Tools for collocation extraction: preferences for active vs. passive

Ulrich Heid     Marion Weller

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

Marrakech, 29-5-2008, LREC-2008
Collocations: definitional elements
Working definition by S. Bartsch 2004:76

Collocations are

lexically and/or pragmatically constrained

recurrent cooccurrences

of at least two lexical items

which are in a direct syntactic relation with each other
Collocations are

*lexically and/or pragmatically constrained*

→ partial idiomatization:
  ○ at lexical-semantic level: choice of collocates
  ○ at morphosyntactic level: (partial) fixedness

*recurrent cooccurrences*

*of at least two lexical items*

*which are in a direct syntactic relation with each other*
Collocations: definitional elements
Working definition by S. Bartsch 2004:76

Collocations are

lexically and/or pragmatically constrained
→ partial idiomatization:
  ○ at lexical-semantic level: choice of collocates
  ○ at morphosyntactic level: (partial) fixedness

recurrent cooccurrences
→ observable by means of association measures

of at least two lexical items

which are in a direct syntactic relation with each other
Collocations: definitional elements
Working definition by S. Bartsch 2004:76

Collocations are

*lexically and/or pragmatically constrained*

→ partial idiomatization:
  ○ at lexical-semantic level: choice of collocates
  ○ at morphosyntactic level: (partial) fixedness

_recurrent cooccurrences_

→ observable by means of association measures

_of at least two lexical items_

→ binary structure: base + collocate, recursion possible

*which are in a direct syntactic relation with each other*
Collocations: definitional elements
Working definition by S. Bartsch 2004:76

Collocations are

*lexically and/or pragmatically constrained*

→ partial idiomatization:
  ○ at lexical-semantic level: choice of collocates
  ○ at morphosyntactic level: (partial) fixedness

*recurrent cooccurrences*

→ observable by means of association measures

*of at least two lexical items*

→ binary structure: base + collocate, recursion possible

*which are in a direct syntactic relation with each other*

→ relational cooccurrence (cf. Evert 2004, e.g.)
  ○ subject + verb: question arises
  ○ verb + object: raise + question
  ○ etc.
Options for collocation extraction (1/4)

Tasks of collocation extraction

Heid/Weller (IMS Stuttgart)
Options for collocation extraction (1/4)

Tasks of collocation extraction

- Identification of known collocations in text
Options for collocation extraction (1/4)

Tasks of collocation extraction

- Identification of known collocations in text
- Identification of new collocation candidates in texts
Options for collocation extraction (1/4)
Tasks of collocation extraction

- Identification of known collocations in text
- Identification of new collocation candidates in texts
- Collection of instances of collocation candidates and overview of morphosyntactic fixedness behaviour
Options for collocation extraction (1/4)

Tasks of collocation extraction

- Identification of known collocations in text
- Identification of new collocation candidates in texts
- Collection of instances of collocation candidates and overview of morphosyntactic fixedness behaviour
Available tool setups

- Statistics-only: association measures (AMs) over word sequences or windows
- Statistics + POS-filter (e.g. Smadja 1993):
  - cooccurrence candidates by statistics
  - filtering with patterns of allowable POS combinations
- POS-based extraction + statistical ranking (Heid 1998, Krenn 2000, Evert 2004, ...):
  - search via POS patterns, ranking via AMs
- Chunking-based extraction + statistical ranking (Ritz 2006, Ritz/Heid 2006)
- Parsing-based extraction + statistical ranking (Villada Moirón 2005, Seretán 2008, Geyken 2008)
Options for collocation extraction (2/4)
Available tool setups

- Statistics-only:
  association measures (AMs) over word sequences or windows
Options for collocation extraction (2/4)

Available tool setups

- **Statistics-only:**
  association measures (AMs) over word sequences or windows
- **Statistics + POS-filter (e.g. Smadja 1993):**
  - cooccurrence candidates by statistics
  - filtering with patterns of allowable POS combinations
Options for collocation extraction (2/4)
Available tool setups

• **Statistics-only:**
  association measures (AMs) over word sequences or windows

• **Statistics + POS-filter (e.g. Smadja 1993):**
  – cooccurrence candidates by statistics
  – filtering with patterns of allowable POS combinations

• **POS-based extraction + statistical ranking**
  – search via POS patterns, ranking via AMs
Options for collocation extraction (2/4)

Available tool setups

- **Statistics-only:**
  association measures (AMs) over word sequences or windows

- **Statistics + POS-filter (e.g. Smadja 1993):**
  - cooccurrence candidates by statistics
  - filtering with patterns of allowable POS combinations

- **POS-based extraction + statistical ranking**
  - search via POS patterns, ranking via AMs

- **Chunking-based extraction + statistical ranking**
  (Ritz 2006, Ritz/Heid 2006)
Options for collocation extraction (2/4)
Available tool setups

- **Statistics-only:**
  association measures (AMs) over word sequences or windows

- **Statistics + POS-filter** (e.g. Smadja 1993):
  - cooccurrence candidates by statistics
  - filtering with patterns of allowable POS combinations

- **POS-based extraction + statistical ranking**
  - search via POS patterns, ranking via AMs

- **Chunking-based extraction + statistical ranking**
  (Ritz 2006, Ritz/Heid 2006)

- **Parsing-based extraction + statistical ranking**
Options for collocation extraction (3/4)
Constraints on collocation extraction from German texts

- German verb placement models

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>VF</th>
<th>LK</th>
<th>MF</th>
<th>RK</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>v-1</td>
<td></td>
<td>Löst</td>
<td>der Mitarbeiter [... das Problem?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional</td>
<td>v-1</td>
<td></td>
<td>Löst</td>
<td>der Mitarbeiter [... das Problem,</td>
<td></td>
<td>so</td>
</tr>
<tr>
<td>Decl. sent.</td>
<td>v-2</td>
<td></td>
<td>Der Mitarbeiter</td>
<td>[... das Problem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subclause</td>
<td>vlast</td>
<td></td>
<td>weil</td>
<td>der Mitarbeiter [... das Problem</td>
<td>löst</td>
<td></td>
</tr>
</tbody>
</table>
Options for collocation extraction (3/4)

Constraints on collocation extraction from German texts

- **German verb placement models**

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>VF</th>
<th>LK</th>
<th>MF</th>
<th>RK</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>v-1</td>
<td></td>
<td>Lösst der Mitarbeiter [...] das Problem?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional</td>
<td>v-1</td>
<td></td>
<td>Lösst der Mitarbeiter [...] das Problem,</td>
<td>so ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decl. sent.</td>
<td>v-2</td>
<td></td>
<td>Der Mitarbeiter löst [...] das Problem</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subclause</td>
<td>vlast</td>
<td></td>
<td>weil der Mitarbeiter [...] das Problem  löst</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

→ More effort to produce extraction patterns, unless parsed data are used
Options for collocation extraction (3/4)
Constraints on collocation extraction from German texts

- German verb placement models

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>VF</th>
<th>LK</th>
<th>MF</th>
<th>RK</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>v-1</td>
<td></td>
<td>Löst</td>
<td>der Mitarbeiter [...] das Problem?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional</td>
<td>v-1</td>
<td></td>
<td>Löst</td>
<td>der Mitarbeiter [...] das Problem,</td>
<td></td>
<td>so ...</td>
</tr>
<tr>
<td>Decl. sent.</td>
<td>v-2</td>
<td>Der Mitarbeiter</td>
<td>löst</td>
<td>[...] das Problem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subclause</td>
<td>vlast</td>
<td></td>
<td>weil</td>
<td>der Mitarbeiter [...] das Problem</td>
<td>löst</td>
<td></td>
</tr>
</tbody>
</table>

→ More effort to produce extraction patterns, unless parsed data are used

- Relatively free constituent order in *Mittelfeld*

→ Risk of low precision on V+PP-collocations, due to object/adjunct problem
Options for collocation extraction (3/4)
Constraints on collocation extraction from German texts

- **German verb placement models**

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>VF</th>
<th>LK</th>
<th>MF</th>
<th>RK</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>v-1</td>
<td></td>
<td>Löst</td>
<td>der Mitarbeiter [...J das Problem?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional</td>
<td>v-1</td>
<td></td>
<td>Löst</td>
<td>der Mitarbeiter [...J das Problem,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decl. sent.</td>
<td>v-2</td>
<td></td>
<td>Der Mitarbeiter</td>
<td>[...] das Problem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subclause</td>
<td>vlast</td>
<td></td>
<td>weil</td>
<td>der Mitarbeiter [...J das Problem löst</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

→ More effort to produce extraction patterns, unless parsed data are used

- **Relatively free constituent order in *Mittelfeld***
  → Risk of low precision on V+PP-collocations, due to object/adjunct problem

- **Case syncretism in German NPs:**
  only 21 % unambiguous (Evert 2004)
  → Risk of lower precision on V+N_{Object}-collocations
Options for collocation extraction (4/4)

Proposed solution

Compromise

- Use of chunked text (available: $\gg 500$ M words):
  
  $\Rightarrow$ no need for large-scale parsing effort:
  
  efficient processing of large amounts of text

- Use of specific sentence types:
  
  The following allow for high precision extraction:
  
  - active + verb-final ($v\text{last}$)
  - passive + verb-1st
  - passive + verb-2nd
  - passive + verb-final

  $\Rightarrow$ Preference for high precision over high recall

  $\Rightarrow$ Detailed data on passives of V+N-collocations

  $\Rightarrow$ But: only approximative data on preferences for passives
Outline architecture
Instance of: chunking-based extraction + statistical ranking

- Preprocessing of corpora
Outline architecture
Instance of: chunking-based extraction + statistical ranking

- Preprocessing of corpora
- Pattern-based extraction
Outline architecture
Instance of: chunking-based extraction + statistical ranking

- Preprocessing of corpora
- Pattern-based extraction
- Intermediate storage in a database
Outline architecture
Instance of: chunking-based extraction + statistical ranking

• Preprocessing of corpora

• Pattern-based extraction

• Intermediate storage
  in a database
• Interpretation, e.g. LogL
Outline architecture
Instance of: chunking-based extraction + statistical ranking

- Preprocessing of corpora
  - Tagging (Schmid 1994), STTS
  - Lemmatization (Schmid 1994)
  - Chunking (Kermes 2003)

- Pattern-based extraction
  - Based on Stuttgart Corpus WorkBench, CWB (Evert 2005)

- Intermediate storage in a database
- Interpretation, e.g. LogL
Outline architecture
Instance of: chunking-based extraction + statistical ranking

- Preprocessing of corpora
  - Tagging (Schmid 1994), STTS
  - Lemmatization (Schmid 1994)
  - Chunking (Kermes 2003)

- Pattern-based extraction
  - based on Stuttgart CorpusWorkBench, CWB (Evert 2005)

- Intermediate storage in a database
- Interpretation, e.g. LogL

Heid/Weller (IMS Stuttgart)  Collocations: active/passive  29-5-08 7 / 24
Outline architecture
Instance of: chunking-based extraction + statistical ranking

- Preprocessing of corpora
  - Tagging (Schmid 1994), STTS
  - Lemmatization (Schmid 1994)
  - Chunking (Kermes 2003)

- Pattern-based extraction
  - based on Stuttgart CorpusWorkBench, CWB (Evert 2005)
  - (Ritz 2006)

- Intermediate storage in a database
- Interpretation, e.g. LogL
Outline architecture
Instance of: chunking-based extraction + statistical ranking

• Preprocessing of corpora
  Tagging (Schmid 1994), STTS
  Lemmatization (Schmid 1994)
  Chunking (Kermes 2003)

• Pattern-based extraction
  based on Stuttgart
  CorpusWorkBench, CWB
  (Evert 2005)

• Intermediate storage
  in a database
  (Ritz 2006)

• Interpretation, e.g. LogL
  (Dunning 1993, Evert 2004)
Extraction details: sample query

MACRO passive_verb-final(0)
1 (  
2 [pos = "(KOU(S|I)|PRELS)"]  
3 []*  
4 <np>  
5 @![pp & !ap & _.np_f not contains "ne" & _.np_f not contains "pron"  
6 & _.np_f not contains "meas" & _.np_h != "@card@"]  
7 ![pp & !ap & _.np_f not contains "ne" & _.np_f not contains "pron"  
8 & _.np_f not contains "meas" & _.np_h != "@card@"]*  
9 </np>  
10 ![np & pos != "\$.|KOUS|VMFIN"]*  
11 [pos = "V.*"]*  
12 [pos = "VVPP"]  
13 [lemma = "(werden|sein)"]  
14 [pos = "V.*"]*  
15 [pos = "\$.|KON"]  
16 )  
17 within s

- verb-final clause: v-participle at the end (12),
  conjunction at the beginning (2)
- NP left of verb complex (4-9)
- removal of unwanted nominals:
  pronouns, proper names, measure items (4-9)
Extraction details: morphosyntactactic features

- noun and verb lemma, and type of determiner (4-9, 12)
- NP number (4-9)
- tense (11/14), modal (11/14) and passive auxiliary (13)
- active/passive and verb placement model: extracted via different named queries
Extraction details: morphosyntactic features

database entry

| n-lemma | v-lemma | prep. | number | fusion | tense | ...
|---------|---------|------|--------|--------|-------|-----
| Diskussion | stehen | zu | singular | yes | present | ...

query matched against text

(features annotated)

preparation: zu + def <DAT>

- fused-prep-article

- n-lemma: Diskussion

- number: singular

- v-lemma: stehen

- participle: present

- vder: yes

partially parsed text

<np>
  die
  <ap> <ap_f |attr|pp|vder|>
    <pp>
      zur
      <np1>
        Diskussion
      </np1>
    </pp>
  </ap_f> <ap>

<np1>
  Problem
</np1>

Probleme
</np>

<ap> [<._ap_f contains "vder"]
<pp>
  <np1> [[]*</np1>

</pp>

)</ap>
Results: data

Corpora used:
- Newspapers (ca. 200 M)
- Juridical Journals (76 M)
- EU texts from JRC:
  Acquis Communautaire (16 M)
Results: data

- Passives: 5.8 – 15.3 % of all occurrences

Corpora used:
- Newspapers (ca. 200 M)
- Juridical Journals (76 M)
- EU texts from JRC:
  Acquis Communautaire (16 M)
Results: data

- Passives: 5.8 – 15.3 % of all occurrences

Corpora used:
- Newspapers (ca. 200 M)
- Juridical Journals (76 M)
- EU texts from JRC: *Acquis Communautaire* (16 M)

- Morphosyntactic preferences of collocations come out clearly: variability vs. fixedness (see example on next slide)
Results: data

- Passives: 5.8 – 15.3 % of all occurrences

Corpora used:
- Newspapers (ca. 200 M)
- Juridical Journals (76 M)
- EU texts from JRC:
  Acquis Communautaire (16 M)

- Morphosyntactic preferences of collocations come out clearly: variability vs. fixedness (see example on next slide)

- Complex-predicate type collocations: no passive under V2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Auffassung vertreten (&quot;be of ... opinion&quot;)</td>
<td>1321</td>
<td>53</td>
<td>97</td>
<td>48</td>
</tr>
<tr>
<td>Bezug nehmen (&quot;make reference&quot;)</td>
<td>783</td>
<td>439</td>
<td>492</td>
<td>0</td>
</tr>
<tr>
<td>Rechnung tragen (&quot;keep track&quot;)</td>
<td>2287</td>
<td>481</td>
<td>492</td>
<td>0</td>
</tr>
<tr>
<td>Gebrauch machen (&quot;make use &quot;)</td>
<td>2095</td>
<td>216</td>
<td>430</td>
<td>0</td>
</tr>
<tr>
<td>Sorge tragen (&quot;care for&quot;)</td>
<td>241</td>
<td>31</td>
<td>43</td>
<td>0</td>
</tr>
</tbody>
</table>
Results: an example case with details

Angst haben ("fear")

<table>
<thead>
<tr>
<th>f</th>
<th>n_lemma</th>
<th>v_lemma</th>
<th>det_sort</th>
<th>num</th>
<th>aktiv_passiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>209</td>
<td>Angst</td>
<td>haben</td>
<td>null</td>
<td>Sg</td>
<td>active</td>
</tr>
<tr>
<td>40</td>
<td>Angst</td>
<td>haben</td>
<td>quant</td>
<td>Sg</td>
<td>active</td>
</tr>
<tr>
<td>6</td>
<td>Angst</td>
<td>haben</td>
<td>def</td>
<td>Sg</td>
<td>active</td>
</tr>
<tr>
<td>2</td>
<td>Angst</td>
<td>haben</td>
<td>null</td>
<td>Pl</td>
<td>active</td>
</tr>
<tr>
<td>1</td>
<td>Angst</td>
<td>haben</td>
<td>indef</td>
<td>Sg</td>
<td>active</td>
</tr>
</tbody>
</table>
Results: an example case with details  
*Konsequenz(en) ziehen* (“draw consequence(s)”)

<table>
<thead>
<tr>
<th>f</th>
<th>n_lemma</th>
<th>v_lemma</th>
<th>det_sort</th>
<th>num</th>
<th>sent_type</th>
<th>aktiv_passiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>null</td>
<td>Pl</td>
<td>v-1</td>
<td>passiv</td>
</tr>
<tr>
<td>5</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>def</td>
<td>Sg</td>
<td>v-1</td>
<td>passiv</td>
</tr>
<tr>
<td>1</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>quant</td>
<td>Pl</td>
<td>v-1</td>
<td>passiv</td>
</tr>
<tr>
<td>11</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>null</td>
<td>Pl</td>
<td>v-2</td>
<td>passiv</td>
</tr>
<tr>
<td>1</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>null</td>
<td>Sg</td>
<td>v-2</td>
<td>passiv</td>
</tr>
<tr>
<td>104</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>null</td>
<td>Pl</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>77</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>def</td>
<td>Pl</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>22</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>def</td>
<td>Sg</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>13</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>quant</td>
<td>Pl</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>11</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>poss</td>
<td>Pl</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>3</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>indef</td>
<td>Sg</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>2</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>dem</td>
<td>Sg</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>1</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>dem</td>
<td>Pl</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>1</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>poss</td>
<td>Sg</td>
<td>vvirsk</td>
<td>aktiv</td>
</tr>
<tr>
<td>16</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>null</td>
<td>Pl</td>
<td>vvirsk</td>
<td>passiv</td>
</tr>
<tr>
<td>3</td>
<td>Konsequenz</td>
<td>ziehen</td>
<td>quant</td>
<td>Pl</td>
<td>vvirsk</td>
<td>passiv</td>
</tr>
</tbody>
</table>
### Results: an example case with details

*Konsequenz(en) ziehen* ("draw consequence(s)")

<table>
<thead>
<tr>
<th>neg</th>
<th>modal</th>
<th>chunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td></td>
<td>Welche Konsequenzen werden aus den Untersuchungen gezogen</td>
</tr>
<tr>
<td>-</td>
<td>muessen</td>
<td>Konsequenzen muessen gezogen werden</td>
</tr>
<tr>
<td>-</td>
<td></td>
<td>Konsequenzen wurden dennoch erst gestern gezogen</td>
</tr>
<tr>
<td>-</td>
<td>muessen</td>
<td>Konsequenzen muessten gezogen werden</td>
</tr>
<tr>
<td>-</td>
<td>muessen</td>
<td>Welche Konsequenzen muessen Ihrer Ansicht nach aus diesem Wahlkampf gezogen werden</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>Konsequenzen wurden aber bisher nicht gezogen</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>Konsequenzen wurden daraus bisher noch nicht gezogen</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>Konsequenzen wurden aus derlei Einsichten freilich nicht gezogen</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>Konsequenzen wurden aber anscheinend daraus nie gezogen</td>
</tr>
<tr>
<td>-</td>
<td>koennen</td>
<td>Konsequenzen koennten aber erst am Ende des Aufklaerungsprozesses gezogen werden</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>Konsequenzen wurden daraus nicht gezogen</td>
</tr>
<tr>
<td>-</td>
<td>koennen</td>
<td>Konsequenz kann aus dem Geschehen in der Front National gezogen werden</td>
</tr>
</tbody>
</table>
Evaluation: precision

Preprocessing

- Chunking: chunk size determination (chu)
- Word order model determination (w.o.)
- Active/passive identification (a/p.)
- Collocation candidates (verb + complement) (v+c.)

<table>
<thead>
<tr>
<th>context type</th>
<th>w.o.</th>
<th>a/p.</th>
<th>chu.</th>
<th>v+c.</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb-second, passive</td>
<td>100.0</td>
<td>100.0</td>
<td>96.0</td>
<td>96.0</td>
</tr>
<tr>
<td>verb-final, active</td>
<td>56.0</td>
<td>98.0</td>
<td>100.0</td>
<td>88.0</td>
</tr>
<tr>
<td>verb-final, passive</td>
<td>100.0</td>
<td>84.0</td>
<td>100.0</td>
<td>80.0</td>
</tr>
<tr>
<td>complete set, average</td>
<td>85.3</td>
<td>94.0</td>
<td>98.7</td>
<td>81.3</td>
</tr>
</tbody>
</table>
Evaluation: precision
Collocation candidate extraction

Categories:
- complex predicates
- collocations: verb + complement
- syntactically valid verb + complement pair
- errors

<table>
<thead>
<tr>
<th>Criteria</th>
<th>set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives + sublang. coll</td>
<td>68.9 %</td>
</tr>
<tr>
<td>– True positives</td>
<td>20.5 %</td>
</tr>
<tr>
<td>– – Complex predicates</td>
<td>2.1 %</td>
</tr>
<tr>
<td>– – Collocations</td>
<td>18.4 %</td>
</tr>
<tr>
<td>– Sublanguage collocations</td>
<td>48.5 %</td>
</tr>
<tr>
<td>True negatives:</td>
<td>31.0 %</td>
</tr>
<tr>
<td>– subject + verb</td>
<td>7.8 %</td>
</tr>
<tr>
<td>– other</td>
<td>23.2 %</td>
</tr>
</tbody>
</table>

Sample: 2338 candidate pair types from Acquis Communautaire
Evaluation: comparison with parsing
Data from juridical journals (78 M words), top 250 candidates per tool

Mini-experiment (F. Fritzinger)

- Compared:
  our system vs. extraction from parsed text (Schiehlen 2003)
- Precision:
  - very high overlap in candidate lists, minimal (ca. 5%) differences are of technical nature
  - parsing allows for better subdivision V+Subj/V+Obj, as it uses a subcategorization dictionary
- Recall ($V+N_{\text{Object}}$): substantial increase with parsing: cf. results by Serețan 2008 for EN and FR

<table>
<thead>
<tr>
<th></th>
<th>types</th>
<th>tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chunking-based</td>
<td>254.930</td>
<td>658.687</td>
</tr>
<tr>
<td>Parsing-based</td>
<td>535.098</td>
<td>1.496.401</td>
</tr>
</tbody>
</table>
Conclusions

We presented

- a chunking + AM-based system for collocation candidate extraction: viable compromise:
  - efficient on large amounts of data
  - good precision: similar to parsing
  - but low recall: less than half of what parsing finds

- a detailed account of morphosyntactic preferences of German V+N-collocations, including passivizability
  ⇒ full picture on flexibility

- correlations between complex predicates and non-passivizability under V-2:
  identification of complex predicates: good precision, but low recall
Next steps

• Combine parsing-based extraction with detailed identification of morphosyntactic features

• Use ambiguity annotation of parser output to separate out:
  – clear evidence vs. possibly incorrect evidence
  – e.g. for Adj+N-collocations:
    *alte Männer und Frauen* (old men and women)

⇒ further increase in precision?

• Analysis of collocation combinations, as e.g. adverbs in collocations are in our intermediate database