



Tagung der Computerlinguistik Studierenden



UNIVERSITÄT
DES
SAARLANDES

Semantik im großen Stil

Wissenserwerb aus großen Korpora und
dem Web

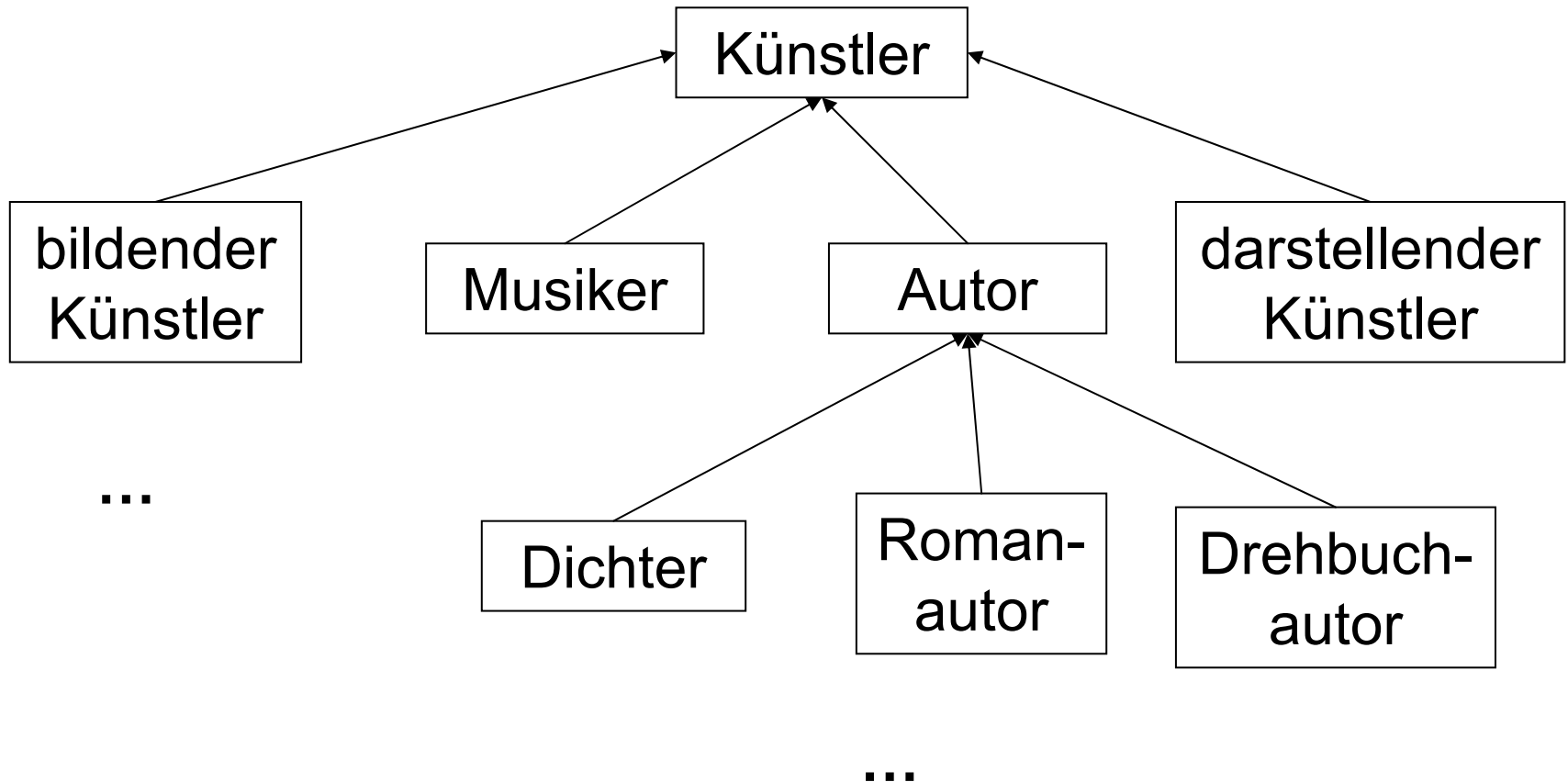
Michaela Regneri

TaCoS 2006

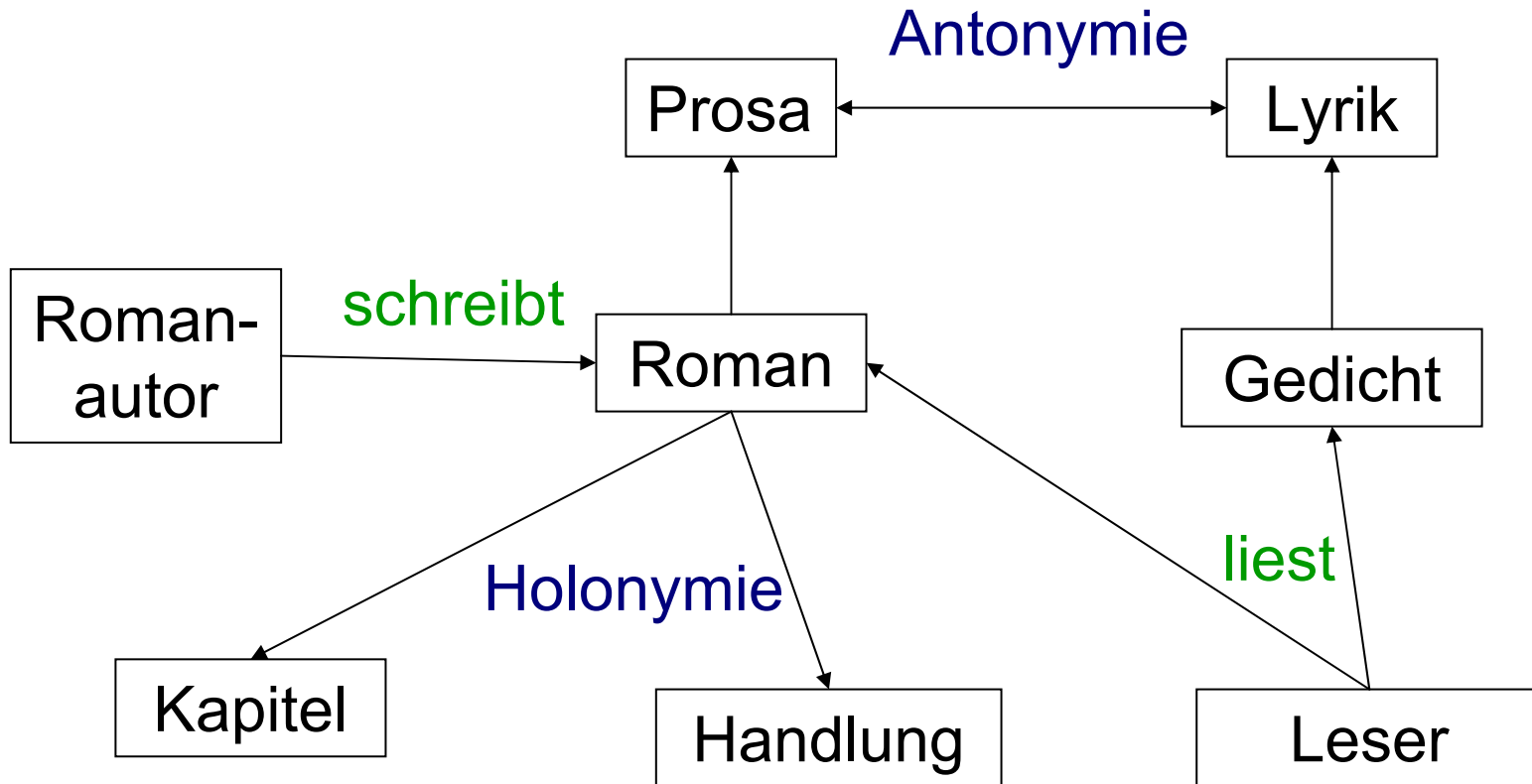
Übersicht

- Wissenserwerb - allgemein
- DIRT – Paraphrasen-Erkennung
- VerbO(c|z)ean:
Verbrelationen – ganz von der Oberfläche

Wissensbasen



Wissensbasen



Wissensbasen

- ...sind nützlich (IR, QA,...)
- ...sind aufwändig und teuer in der (manuellen) Erstellung
- ...könnten vielleicht automatisch gelernt werden?

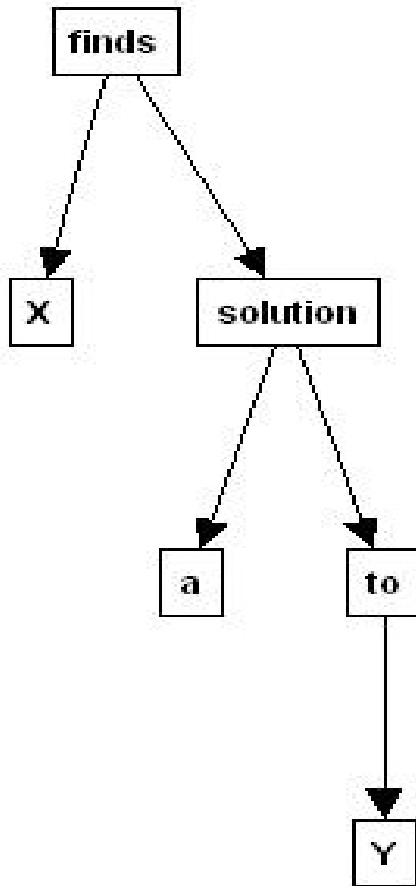
Lernansätze

- Nomen-Ähnlichkeit (Hindle 90, Lin 98)
- Hyponomie (Hearst 98, Ruiz et al. 05)
- Meronomie (Girju et al. 05)
- Verb-Klassen (Stevnson & Joanis 03)
- Inferenzregeln (Marcu & Popescu 05)
- Paraphrasen (Lin & Pantel 01, Pado & Erk 05)
- ...

DIRT (Lin & Pantel 2001)

- „Deriving Inference Rules From Text“ – automatisches Lernen von Paraphrasen
- *Distributional Hypothesis* (Harris 1954): Dinge, die in ähnlichem Kontext auftauchen, haben ähnliche Bedeutung
- „Kontexte“ sind hier Pfade im Dependenzbaum

DIRT (Lin & Pantel 2001)



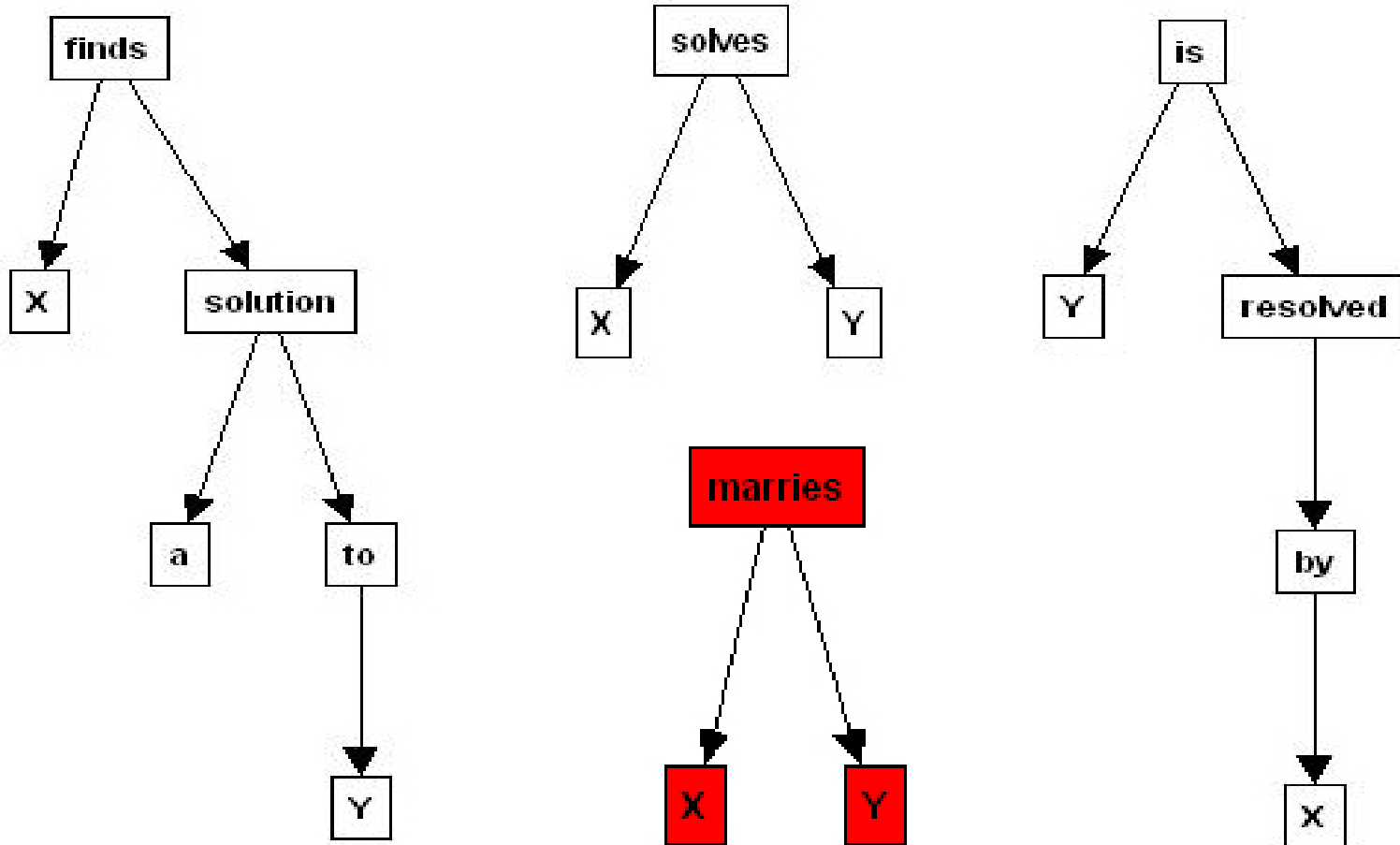
<i>„X finds a solution to Y“</i>	
X	Y
comission	strike
committee	civil war
committee	crisis
government	crisis
government	problem
sheriff	dispute

DIRT (Lin & Pantel 2001)

<i>„X finds a solution to Y“</i>	
X	Y
comission	strike
committee	civil war
committee	crisis
government	crisis
government	problem
sheriff	dispute

<i>„X solves Y“</i>	
X	Y
clout	crisis
government	problem
researcher	mystery
she	problem
sheriff	murder
petition	woe

DIRT (Lin & Pantel 2001)



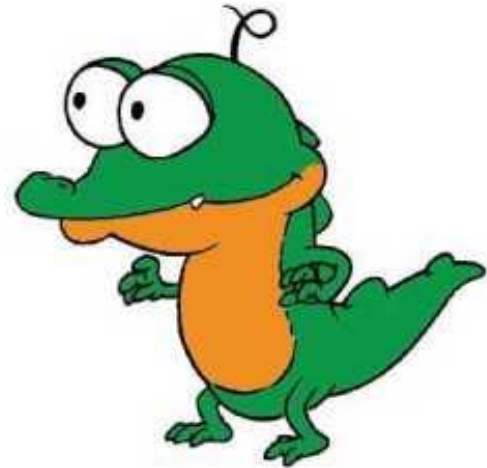
DIRT (Lin & Pantel 2001)

Die „Top 22“ der Paraphrasen für „X solves Y“:

- Y is solved by X
- Y is resolved in X
- X resolves Y
- Y is solved through X
- X finds a solution to Y
- X rectifies Y
- X tries to solve Y
- X copes with Y
- X deals with Y
- X overcomes Y
- X makes Y worse
- Y is resolved by X
- X eases Y
- X addresses Y
- X tackles Y
- X seeks a solution to Y
- X alleviates Y
- X do something about Y
- X corrects Y
- X solution to Y
- X is a solution to Y
- X irons out Y

VerbOcean - VerbOzean

(Chklovski & Pantel 2005)

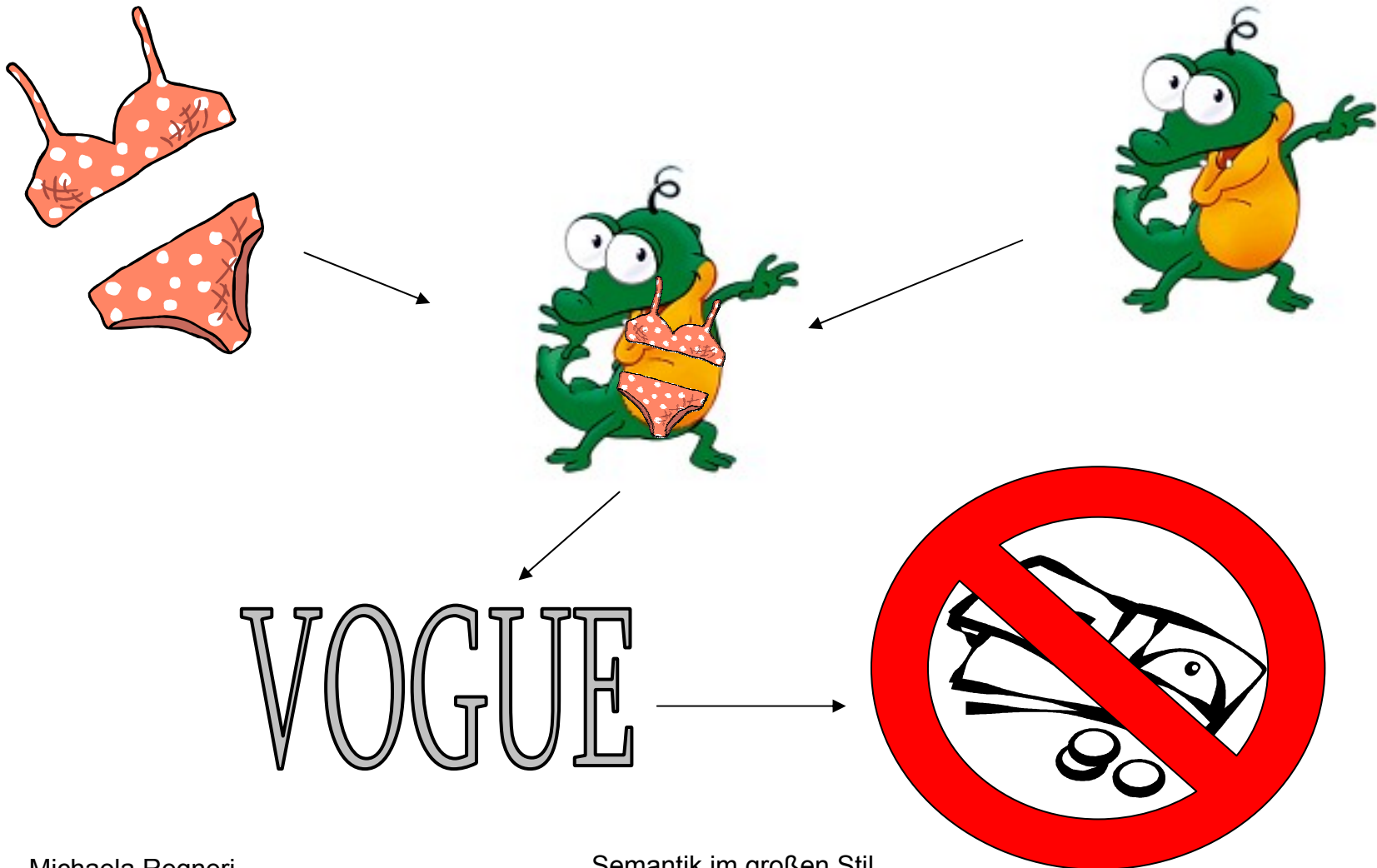


<http://semantics.isi.edu/ocean/>

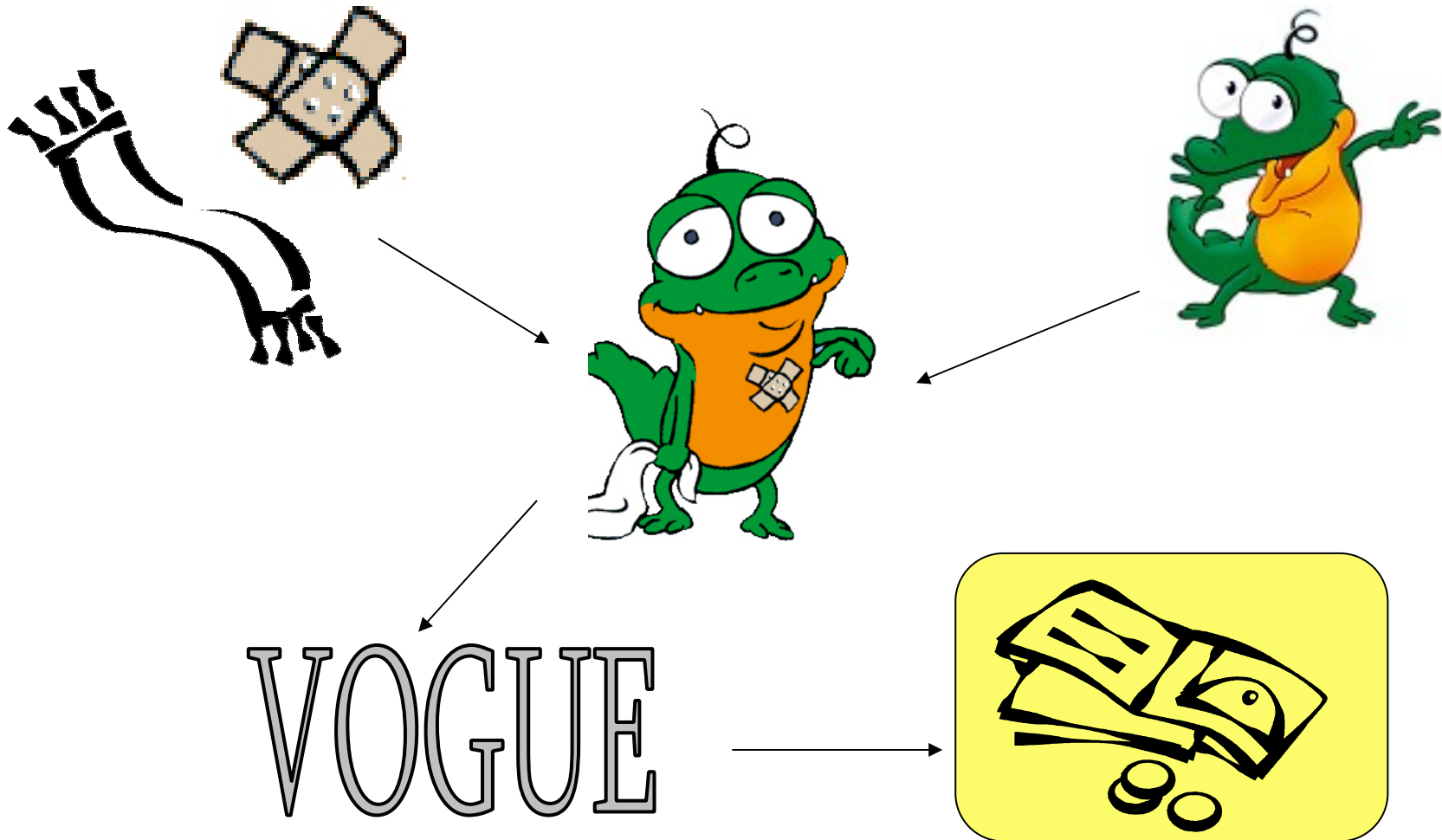
Konzept

- Ziel: Feststellen, ob zwei Verben in einer bestimmten Relation stehen
- Sammeln von stark assoziierten Verbpaaren
- Entwerfen von Pattern für bestimmte Verbrelationen (z.B. „zuerst v1 und dann v2“)
- Prüfen, ob ein Verbpaar in dem Pattern öfter auftaucht als zufällig

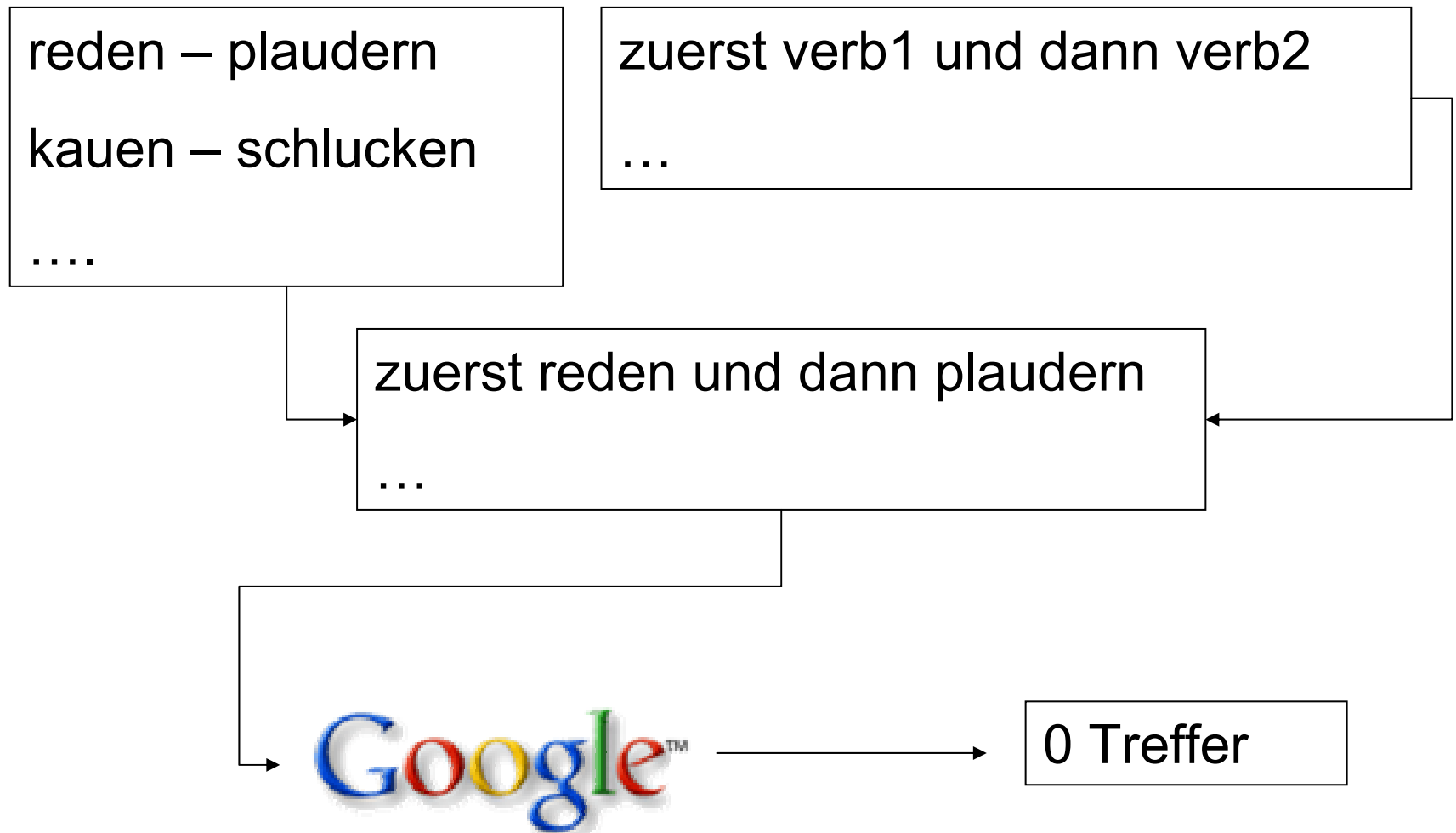
Zweiteiler = „Krokodil-Dress“?



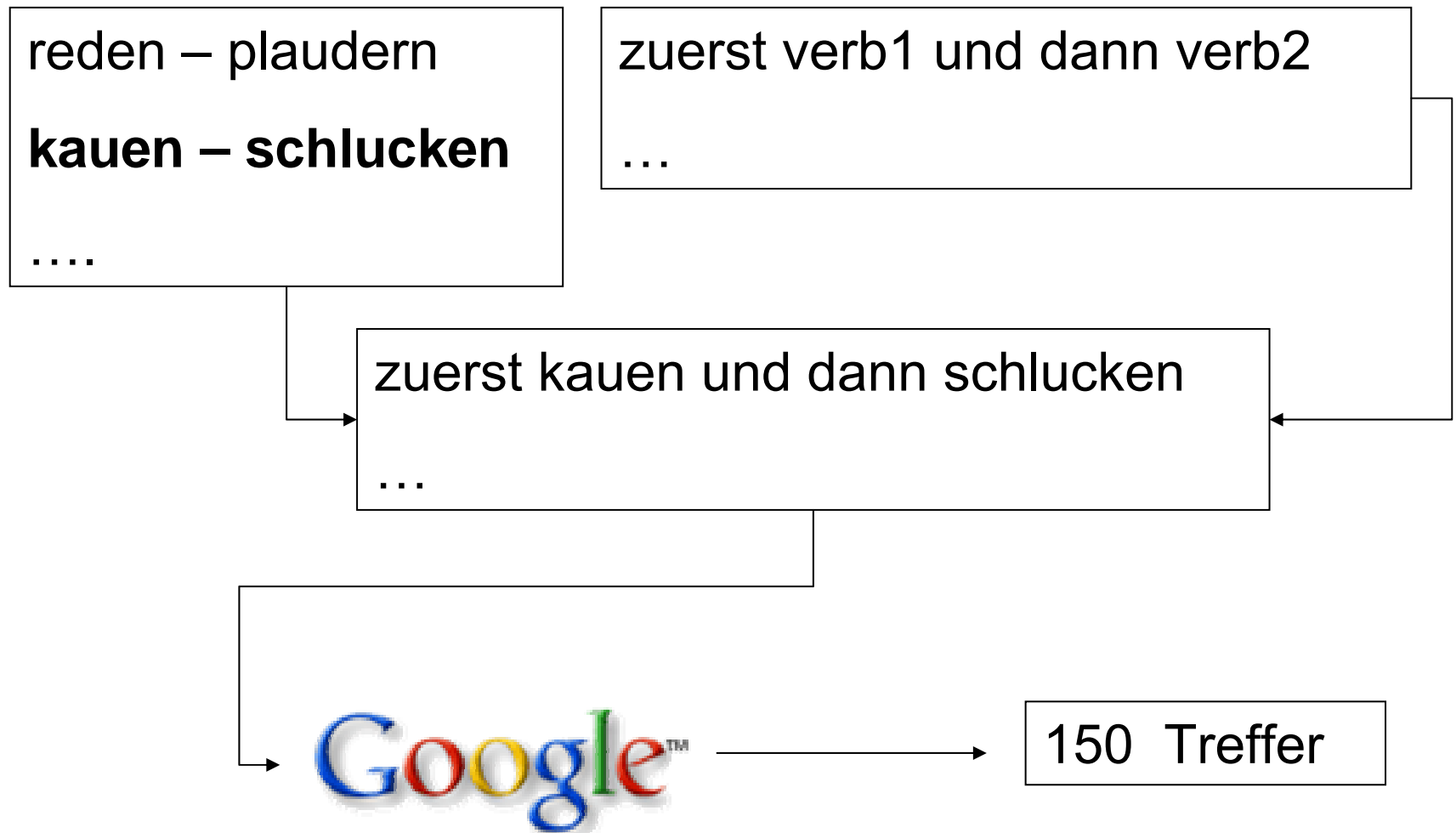
Zweiteiler = „Krokodil-Dress“?



Verbpaar = „Zeitliches Folgen“?



Verbpaar = „Zeitliches Folgen“?



Extraktion von Verbpaaren

VerbOcean

-

VerbOzean

- Paraphrasen mit DIRT (aus 1.5 GB Korpus)
- Verben aus jedem Paraphrasenpaar → Verbpaar
- 29165 Paare

- Paare aus Verbassoziations-Experiment (Schulte im Walde & Melinger 2005)
- 4824 Paare

Verbrelationen

VerbOcean

- narrow similarity
- broad similarity
- antonymy
- strength
- enablement
- happens-before

-

VerbOzean

- vorerst nur „zeitliches Folgen“ (under construction 😊)
- Kandidaten:
 - ☐ Antonymie
 - ☐ Skript-Schwester
 - ☐ Reversivität
 - ☐ ... ?

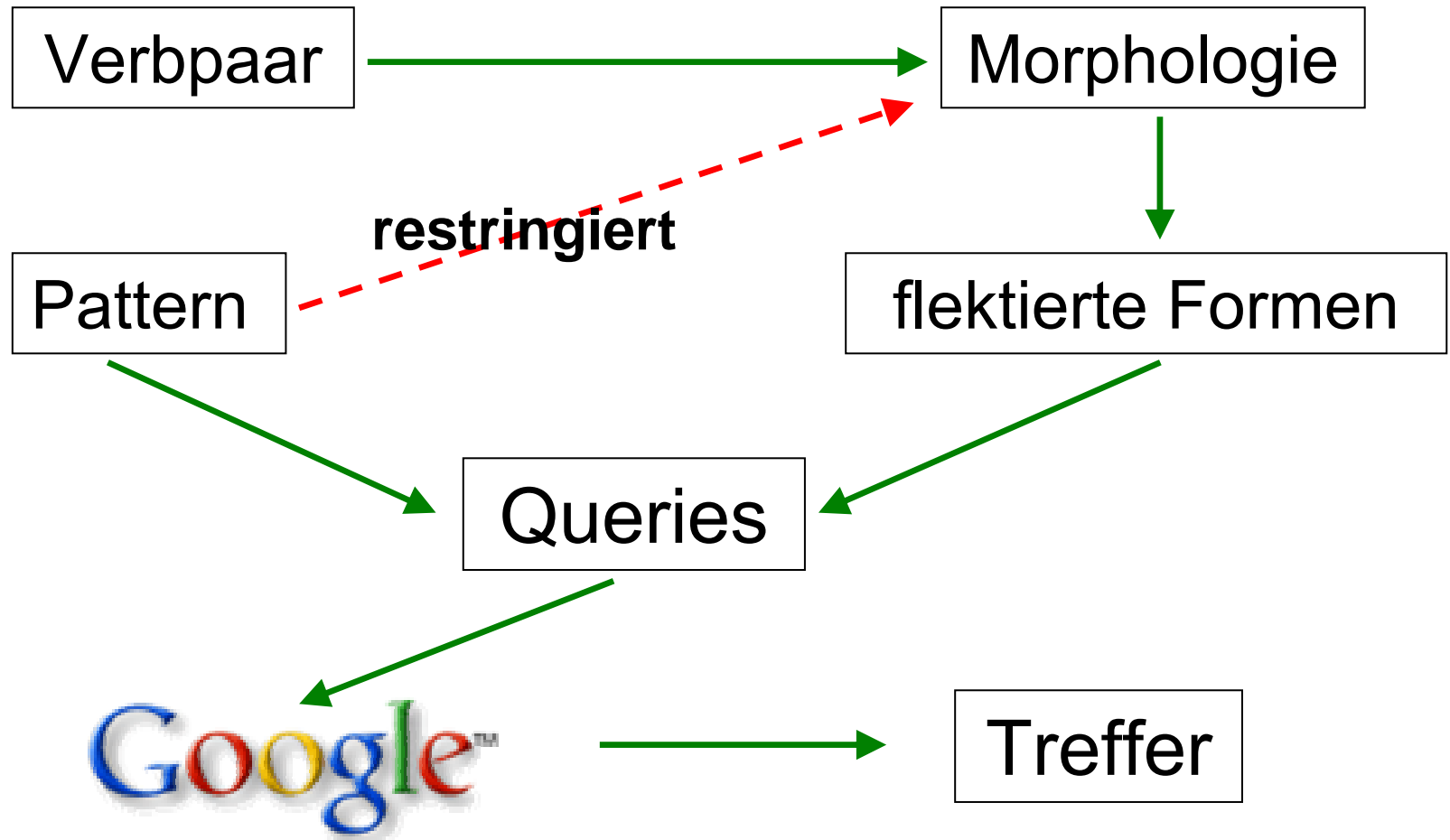
Patterns

- Suche nach Verben in bekannter Relation im Korpus
(z.B. „(.){1-5} kauen (.){1-5} schlucken (.){1-5})
- intuitive Überlegungen → auf Korpus testen
- Auswahl von Patterns mit möglichst vielen Treffern und möglichst wenig Noise

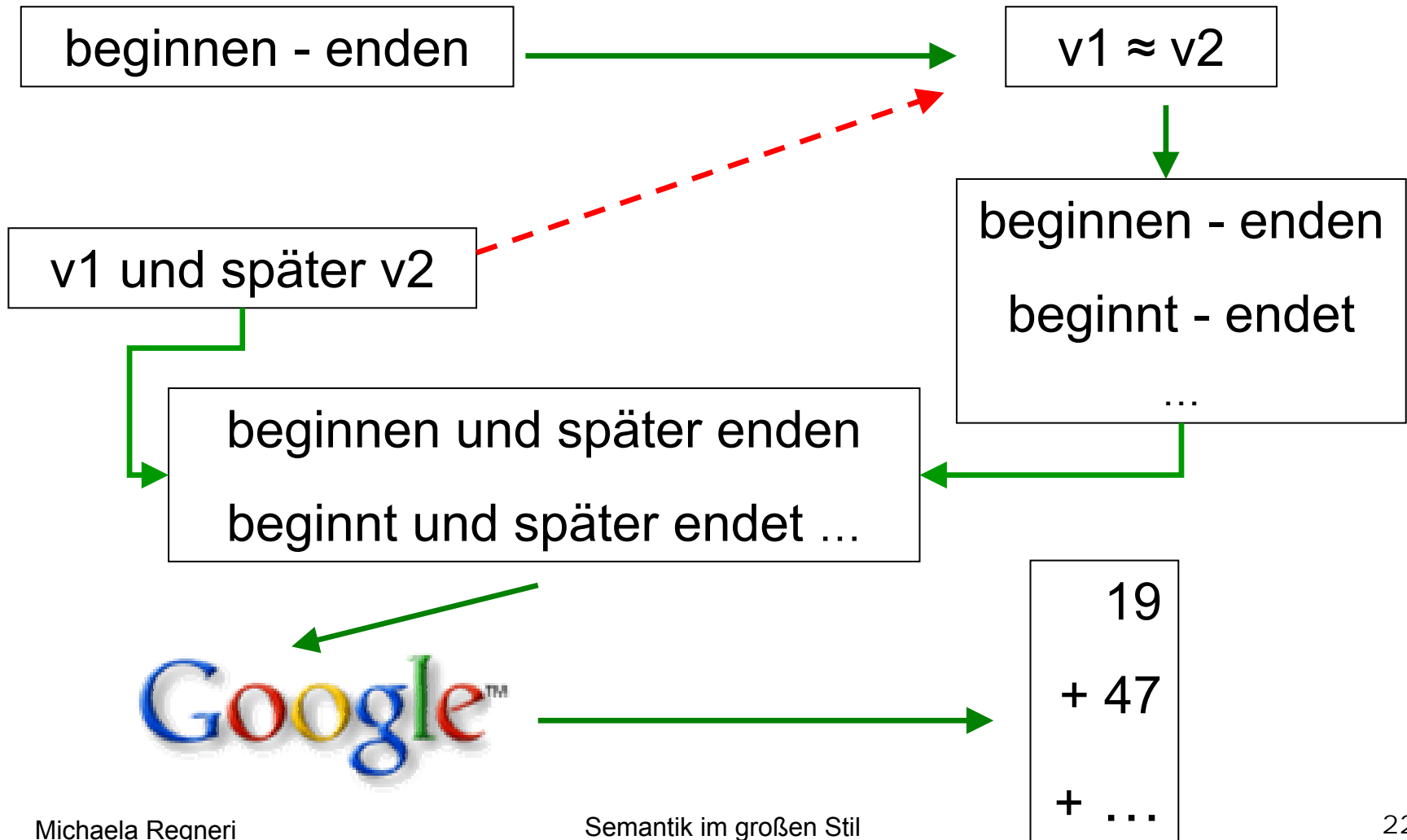
Patterns (für zeitliches Folgen)

- "v1 und bereits v2"
- "v1 und schließlich v2"
- "v1 und bald v2"
- "v1 und später v2"
- "zuerst v1 und dann v2"
- "v1 und sofort v2"
- "v1 und anschließend v2"
- "v1 und dann v2"

Sammeln von Google-Treffern



Sammeln von Google-Treffern



Google-Anfragen

- Ca. 500.000 Anfragen pro Durchlauf
 - Google-API: Anfrage-Limit von 10.000/Tag
 - Google-Anfragen über URL (User-Agent!)
- Präzisere Suche durch Ausschluss eines Wortes ohne Treffer

(→ - )

Relationszugehörigkeit - MI

- Häufige Verben ergeben viele Hits

- Mutual Information:

$$\frac{p(\text{verb1}, \text{pattern}, \text{verb2})}{p(\text{verb1}) * p(\text{pattern}) * p(\text{verb2})}$$

- Cut-Off: MI für ein Verbpaar muss einen bestimmten Wert überschreiten → Relationszugehörigkeit

DeWac-Korpus (Adam Kilgarriff)

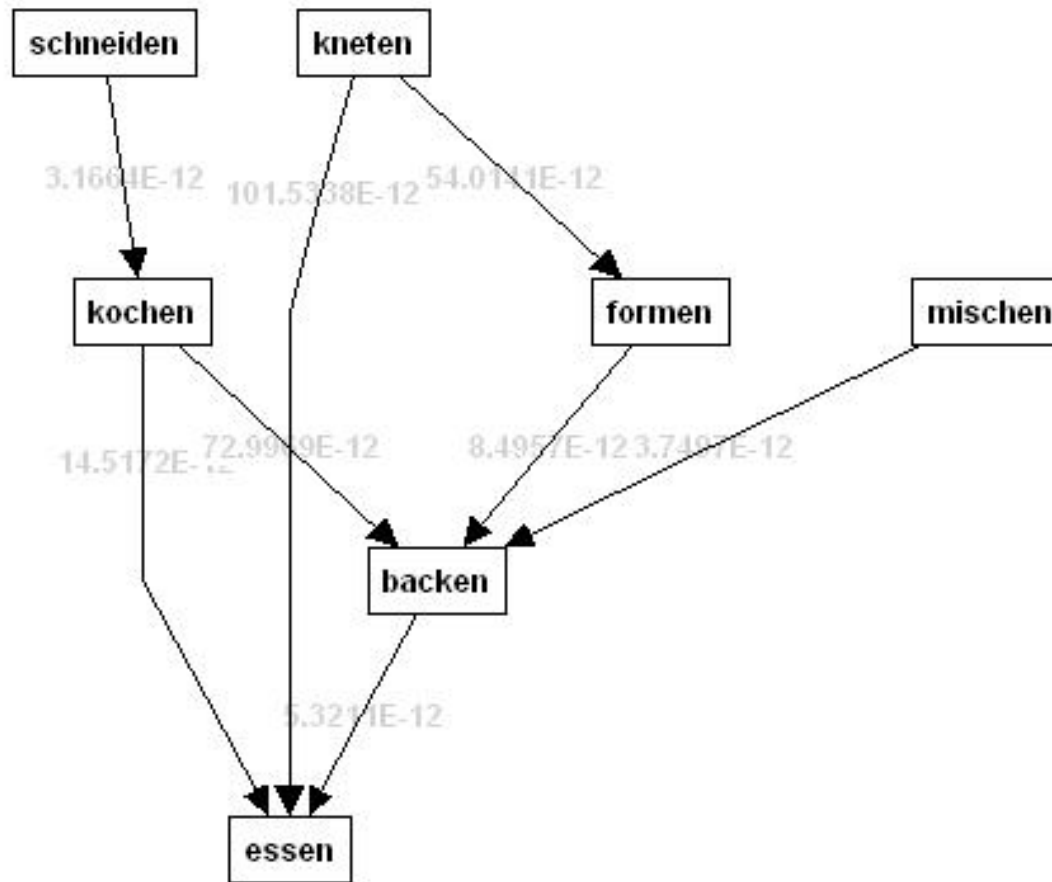
- Google-Suche zur Bestimmung von Pattern- und Verbhäufigkeiten ungeeignet
- Kommerzielles Korpus
- 1.7 mrd. Wörter
- POS-getaggt, lemmatisiert
- balanciert

Top 15 „Zeitliches Folgen“

verloben	verheiraten
einfrieren	auftauen
verdauen	ausscheiden
waschen	trocknen
erhitzen	abkühlen
kauen	schlucken
pachten	kaufen

trennen	scheiden
kneten	essen
<i>schwimmen</i>	<i>paddeln</i>
entladen	laden
leihen	zurückgeben
kochen	backen
mieten	kaufen
rauchen	aufhören

Beispiel-“Frame“



Evaluation

- Annotation durch fünf Annotatoren
- Testset:
 - 30 erkannte Paare
 - 15 mit Google-Treffern, aber unter Cut-Off
 - 15 ohne Google-Treffer
- Auswertung läuft...

Komplexere Aufgabenstellungen

- Sparse Data
 - Verbfrequenzen
 - Finden von (mehr) Verbpaaren
- Morphologie im Deutschen
- Präzision von Google
- Recall (mehr Patterns)

Zusammenfassung

- Interessante Ansätze zum automatischen Wissenserwerb
- Diverse Möglichkeiten und Einschränkungen, je nach Korpus (Web?)
- Noch viel Raum für Kreativität 😊