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: 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

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

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

Options for collocation extraction (2/4)

Available tool setups

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
  (Heid 1998, Krenn 2000, Evert 2004, . . .):
  - 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
  (Heid 1998, Krenn 2000, Evert 2004, . . .):
  - 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
  (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, Sereţan 2008, Geyken 2008)

Options for collocation extraction (3/4)

Constraints on collocation extraction from German texts

- German verb placement models

Type	Model	VF	LK	MF	RK	NF
Question	v-1		Löst	der Mitarbeiter [...] das Problem?		
Conditional	v-1		Löst	der Mitarbeiter [...] das Problem,		so ...
Decl. sent.	v-2	Der Mitarbeiter	löst	[...] das Problem		
Subclause	vlast		weil	der Mitarbeiter [...] das Problem	löst	

Options for collocation extraction (3/4)

Constraints on collocation extraction from German texts

- German verb placement models

Type	Model	VF	LK	MF	RK	NF
Question	v-1		Löst	der Mitarbeiter [...] das Problem?		
Conditional	v-1		Löst	der Mitarbeiter [...] das Problem,		so ...
Decl. sent.	v-2	Der Mitarbeiter	löst	[...] das Problem		
Subclause	vlast		weil	der Mitarbeiter [...] das Problem	löst	

→ 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

Type	Model	VF	LK	MF	RK	NF
Question	v-1		Löst	der Mitarbeiter [...] das Problem?		
Conditional	v-1		Löst	der Mitarbeiter [...] das Problem,		so ...
Decl. sent.	v-2	Der Mitarbeiter	löst	[...] das Problem		
Subclause	vlast		weil	der Mitarbeiter [...] das Problem	löst	

  → 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

Type	Model	VF	LK	MF	RK	NF
Question	v-1		Löst	der Mitarbeiter [...] das Problem?		
Conditional	v-1		Löst	der Mitarbeiter [...] das Problem,		so ...
Decl. sent.	v-2	Der Mitarbeiter	löst	[...] das Problem		
Subclause	vlast		weil	der Mitarbeiter [...] das Problem	löst	

  → 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):

  ⇒ 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 (`vlast`)
  - passive + verb-1st
  - passive + verb-2nd
  - passive + verb-final

  ⇒ Preference for high precision over high recall
  ⇒ Detailed data on passives of V+N-collocations
  ⇒ 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

- 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

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

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
```

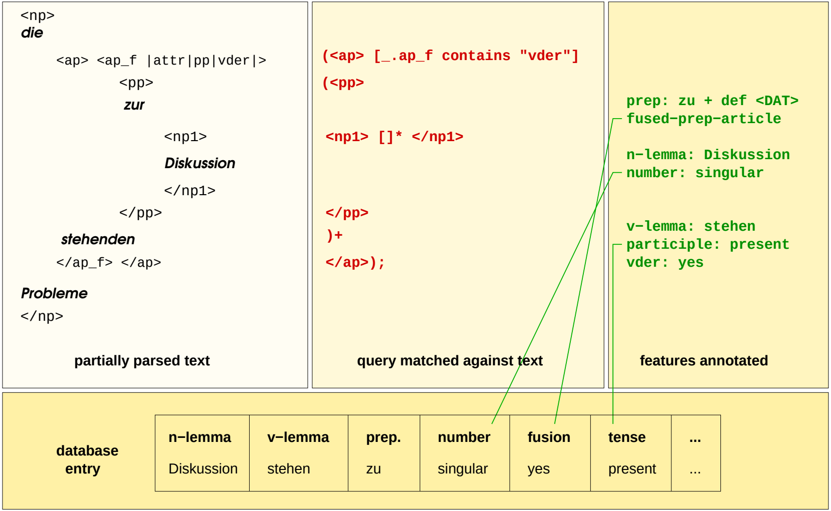
- verb-final clause: v-participle at the end (12),
  conjunction at the beginning (2)
- NP left of verb complex (4-9)
- removal of unwanted nominals:
  pronouns, proper names, measure items (4-9)

Extraction details: morphosyntactic features

```
     MACRO passive_verb-final(0)
1    (
2    [pos = "(KOU(S|I)|PRELS)"]
3    []*
4    <np>
...
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
```

- 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



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

partially parsed text

```
(<ap> [_.ap_f contains "vder"]
(<pp>

 <np1> []* </np1>


 </pp>
 )+
 </ap>);
```

query matched against text

```
prep: zu + def <DAT>
fused−prep−article

n−lemma: Diskussion
number: singular

v−lemma: stehen
participle: present
vder: yes
```

features annotated

database entry

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

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

Candidate	A:V-L	P:V-1	P:V-L	P:V-2
Auffassung vertreten (“be of . . . opinion”)	1321	53	97	48
Bezug nehmen (“make reference”)	783	439	492	0
Rechnung tragen (“keep track”)	2287	481	492	0
Gebrauch machen (“make use ”)	2095	216	430	0
Sorge tragen (“care for”)	241	31	43	0

Results: an example case with details

Angst haben (“fear”)

```
f   | n_lemma | v_lemma | det_sort| num  | aktiv_passiv
------------------------------------------------
209 | Angst   | haben   | null    | Sg   | active
 40 | Angst   | haben   | quant   | Sg   | active
  6 | Angst   | haben   | def     | Sg   | active
  2 | Angst   | haben   | null    | Pl   | active
  1 | Angst   | haben   | indef   | Sg   | active
```

Results: an example case with details

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

```
  f  |  n_lemma   | v_lemma | det_sort | num | sent_type | aktiv_passiv
-----+------------+---------+----------+-----+-----------+--------------
  13 | Konsequenz | ziehen  | null     | Pl  | v-1       | passiv
   5 | Konsequenz | ziehen  | def      | Sg  | v-1       | passiv
   1 | Konsequenz | ziehen  | quant    | Pl  | v-1       | passiv
  11 | Konsequenz | ziehen  | null     | Pl  | v-2       | passiv
   1 | Konsequenz | ziehen  | null     | Sg  | v-2       | passiv
 104 | Konsequenz | ziehen  | null     | Pl  | vvirsk    | aktiv
  77 | Konsequenz | ziehen  | def      | Pl  | vvirsk    | aktiv
  22 | Konsequenz | ziehen  | def      | Sg  | vvirsk    | aktiv
  13 | Konsequenz | ziehen  | quant    | Pl  | vvirsk    | aktiv
  11 | Konsequenz | ziehen  | poss     | Pl  | vvirsk    | aktiv
   3 | Konsequenz | ziehen  | indef    | Sg  | vvirsk    | aktiv
   2 | Konsequenz | ziehen  | dem      | Sg  | vvirsk    | aktiv
   1 | Konsequenz | ziehen  | dem      | Pl  | vvirsk    | aktiv
   1 | Konsequenz | ziehen  | poss     | Sg  | vvirsk    | aktiv
  16 | Konsequenz | ziehen  | null     | Pl  | vvirsk    | passiv
   3 | Konsequenz | ziehen  | quant    | Pl  | vvirsk    | passiv
```

Results: an example case with details

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

```
 neg | modal   |                                        chunk
-----+---------+------------------------------------------------------------------------------------
 -   |         | Welche Konsequenzen werden aus den Untersuchungen gezogen
 -   | muessen | Konsequenzen muessen gezogen werden
 -   |         | Konsequenzen wurden dennoch erst gestern gezogen
 -   | muessen | Konsequenzen muessten gezogen werden
 -   | muessen | Welche Konsequenzen muessen Ihrer Ansicht nach aus diesem Wahlkampf gezogen werden
 +   |         | Konsequenzen wurden aber bisher nicht gezogen
 +   |         | Konsequenzen wurden daraus bisher noch nicht gezogen
 +   |         | Konsequenzen wurden aus derlei Einsichten freilich nicht gezogen
 +   |         | Konsequenzen wurden aber anscheinend daraus nie gezogen
 -   | koennen | Konsequenzen koennten aber erst am Ende des Aufklaerungsprozesses gezogen werden
 +   |         | Konsequenzen wurden daraus nicht gezogen
 -   | koennen | Konsequenz kann aus dem Geschehen in der Front National gezogen werden
```

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.`)

context type	w.o.	a/p.	chu.	v+c.
verb-second, passive	100.0	100.0	96.0	96.0
verb-final, active	56.0	98.0	100.0	88.0
verb-final, passive	100.0	84.0	100.0	80.0
complete set, average	85.3	94.0	98.7	81.3

Evaluation: precision

Collocation candidate extraction

Categories:

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

Criteria	set 2
True positives + sublang. coll	68.9 %
– True positives	20.5 %
– – Complex predicates	2.1 %
– – Collocations	18.4 %
– Sublanguage collocations	48.5 %
True negatives:	31.0 %
– subject + verb	7.8 %
– other	23.2 %

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_{Object}): substantial increase with parsing:
  cf. results by Sereţan 2008 for EN and FR

	types	tokens
Chunking-based	254.930	658.687
Parsing-based	535.098	1.496.401

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