

# Foundations of co-occurrences and their applications

Text Mining for Classical Studies  
Leipzig, 2009/10/01

Marco Büchler  
Natural Language Processing Group  
Department of Computer Science  
University of Leipzig

---

- Motivation / Definition
- Foundations
  - Graph
  - Visualisation
  - Small worlds
  - Significances
- Applications
  - Categorised co-occurrences
  - Chinese Whispers (based on PHD of Dr. Chris Biemann)
  - Semantic similarity (based on PHD of Dr. Stefan Bordag)

Co-occurrences: Motivation

- Definition:
  - Common occurrence of at least two objects/events within a dedicated window
  - Possible windows in Classical Studies: line, sentence, paragraph, document, author, century
- Motivation:
  - Psycholinguistic experiments: Given a word: What is the first word test persons answer?

Stimulus	Antwort VP	Anz. VPs	Kollokation	Signifikanz
Butter	Brot	60	Brot	51
	weich	40	Käse	49
	Milch	32	Zucker	29
	Margarine	27	Milch	23
	Käse	20	Margarine	22
	Fett(e)	16	Mehl	18
	gelb	14	Eier	16
	Butterbrot	8	Pfund	14
	Dose	6	zerlassener	13
	essen	6	Fleisch	13

- Example sentence:

Die böse Katze jagt die kleine Maus.

- Selection of the best window for the task:
  - **Sentence co-occurrences:** <Die, böse>, <Die, Katze>, <Die, jagt> ....  
<jagt, Die>, ... <jagt, Maus>
  - **Left sentence co-occurrences:** <böse, Die>, <Katze, Die>, <Katze, böse>, <jagt, Die> ...
  - **Right sentence co-occurrences:** <Die, böse>, <Die, Katze>, <Die, jagt>, ...
  - **Fixed window size of e. g. 3:** <Die, böse>, <böse, Die>, <böse, Katze>, <Katze, böse>
- Further options (Normalisation):
  - Baseform reduction
  - Capitalisation
  - Diacritics
  - ...

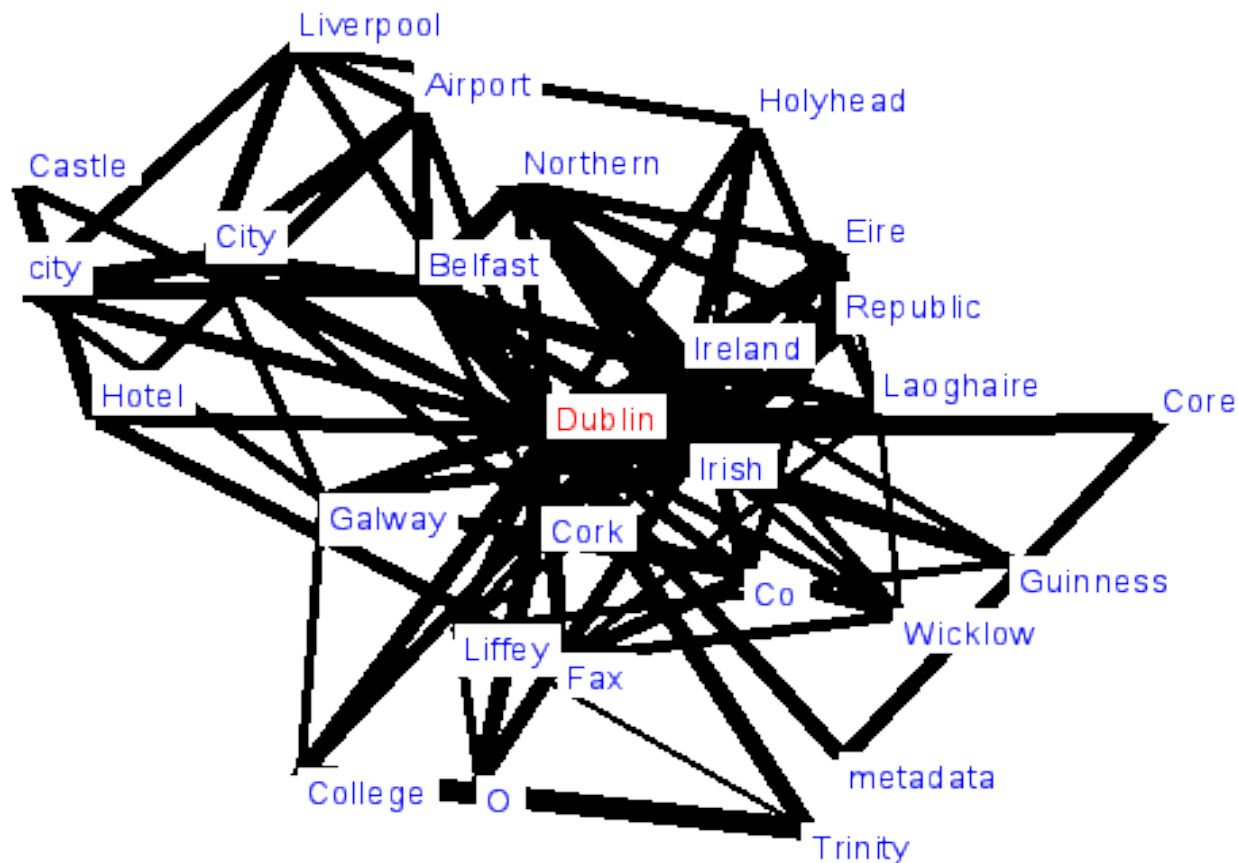
Co-occurrences: Foundations

- Definition (Wikipedia):
  - pairwise relations between objects from a certain collection
  - a collection of vertices or 'nodes' and a collection of edges that connect pairs of vertices
  - Formal: Graph  $G=(V,E)$   $V$ =collection of vertices,  $E$ =collection of edges
- Example sentence (again):

Die böse Katze jagt die kleine Maus.

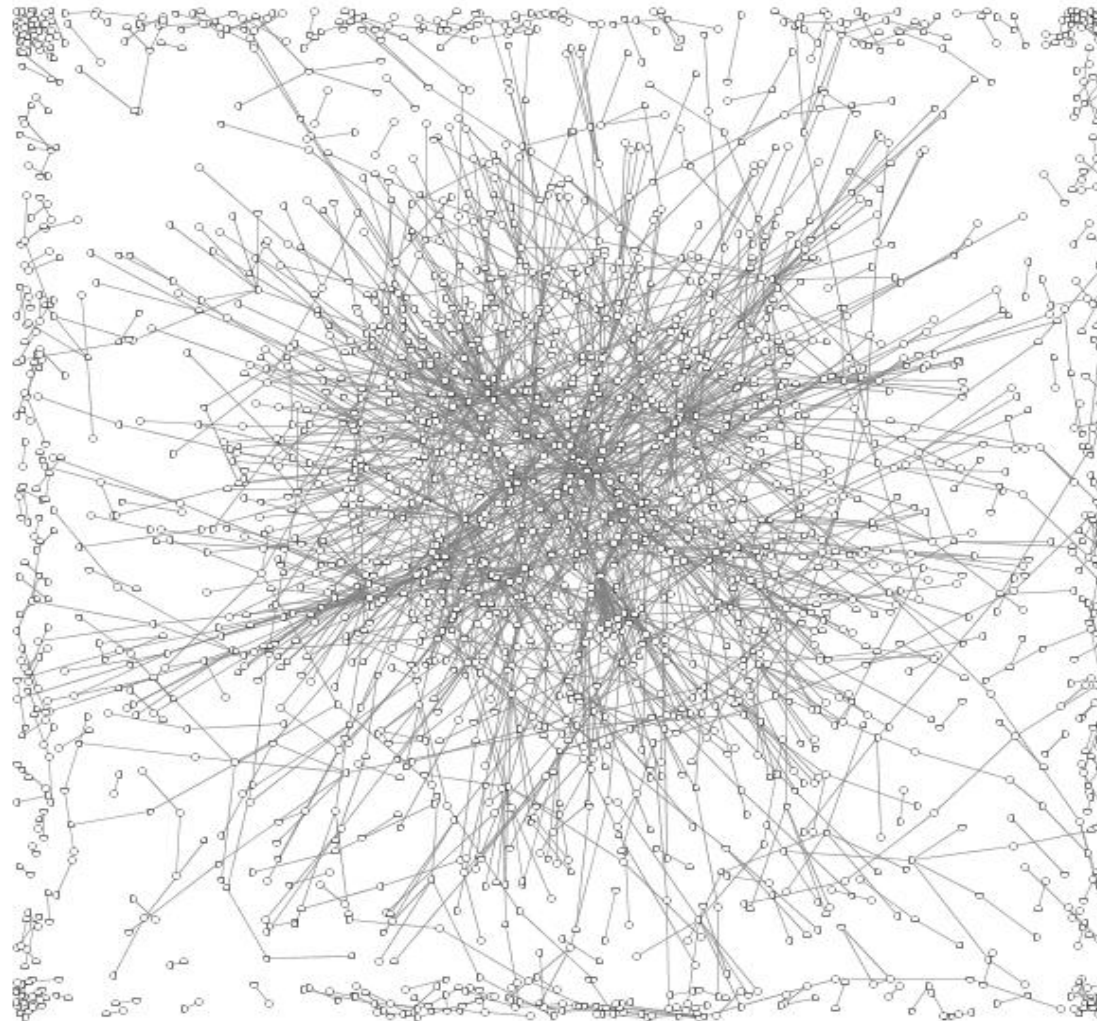
- $V = \{\text{böse, Die, die, jagt, Katze, kleine, Maus}\}$
- $E_{\text{sent-cooc}} = \{(\text{Die, böse}), (\text{Die, Katze}), \dots, (\text{Maus, kleine})\}$
- Co-occurrences are an untyped graph.

Graph v. 1.5 für Dublin



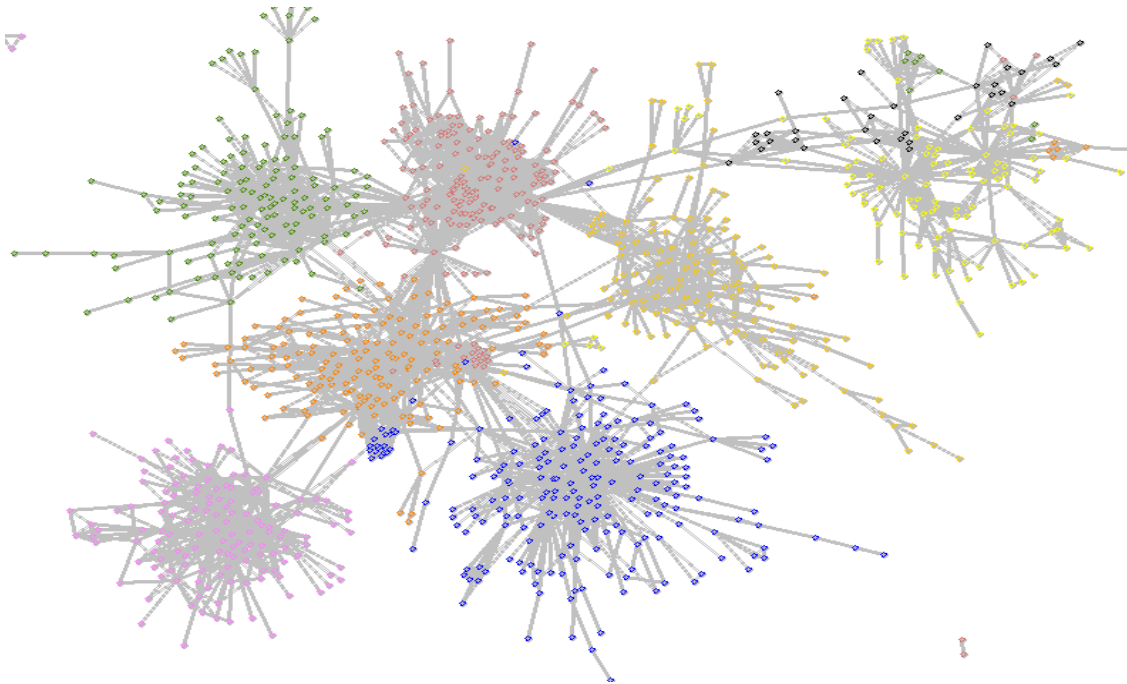
- Will be replaced by **prefuse** library
  - Better visualisation
  - Gravity between all words are used for alignment

graph:  
nodes: 2177  
edges: 2066  
iteration: 0  
option: top



ASV-TOOLBOX 2005

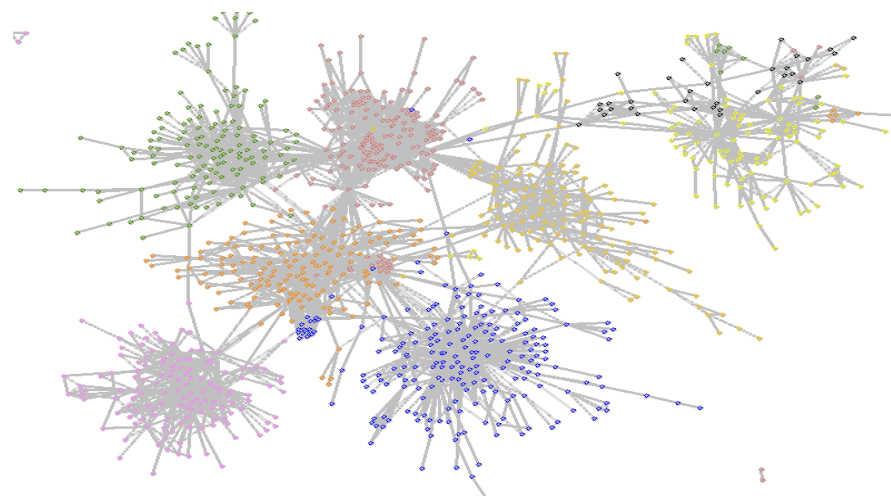
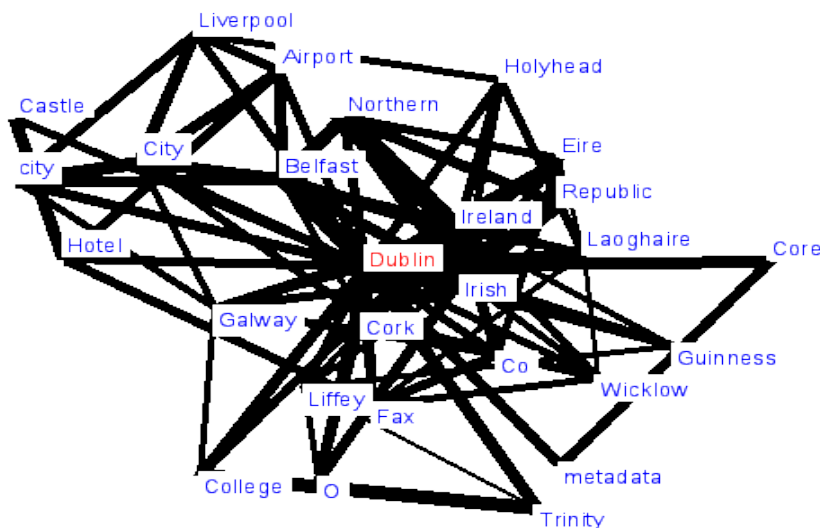
*(Source: chat texts)*



- Expected clusters:
  - Centuries
  - Dialects
  - Literary classifications
  - Epithets
  - Geographical regions (cities, landscapes)
  - Maybe authors
  - Important eAQUA works: bachelor thesis of Marcus Puchalla, AG NER

- “Definition/Motivation”:
  - What's the average path length in a graph?

Graph v. 1.5 für Dublin



- Average path length is typically not larger than 7.
- Simple proof of concept (Using XING):
  - Every person of my contacts has in average about 73 contacts (1. and 2. level)
  - $\log_{73}(6,800,000,000) = 5,28$



- **Relevance to subproject 4.1 - Atthidographers (later)**
- **Relevance to argumentation trails (TMS 2009, March 2009)**

- Theoretical motivation 1 (Based on Aristoteles' law of association) ([Zimb92], [MH02]):
  - Similarity
  - Contrast/Dissimilarity
  - **Contiguity (human learning)**
- Contiguity:
  - An association of at least two objects/events will be established if they occur closely together in location and time.
  - Well-known example: Pawlov experiments
- Theoretical motivation 2 (Based on Harris 1953):
  - The Distributional Hypothesis in Linguistics is that words that occur in the same contexts tend to have similar meanings.

- Realisation:
  - Compute common observation frequency  $\mathbf{O}$  of two words
  - Closely together: Realised by choosing relevant window size
- Motivation:
  - Observing two concepts closely together is a working method of Classical Studies. Right?
  - But are they relevant?
- Critics:
  - Not every pair of objects/words occurring frequently in a common window have a significant relevance

- Example 1:
  - Object<sub>1</sub>: **traffic sign** Object<sub>2</sub>: **building**
  - **O** will be very large (look out of the window)
  - Relevance: non relevant – low relevant
- Example 2:
  - Event<sub>1</sub>: **move hand on hot plate** Event<sub>2</sub>: **hot plate is switched on.**
  - **O** will be very small (Don't try it!)
  - Relevance: highly relevant
- Result:
  - Observing a large common frequency **O** of two concepts can't measure relevance

- Example 3:
  - Experiments with 2 dices
  - $6 \times 6 = 36$  possible events
  - e. g.  $P(1,2) = 1/36$
  - Assumption:  $n = 3600$  trials
  - Expectation **E** of dicing a 1 and a 2 is about 100
  - Question: Is a common occurrence of e. g. 100 relevant?
  - Answer: No, it's what you would expect. It's random.
- Solution:
  - Compare observation **O** with expectation **E**!
  - Using **O-E**: Critical since a large O and E tend to have a larger difference. Example:  $1000 - 100 = 900$  vs.  $100 - 10 = 90$ . Failed.
  - Using **O/E**: Example:  $1000/100 = 10$  vs.  $100/10 = 10$ . Good.
  - Using **log(O/E)** brings further benefits:
    - Small differences will be scaled down.
    - Positive values mean observation is **larger** than expectation.
    - Negative values mean observation is **smaller** than expectation.

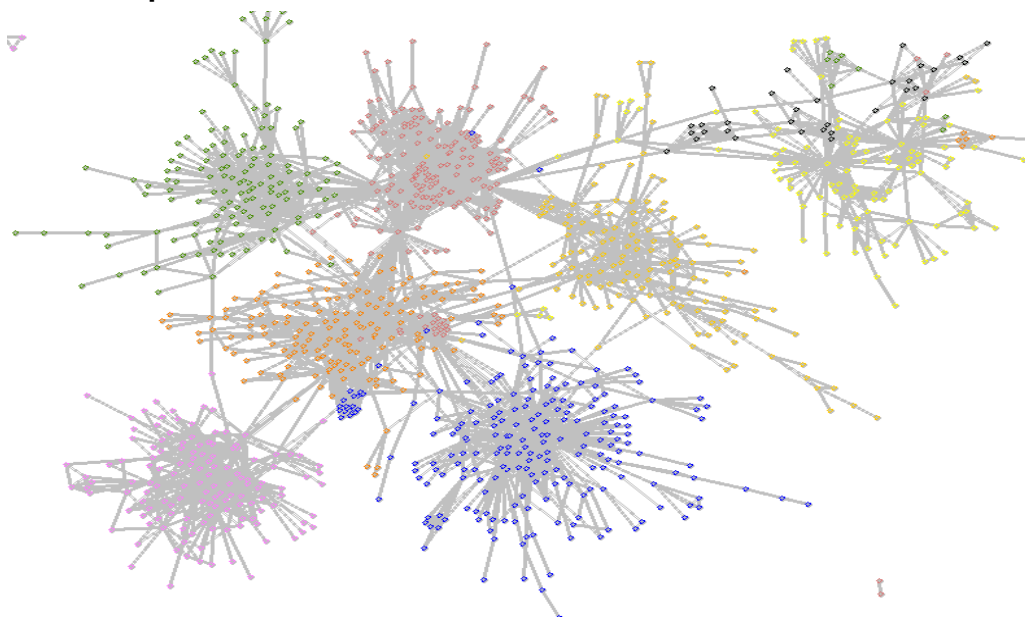
$$\begin{array}{ccc} & sig = \log\left(\frac{O}{E}\right) & \\ & \swarrow \quad \searrow & \\ sig_{MI} = \log\left(\frac{P(W_1, W_2)}{P(W_1) * P(W_2)}\right) & & sig_{LGL} \approx \log\left(\frac{H_A}{H_0}\right) \end{array}$$

- Mutual Information (Church 89):
  - Derivation from stochastic independence
  - Derivation also possible from information theory
- Log-likelihood (Dunning 93):
  - Approximating probabilities by binomial distributions

Co-occurrences: Applications

- Computing labeled significant terms
  - Literary classification specific terms,
  - Century specific terms,
  - Geographic specific terms,
  - Gender specific terms,
  - Author specific terms,
  - Epithetic specific terms and
  - Work specific terms.
- Algorithm:
  - One word will be replaced by a text's category
  - Computing co-occurrences between a category and words of a sentence
- Control questions:
  - What are the differences between TF\*IDF, difference analysis and categorised co-occurrences?

- Graph based cluster algorithm:
  - Step 1: Initialize nodes randomly with any colour
  - Step 2: Inherit colour from predominant node
  - Step 3: Iterate this process several times



- Words within the same cluster have the same colour
- Usage in eAQUA:
  - Completion significant terms of typed graphs by replacing step 1 by own colour set (significant term classes: cities, centuries, dialects, etc.)

- Basic assumption:
  - Two words being semantic similar have the same co-occurrences
- Example:
  - computer: {mouse, battery, display, portable}
  - laptop: {mouse, keyboard, display, mainframe}
- Algorithm (simplification):
  - Compute co-occurrences for all words
  - Compare co-occurrence profiles of two words
  - Compute intersection **I** of co-occurrence profiles of both words
  - Compute union **U** of co-occurrence profiles of both words
  - Compute ratio  $\text{sim} = \text{card}(\mathbf{I}) / \text{card}(\mathbf{U})$
- Example (continued):
  - $\mathbf{I} = \{\text{mouse, display}\}$
  - $\mathbf{U} = \{\text{mouse, battery, display, portable, keyboard, mainframe}\}$
  - $\text{sim} = \text{card}(\mathbf{I}) / \text{card}(\mathbf{U}) = 2/6 = 1/3$

- A little bit more complex:
  - Removing  $\text{card}(\mathbf{I})/\text{card}(\mathbf{U})$  by measuring the cosine value (angle) of two word's co-occurrence vectors

Co-occurrences: Summary

- Co-occurrences are formally an untyped graph.
- Nodes of a co-occurrence graph can be typed by classes (centuries, cities, authors, literary classifications) of significant terms or NER types.
- There exists different kinds of co-occurrences depending on window size.
- It's not enough to investigate the observed frequency.
- Words in a co-occurrence graph are clustered.
- The average path length from one word to another is not large than 7 (small world property).
- Similar used words have similar co-occurrence profiles (semantic similarity).
- Depending on the application different kinds of normalisations are required.