Automatic identification of construction candidates for a Swedish constructicon

Linnea Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice, and Emma Sköldberg

Språkbanken
University of Gothenburg

2013-04-19
The project: a Swedish constructicon (SweCxn)

- A *Swedish Constructicon* (SweCxn) is a project with the goal to collect, analyze, and describe Swedish constructions.
- The project is a collaboration between LT researchers, grammarians, lexicographers, phraseologists, semanticists, and L2-researchers.
- The development version is publicly available here (updated every night):
  
  http://språkbanken.gu.se/eng/resource/konstruktikon
Partially schematic constructions

▶ The focus of this talk is on *partially schematic constructions* = constructions with both lexical and (typed) variable constituents.

▶ Example:

**VB PN way**, e.g., we dined our way through the south of France.

▶ SweCxn example:

**X och X** ‘X and X’

<table>
<thead>
<tr>
<th>reaktiv_x_och_x</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>definition</strong></td>
</tr>
<tr>
<td><strong>structure</strong></td>
</tr>
</tbody>
</table>

- Många hittar till min sida nu för tiden, eller [ [många]_{item} [och]_{Koni} [många]_{item} ]_{reaktiv_x_och_x} allt är ju relativt förstås.
Methodology

- One of the goals of the project is to give as complete description of Swedish partially schematic constructions as possible.

- But how do we do that?
  - Cross language transfer – look into existing Constructicons
  - Scientific publications about Swedish constructions
  - Dictionaries and grammars (no systematic description exists though)
  - Investigate families of constructions – ‘friends’, e.g., reflexive constructions.
  - Annotate corpora (requires a highly trained eye to be able generalize construction instances to partially schematic constructions).
  - LT methods (this talk)
Automatic identification of constructions

- The LT work within the project focuses on identification of constructions in a double meaning:
  - to detect partially schematic constructions in large amount of texts;
  - to locate the SweCxn constructions in large amount of texts.
- Detecting partially schematic constructions = creating candidates list – suggestions that sometimes hit the spot, sometimes just direct your attention.
- Locating SweCxn constructions: What kind of formal information do we need to be able to precisely identify the SweCxn construction automatically?
Starting point

- The experiment is built upon the Swedish LT research infrastructure of Språkbanken ‘Language bank of Sweden’.
- Korp ’raven’ – corpora infrastructure (search interface; annotation pipeline; web services; downloadable annotated corpora)
- Karp ’carp’ – lexical infrastructure (search interface; web services; downloadable linked lexicons)
- A good starting point for the following experiment.
The corpus: SUC 2.0

- Corpus: SUC 2.0, a balanced corpus for Swedish that has been manually annotated with lemmas and MSD.
- A rather small material: 1,17 million tokens.

<table>
<thead>
<tr>
<th>word</th>
<th>msd</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hur</td>
<td>HA</td>
<td>hur</td>
</tr>
<tr>
<td>är</td>
<td>VB. PRS. AKT</td>
<td>vara</td>
</tr>
<tr>
<td>det</td>
<td>PN. NEU. SIN. DEF. SUB+OBJ</td>
<td>den</td>
</tr>
<tr>
<td>då</td>
<td>AB</td>
<td>då</td>
</tr>
<tr>
<td>i</td>
<td>PP</td>
<td>i</td>
</tr>
<tr>
<td>Mellanöstern</td>
<td>PM. NOM</td>
<td>Mellanöstern</td>
</tr>
<tr>
<td>?</td>
<td>MAD</td>
<td></td>
</tr>
</tbody>
</table>

‘How about the Middle East?’
Hybrid n-gram

- The experiment builds upon the work on StringNet.
- A fundamental concept in StringNet is hybrid n-grams.
- Hybrid n-grams are a generalization of n-grams where we also include the information in the annotation layers.
- We only consider lemmas and PoS in this experiment, but many other alternatives are of course possible.
- Hur är ⇒ hur vara, hur VB, HA vara, and HA VB.
Generation of hybrid n-grams

- 2-, 3-, 4-grams
- Since we want to catch partially schematic constructions, we filter out the schematic (e.g., HA VB) and lexical (e.g., hur vara) ones.
- We also filter out hybrid n-grams containing one of the following PoS: MID, MAD, PAD och UO (punctuations and foreign words).
- End result: 16 million hybrid n-grams of which 8.8 millions are unique.
The next step is to rank all hybrid n-grams.
This is done with an associative measure.
PMI – point-wise mutual information
PMI has a known problem: the measure prefers low-frequency items.
A known remedy is to multiply with the absolute frequency.
Another problem: boiler-plate text, e.g., *For subscription enquiries e-mail:...*

UIF: unique instance frequency – repeated instances are only counted once.

The formula:

\[
\text{PMI-UIF}(H) = \text{UIF} \times \log_2 \left( \frac{P(H)}{\prod_{x \in H} P(x)} \right)
\]
The problem of redundancy

- The problem: many hybrid n-grams are subsequences of other hybrid n-grams. This results in massive redundancy.
- The solution: filter out the hybrid n-grams that are subsequences of other n-grams with higher PMI-UIF.
- But this still suffers from redundancy, since a PoS and a lemma sometimes correspond to the same set of words (e.g., the infinitive mark $att_{IE}$ ‘to’).
- Solution: lemma and PoS are mergeable, e.g., $att_{IE} = IE$. 
The candidate list

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Most Frequent Instance</th>
<th>Absolute Frequency</th>
<th>Relative Frequency</th>
<th>UIF-PMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>vara VB ute AB och KN VB</td>
<td>är ute och letar (3)</td>
<td>15</td>
<td>0.93</td>
<td>52.24</td>
</tr>
<tr>
<td>vara VB JJ för PP att IE</td>
<td>är viktiga för att (2)</td>
<td>26</td>
<td>1.61</td>
<td>52.83</td>
</tr>
<tr>
<td>stänga VB av PL NN</td>
<td>stängt av motorn (1)</td>
<td>11</td>
<td>0.68</td>
<td>52.25</td>
</tr>
</tbody>
</table>

- Fields: the candidate, most frequent instance, absolute frequency, relative frequency, UIF-PMI
- Top-2500
- http://spraakbanken.gu.se/eng/resource/konstrukktikon/candidates
- Other candidate lists have been produced using automatically annotated corpora (run through the Korp annotation pipeline).
Korp instances

CQP query:

[lemma contains "vara" & pos = "VB"] [lemma contains "ute" & pos = "AB"] [lemma contains "och" & pos = "KN"] [pos = "VB"]

Search

KWIC: hits per page: 25 sort within corpora: not sorted Statistics: compile based on: word

Corpus

SUC 2.0

text attributes
text: kk82

word attributes

part-of-speech: verb
baseform: vara
lemgram: vara (verb)
sense: vara
initial part: [empty]
final part: [empty]
dependency relation: ROOT
msd: VB

Show Dependency Tree
Analysis of the construction candidates

- Of 2500 candidates, 50 was accepted as actual candidates.
- An indication that the method capture what we want is that we also get constructions already in the Swedish constructicon.
- Example:

  \text{RG NN \textit{per}_{PP} NN}

- \textit{en gång per dygn} ‘one time per day’, \textit{500 kronor per månad} ‘500 crowns per month’
Analysis of the construction candidates

- Another example: time expressions
  \[\text{den}_{DT} \ RO \ NN\]
  - \textit{den 1 juli} ‘(the) July 1’
  - \textit{den tionde mars} ‘(the) tenth of March’

- Compare with PP NN:
  - \textit{i mars} ‘in March’,
  - \textit{på morgonen} ‘in the morning’
  - \textit{på eftermiddagen} ‘in the afternoon’.
Analysis of the construction candidates

- Examples of constructions not in SweCxn:
  - RG år\textsubscript{NN\textsubscript{GEN}} ålder\textsubscript{NN}
    - sju års ålder 'seven years of age'
  - varken\textsubscript{KN} NN eller\textsubscript{KN} NN
    - varken uppehållstillstånd eller arbetstillstånd 'neither residence permit nor work permit'
  - vara\textsubscript{VB} sig\textsubscript{PN} NN eller\textsubscript{KN} (NN)
    - vare sig fotboll eller ishockey 'neither football or ice hockey'
  - vara\textsubscript{VB} sig\textsubscript{PN} PN VB (eller\textsubscript{KN} inte\textsubscript{AB})
    - vare sig vi vill eller inte 'regardless if we want or not'
Analysis of the construction candidates

▶ \( \textit{vara}_{VB} \ \textit{ute}_{AB} \ \textit{och}_{KN} \ \textit{VB} \)
  ▶ \textit{vara ute och jaga} ‘be out hunting’
▶ But also metaphorically:
  ▶ \textit{vara ute och cykla} ‘be out biking’
  ▶ \textit{vara ute och segla} ‘be out sailing’
  ▶ \textit{vara ute och snurra} ‘be out spinnig’
Error analysis

- Too syntactic:
  - \textit{den_{DT} JJ NN VB}
    
    \textit{de nordiska länderna är ‘The Nordic countries are’}
  - \textit{SN PN VB en_{DT}}
    
    \textit{att det var en ‘that there was a’}

- Fragmented:
  - \textit{vara_{VB} sig_{PN} NN eller_{KN} (NN)}

- Too lexical:
  - \textit{i_{PP} all_{DT} fall_{NN} VB}
  - \textit{över_{PP} huvud_{NN} tagen_{PC} VB}
Result

- From a methodological standpoint the experiment gave good results since we were able to identify previously undescribed partially schematic constructions.
- The precision is not impressive (2%), but as a push-button-tool for construction discovery it has proven to be an important contribution to the work on a Swedish constructicon.
Problems and future work

- Problems:
  - too lexical
  - too syntactic
  - fragmented

- Future work:
  - to use the lexical resources of Språkbanken in combination with the syntactic analysis of the Korp pipeline to try to improve the precision.
  - move towards more syntactic descriptions (sometimes using subtrees instead of tokens)
  - to investigate how to make the candidate sizes more flexible to capture larger units.
A quick demo

- If the internet gods are willing:
  http://spraakbanken.gu.se/eng/resource/konstruktikon/candidates