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?

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Mandarin</th>
<th>Dutch</th>
<th>Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horiz., no contact</td>
<td>Lamp above table</td>
<td>&lt; above</td>
<td>shang</td>
<td>boven yläpuolella &gt;</td>
</tr>
<tr>
<td>Stable support</td>
<td>Cup on table</td>
<td>&lt; on</td>
<td>shang</td>
<td>op -ssa &gt;</td>
</tr>
<tr>
<td>Tenuous support</td>
<td>Coat on hook</td>
<td>&lt; on</td>
<td>shang</td>
<td>aan -lla &gt;</td>
</tr>
<tr>
<td>Containment</td>
<td>Apple in bowl</td>
<td>&lt; in</td>
<td>li</td>
<td>in -lla &gt;</td>
</tr>
</tbody>
</table>
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)

[SP] Specific
I have something to tell you.

[NS] Irrealis non-specific
Call someone else!

[QU] Question
Is somebody home?

[CD] Conditional
If you see anything...

[IN] Indirect negation
I don't think anything will happen.

[CP] Comparison
She can run faster than anybody.

[DN] Direct negation
There's nobody home.

[FC] Free Choice
You can pick anything!
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

<table>
<thead>
<tr>
<th>Utterance</th>
<th>en</th>
<th>nl</th>
<th>es</th>
<th>sr</th>
<th>function</th>
</tr>
</thead>
<tbody>
<tr>
<td>someone is here</td>
<td>someone</td>
<td>iemand</td>
<td>alguien</td>
<td>neko</td>
<td>SP</td>
</tr>
<tr>
<td>anyone got 5 billion?</td>
<td>anyone</td>
<td>iemand</td>
<td>alguien</td>
<td>neko</td>
<td>QU</td>
</tr>
<tr>
<td>she could beat anyone</td>
<td>anyone</td>
<td>iedereen</td>
<td>qualquer</td>
<td>neko</td>
<td>FC</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q1: Frequency of functions

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

<table>
<thead>
<tr>
<th></th>
<th>SP</th>
<th>NS</th>
<th>CD</th>
<th>IN</th>
<th>DN</th>
<th>CP</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>people</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>20</td>
<td>07</td>
<td>16</td>
<td>05</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td><strong>things</strong></td>
<td>28</td>
<td>15</td>
<td>05</td>
<td>09</td>
<td>02</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td><strong>overall</strong></td>
<td>24</td>
<td>17</td>
<td>06</td>
<td>11</td>
<td>03</td>
<td></td>
<td>33</td>
</tr>
</tbody>
</table>

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
**Q1: Frequency of functions**

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

<table>
<thead>
<tr>
<th></th>
<th>SP</th>
<th>NS</th>
<th>CD</th>
<th>QU</th>
<th>IN</th>
<th>DN</th>
<th>CP</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEOPLE</td>
<td>.16</td>
<td>.20</td>
<td>.07</td>
<td>.16</td>
<td>.05</td>
<td>.28</td>
<td>.01</td>
<td>.08</td>
</tr>
<tr>
<td>THINGS</td>
<td>.28</td>
<td>.15</td>
<td>.05</td>
<td>.09</td>
<td>.02</td>
<td>.36</td>
<td>.00</td>
<td>.06</td>
</tr>
<tr>
<td>Overall</td>
<td>.24</td>
<td>.17</td>
<td>.06</td>
<td>.11</td>
<td>.03</td>
<td>.33</td>
<td>.00</td>
<td>.06</td>
</tr>
</tbody>
</table>

**Table:** Distribution of functions given ontological category.

SP :: specific  
NS :: non-spec.  
CD :: conditional  
QU :: question  
IN :: indirect neg.  
DN :: direct negation  
CP :: comparison  
FC :: free choice
Frequent functions are prototypes of Haspelmath’s map.

<table>
<thead>
<tr>
<th></th>
<th>SP</th>
<th>NS</th>
<th>CD</th>
<th>QU</th>
<th>IN</th>
<th>DN</th>
<th>CP</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEOPLE</td>
<td>.16</td>
<td>.20</td>
<td>.07</td>
<td>.16</td>
<td>.05</td>
<td>.28</td>
<td>.01</td>
<td>.08</td>
</tr>
<tr>
<td>THINGS</td>
<td>.28</td>
<td>.15</td>
<td>.05</td>
<td>.09</td>
<td>.02</td>
<td>.36</td>
<td>.00</td>
<td>.06</td>
</tr>
<tr>
<td>Overall</td>
<td>.24</td>
<td>.17</td>
<td>.06</td>
<td>.11</td>
<td>.03</td>
<td>.33</td>
<td>.00</td>
<td>.06</td>
</tr>
</tbody>
</table>

![Diagram showing the frequencies of various functions]

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

<table>
<thead>
<tr>
<th>( k )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEOPLE</td>
<td>.20</td>
<td>.25</td>
<td><strong>.41</strong></td>
<td>.35</td>
<td>.34</td>
<td>.34</td>
<td>.32</td>
<td>.30</td>
<td>.32</td>
</tr>
<tr>
<td>THINGS</td>
<td>.30</td>
<td>.38</td>
<td><strong>.47</strong></td>
<td>.36</td>
<td>.35</td>
<td>.35</td>
<td>.33</td>
<td>.39</td>
<td>.33</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SP</th>
<th>NS</th>
<th>CD</th>
<th>QU</th>
<th>IN</th>
<th>DN</th>
<th>CP</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>24</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Q2: Granularity of functions

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

- Too fine-grained

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SP</th>
<th>NS</th>
<th>CD</th>
<th>QU</th>
<th>IN</th>
<th>DN</th>
<th>CP</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>24</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Beekhuizen, Watson & Stevenson
Semantic Typology and Parallel Corpora
CogSci 2017 10 / 14
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

![Diagram showing semantic typology and parallel corpora]
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!