Between collocation and construction: Lexical preferences in non-idiomatic word combinations

Stefan Evert\textsuperscript{1}, Ulrich Heid\textsuperscript{2}

\textsuperscript{1}FAU Erlangen–Nürnberg  
\textsuperscript{2}Universität Hildesheim

EUROPHRAS 2019
Collocations often understood as word pairs (Hausmann 2004, Mel'čuk 2003)

- pay + attention
- deeply grateful
- strong objections
- criticise severely

Collocation as a syntactic phenomenon (Bartsch 2004)

Longer word combinations

- X pays {particular, special, close, ...} attention
- X {raises, has} strong objections
- X earns Y respect
- DE X übt heftige Kritik (‘criticises severely’)
Research programme

Claims

★ Collocations at the centre of the syntax-lexicon continuum
★ Longer combinations as collo-constructions (Herbst 2018)

Research questions

★ Can corpus data help delineate the status of word combinations?
★ Could a classification support lexicographic presentation?
★ Is the compilation of a comprehensive Collo-Constructicon possible?

Methodological prerequisite

★ Suitable methods for the quantitative analysis of lexico-grammatical patterns beyond word pairs in large corpora
Two small case studies

**DE** X übt {heftige, scharfe, massive, harsche, …} Kritik

interpreted as a combination of collocations (Zinsmeister & Heid 2003)

**EN** Y earns {living, money, wages, income, salary, …} vs. X earns Y {respect, nickname, title, …}

interpreted as (constructional) lexical preferences (cf. Herbst 2018)
### Corpus data: Kritik üben

#### German news corpus, 205 M words, 1990s

#### Kritik + ADJ

<table>
<thead>
<tr>
<th>ADJ</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>heftig</td>
<td>1069</td>
</tr>
<tr>
<td>scharf</td>
<td>1006</td>
</tr>
<tr>
<td>harsch</td>
<td>417</td>
</tr>
<tr>
<td>massiv</td>
<td>389</td>
</tr>
<tr>
<td>öffentlich</td>
<td>357</td>
</tr>
<tr>
<td>hart</td>
<td>283</td>
</tr>
</tbody>
</table>

#### VERB (+ Prep) + Kritik

<table>
<thead>
<tr>
<th>Verb + Preposition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>üben</td>
<td>455</td>
</tr>
<tr>
<td>stoßen auf</td>
<td>223</td>
</tr>
<tr>
<td>es gibt</td>
<td>108</td>
</tr>
<tr>
<td>reagieren auf</td>
<td>78</td>
</tr>
<tr>
<td>äußern</td>
<td>70</td>
</tr>
</tbody>
</table>

#### Examples:

- **Hempel äußerte scharfe Kritik**
- **Brandbriefe mit scharfer Kritik**
- **Gewerkschaften reagieren mit scharfer Kritik**
- **von der Parteilinken kam scharfe Kritik**
Corpus data: Kritik üben

German news corpus, 205 M words, 1990s

No strong associations between these verbs & adjectives

Kritik üben + ADJ:
- scharf, heftig, hart, harsch, massiv, deutlich, konstruktiv, herb, ...

VERB + \{scharfe, heftige, massive\} Kritik:
- üben, stoßen auf, äußern, es gibt, reagieren auf, auslösen, ernten, ...

Proposal for description: combination of binary collocations

- \{scharfe, heftige, massive, \ldots\} Kritik
  + Kritik \{üben, stoßen auf, \ldots\}
Corpus data: *earn*

British National Corpus, 100 M words, 1990s

🌟 Analysis of two syntactic patterns

1. *Y earns sth.*
2. *X earns Y sth.*

🌟 Focus on lexical realization of direct object

🌟 Pattern 1

- *sbdy earns <n> pounds*
- *sbdy earns {money, interest, profits, …}*
- *sbdy earns {salary, wages, revenue, …}*
- *sbdy earns {a, his, her, …} living*
Corpus data: earn

British National Corpus, 100 M words, 1990s

Pattern 2

- **sth. earns Y respect**
- **sth. earns Y {reputation, fame, recognition, award, praise, …}**
- **sth. earns Y the {nickname, title, sobriquet, epithet, …} NOUN**
- [sports] **sth. earns Y {a place, …, points, …, championship, …}**
- [rare] **sth. earns Y {hatred, enemies, derision, …}**
- [very rare] **sth. earns Y {extra cash, money, fees, gold bars, …}**

less than 4% of retrieved examples

Proposal for description:

- The valency pattern (2) comes with semantic and/or lexical preferences which are different from those of pattern (1)
- Constructional interpretation: Valency pattern and lexical preferences go together as collo-construction (Herbst 2018)
Consequences

Corpus linguistic tasks:

🌟 Identifying collo-constructional phenomena:
   How many and which components belong together?

🌟 Possibly separating collo-constructions and collocations

▷ within a dependency-based framework

Lexicographic task:

🌟 Describing valency and collo-constructional data in an integrated way, especially for text production dictionaries
In an *open barouche* [...] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [...]
In an *open barouche* [...] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [...]
Co-occurrence as cross-classification

**Item = instance of adjective-noun dependency relation**

In an open barouche [...] stood a stout old gentleman, in a blue coat and bright buttons, corduroy breeches and top-boots; two young ladies in scarfs and feathers; a young gentleman apparently enamoured of one of the young ladies in scarfs and feathers; a lady of doubtful age, probably the aunt of the aforesaid; and [...] 

\[
f(\text{young, gentleman}) = 1 \\
\text{sample size N = 9}
\]
Contingency tables & association measures

See Evert (2008) for details | http://www.collocations.de/

<table>
<thead>
<tr>
<th></th>
<th>( w_2 )</th>
<th>( \neg w_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_1 )</td>
<td>( O_{11} )</td>
<td>( O_{12} )</td>
</tr>
<tr>
<td>( \neg w_1 )</td>
<td>( O_{21} )</td>
<td>( O_{22} )</td>
</tr>
</tbody>
</table>

\[ = C_1 \]
\[ = C_2 \]
\[ = N \]

\( w_1 \)
\[ E_{11} = \frac{R_1 C_1}{N} \]
\[ E_{12} = \frac{R_1 C_2}{N} \]

\( \neg w_1 \)
\[ E_{21} = \frac{R_2 C_1}{N} \]
\[ E_{22} = \frac{R_2 C_2}{N} \]

**observed**

**expected**
Statistical association measures (AM)

See Evert (2008) for details | http://www.collocations.de/

\[ \text{MI} = \log_2 \frac{O}{E} \]
\[ \text{MI}^k = \log_2 \frac{O^k}{E} \]
\[ \text{local-MI} = O \cdot \log_2 \frac{O}{E} \]

\[ \text{z-score} = \frac{O - E}{\sqrt{E}} \]
\[ \text{t-score} = \frac{O - E}{\sqrt{O}} \]
\[ \text{simple-ll} = 2 \left( O \cdot \log \frac{O}{E} - (O - E) \right) \]

\[ \text{chisquared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]
\[ \text{chisquarecorr} = \frac{N \left( |O_{11}O_{22} - O_{12}O_{21}| - N/2 \right)^2}{R_1R_2C_1C_2} \]

\[ \text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \]
\[ \text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}} \]

\[ \text{Dice} = \frac{2O_{11}}{R_1 + C_1} \]
\[ \text{odds-ratio} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \]
\[ \Delta P_{2|1} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2} \]
Moving beyond word pairs

- Incremental extension of n-grams & technical terms (e.g. LocalMaxs, da Silva et al. 1999)
- Generalize expected frequencies and association measures to word triples (Lin 1998, Zinsmeister & Heid 2003)
- Hypothesis tests in n-dimensional contingency tables (Blaheta & Johnson 2001)
- Various heuristic techniques (e.g. C-value/NC-value, Frantzi et al. 2000; Rogers 2017)
The “slot” model: *earn* (pattern 1)

**Based on fragment of syntactic dependency graph**

<table>
<thead>
<tr>
<th>Subj</th>
<th>V</th>
<th>Adj</th>
<th>Obj</th>
</tr>
</thead>
<tbody>
<tr>
<td>they, he</td>
<td>earn</td>
<td>—</td>
<td>money</td>
</tr>
<tr>
<td>manager</td>
<td>earn</td>
<td>—</td>
<td>chance</td>
</tr>
<tr>
<td>I</td>
<td>earn</td>
<td>first</td>
<td>salary</td>
</tr>
<tr>
<td>labourer</td>
<td>earn</td>
<td>more</td>
<td>more money</td>
</tr>
<tr>
<td>Jane</td>
<td>earn</td>
<td>much</td>
<td>sympathy</td>
</tr>
<tr>
<td>Doris</td>
<td>earn</td>
<td>—</td>
<td>salary</td>
</tr>
<tr>
<td>you</td>
<td>buy</td>
<td>fresh</td>
<td>food</td>
</tr>
<tr>
<td>they</td>
<td>buy</td>
<td>—</td>
<td>something</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>nationalist</td>
<td>support</td>
</tr>
</tbody>
</table>
The “slot” model: *earn* (pattern 1)

Based on fragment of syntactic dependency graph

Different fixed & open-slot MWC within this frame:

- ⭐ ● – earn – ● – money
- ⭐ worker – earn – ● – ●
- ⭐ ● – earn – good – money
- ⭐ ● – earn – ✗ – keep
- ⭐ company – earn – huge – profit
- ⭐ Pron – earn – A – support
- ⭐ worker – earn – ● – [MONEY]
Reduction to pairwise hypotheses

→ significance tests & association measures

Subj — V — Adj — Obj

- • - • - earn - • - • - adj - • - • - obj - • - • - good - • - • - money

- • - • - earn - • - • - adj - • - • - obj - • - • - good - • - • - money

- • - • - earn - • - • - adj - • - • - obj - • - • - good - money

- • - • - earn - • - • - adj - • - • - obj - • - • - good - money
Reduction to pairwise hypotheses

→ significance tests & association measures

Subj — V — Adj — Obj

• – • – good – •

• – earn – • – •

• – • – good – money

• – • – good – • – money

• – earn – • – money

• – earn – good – money
Reduction to pairwise hypotheses

→ significance tests & association measures

Subj ——— V ——— Adj ——— Obj

worker – ● – ● – ●

● – earn – ● – ●

● – ● – good – ●

● – ● – ● – ● – money

● – earn – good – money

worker – earn – good – money
Reduction to pairwise hypotheses

→ collocational paradigms & type-token distribution

Subj ——► V ——► Adj ——► Obj

worker — earn — good — money

... decent big extra much good more

... salary living wages money

• – earn – good – [MONEY]

worker – earn – good – money

• – earn – [POS] – money
Reduction to pairwise hypotheses

→ collocational paradigms & type-token distribution

- worker - earn - good - money
- earn - A - N
- earn - [POS] - [MONEY]
- earn - good - [MONEY]
- earn - [POS] - money

worker - earn - good - money
Methodological principles

🌟 **Statistical association** is multi-faceted
- frequency (→ familiarity, log-likelihood)
- salience (→ conservative MI)
- predictability (→ conditional probability, ΔP)

🌟 **AM for pairwise hypotheses** (syntagmatic)
- determine whether MWC is part of larger MWC or independent
- structure of complex MWC (V–A–Obj vs. V–Obj + A–Obj)
- challenge: what are appropriate decision criteria?

🌟 **Type-token distributions** (paradigmatic hypotheses)
- cf. Diwersy/Evert/Heinrich/Proisl (yesterday)
- challenge: include distribution of AM scores

🌟 **Semantic patterns** (→ thesaurus or word embeddings)
- also: distinguish semantic preference vs. lexical collocation
Work in progress …

Thank you

This is an ongoing research programme.
Please ask questions!
References


References


