Parallel corpora and semantic typology

Barend Beekhuizen
barend@cs.toronto.edu
University of Toronto
(joint work with Suzanne Stevenson)

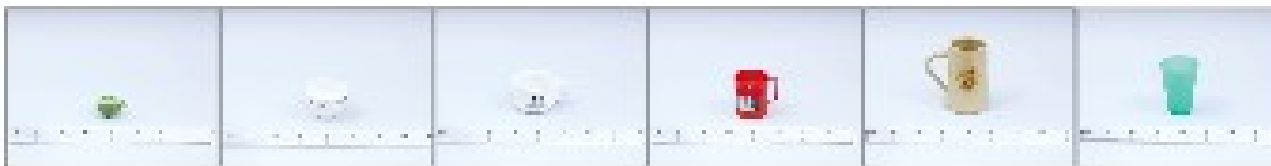
Computational Linguistics Seminar @ UvA
November 1, 2016

Semantic typology

- Languages vary widely in how they carve up the space of possible meanings
- But there are also strong biases: only a small subset of all possible variations are attested
- Semantic typology: describing and explaining the types of semantic categorization systems

- Given this: how to do meaning in CL in a language-independent way?

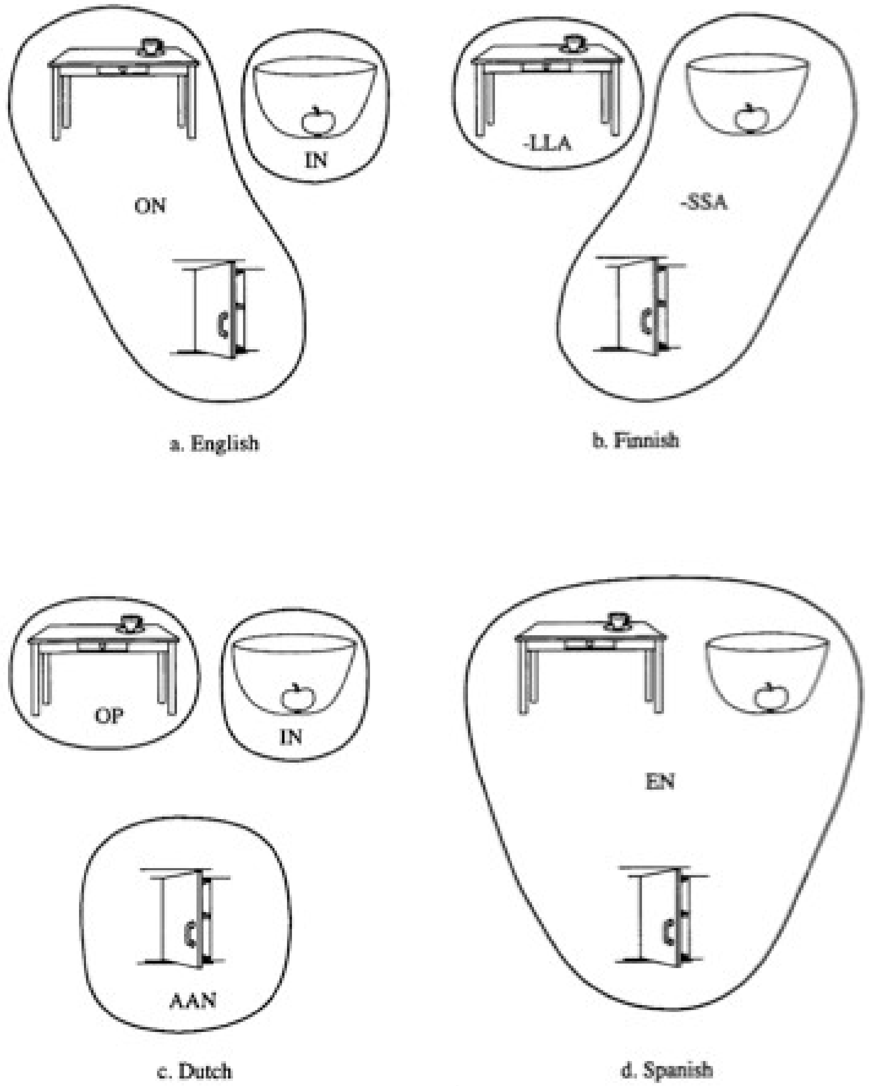
Semantic typology



<table>
<tr><td>Faroese</td><td colspan="12">koppur</td></tr>
<tr><td>Dutch (BE)</td><td colspan="10">tas</td><td colspan="2">beker</td></tr>
<tr><td>Frisian</td><td colspan="4">kopke</td><td colspan="8">beker</td></tr>
<tr><td>Danish</td><td colspan="8">kop</td><td colspan="4">krus</td></tr>
<tr><td>Norwegian</td><td colspan="8">kopp</td><td colspan="2">krus</td><td colspan="2">glass</td></tr>
<tr><td>Icelandic</td><td colspan="8">bolli</td><td colspan="2">krús</td><td colspan="2">glas</td></tr>
<tr><td>Luxembourgish</td><td colspan="8">Taass</td><td colspan="2">Béierkrou</td><td colspan="2">Becher</td></tr>
<tr><td>German</td><td colspan="8">Tasse</td><td colspan="2">Krug</td><td colspan="2">Becher</td></tr>
<tr><td>Schwyzerdütsch</td><td colspan="8">tassli</td><td colspan="2">humpè</td><td colspan="2">bächèr</td></tr>
<tr><td>Swedish</td><td colspan="6">kopp</td><td colspan="4">mugg</td><td colspan="2">glass</td></tr>
<tr><td>English</td><td colspan="5">cup</td><td colspan="5">mug</td><td colspan="2">cup</td></tr>
<tr><td>Dutch (NL)</td><td colspan="4">kopje</td><td colspan="2">kom</td><td colspan="4">mok</td><td colspan="2">beker</td></tr>
</table>

Majid et al. (2014)

Semantic typology

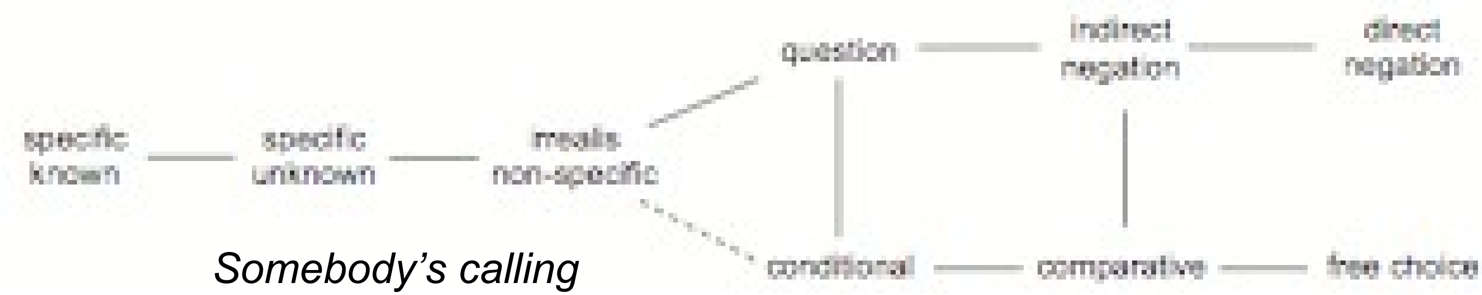


a. English
b. Finnish
c. Dutch
d. Spanish

Bowerman (1996)

Semantic typology

Did anybody see my dog?



Somebody's calling

Mary can run faster than anybody

Haspelmath (1997)

Data and methods

- Sources of data:
  - Grammars/dictionaries (Haspelmath 1997, Youn 2016)
  - Elicitation (Bowerman 1996, Lang 2001, Majid et al. 2008, 2014)
  - Corpora (Mayer & Cysouw 2012)
- Method of analysis:
  - Visualization (manual, automatic; Croft & Poole 2008, Majid et al. 2008, 2014)
  - Implicational semantic maps (Haspelmath 1997, Ito & Narrog 2009, Regier et al. 2013)

Issues with data/methods

- Elicitation is resource-intensive
- Decisions, hence room for bias in:
  - Functions
    - what delimits a ‘recipient’
  - Domains
    - where does ‘the dative region’ end?
  - Correspondence
    - are ‘recipients’ in two languages the same function?
  - Analysis
    - where to place functions on map, where to draw edges?

The view from CL

- Word alignments from parallel corpora:
  - Brown et al. (1993), Liang et al. (2006)
  - But only bitext (exceptions: Östling 2012, Mayer & Cysouw -- however don't scale)
- Why not use word embeddings?
  - Monolingual (e.g., Word2Vec - Mikolov et al 2012)
    - Project onto each other for bitext (Faruqui & Dyer 2014)
  - ... or multilingual (Hermann & Blunsom 2014, Vulic & Moens 2014, Upadhyay et al 2016)
  - But:
    - only one embedding per word type
    - question of scalability to *n* languages
  - Otherwise interesting -- still exploring

Our contribution

- Start from sentence-aligned translated texts as a source of analysis for semantic typology (cf. Mayer & Cysouw 2012)
- Working at the token/exemplar level: no functions, just ‘clouds’ of instances (Croft ms.)
- Allowing for full-corpus coverage: domains simply larger ‘clouds’ of instances
- (Automatic graph inference -- for another day)

A pipeline for corpus-based semantic typology

- Sentence-aligned translations
- Extract symmetrical pairwise alignments
- Per utterance:
  - Create a graph of all pairwise alignments
  - Use graph clustering to extract sets of strongly mutually aligned words
  - Use dimensionality reduction techniques over these to create a semantic space of word usages

Sentence-aligned translations

- [Eng] it rained yesterday
- [Dut] het regende gisteren
- [Ger] es regnete gestern
- [Spa] ayer llovió

Word alignment

- [Eng] it rained yesterday
- [Dut] het regende gisteren
- [Ger] es regnete gestern
- [Spa] ayer llovió

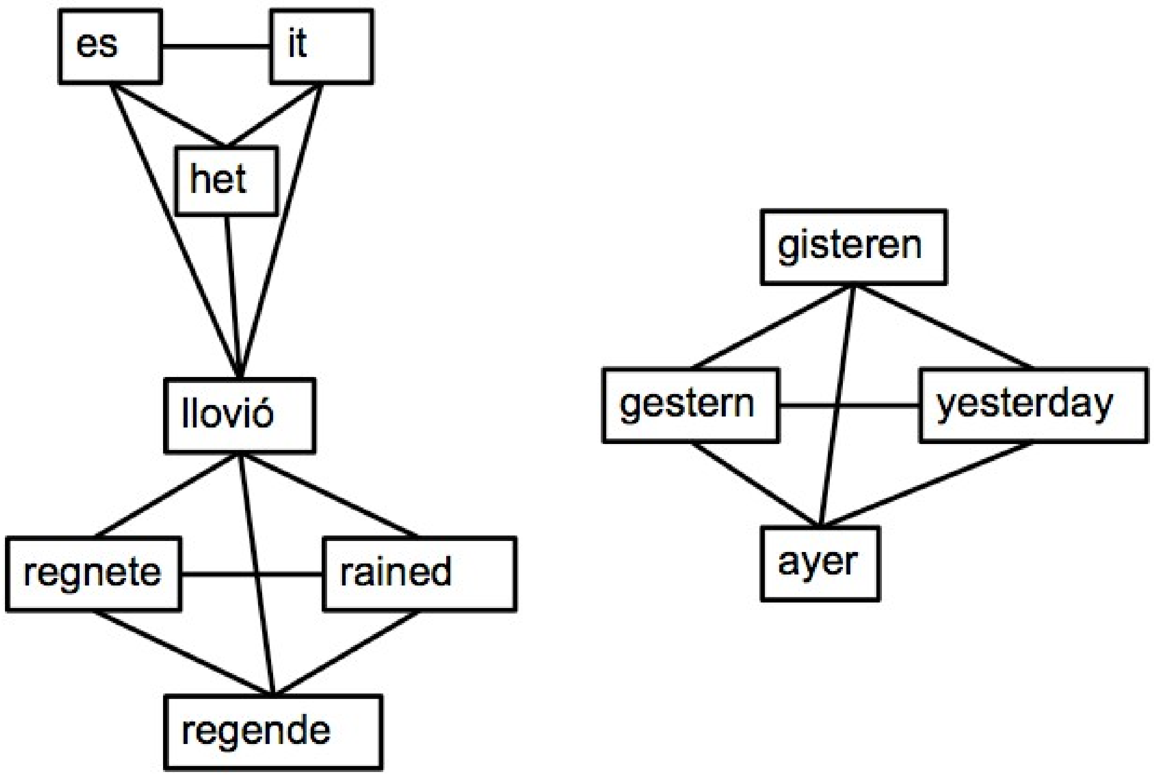
- Using Liang et al. (2006) symmetrical alignment method

Word alignment

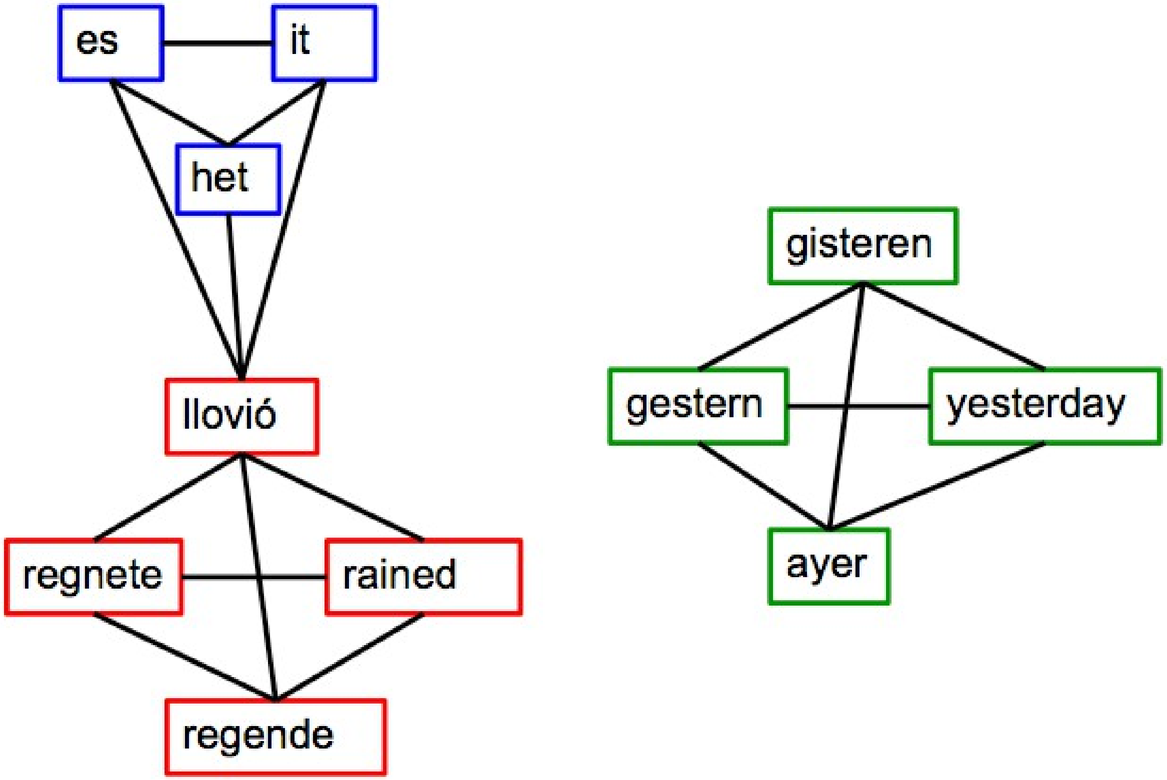
- [Eng] it rained yesterday
- [Dut] het regende gisteren
- [Ger] es regnete gestern
- [Spa] ayer llovió

- Using Liang et al. (2006) symmetrical alignment method

Utterance graph



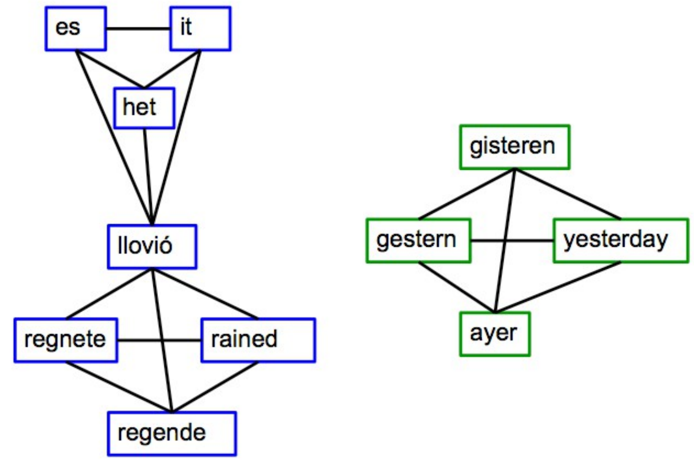
Finding mutually aligned words

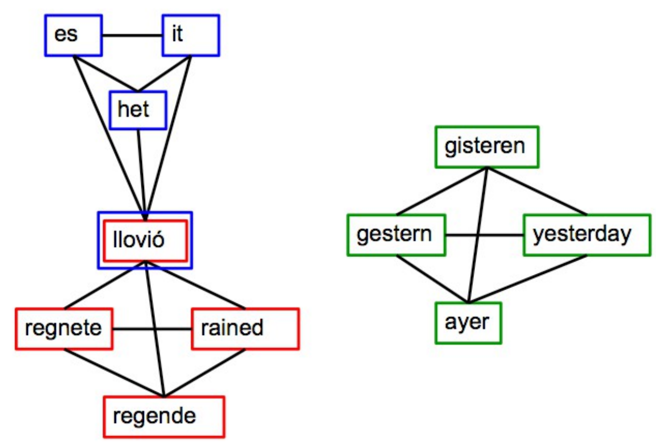


Graph clustering

- Two extremes of graph clustering:
  - all connected components (top)
  - all maximal cliques (bottom)
- connected components
  --> too underconstrained
- maximal cliques
  --> too strict
- solution: graph clustering
  - finding sweet spot in between

    components and cliques

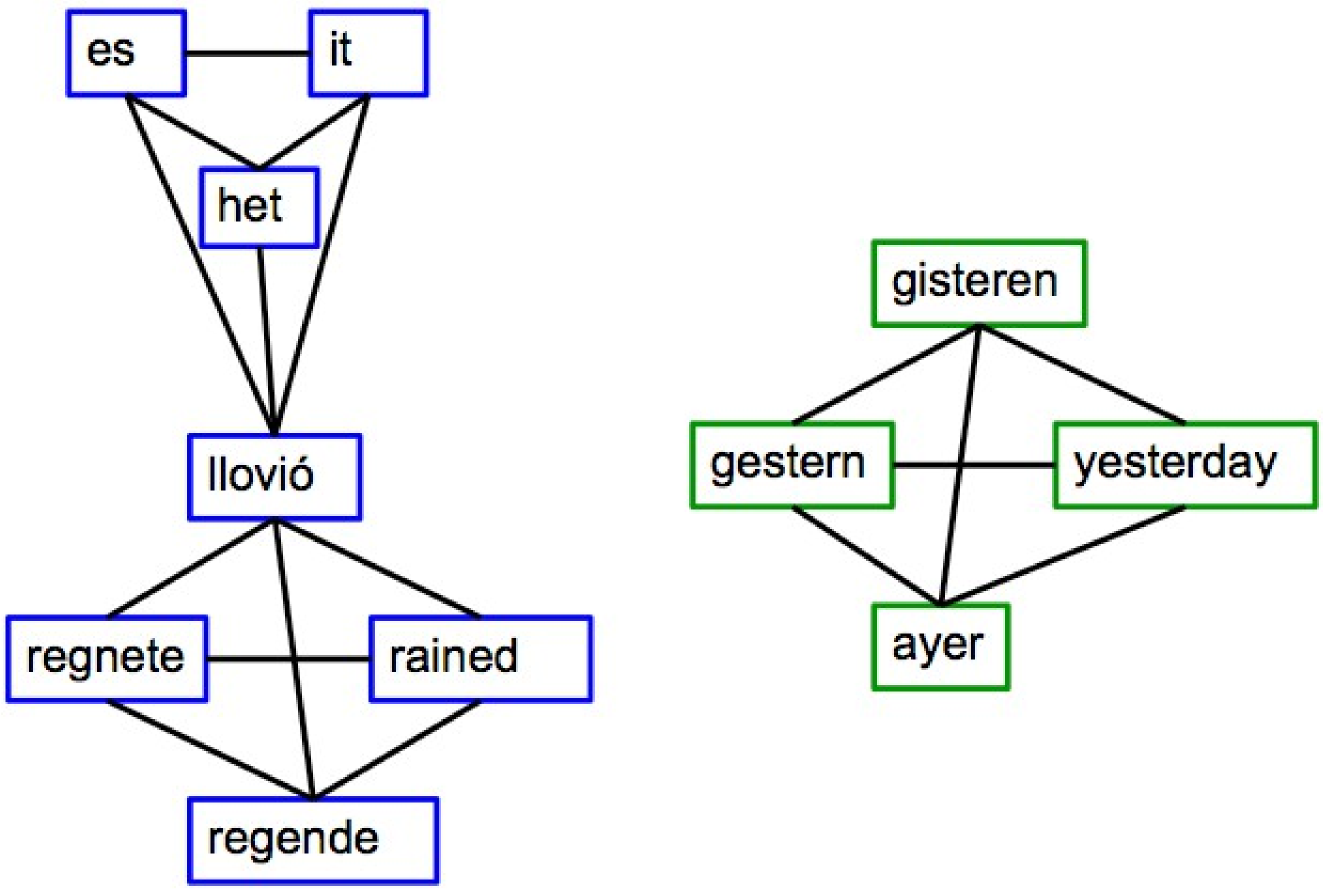




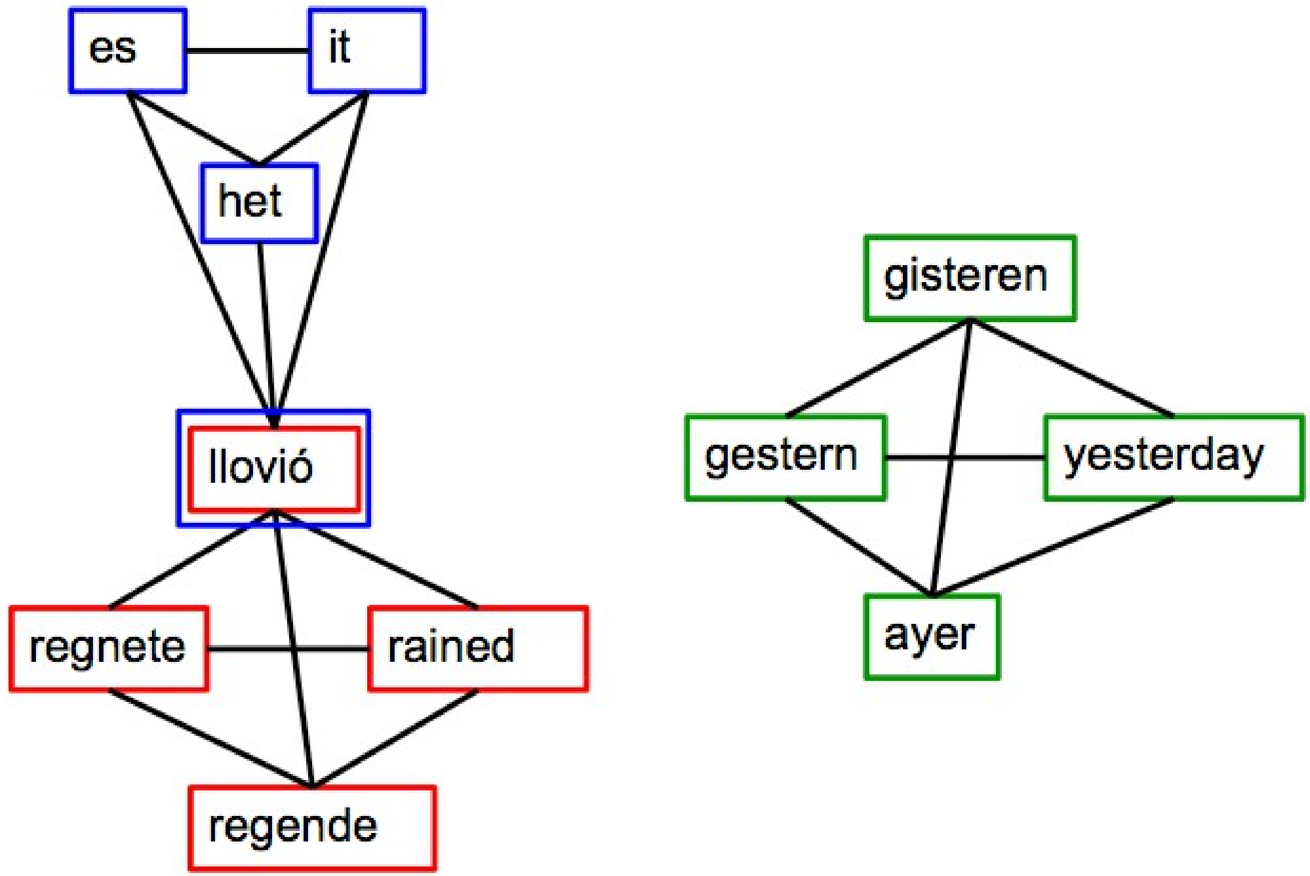
K-clique clustering/percolation

- Get all *k*-cliques (cliques of size *k*)
- Get adjacencies between *k*-cliques
  - *k*-cliques are adjacent if they share *k* - 1 nodes
- Clusters ('communities') are the maximal unions of adjacent *k*-cliques

$k = 2$


es
it
het
llovió
regnete
rained
regende
gisteren
gestern
yesterday
ayer

$k = 3$



es
it
het
llovió
regnete
rained
regende
gisteren
gestern
yesterday
ayer

Dimensionality reduction

- *K*-cliques to table
- Idea: same kind of table as elicitations: how to express a particular bit of meaning

Eng	*Dut*	*Ger*	*Spa*
it	het	es	llovió
rained	regende	regnete	llovió
yesterday	gisteren	gestern	ayer
...	...	...	...

Dimensionality reduction

- 1] random projection (Dasgupta 2000):
  - Vectorize words so that we have *ud* matrix (*u* usages, *d* vectorized words)
  - Obtain *uk* (where $k \ll d$) matrix: form a random matrix $R = kd$ and project *ud* through *R* onto *uk*
  - Good for global inspection and evaluation
- 2] Optimal Classification (Croft & Poole 2008)
  - Scale *ud* into *k* dimensions
    - by trying to optimally place all usages *u* in the *k* dimensional space, s.t. for every word, there is a cutting plane in the space that divides the space into instances of that word and non-instances ('classification')
  - Good for visualization in case studies

Some preliminary studies: evaluation

- Global:
  - Quality of multilingual alignment: Strong’s numbers
  - Quality of derived vector spaces: word-similarity task
- Case studies:
  - Procedure for discovering biases in crosslinguistic variation

Study I: parallel bibles

- Bible is available in >900 languages.
- But small data (6K - 27K lines)

- Strong's numbers (annotation of Hebrew/Greek source words) --> clusters of translation-equivalent words
- We use these as a gold standard and see how well our clusters approximate them

Study I: parallel bibles

- Stong’s numbers available for 9 bibles (2 German, 2 English, 1 Dutch, 1 Indonesian, 1 Portuguese, 1 French, 1 Rusian)
- ~ 6000 lines
- Cluster quality evaluated by F_1 score per gold cluster, with cluster Precision and Recall:
  - $P = \max_{c_u} |c_g ! \quad c_u| \ / \ |c_u|$
  - $R = \max_{c_u} |c_g ! \quad c_u| \ / \ |c_g|$
- (Östling (2012) proposes similar task, but scores are incomparable)

Study I: parallel bibles

k	P	R	F_1 micro	F_1 macro
k=2	.20	**.98**	.25	.33
k=3	.69	.95	.75	.80
k=5	.89	.90	.89	.89
components	.21	**.98**	.25	.35
cliques	**.98**	.89	**.93**	**.93**

- components-clusters too large (lowest P, highest R)
- *k*-clique with low *k* similar
- clique-clusters better scores (high P, highest R)
- but have hard time with MWUs (Ind. *di dalam* ‘in’)
- higher *k*s make *k*-clique method approach clique
- <u>and</u> get MWUs

Study I: parallel bibles

- Cliques give best clusters
- But fail to capture many-to-one mappings
- Finding *k* ...
- Ideally, parameter-free clustering (High-Density Clustering)

Studies II & III: subtitles

- Bible many languages, but: small corpus & v. particular genre norms
- Other massively-parallel corpus: subtitles
- OPUS (Lison & Tiedemann 2016), based on www.opensubtitles.org
- From bitext to multitext (parallel stcs across all lgs): (~27K lines subtitles in 30 languages)
- Some diversity (East Asian (5 lg. fams), Semitic, many Indo-Eur. languages)
- Somewhat naturalistic language -- film dialogue

Study II: subtitles

- Other evaluation: how well do the usage clusters reflect human-rated word similarities
- Word similarity rating
  - SimLex 999 (999 word pairs; Hill et al. 2014)
  - Rated similarity between 0 and 10
- Method: <u>usage</u> clusters to <u>usage</u> vectors with Random Projection (diff. from <u>word</u> vector)
- Model word similarity as nearest-neighbors of two word's usages:
  - $\mathrm{sim}(w_1, w_2) = 1 - \min_{u_1,u_2} \cos(u_1,u_2)$
- Compare model word similarity to human word sim. rating with Spearman's rho.
- Only looking at words for which $n > 10$

Study II: similarity

this model	*k = 9, d = 128*	.34 [.31-.42]	163 words
	k = 9, d = 256	.32 [.28-.38]	
	k = 15, d = 128	.32 [.29-.38]	190 words
	k = 15, d = 256	.34 [.31-.38]	
	k = 21, d = 128	.36 [.28-.47]	114 words
	k = 21, d = 256	.37 [.32-.49]	
Word2Vec	(Mikolov et al. 2013)	.37	
Best resource-free	(Schwartz et al. 2016)	.56	

Study II: similarity

- Vectors from clusters give reasonable performance
- Much room for improvement in dim. reduction
- Future: word <u>usage</u> similarity (Erk et al. 2009)
  - *This **bank** was built in 1816*
  - *My **bank** doesn’t charge fees*
  - *We were fishing from the **bank** of the river*
  - *I like going to the **shore** to fish*

Study III: typology

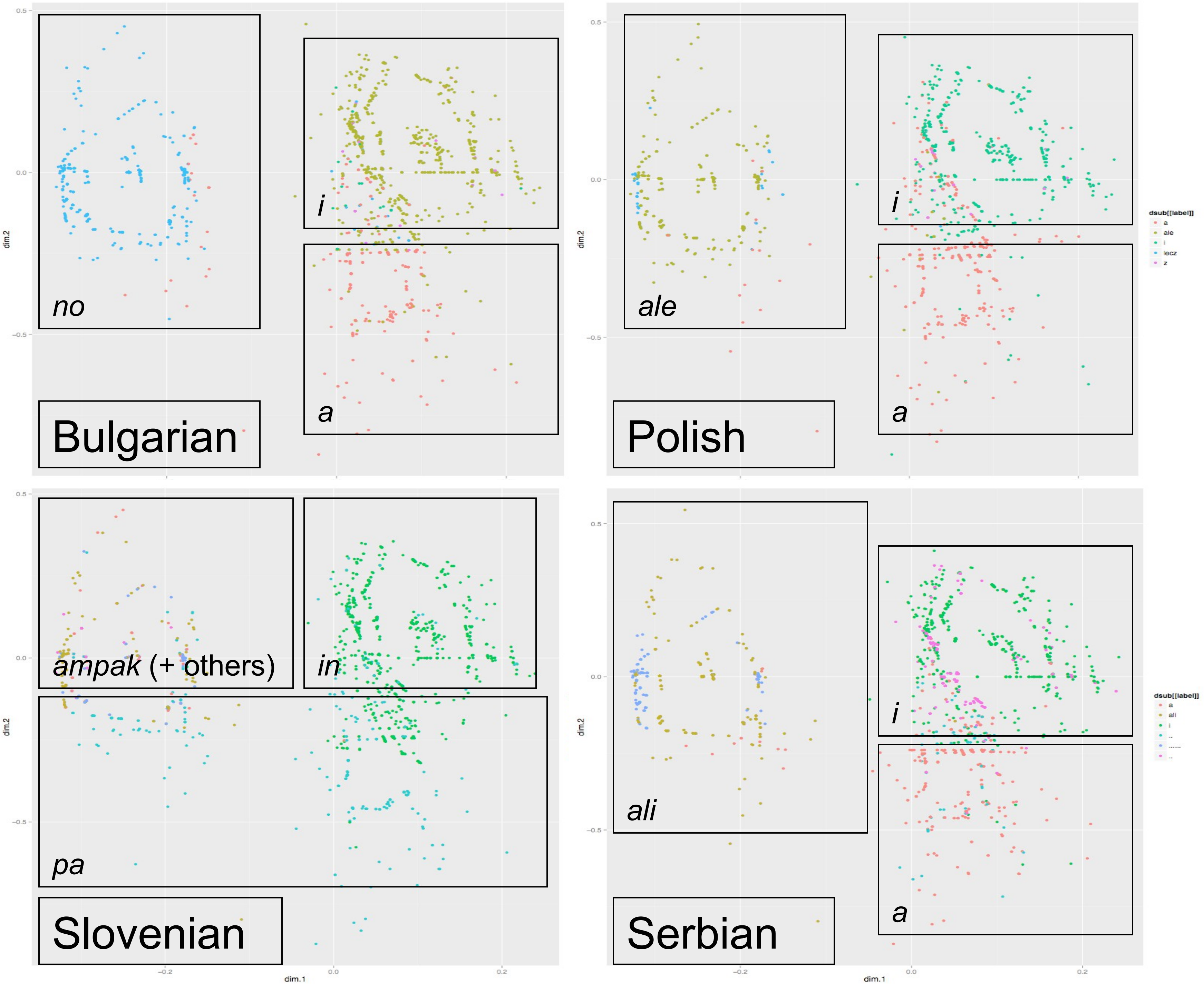
- Patterns of lexical cuts (where does the 'cup' end and the 'mug' start)
- Do we find types of languages?

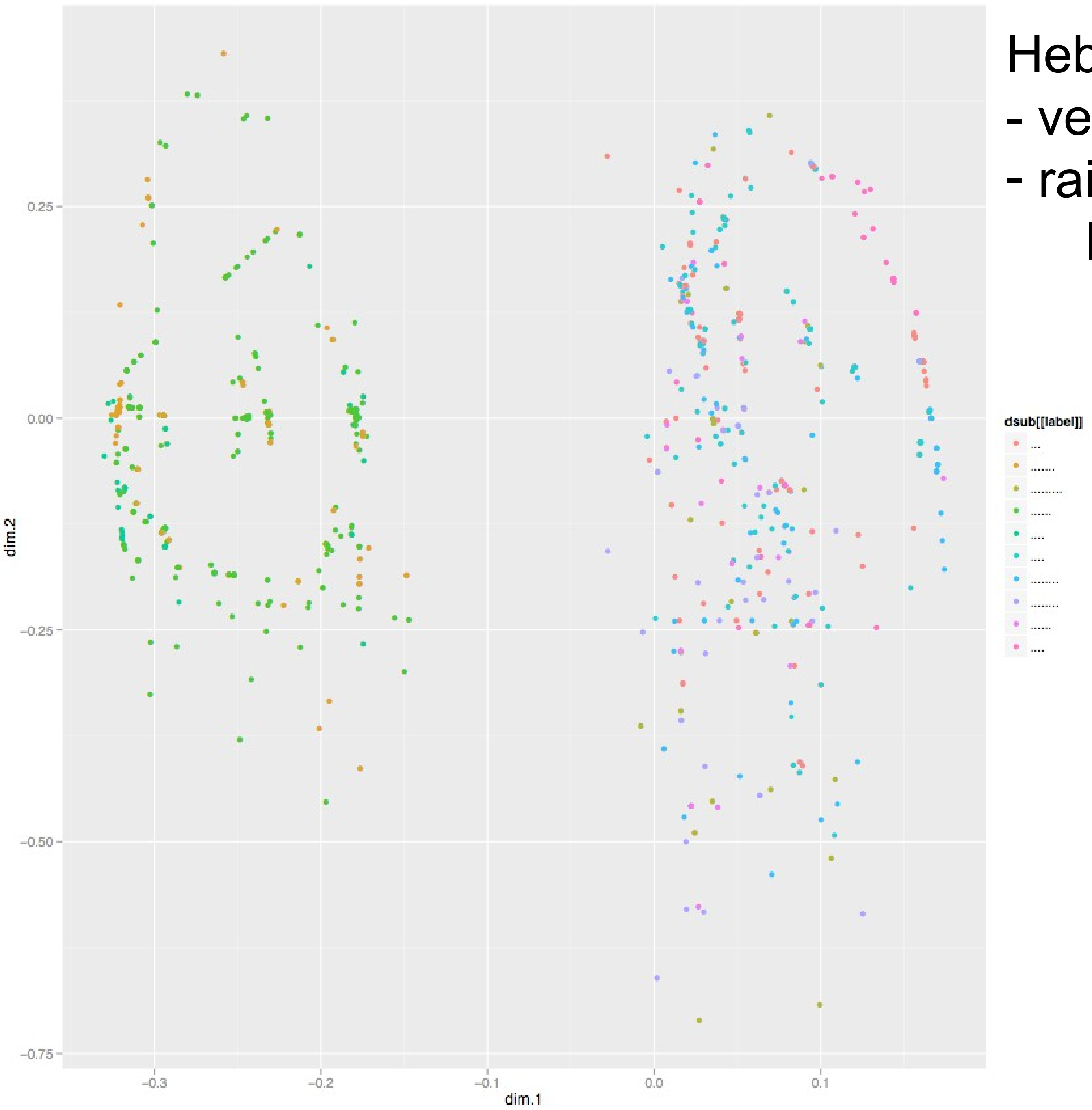
Study III: typology

- Method:
  - Same corpus (~27K lines subtitles in 30 languages)
  - Take a field of words in a particular language (e.g. coordinating conjunctions -- *and* and *but*)
  - Extract all clusters containing any of those words
  - Apply dimensionality reduction (MDS, PCA)
    - In our case: Croft and Poole’s (2008) Opimal Classification MDS

Conjunctions

- English
  - *and* (right)
  - *but* (left)


Bulgarian
no
i
a
Polish
ale
i
a
dsub[[label]]
a
ale
i
lecz
z
Slovenian
ampak (+ others)
in
pa
Serbian
ali
i
a
dsub[[label]]
a
ali
i
dim.1
dim.2



Hebrew:
- very mixed clusters
- raises the question:
  Indo-Eur. bias?

Study III: typology

- Discovering crosslinguistically common (English, most Slavic), and rare (Slovenian) ways of expressing some meaning
- Future: more analysis of differences
- Future: comparing clustering of usages against ‘manual’ groupings
- Future: predicting naturalness of categorization (is Slovenian conjunction system harder to learn)
- Future: using morphological segmentation/stemmer in pipeline

Conclusion

- Semantic typology: structure in the word-meaning inventory of the world's languages
- Mostly manual
- Pipeline: alignment -- graph clustering -- dimensionality reduction
  - Lots of room for further exploration: alignments vs. embeddings, various clustering algorithms and dimensionality reduction procedures
- Yields good clusters that reflect human semantic similarity judgments reasonably and can be used to study variation in expression and how common particular systems are

Thank you!