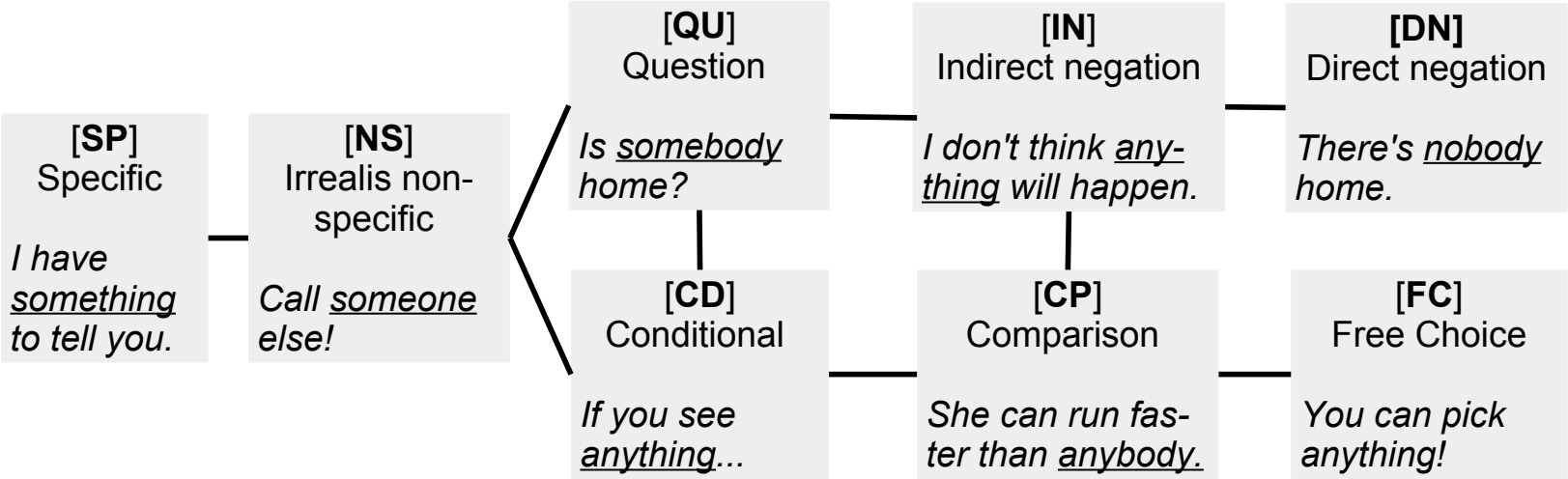
- Elicitation (Berlin & Kay 1969),
- Secondary sources (Haspelmath 1997),
- Primary text (Cysouw & Wälchli 2009)
  - translated parallel data (subtitles, bibles)
  - reflects actual usage patterns
  - can be used for more abstract domains

Our goals:

- contributing to pipeline of extracting verbalization in many languages from parallel text
- compare text-based representations to representations from secondary sources

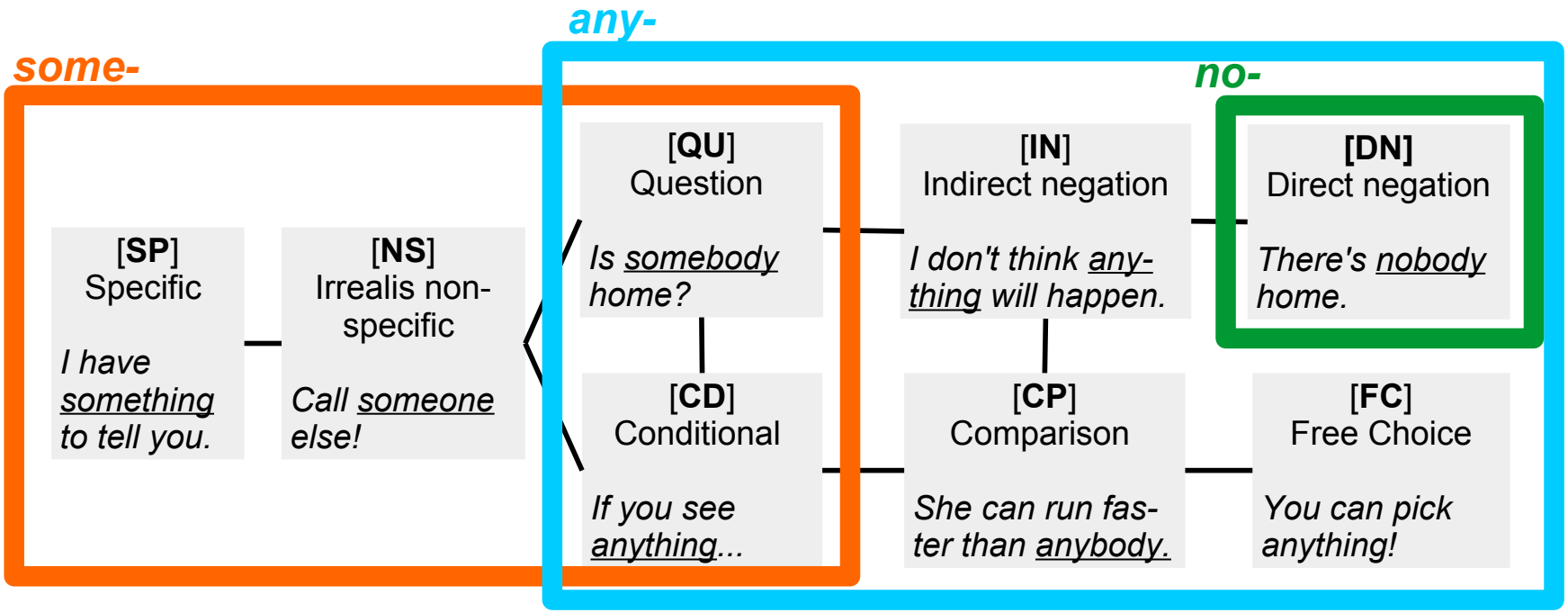
Case study: indefinite pronouns

- Cross-linguistic variation in term extensions
- Formalized using semantic map method (Haspelmath 1997)



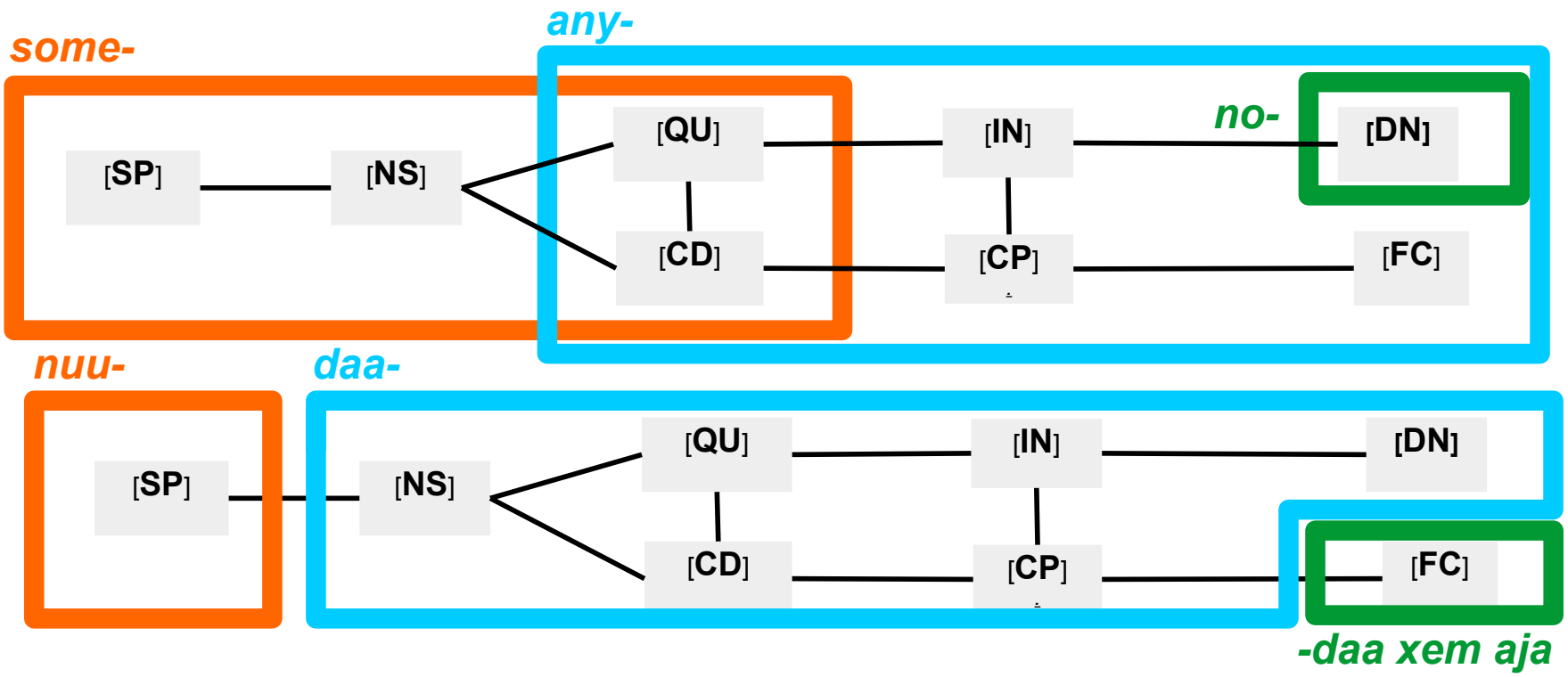
Case study: indefinite pronouns

- Cross-linguistic variation in term extensions
- Formalized using semantic map method (Haspelmath 1997)



Case study: indefinite pronouns

- Cross-linguistic variation in term extensions
- Formalized using semantic map method (Haspelmath 1997)

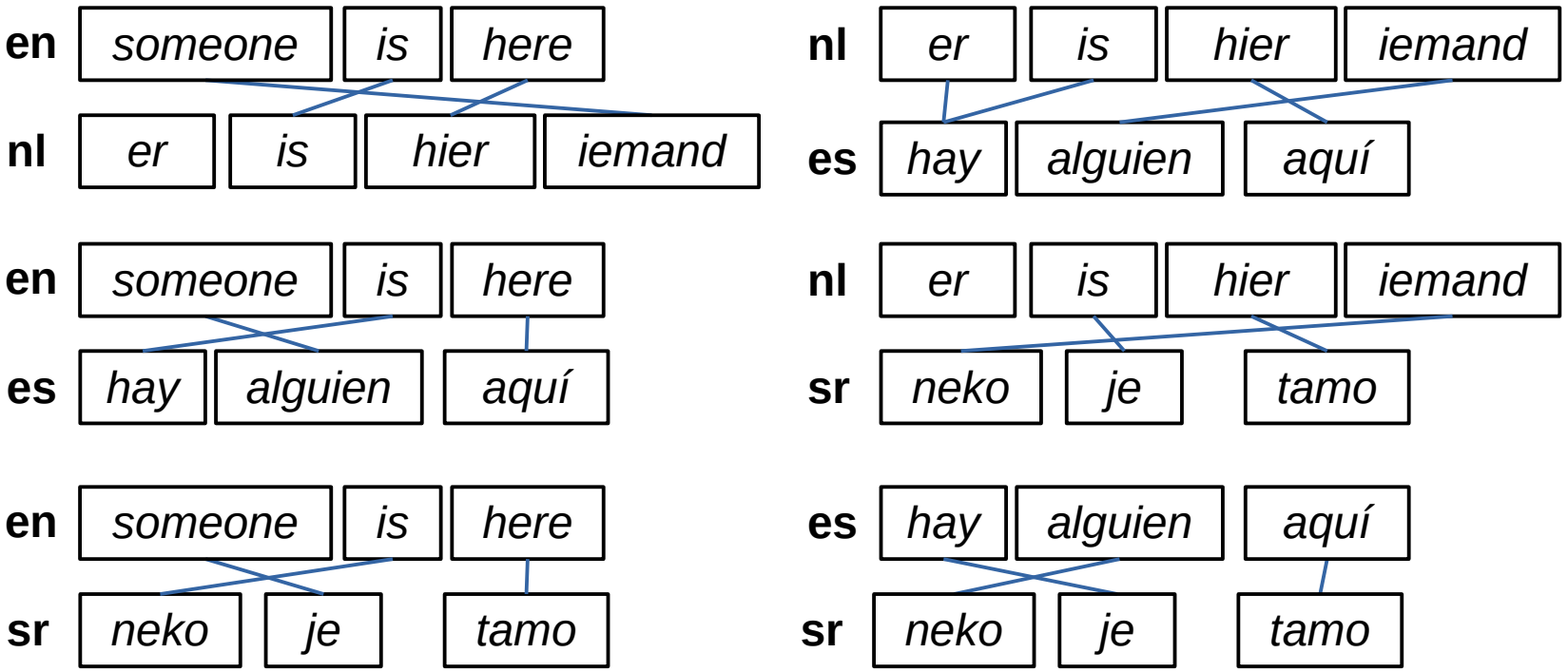


Questions

- Despite great insight, limitations of approach
- Questions better answered with primary texts:
  - Q1 Are all functions equally frequent?
  - Q2 Are functions defined at the right level of granularity?
  - Q3 Do functions display discrete or fuzzy boundaries?
  - Q4 Are functions internally homogenous or do they display further internal structure?

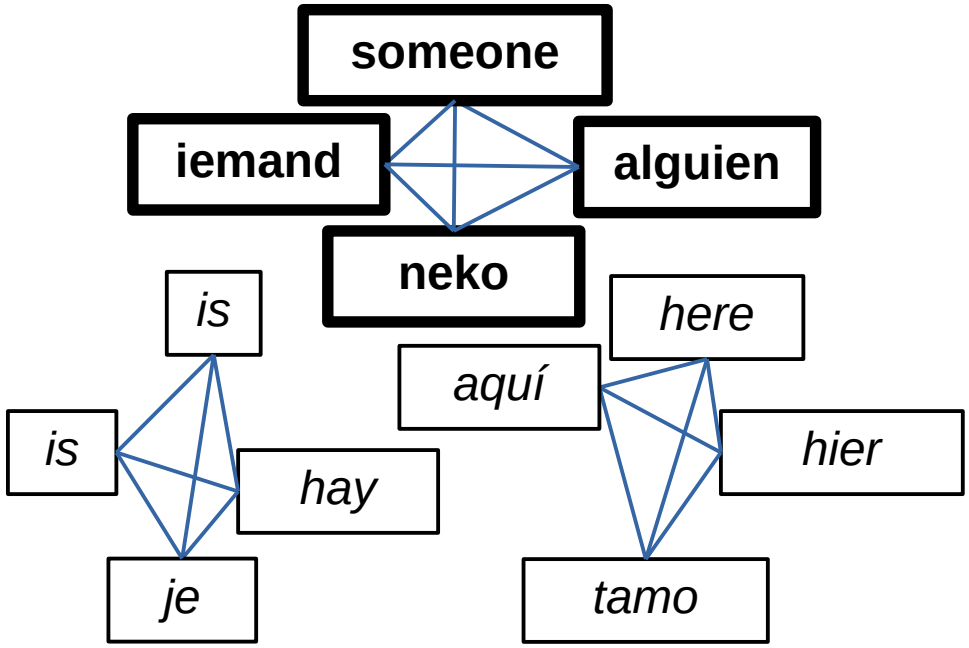
Method

- Subtitles in 30 languages (9 families); parallelized and aligned



Method

- Subtitles in 30 languages (9 families); parallelized and aligned
- Extracted clusters of mutually aligned words



Method

- Subtitles in 30 languages (9 families); parallelized and aligned
- Extracted clusters of mutually aligned words
- Linearized clusters and annotated functions

Utterance	en	nl	es	sr	function
someone is here	**someone**	**iemand**	**alguien**	**neko**	**SP**
anyone got 5 billion?	**anyone**	**iemand**	**alguien**	**neko**	**QU**
she could beat anyone	**anyone**	**iedereen**	**qualquier**	**neko**	**FC**
....					

Q1: Frequency of functions

- Split over PEOPLE (e.g., *anyone*, *somebody*) and THINGS (e.g., *nothing*, *anything*)
- What is the relative frequency per function?

Q1: Frequency of functions

- Split over PEOPLE (e.g., *anyone, somebody*) and THINGS (e.g., *nothing, anything*)
- What is the relative frequency per function?

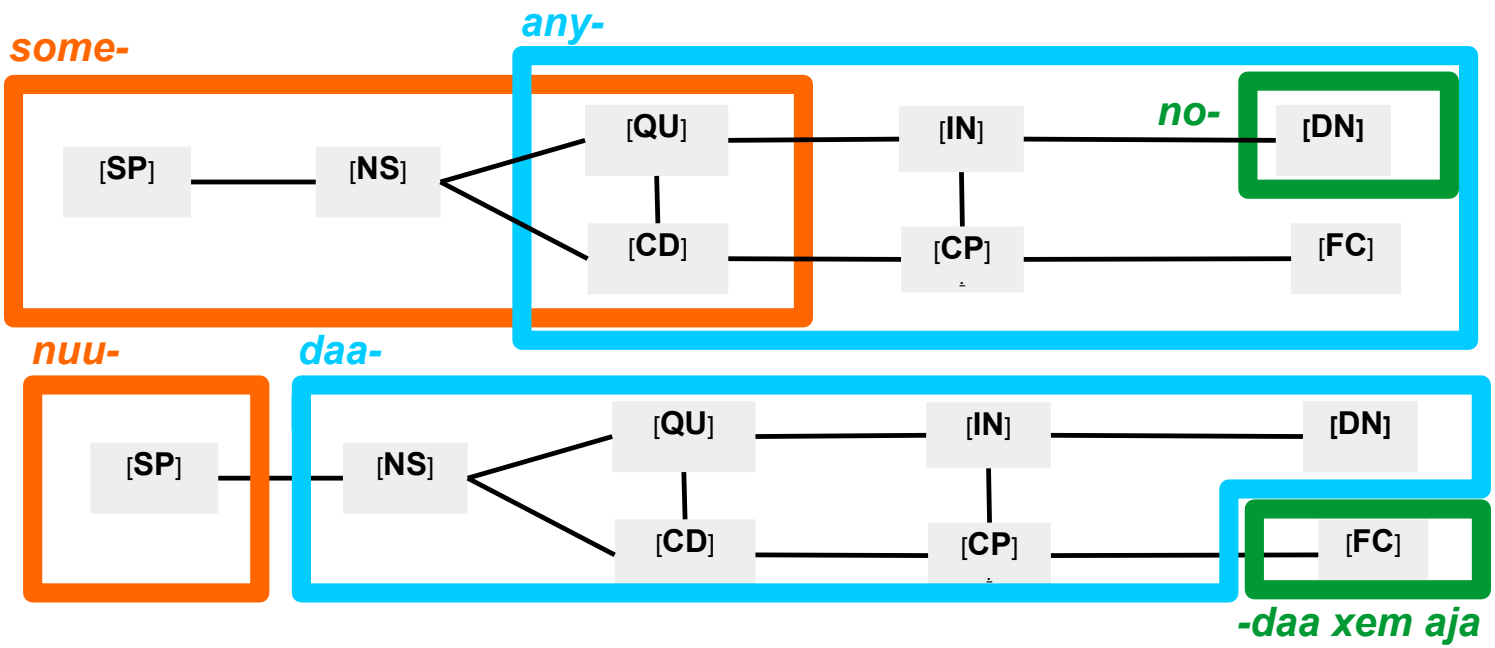
	SP	NS	CD	QU	IN	DN	CP	FC
PEOPLE	**.16**	.20	.07	.16	.05	**.28**	.01	.08
THINGS	**.28**	.15	.05	.09	.02	**.36**	.00	.06
Overall	**.24**	.17	.06	.11	.03	**.33**	.00	.06

Table: Distribution of functions given ontological category.

SP	specific	CD	conditional	IN	indirect neg.	CP	comparison
NS	non-spec.	QU	question	DN	direct negation	FC	free choice

☞ Frequent functions are prototypes of Haspelmath's map.

	SP	NS	CD	QU	IN	DN	CP	FC
PEOPLE	**.16**	.20	.07	.16	.05	**.28**	.01	.08
THINGS	**.28**	.15	.05	.09	.02	**.36**	.00	.06
Overall	**.24**	.17	.06	.11	.03	**.33**	.00	.06



Q2: Granularity of functions

- Is 8 the right number of functions?
- Evaluate with automatic clustering:
  - compare k-means clustering against annotated data

Q2: Granularity of functions

- Is 8 the right number of functions?
- Evaluate with automatic clustering:
  - compare k-means clustering against annotated data

	$k =$	2	3	**4**	5	6	7	8	9	10
PEOPLE		.20	.25	**.41**	.35	.34	.34	.32	.30	.32
THINGS		.30	.38	**.47**	.36	.35	.35	.33	.39	.33

Q2: Granularity of functions

- For $k = 4$, what do the clusters look like?

Q2: Granularity of functions

- For $k = 4$, what do the clusters look like?

☞ Too fine-grained

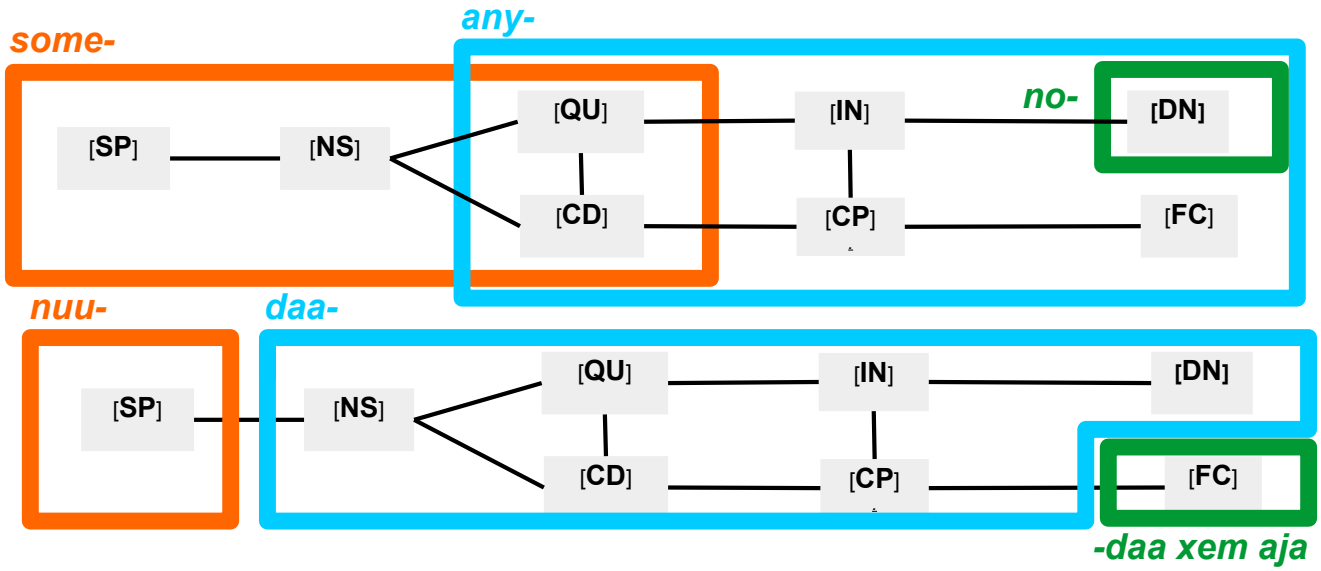
Cluster	SP	NS	CD	QU	IN	DN	CP	FC
1	**18**	**24**	**6**	3	0	2	0	0
2	1	0	2	**15**	1	4	0	2
3	0	0	1	0	**5**	**27**	0	0
4	0	0	0	0	0	0	**1**	**7**

Q2: Granularity of functions

- For $k = 4$, what do the clusters look like?

☞ Too fine-grained

Cluster	SP	NS	CD	QU	IN	DN	CP	FC
1	**18**	**24**	**6**	3	0	2	0	0
2	1	0	2	**15**	1	4	0	2
3	0	0	1	0	**5**	**27**	0	0
4	0	0	0	0	0	0	**1**	**7**



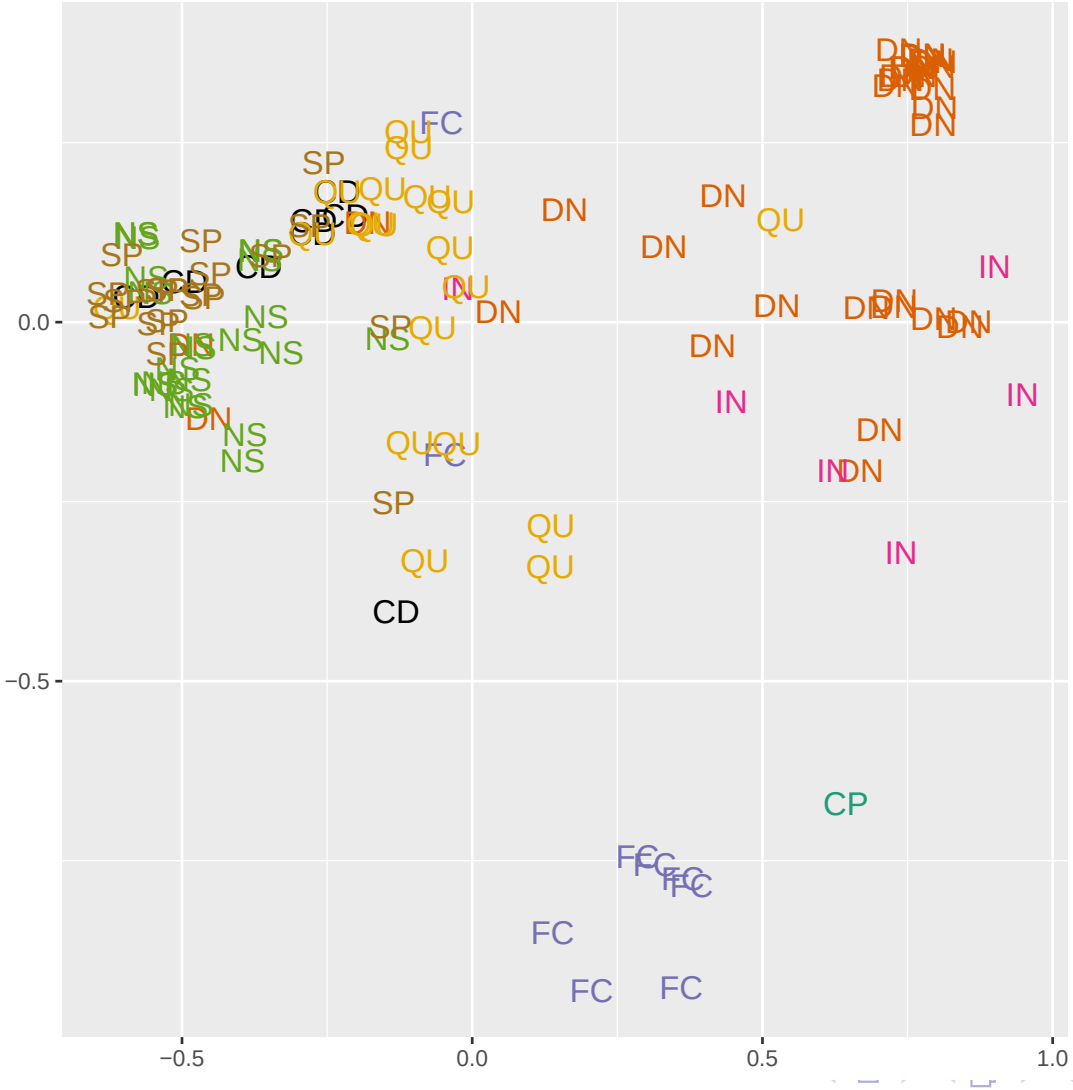
Q3: boundaries between clusters

- Optimal Classification MDS (Croft & Poole 2008)

Q3: boundaries between clusters

- Optimal Classification MDS (Croft & Poole 2008)

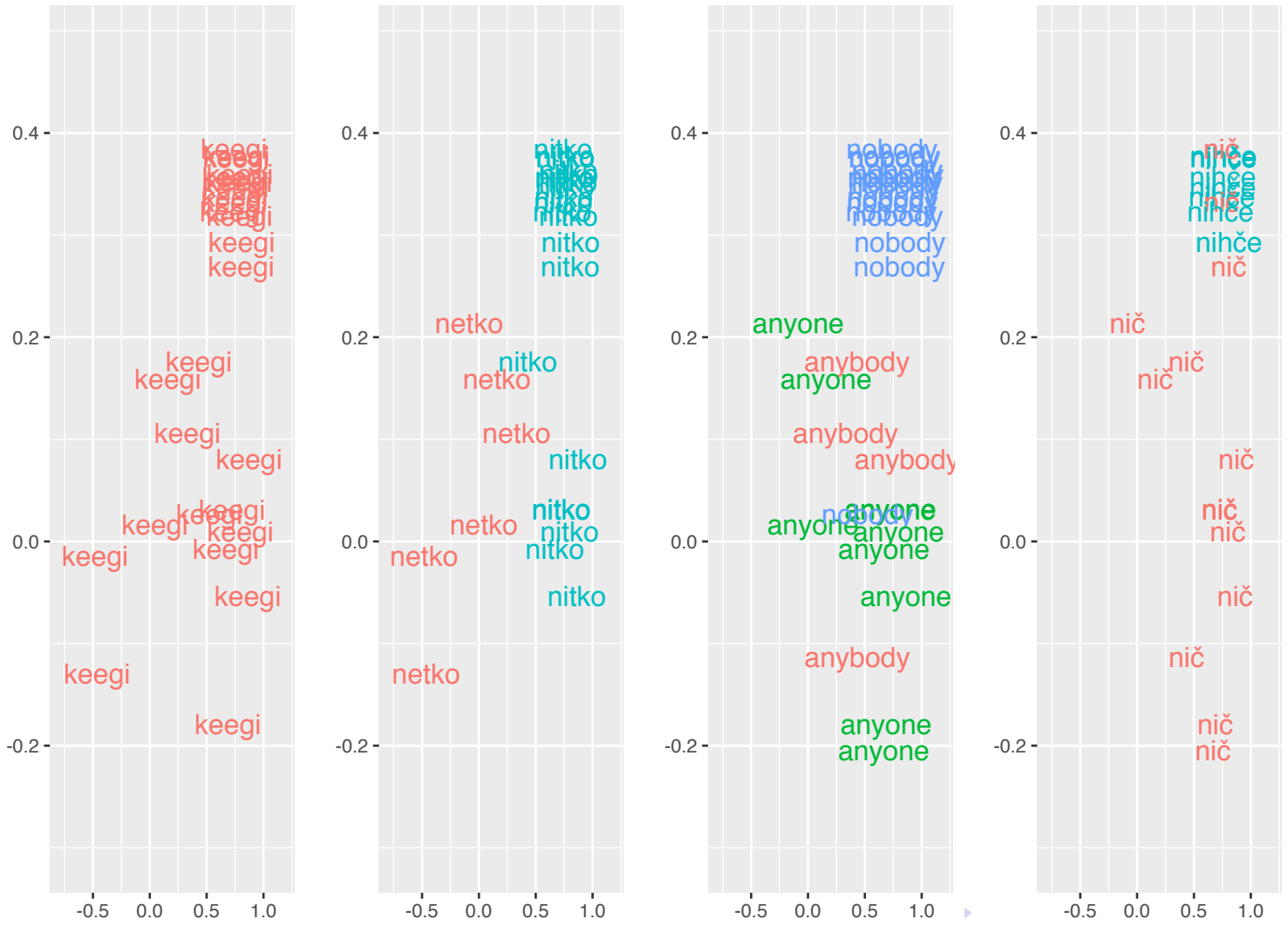
☞ Clear clusters, but with 'bridges' between them



Q4: internal homogeneity

- Direct negation for PEOPLE in Estonian, Croatian, English, Slovene.

☞ Internal scale: Emphatic > Subjects > Other functions



Recap

- cross-linguistic patterns of co-categorization – cognitive representation
- studies indefinite pronouns in parallel usage data (subtitles)
- handcrafted model is both too fine-grained and too coarse grained
- usage data allows for fine-grained exploration of semantic contrasts

Recap

- cross-linguistic patterns of co-categorization – cognitive representation
- studies indefinite pronouns in parallel usage data (subtitles)
- handcrafted model is both too fine-grained and too coarse grained
- usage data allows for fine-grained exploration of semantic contrasts

Technical extensions

- Scalability: pairwise alignments
- Use of non-parallel text (translationese!)

Recap

- cross-linguistic patterns of co-categorization – cognitive representation
- studies indefinite pronouns in parallel usage data (subtitles)
- handcrafted model is both too fine-grained and too coarse grained
- usage data allows for fine-grained exploration of semantic contrasts

Technical extensions

- Scalability: pairwise alignments
- Use of non-parallel text (translationese!)

Cognitive plausibility

- E.g., ease of acquisition/order of acquisition
- Similarity/acceptability judgments of language users
- ...

Thank you!