

STS 3 Preparing Candidate Data

Brigitte Krenn

ÖFAI

Vienna

Collocations

Terminology & Definitions

- Firth's Notion of Collocation

``Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words."''

``One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, its collocation with *night*."''

Terminology
(definition of collocations)
versus
Defining characteristics
(description of properties)

Terminology

- idioms, preferably used in the English literature, e.g. Bar-Hillel:55, Hockett:58, Katz;Postal:63, Healey:68, Makkai:72.
- phraseological units, (Ge.: Phraseologismus) is a widely used generic term in the German literature, e.g. BurgerEA:82, Fleischer:82.
- light-verb constructions, support-verb constructions, refer to very particular phenomena, cross-categorisation with idioms

Terminology

- multi-word lexemes, e.g. Tschichold:97, BreidtEA:96.
- multi-word expressions, e.g. Segond; Tapanainen:95
- non-compositional compounds, e.g. Melamed:97
- etc.

Terminology

- influenced by
 - different linguistic traditions
 - computational linguistics: multi-word units/expressions/lexemes
- What are the phenomena?
 - lexically determined word co-occurrences
 - multi-words, multi-units, phrases

Defining Characteristics of Collocations

- Lexical Selection
- Syntactic rigidity
- Word formation processes
- Recurrence
- ? Semantics (idiomaticity)
- ? Pragmatic function

Lexical Selection

Word co-occurrence is determined by lexical rather than by semantic criteria (cf. Firth's notion of collocation)

As a consequence, the lexically selected words cannot be replaced by other semantically and morphosyntactically equivalent ones, cf. "lexical stability" in [Fleischer:82]

Restrictions in Syntactic Generativity

- Collocations range from completely fixed to syntactically flexible constructions.
- Syntactic restrictions usually coincide with semantic restrictions and thus are indicators for the degree of lexicalization of a particular word combination.
- Particular word combinations are associated with specific restrictions that cannot be inferred from standard rules of grammar and thus need to be stored together with the collocation.

Recurrence

- Within corpora, the proportion of collocations is larger among highly recurrent word combination than among infrequent ones.

Idiomatcity

- Idiomatcity is a frequently mentioned characteristic of lexicalizations.
- Idiomatcity usually is defined by **semantic noncompositionality**, i.e., the meaning of an idiomatic word combination is not a function of the semantics of the individual words, but is associated to the word combination as a whole.

Idiomatcity

- Semantic opacity, however, is not sufficient for the definition of collocations as there exists a variety of conventionalized word combinations that range from
 - fully compositional ones like *Hut aufsetzen* ('put on a hat'), *Jacke anziehen* ('put on a jacket')
 - to
 - semantically opaque ones like *{\it ins Gras beißen}* ('bite into the grass' literal meaning, 'die' idiomatic meaning).

Words, Multi-words or Phrases

- Collocations can be
 - word level phenomena (?multi-word unit)
 - phrase level phenomena (collocation phrase)
- Collocation phrases consist of the lexically determined words (collocates) only or contain additional lexically underspecified material.

Word-level Collocations

- Adjective- and Adverb-Like Collocations
 - *nichts desto trotz* ('nonetheless') adverb
 - *fix und fertig* ('exhausted') adjective
- Preposition-Like Collocations
 - *im Lauf(e), im Zuge* ('during')
 - *an Hand* ('with the help of')

Word-level Collocations

- Noun-Like Collocations
 - *Rotes Kreuz* (Red Cross)
 - *Wiener Sängerknaben* (Vienna choir boys)
 - *Hinz und Kunz* ('every Tom, Dick and Harry')
- Sequences where the nouns are duplicated
 - *Schulter an Schulter* (shoulder to shoulder),
 - *Kopf an Kopf* (neck and neck)

Word-level Collocations

- Modal constructions
 - *sich (nicht) lumpen lassen* ('to splash out')
- Verb-object combinations
 - *übers Ohr hauen* ('take somebody for a ride')
 - *unter die Lupe nehmen* ('take a close look at')
 - *zum Vorschein bringen* ('bring something to the light')
 - *des Weges kommen* ('to approach')
 - *Lügen strafen* ('prove somebody a liar')

Word-level Collocations

- Copula constructions
 - *guten Glaubens sein* ('be in good faith')
 - *auf Draht sein* ('be on the ball')
- Proverbs
 - *Morgenstund hat Gold im Mund* (morning hour has gold in the mouth)
 - *wissen, wo der Barthel den Most holt* (know where the Barthel the cider fetches, 'know every trick in the book')

Summing up,

- Structural dependency

the collocates of a collocation are syntactic dependents, thus knowledge of syntactic structure is a precondition for accurate collocation identification.

- Syntactic context

may help to discriminate literal and collocational readings, see for instance *im Lauf*, *im Zug* where a genitive to the right is a strong indicator for collocational reading.

Summing up,

- **Markedness**

morphologically or syntactically marked constructions like seemingly incomplete syntactic structure or archaic e-suffix are suitable indicators for collocations, see *im Laufe*, *im Zuge* for e-suffix and *zu Recht*, *an Hand* for incomplete syntactic structures.

- **Single-word versus multi-word units**

single-word occurrences of word combinations indicate word-level collocations, see for instance *zu Recht*, *zurecht*.

- **Syntactic rigidity**

is an important indicator for collocations see for instance *Hinz und Kunz*, *an und für sich*, *fix und fertig*, *Kopf an Kopf*.

3 Defining Characteristics of Collocations

- over proportionally high recurrence of collocational word combinations compared to noncollocational word combinations in corpora;
- grammatical restrictions in the collocation phrases;
- lexical determination of the collocates of a collocation.

Collocations as N-grams

- Represent a collocation by its collocates!
- AMs (association measures) are typically bi-gram statistics.
- Numeric versus syntactic span?

Numeric Span

Def:

- The numeric span delimits the lexical context within which collocation partners (collocates) are found.

w_i, w_j are to be found, with $|j - i| + 1 \leq r$

Numeric Span

Serious drawback: Definition of Span Size

- If the **span size** is kept **small**, it is **unlikely to properly cover** nonadjacent collocates of **structurally flexible collocations**.
- **Enlarging the span size** leads to an increase of candidate collocations including an **increase of noisy data** which need to be discarded in a further processing step.

Other weaknesses to be worked around

- Over-proportional frequency of function words within texts
 - use stop word lists
- Insensitivity to punctuation
 - ! use a sentence as the largest unit within which the collocates of a collocation may occur
- Insensitivity to parts-of-speech
 - ! knowing parts-of-speech allows a large number of syntactically invalid n-grams to be excluded beforehand

More Weaknesses

- Insensitivity to syntactic structure
 - ! Further improvement of the appropriateness of the collocation candidates selected is achieved by the availability of structural and/or dependency information.

Proposal

- Step by step/gradual replacement of
 - the notion of **numeric span**by
 - the notion of **syntactic span**.
- What does it imply?
- Do we really want/need it?

Distribution of Words and Word Combinations in Text

- Zipf's law
- $n_c > n_{c+1}$, n_c the number of words occurring c -times
- i.e., with increasing count c the number of words occurring c -times decreases.

Extraction Strategies

- A simple Procedure for PN- and PNV-Extraction
 - extraction of PN-combinations from PPs
 - extraction of main verbs
 - combination of PN-pairs and verbs co-occurring in a sentence
- Result
 - a theoretical maximum of PNV combinations, i.e.,
 - verbs are duplicated in sentences that contain more than one PP,
 - PPs are duplicated in sentences where more than one main verb is found.

Extraction Strategies

- required:
 - PoS-tagging
 - basic phrase chunking
 - infinitives with *zu* (to) are treated like single words,
 - separated verb prefixes are reattached to the verb

Extraction Strategies

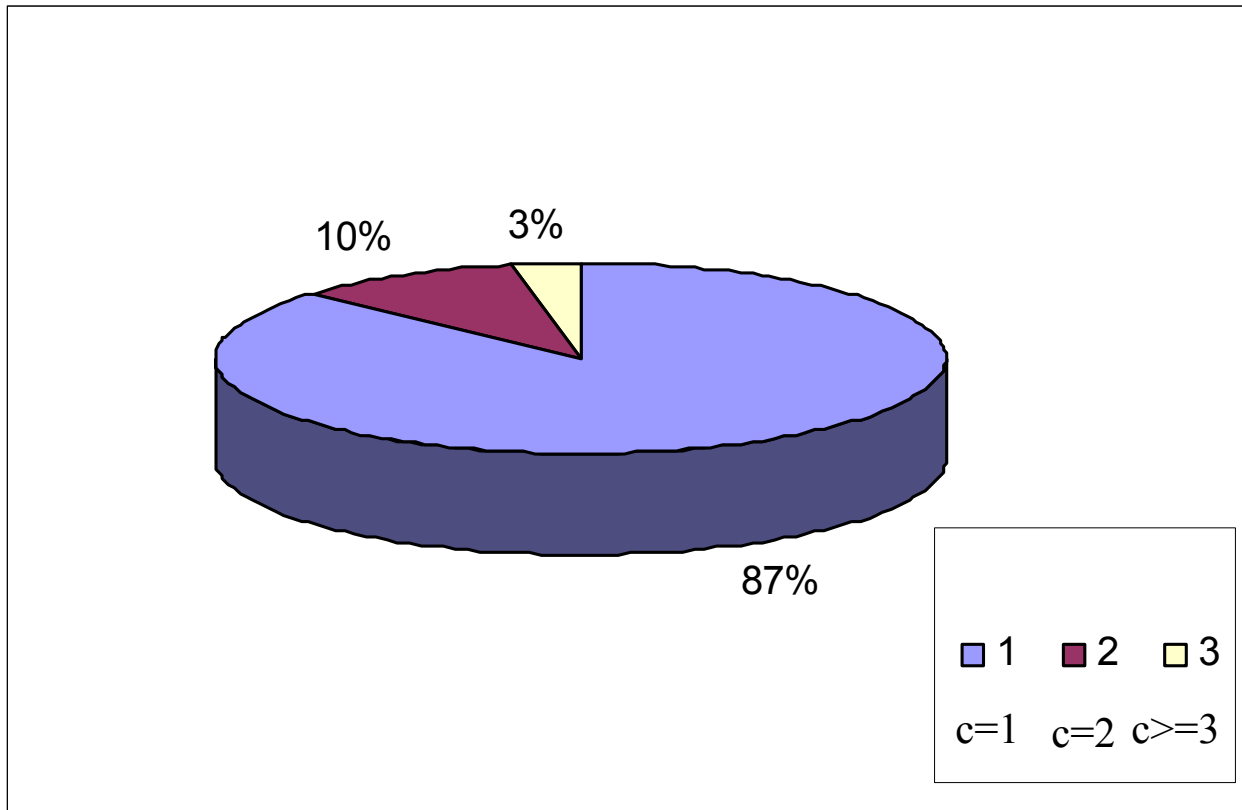
- Full forms or base forms ?
 - depends on language and collocation type
- required:
 - morphological analysis

An Example

- corpus size: 8 million words of the Frankfurter Rundschau corpus
- 569,310 PNV-combinations (types) have been selected from the extraction corpus including main verbs, modals and auxiliaries. (theoretical maximum)
- Considering only combinations with main verbs, the number of PNV-types reduces to 372~212 (full forms).
- multiplication of the types by their ranks results in 454~088 PNV-instances

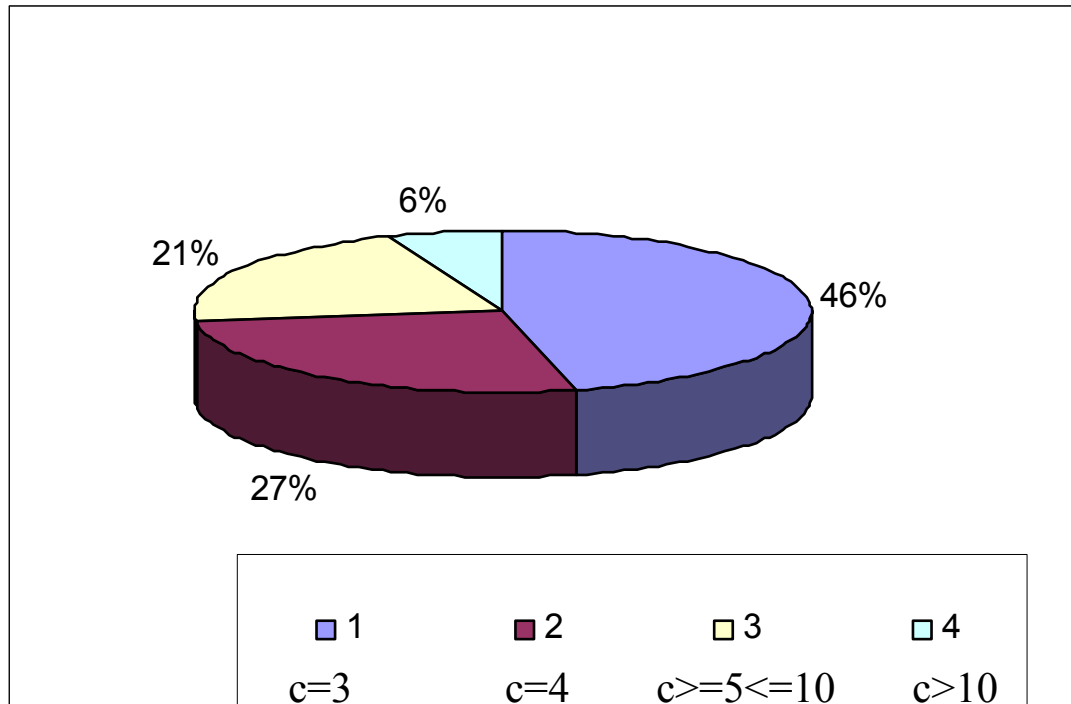
Distribution of PNV types according to rank

Base: 372,212 ranked full form PNV types



Distribution of PNV types according to rank

Base: 10,430 PNV types with $c \geq 3$



3 Extraction Strategies

- Strategy 1: Retrieval of n-grams from word forms only (w_i)
- Strategy 2: Retrieval of n-grams from part-of-speech annotated word forms (wt_i)
- Strategy 3: Retrieval of n-grams from word forms with particular parts-of-speech, at particular positions in syntactic structure ($wt_i c_j$)

Spans tested

$$W_i \ W_{i+1}$$

$$W_i \ W_{i+1} \ W_{i+2}$$

$$W_i \ W_{i+2} \ W_{i+3}$$

$$W_i \ W_{i+3} \ W_{i+4}$$

Results of Strategy 1

- Retrieval of PP-verb collocations from word forms only is clearly inappropriate as function words like articles, prepositions, conjunctions, pronouns, etc. outnumber content words such as nouns, adjectives and verbs.
- Blunt use of stop word lists leads to the loss of collocation-relevant information, as accessibility of prepositions and determiners may be crucial for the distinction of collocational and noncollocational word combinations.

Results of Strategy 1

- most useful/informative span: $w_i w_{i+1} w_{i+2}$
- examples

bis & 17 & Uhr 2222

FRANKFURT & A. & M. 949

in & diesem & Jahr 915

um & 20 & Uhr 855

Di. & bis & Fr 807

10 & bis & 17 779

Tips & und & Termine 597

in & der & Nacht 582

we have learned

- useful/informative span size is language specific
- we find a number of different constructions
- e.g.
 - NP, PP, ...
 - names, time phrases, conventionalized constructions, ...

Results of Strategy 2

wt_i wt_{i+1} with preposition t_i and noun t_{i+1}

- PPs with arbitrary preposition-noun co-occurrences such as
 - *am Samstag* (on Saturday),
 - *am Wochenende* (at the weekend),
 - *für Kinder* (for children)
- Fixed/conventionalized? PPs such as
 - *zum Beispiel* (for example)

Results of Strategy 2

wt_i wt_{i+1} with preposition t_i and noun t_{i+1}

- PPs with a strong tendency for particular continuation such as
 - *nach Angaben* + NP_{gen} ('according to'),
 - *im Jahr* + Card (in the year).
- Potential PP-collocates of verb-object collocations such as
 - *zur Verfügung* (at the disposal)

Results of Strategy 2

$wt_i wt_{i+2}$ with preposition t_i and noun t_{i+1}

- typically cover PPs with pre-nominal modification

Cardinal, for instance, is the most probable modifier category co-occurring with

bis ... Uhr (until o'clock)

- Adjective is the predominant modifier category related to

im ... Jahr (1272 of 1276 cases total),

vergangenen (Adj, last, 466 instances)

Results of Strategy 2

$wt_i wt_{i+3}$ with preposition t_i and noun t_{i+1}

- typically exceeds phrase boundaries

im Jahres ($in_{\text{dat}} \text{year}_{\text{gen}}$), for instance, originates from PP NP_{gen}

e.g. *im September dieses Jahres* (in the September of this year)

Results of Strategy 2

$wt_i \quad wt_{i+1} \quad wt_{i+2}$

with preposition t_i and noun t_{i+1} and verb t_{i+2}

- Frequent **preposition-noun-participle** or **-infinitive** sequences are good indicators for PP-verb collocations, especially for **collocations that function as predicates** such as support-verb constructions and a number of figurative expressions.
 - zur Verfügung gestellt (made available)
 - in Frage gestellt (questioned)
 - in Verbindung setzen (to contact)

Results of Strategy 2

$wt_i \quad wt_{i+2} \quad wt_{i+3}$

$wt_i \quad wt_{i+3} \quad wt_{i+4}$

with preposition t_i and noun t_{i+2} and verb t_{i+3}

with preposition t_i and noun t_{i+3} and verb t_{i+4}

- a variety of PPs with prenominal modification are covered
- but also phrase boundaries are more likely to be exceeded
 - durch Frauen helfen \rightarrow durch X (Y) Frauen helfen

Results of Strategy 3

$wt_i c_k$ $wt_j c_k$ $wt_l c_m$

PP-Collocate	V-Collocate	Right Neighbour	Co-occurring Main Verb
zur Verfügung	stehen	189	404
zur Verfügung	stellen	240	457
in Kraft	treten	99	126
in Kraft	setzen	12	23
in Kraft	bleiben	0	5

Conclusion

- There is **no single best strategy** to extract an optimal set of candidate data from a corpus.
- You need to **know** a least some **structural** and **distributional properties** of the phenomena you are searching for.
- Preparation of candidate data influences distributions.
- Distributional properties determine the outcome of AMs.
- **Know** the distributional assumptions underlying **the AMs you use**.