

# **Collocations and How to Find Them**

Tanja Gaustad

22 May 2002

## Examples of Collocations

- He **kicked** the **bucket** yesterday
- I **heard** it **through** the **grapevine**
- She is **under** a lot of **pressure**
- They **made** it **up** to him
- All of this happened in **broad daylight** ( $\neq$  bright daylight)
- I like **strong tea** ( $\neq$  powerful tea)

## Examples of Collocations

- Onze buurman heeft **de pijp** aan Maarten gegeven
- Dat **springt wel in het oog**
- Zij heeft zonder commentaar **het veld** geruimt
- Hij **gaat problemen** altijd **uit de weg**
- De regering **neemt een belangrijk besluit** ( $\neq$  maken)
- Ik **neem een photo** van de koningin (?maken)

# Collocations

- Lexicalised phrases: partially idiosyncratic syntax or semantics
  - fixed expressions: fully lexicalised (*by and large, in de gaten houden*)
  - semi-fixed expressions: constraint on word order and composition (*kick the social bucket, een goed/slecht figuur slaan*)
  - syntactically flexible expressions: e.g. support verb constructions (*make it up to someone, met iets in je maag zitten*)
- Institutionalised phrases: syntactically and semantically compositional, but occurring with markedly high frequency, “statistically idiosyncratic” (*strong tea, sterke thee*)

# Problems

- No clear cut definition
- Where should the boundary be drawn between institutionalised phrases and other frequent word combinations?

## Linguistic Tests to Identify Collocations (Sailer 2000)

- 4 Semantic criteria
- 6 Syntactic criteria

N.B.: These linguistic tests are only an indication. An idiomatic expression can fulfill some of the requirements stated in these tests and still be an idiom!

Examples used for illustration:

*kick the bucket* ('die') and *spill the beans* ('reveal a secret')

## Linguistic Tests to Identify Collocations: Semantic Criteria

1. Every element of the collocation can be given a meaning with which it is also found outside that collocation
  - *kick the bucket* means 'die'. Even though all words inside the collocation can appear independently, none of them does so in any meaning that could be considered part of the collocation.
  - In principle, the idiomatic meaning could be split into its component parts (i.e. *spill* means 'reveal' and *the beans* means 'the secret'). Event though *spill* is sometimes used with the 'reveal' meaning in slang, *the beans* do not occur with the idiomatic meaning outside this collocation.

## Linguistic Tests to Identify Collocations: Semantic Criteria

2. The meaning can be found in a compositional way (by combining the meaning of its parts)
  - Since the meaning 'die' is assigned to the idiom *kick the bucket* as a unit, it follows that the overall meaning of the idiom cannot be computed by regular means from that of its components.
  - Once we know that *spill the beans* means 'reveal a secret' we could split the idiomatic meaning of the idiom into its component words. This idiom meets this 2nd semantic criterion.

# Linguistic Tests to Identify Collocations: Semantic Criteria

3. Parts of the collocation can be semantically modified
  - Pat kicked the **proverbial** bucket
  - Pat spilled the **well-guarded** beans
  
4. If the collocation contains a NP, a pronoun can refer to the NP
  - \* Pat kicked the bucket and Harry kicked **it** too
  - I was worried that the beans might be spilled, but **they** weren't

# Linguistic Tests to Identify Collocations: Syntactic Criteria

1. Every element in the collocation occurs in the same form in some other combination
  - Peter *kicked* a ball; Mary read *the* book; John carried the heavy *bucket*
  - Pat *spilled* some water; Give me *the* book; Harry doesn't like *beans*
2. The collocation is syntactically regularly built
  - Both *kick the bucket* and *spill the beans* have a regular syntactic structure.

## Linguistic Tests to Identify Collocations: Syntactic Criteria

3. If it is a VP-XP combination, the fixed constituent XP can be modified syntactically
  - Pat kicked the **proverbial** bucket
  - Pat spilled the **well-guarded** beans
4. If it is a VP-XP combination, it can be passivised
  - \* The bucket was kicked by Pat
  - The beans were spilled in this article

## Linguistic Tests to Identify Collocations: Syntactic Criteria

5. If it is a VP-XP combination, the fixed constituent XP can be topicalised
  - \* The social bucket, Pat really kicked with his dumb remark at he party last night
  - \* The beans John spilled
6. If it is a VP-XP combination, the fixed constituent XP can be a relative pronoun
  - \* The old lady kicked the bucket **that** the murderer had planned for her
  - \* The beans **that** the alleged arms dealer spilled made the party leader resign

# Importance

- Natural language generation
- Computational lexicography
- Parsing
- Information retrieval
- Corpus linguistics
- Sociolinguistics

# Statistical Corpus-Based Approaches

- Frequency
- Hypothesis testing:  $\chi^2$ , log-likelihood
- Mutual information
- Phrasal entropy

## Statistical Corpus-Based Approaches

- All tests applied to **collocation candidates** extracted from the POS-tagged Eindhoven corpus (250,000 occurrences)
- Only patterns occurring at least 10 times were considered (frequency cutoff)
- Extracted constructions: **P NP P** (*onder leiding van, in plaats van*)
- N.B.: Most formulas apply to *bigrams*, but the extracted constructions contain *three* words  
Solution: Treat two of the three words of as a unit and test both “bigram” combinations

# Frequency

- Simplest method
- Counting co-occurrences
- Works well for fixed phrases
- To filter some undesired output:
  - Use POS Filter to filter out function words (*of the, is a*)
  - Discard all NPs containing proper names or numbers

## Top 10 Results for Frequency

| rank | collocation          | frequency |
|------|----------------------|-----------|
| 1    | in plaats van        | 1253      |
| 2    | op basis van         | 816       |
| 3    | onder leiding van    | 710       |
| 4    | op het gebied van    | 659       |
| 5    | aan het eind van     | 609       |
| 6    | ten opzichte van     | 579       |
| 7    | in tegenstelling tot | 549       |
| 8    | op grond van         | 541       |
| 9    | na afloop van        | 520       |
| 10   | aan de hand van      | 511       |

## $\chi^2$

- Compares observed frequencies with frequencies expected for independence and is applied to tables

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where  $i$  ranges over rows of the table,  $j$  ranges over columns,  $O_{ij}$  is the observed value for cell  $(i,j)$  and  $E_{ij}$  is the expected value.

- Look up critical value at given probability level (e.g.  $\alpha = 0.5$ )
- If  $\chi^2$  value  $<$  critical value, the two words tested do not occur independently of each other and are thus a good candidate for a true collocation

## Top 10 Results for $\chi^2$

| rank | collocation         | rank 1 | rank 2 |
|------|---------------------|--------|--------|
| 1    | onder leiding van   | 10     | 76     |
| 2    | ten opzichte van    | 3      | 91     |
| 3    | op basis van        | 63     | 70     |
| 4    | op weg naar         | 113    | 22     |
| 5    | aan het eind van    | 56     | 83     |
| 6    | naar aanleiding van | 40     | 103    |
| 7    | in plaats van       | 106    | 52     |
| 8    | op het gebied van   | 92     | 82     |
| 9    | ten koste van       | 6      | 175    |
| 10   | op grond van        | 93     | 92     |

# Log-likelihood

- Compares probability of two hypotheses:  $H_1$ , which assumes that the words in the bigram are independent, and  $H_2$ , which assumes that they are dependent

Hypothesis 1 (independence) :  $P(w_2|w_1) = P(w_2|\neg w_1)$

Hypothesis 2 (dependence):  $P(w_2|w_1) \neq P(w_2|\neg w_1)$

Log-likelihood ratio  $\lambda$ :

$$\lambda = -\log \frac{L(H_1)}{L(H_2)}$$

- The higher  $\lambda$ , the more likely it is that the two words are dependent.

## Top 10 Results for Log-likelihood

| rank | collocation          | rank 1 | rank 2 |
|------|----------------------|--------|--------|
| 1    | in plaats van        | 2      | 1      |
| 2    | onder leiding van    | 1      | 4      |
| 3    | op basis van         | 4      | 3      |
| 4    | ten opzichte van     | 3      | 8      |
| 5    | op het gebied van    | 7      | 6      |
| 6    | aan het eind van     | 6      | 7      |
| 7    | in tegenstelling tot | 12     | 2      |
| 8    | op weg naar          | 14     | 5      |
| 9    | op grond van         | 11     | 9      |
| 10   | naar aanleiding van  | 9      | 12     |

## Mutual Information

- Tells us about the amount of information a random variable contains about another, in other words: it is a measure of the common information in two variables, a measure of independence between variables

$$MI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

- If  $MI = 0$ , the words are independent  
If  $MI > 0$ , the words might be a good collocation candidate

## Top 10 Results for Mutual Information

| rank | collocation                            | rank 1 | rank 2 |
|------|--|--------|--------|
| 1    | vanaf hun aankomst tot                 | 178    | 534    |
| 2    | ten strijde tegen                      | 300    | 830    |
| 3    | per persoon per                        | 359    | 869    |
| 4    | over de artistieke uitwisseling tussen | 999    | 251    |
| 5    | wegens betrokkenheid bij               | 601    | 717    |
| 6    | tussen de bedrijven door               | 469    | 863    |
| 7    | via een onderaards geweld met          | 189    | 1165   |
| 8    | wegens gebrek aan                      | 594    | 812    |
| 9    | na de nederlaag tegen                  | 876    | 578    |
| 10   | uit protest tegen                      | 1068   | 409    |

# Entropy

- Entropy reflects the amount of information we have about a random variable, its self-information

$$H(p) = H(x) = - \sum_{x \in X} p(x) \log_2 p(x)$$

where  $p(x)$  is the probability of a random variable  $x \in X$  and  $X$  is the alphabet

## Phrasal Entropy

- Phrasal Entropy calculates the entropy observed in phrases that may be collocates inside a larger phrase.

$$\text{PE}(P, N) = - \sum_{i=1, j=1}^{m, k} \frac{f(P P_{instance_i P N_j})}{f(P N_j)} \log \frac{f(P P_{instance_i P N_j})}{f(P N_j)}$$

where  $f(P P_{instance_i}) = m$  and  $m$  corresponds to the number of occurrences of  $P P_{instance_i}$  in the extraction corpus, and  $f(P N_j) = k$  and  $k$  is the total number of tuples ( $P N_j$ ) in the extraction corpus.

- The lower the phrasal entropy, the more rigid/fixe the combination of  $P$  and  $N$

## Phrasal Entropy Example

| PE    | tuple ( $PN_j$ )                  |
|-------|-----------------------------------|
| 0.077 | in tegenstelling                  |
| <hr/> |                                   |
| $m$   | $P_{instance_j}$                  |
| 549   | in tegenstelling tot              |
| 2     | in scherpe tegenstelling tot      |
| 1     | in schrille tegenstelling tot     |
| 1     | in tegenstelling met              |
| 1     | in opmerkelijke tegenstelling tot |
| 1     | in de tegenstelling tussen        |

## Top 10 Results for Phrasal Entropy

| rank | collocations         | PE | most frequent instance of tuple |
|------|----------------------|----|---------------------------------|
| 1    | van tal              | 0  | van tal van (12)                |
| 2    | door vertrek         | 0  | door het vertrek van (22)       |
| 3    | met strijkkwartetten | 0  | met strijkkwartetten (25)       |
| 4    | over aard            | 0  | over de aard van (21)           |
| 5    | ter gelegenheid      | 0  | ter gelegenheid van (221)       |
| 6    | na verloop           | 0  | na verloop van (121)            |
| 7    | tot kern             | 0  | tot de kern van (16)            |
| 8    | uit koker            | 0  | uit de koker van (19)           |
| 9    | uit winterpaleis     | 0  | uit het winterpaleis (12)       |
| 10   | in vlaag             | 0  | in een vlaag van (13)           |

## Comparison of Statistical Tests

| test     | n       | nbest |     | all |
|----------|---------|-------|-----|-----|
|          |         | 100   | 300 |     |
| raw freq | 248,683 | 50    | 65  | 84  |
| $\chi^2$ | 2,084   | 52    | 69  | 77  |
| MI       | 2,084   | 23    | 39  | 77  |
| LL       | 2,084   | 53    | 67  | 77  |
| PE       | 3,363   | 10    | 35  | 66  |

- Results of applying statistical tests is compared with initially extracted list of collocation candidates
- 100 and 300 best items found were compared to the “gold standard” list from Van Dale (88 occurrences)
- The last row shows that some collocations occur less than 10 times of not at all

## Strengths and Weaknesses of Statistical Tests

- Mutual information tends to assign very high scores to low frequency data (leads to bad performance)
- Log-likelihood and  $\chi^2$  perform equally well, also with low frequency data.
- Phrasal entropy is very sensitive to sparse data problem
- Raw frequency works surprisingly well, but contains a lot of noise. Statistical tests do not work a lot better, but return more accurate lists of collocations

# Evaluation

- No standard list of true collocations
- Human judgment
  - Advantage: gradually increase list of true collocations
  - Problem: Requirements not clear enough, personal variation in judgment (10% agreement)
- Circular problem: If we would know how to extract true collocations, we would have a “gold standard” list