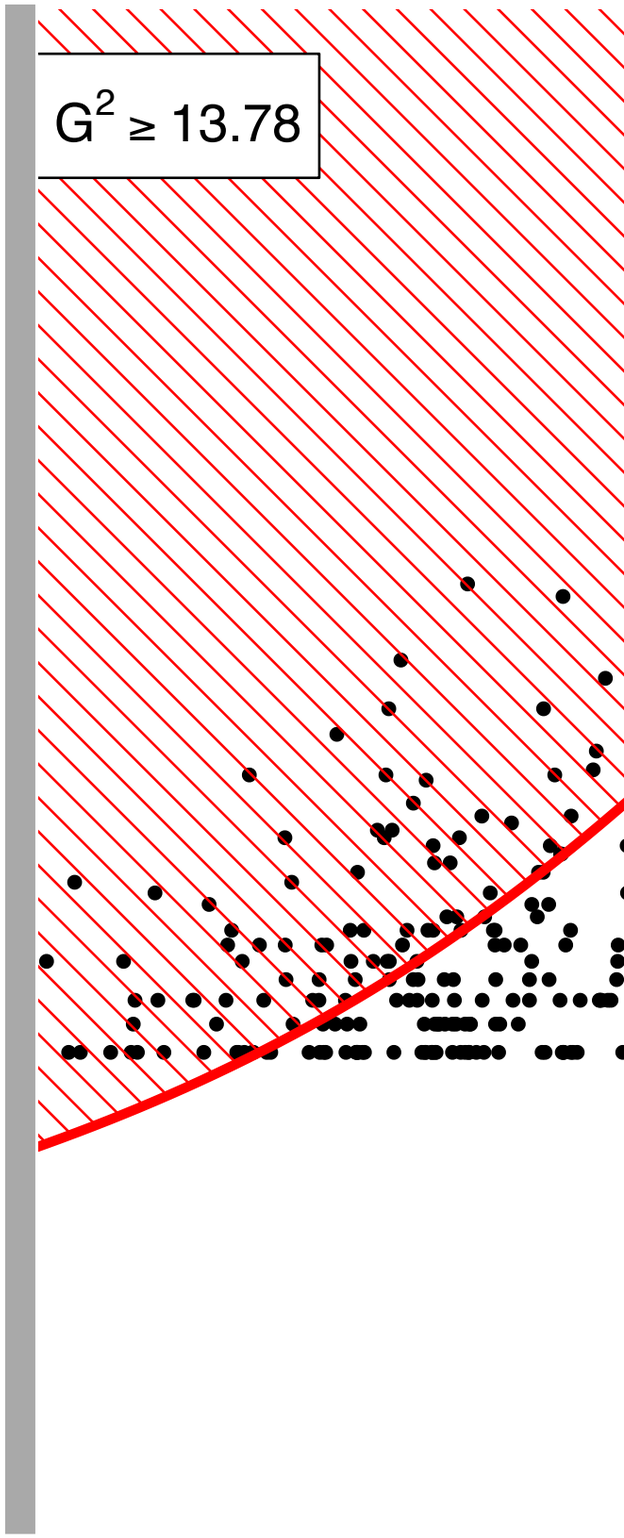
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Stiftung Universität Hildesheim
2003



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

Stefan Evert[1], Ulrich Heid[2]

[1]FAU Erlangen-Nürnberg
[2]Universität Hildesheim

EUROPHRAS 2019

Introduction: non-idiomatic MWE

- 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	8260
heftig	1069
scharf	1006
harsch	417
massiv	389
öffentlich	357
hart	283

VERB (+ Prep) + *Kritik*	
üben	455
stoßen auf	223
es gibt	108
reagieren auf	78
äußern	70

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, …} Kritik*
    + *Kritik {üben, stoßen auf, …}*

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

Syntactic co-occurrence

A simple example: adjectival noun modification (prenominal adjectives)

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(young, gentleman) = ?

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 [. . .]

→

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

f(young, gentleman) = ?

	•|gent.	•|¬gent	
young|•	O_{11}	O_{12}	R_1
¬young|•	O_{21}	O_{22}	R_2
	C_1	C_2	N

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 [. . .]

→

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

f(young, gentleman) = 1
sample size N = 9

	•|gent.	•|¬gent	
young|•	1	2	3
¬young|•	2	4	6
	3	6	9

Contingency tables & association measures

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

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

observed

	w_2	$\neg w_2$
w_1	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$\neg w_1$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

expected

Statistical association measures (AM)

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

$$\text{MI} = \log_2 \frac{O}{E}$$

$$\text{MI}^k = \log_2 \frac{O^k}{E}$$

$$\text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}}$$

$$\text{t-score} = \frac{O - E}{\sqrt{O}}$$

$$\text{simple-ll} = 2 \left(O \cdot \log \frac{O}{E} - (O - E) \right)$$

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{chi-squared}_{\text{corr}} = \frac{N \left(|O_{11} O_{22} - O_{12} O_{21}| - N/2 \right)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

$$\text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1}$$

$$\text{odds-ratio} = \log \frac{\left(O_{11} + \frac{1}{2}\right)\left(O_{22} + \frac{1}{2}\right)}{\left(O_{12} + \frac{1}{2}\right)\left(O_{21} + \frac{1}{2}\right)}$$

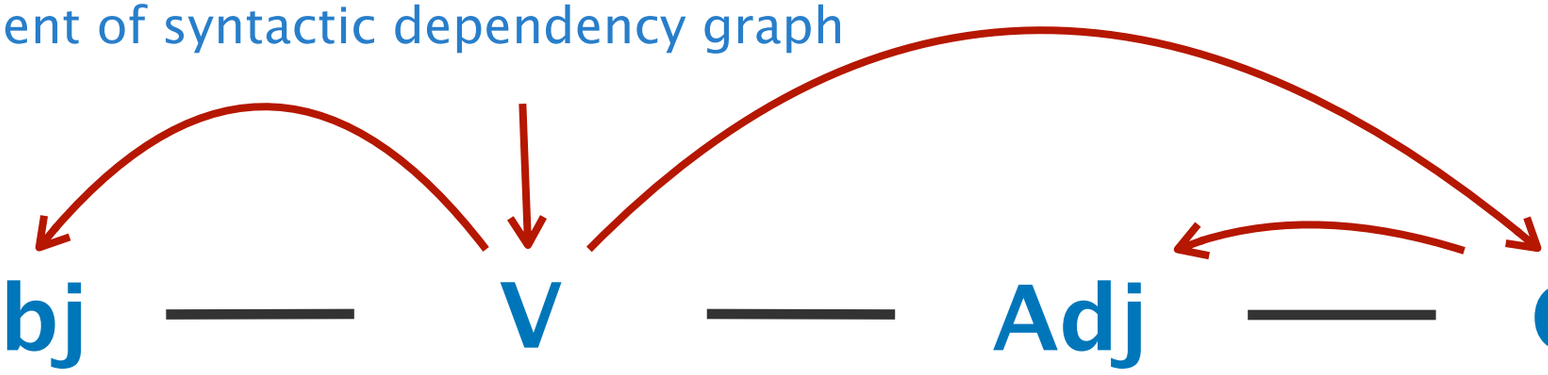
$$\Delta P_{2|1} = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

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)

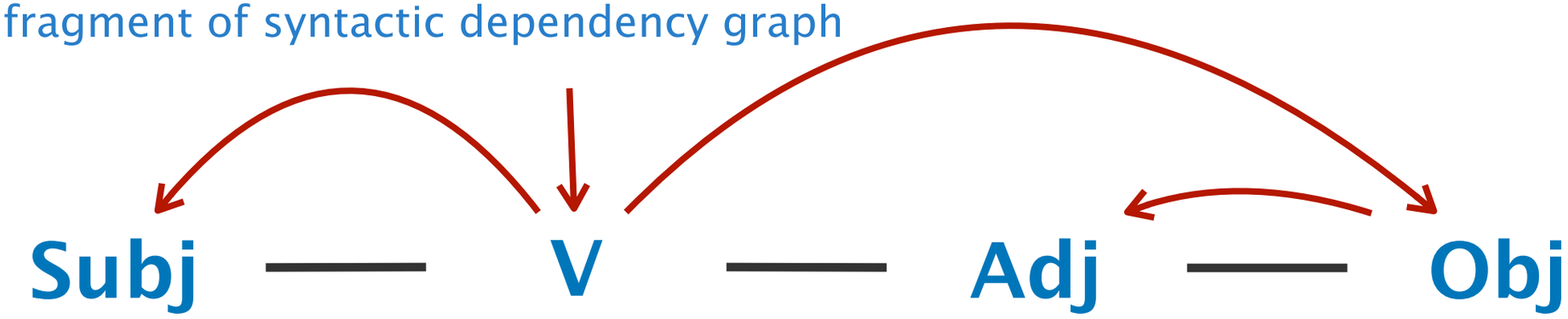
based on fragment of syntactic dependency graph



Subj	V	Adj	Obj
they	earn	—	money
he	earn	—	chance
manager	earn	—	—
I	earn	first	salary
labourer	earn	—	more
—	earn	more	money
Jane	earn	much	sympathy
—	earn	—	salary
Doris	buy	fresh	food
—	buy	—	something
you	buy	—	it
they	buy	nationalist	support
…	…	…	…

The “slot” model: *earn* (pattern 1)

based on fragment of syntactic dependency graph

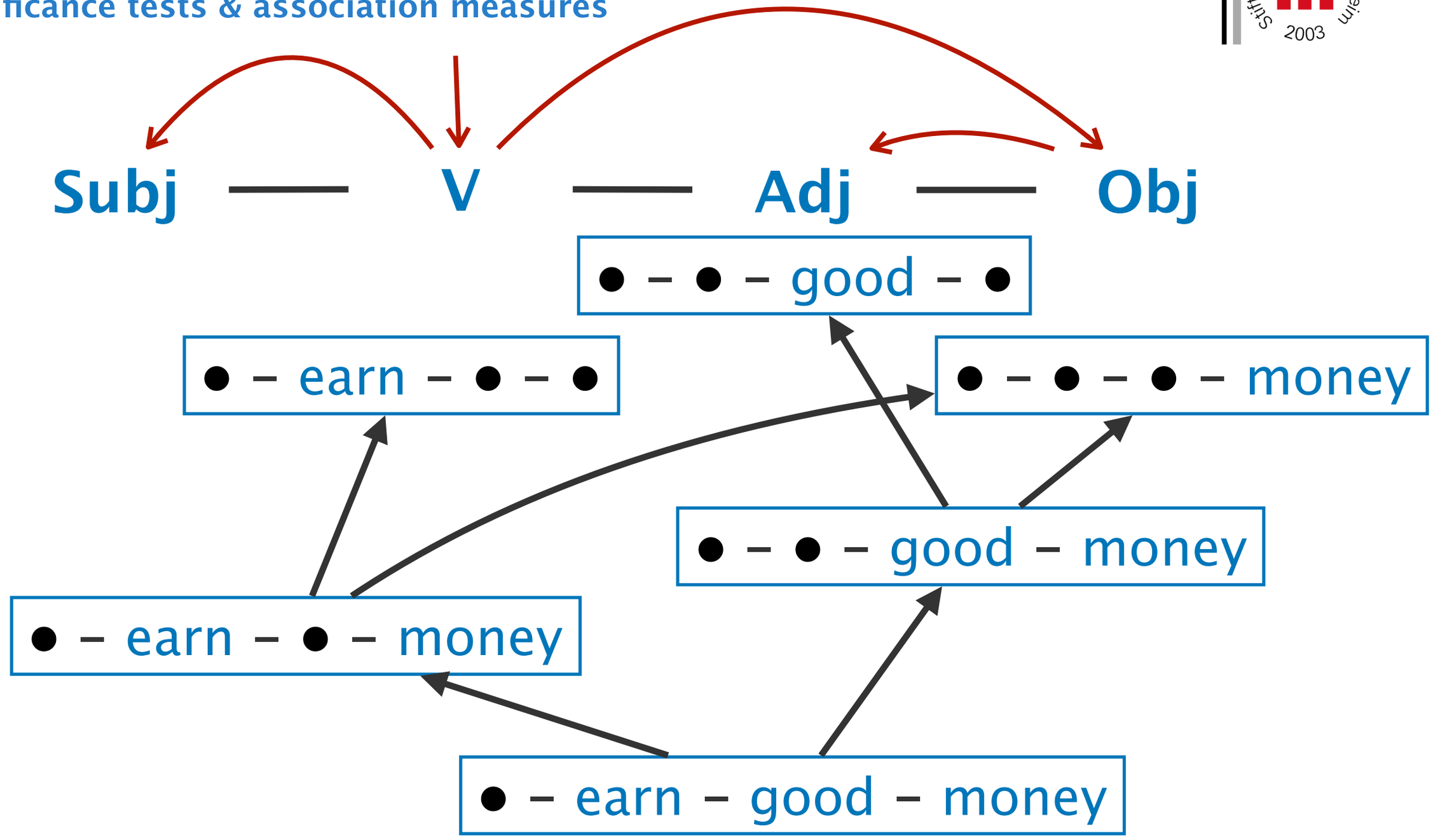


Different fixed & open-slot MWC within this frame:

- ● – earn – ● – money
- worker – earn – ● – ●
- ● – earn – good – money
- ● – earn – × – keep
- company – earn – huge – profit
- **Pron** – earn – **A** – support
- worker – earn – ● – [MONEY]

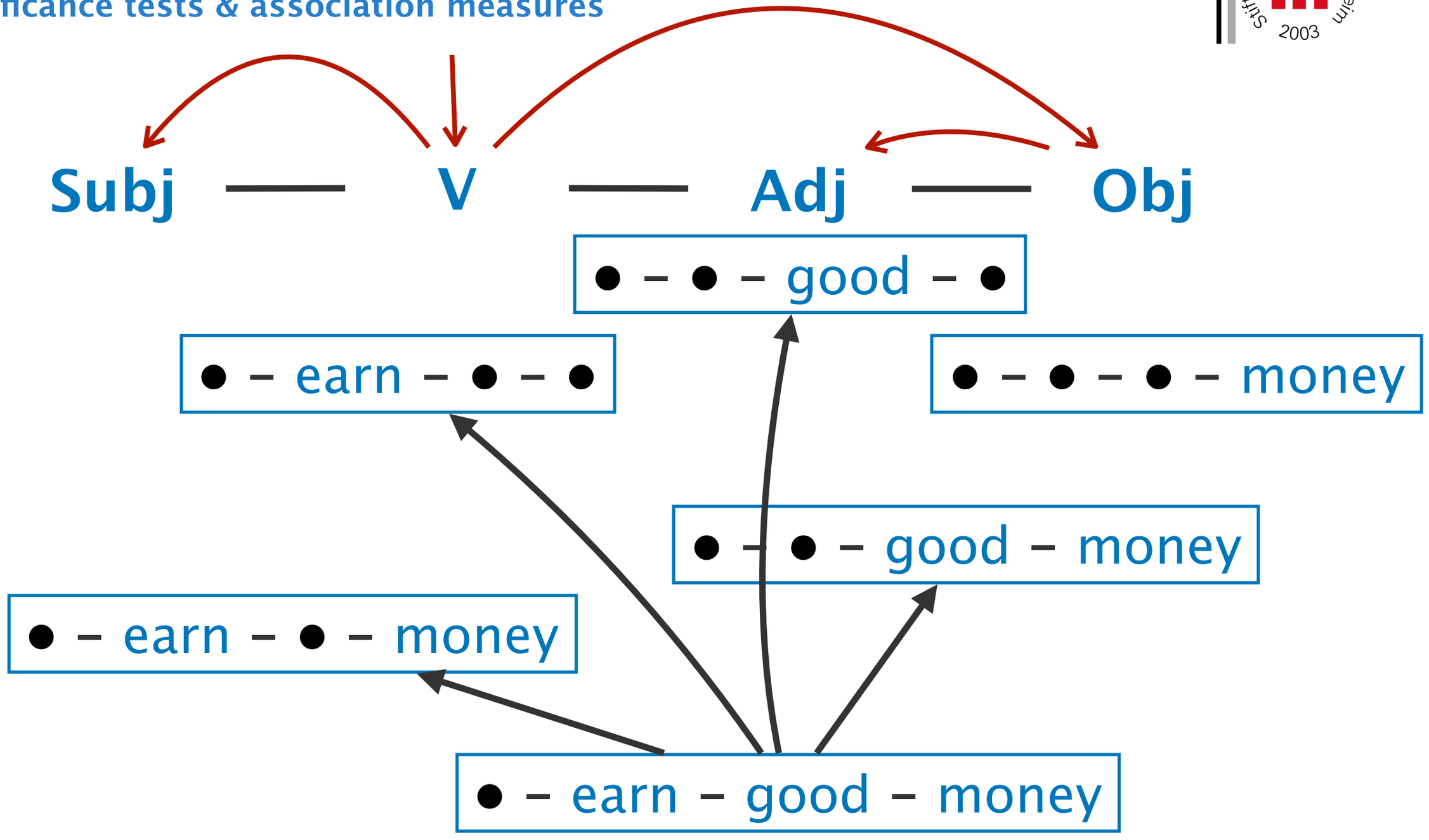
Reduction to pairwise hypotheses

→ significance tests & association measures



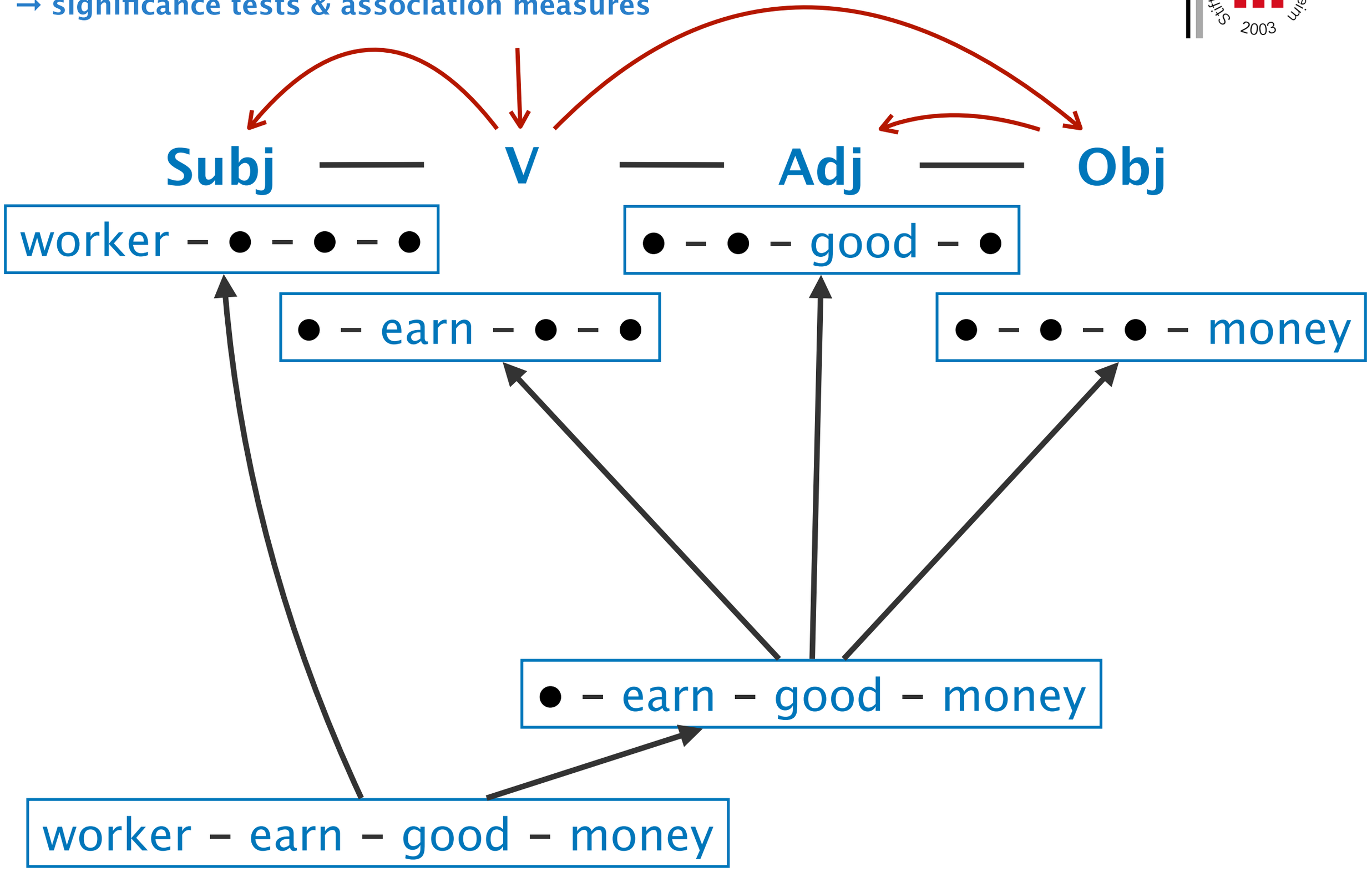
Reduction to pairwise hypotheses

→ significance tests & association measures



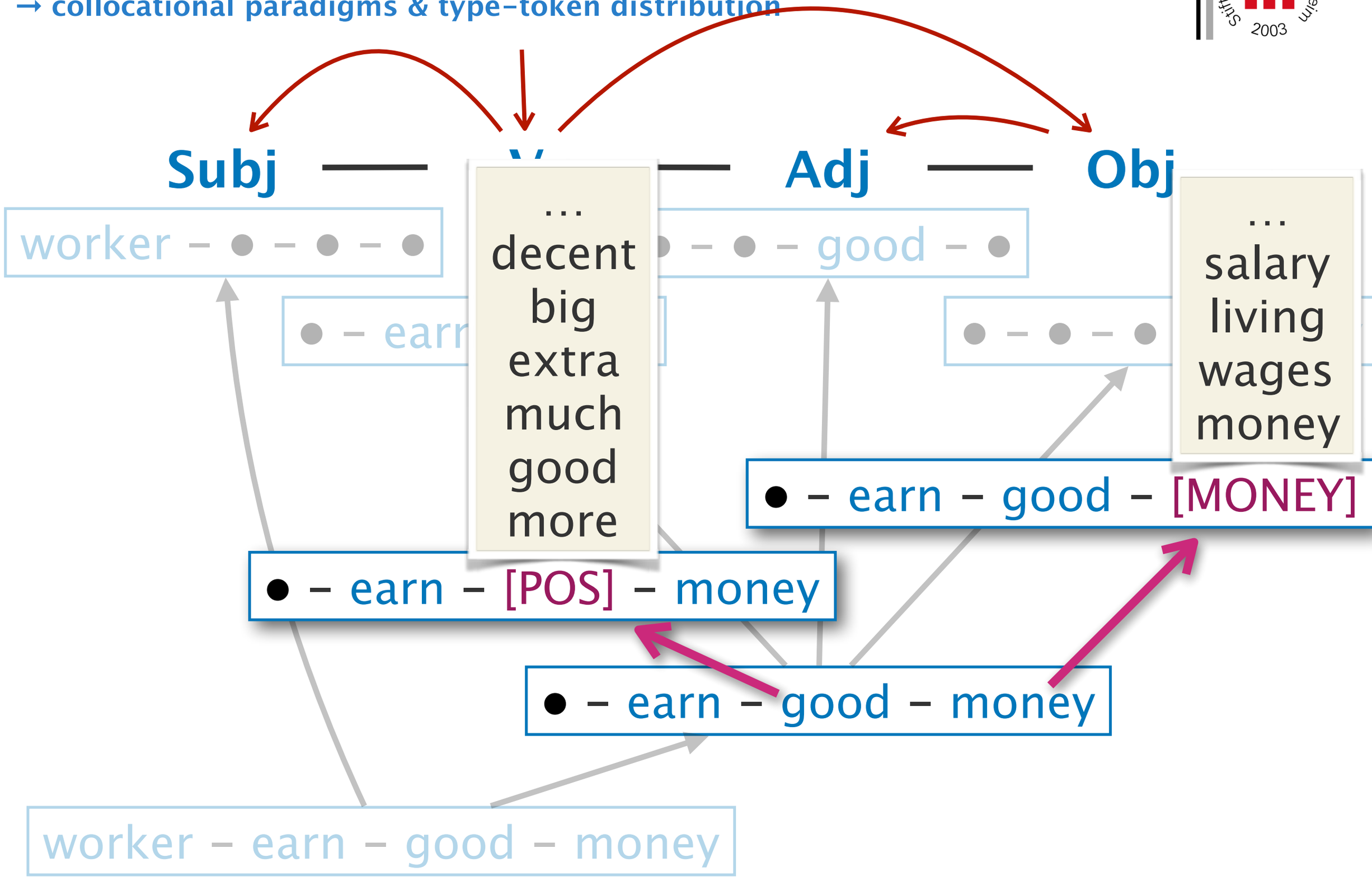
Reduction to pairwise hypotheses

→ significance tests & association measures



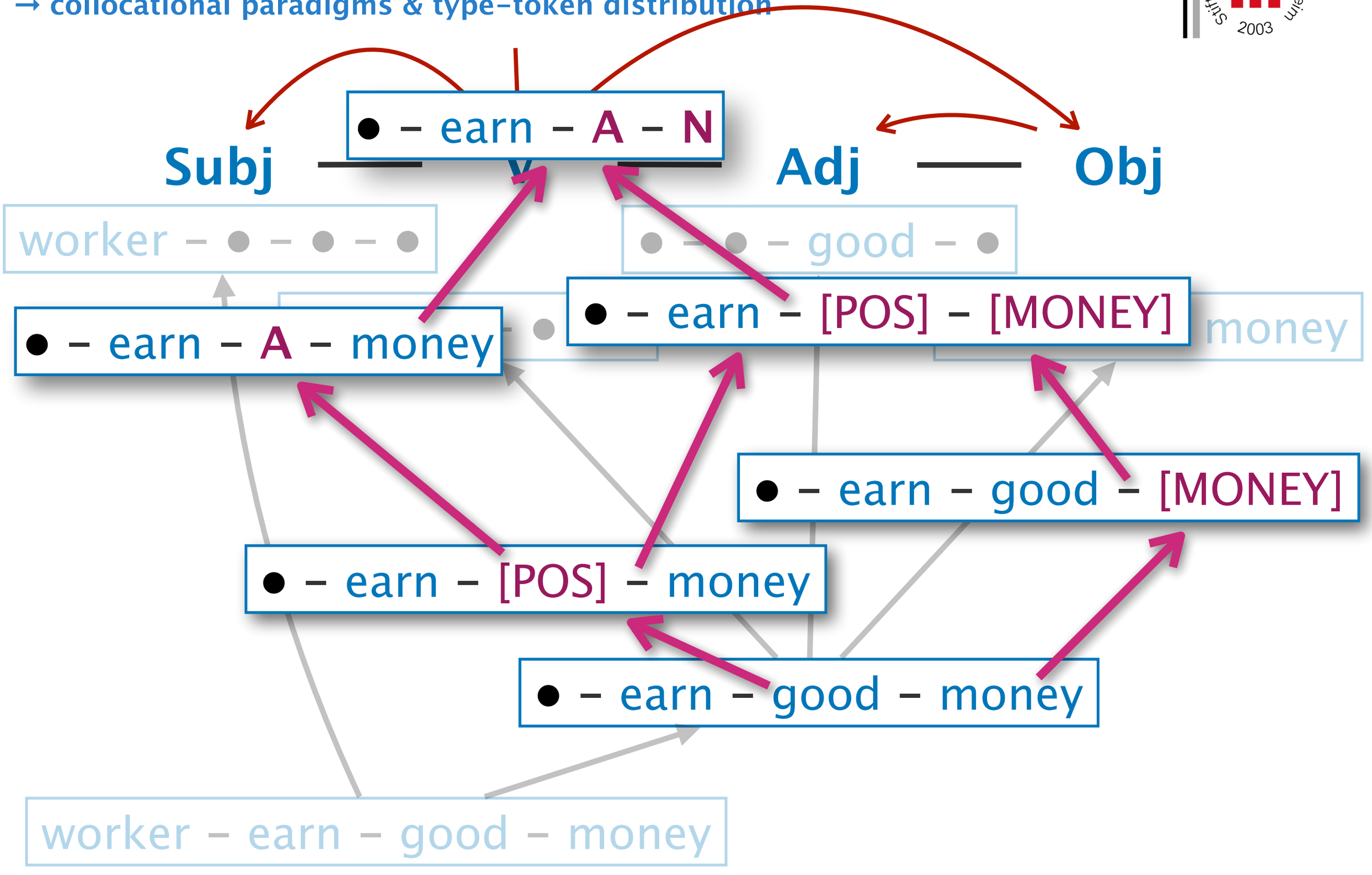
Reduction to pairwise hypotheses

→ collocational paradigms & type-token distribution



Reduction to pairwise hypotheses

→ collocational paradigms & type-token distribution



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

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English*. Narr, Tübingen.

Blaheta, D. and Johnson, M. (2001). Unsupervised learning of multi-word verbs. In *Proceedings of the ACL Workshop on Collocations*, pages 54–60, Toulouse, France.

da Silva, J. F., Dias, G., Guilloré, S., and Lopes, G. P. (1999). Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Progress in Artificial Intelligence, volume 1695 of Lecture Notes in Artificial Intelligence*, pages 113–132. Springer-Verlag.

Evert, S. (2008). Corpora and collocations. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, New York.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3:115–130.

Hausmann, F. J. (2004). Was sind eigentlich Kollokationen? In Steyer, K., editor, *Wortverbindungen – mehr oder weniger fest, Jahrbuch des Instituts für Deutsche Sprache 2003*, pages 309–334. de Gruyter, Berlin.

References

Herbst, T. (2018). Is language a collostructicon? A proposal for looking at collocations, valency, argument structure and other constructions. In Cantos-Gómez, P. and Almela-Sánchez, M., editors, *Lexical Collocation Analysis: Advances and Applications*, pages 1–22. Springer International, Cham.

Lin, D. (1998). Extracting collocations from text corpora. In P*roceedings of the First Workshop on Computational Terminology*, pages 57–63, Montreal, Canada.

Mel'cuk, I. A. (2003). Collocations: définition, rôle et utilité. In Grossmann, F. and Tutin, A., editors, *Les Collocations: analyse et traitement*, pages 23–31. De Werelt, Amsterdam.

Rogers, J. (2017). An objective method of identifying teachworthy multi-word units for second language learners. In *Proceedings of EUROPHRAS 2017*, pages 148–153, London, UK.

Zinsmeister, H. and Heid, U. (2003). Significant triples: Adjective+noun+verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003)*, pages 92–101, Budapest, Hungary.