Heidelberg Institute for
Theoretical Studies

HITS

Multiword Expressions

WS 2019/20
2. Sitzung (29.10.2019)

Institut für Computerlinguistik, Universität Heidelberg

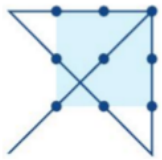
Mark-Christoph Müller

Natural Language Processing Group

Heidelberg Institute for Theoretical Studies gGmbH

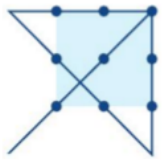
mark-christoph.mueller@h-its.org

Organisatorisches



- Gewichtung für die Notenvergabe
  - Mitarbeit **30%**
  - Referat **40%**
  - Hausarbeit / Programmierprojekt **30%**
- Literatur für die Referate
  - Es sind jeweils **alle angegebenen** Papiere zu berücksichtigen
  - Bei mehreren Referenten und mehreren Papieren
    - Querverbindungen herstellen!
    - **Alle** Referenten tragen **alle** Papiere vor!

Organisatorisches



- Referate
  - siehe Kursseite https://www.cl.uni-heidelberg.de/courses/ws19/expr/
  - **Entfall** am 26.11.2019 wg. Klimastreik
  - Verschiebung der Themen 26.11. bis 17.12. um eine Woche
  - Streichung von *N-N Compound Analysis III* vom 07.01.2020

Was sind *Multiword Expressions*?

- Keine allgemein akzeptierte Definition, z.B.

> - "a multiword unit or a collocation of words that co-occur together statistically more than chance" (Carpuat and Diab 2010)
> - "a sequence of words that acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002)
> - "idiosyncratic interpretations that cross word boundaries" (Sag et al. 2002)
> - "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim 2010)

Constant et al. (2017)

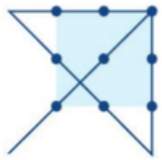
- Wichtige Aspekte

  1) **Idiosynkratisch** (d.h. ungewöhlich, unregelmäßig),

  2) **komplex** (= zusammengesetzt), d.h. (theoretisch) zerlegbar,

  3) statistisch **signifikant häufig**

1) Semantische Kompositionalität



- Eine Sprache zu beherrschen bedeutet, in der Lage zu sein
  - komplexe (i.S.v. *zusammengesetzte*) Ausdrücke zu **bilden**, um best. Bedeutungen zu vermitteln, sowie
  - die Bedeutung komplexer Ausdrücke zu **analysieren**.

- "The Principle of **Semantic Compositionality** is the principle that the meaning of an expression is a **function** of, and only of, the **meanings of its parts** together with the **method** by which those parts are **combined**." (Pelletier 1994, Hervorhebung MCM)

- Beispiel: Modifikation eines Nomens mit einem Adjektiv

1) Semantische Kompositionalität

Objekt: H_2O → NOMEN

Eigenschaft: Temp <= 5° Celsius → ADJEKTIV

kompositionelle *Bedeutung* ←

Sprach-spezifische Kombination

ADJEKTIV + NOMEN **NOMEN + ADJEKTIV**

“kaltes Wasser”

“soğuk su”

“cold water”

“холодная вода”

“agua fria”

“eau froide”

1) Wenn die Kompositionalität fehlt ...

Objekt: Canis lup. fam. → NOMEN

Eigenschaft: Temp > 30° Celsius → ADJEKTIV

Sprach-spezifische Kombination

ADJEKTIV + NOMEN **NOMEN + ADJEKTIV**

kompositionelle Bedeutung

nicht-kompositionelle (idiomatische) Bedeutung

“hot dog”

Betonung: 'hot dog **vs.** hot 'dog

1) Beobachtung: Kompositionalität & Syntax

- Regeln für die Modifikation eines Nomens mit einem Adjektiv sind Teil der Syntax der Sprache (“Grammatik”)

- **Kompositionelle** Bedeutung
  - “<NOMEN> hat die Eigenschaft <ADJEKTIV>”
    - --> “cold water” (und die meisten anderen)

- **Tatsächliche** Bedeutung kann nicht-kompositionell / **idiomatisch** sein
  - --> “hot dog”

1) Was, wenn es keine Syntax gibt?

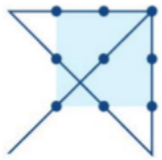
- Komplexe Ausdrücke können auch durch einfache Verkettung gebildet werden
  - Nominal-Kompositum / Nominal Compound
  - $\text{Nomen}_1 + \text{Nomen}_2 = \text{Nomen}_3$
- ***Sehr*** produktiv
- ***Extrem viele*** semant. Relationen
  - Keine kompositionelle Standard-Lesart

N1 CAUSE N2	sex scandal, withdrawal symptom
N2 CAUSE N1	tear gas, shock news
N1 HAVE N2	lemon peel, school gate
N2 HAVE N1	camera phone, picture book,
N1 MAKE N2	court order, snowball
N2 MAKE N1	computer industry, silk worm
N2 USE N1	steam iron, wind farm
N2 BE N1	island state, soldier ant
N2 IN N1	field mouse, letter bomb
N2 FOR N1	arms budget, steak knife
N2 FROM N1	business profit, olive oil
N2 ABOUT N1	tax law, love letter

‘pig iron’?

- Bedeutung muss stattdessen
  - **bekannt** sein, oder
  - durch (Welt-)Wissen **erschlossen** werden (bei Neubildungen)

1) Semantische Kompositionalität



- Fragen?

Was sind *Multiword Expressions*?

- Keine allgemein akzeptierte Definition, z.B.

> - "a multiword unit or a collocation of words that co-occur together statistically more than chance" (Carpuat and Diab 2010)
> - "a sequence of words that acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002)
> - "idiosyncratic interpretations that cross word boundaries" (Sag et al. 2002)
> - "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim 2010)

Constant et al. (2017)

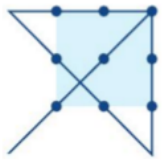
- Wichtige Aspekte

  1) **Idiosynkratisch** (d.h. ungewöhlich, unregelmäßig),

  2) **komplex** (= zusammengesetzt), d.h. (theoretisch) zerlegbar,

  3) statistisch **signifikant häufig**

2) Strukturelle Komplexität von MWEs I



- Was ist überhaupt ein **Wort**?
- Fokus auf geschriebene Sprache, oft **typographisch** definiert:
  - Folge von Buchstaben (ggfs. Zahlen) zwischen Leerzeichen etc.
  - Entspricht dem 'token' in NLP
    - RegExp: \w = 'word character', \b = 'word boundary'
- Vorteile:
  - Universell (\b ist wahrsch. in jeder Sprache eine Wortgrenze)
- Probleme:

  1) Für best. MWEs ist Schreibung variabel

  2) Verschiedene Sprachen haben verschiedene Regeln

2) Strukturelle Komplexität von MWEs II

- Innerhalb einer Sprache kann Schreibung variabel sein
- **Nominal-Komposita:** $\text{Nomen}_1 + \text{Nomen}_2 = \text{Nomen}_3$
  - Englisch

  - *health care* (***open*** *compound)* ***Default* bei Neubildung!**
  - *health-care* *(closed / **hyphenated** compound)*
  - *healthcare* *(**closed** compound)*

  - *computer specialist*
  - *computer-specialist*
  - ***computerspecialist***

2) Strukturelle Komplexität von MWEs III

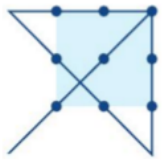
- Verschiedene Sprachen haben verschiedene Regeln
- Deutsch
  - **Computer Spezialist*
  - *Computer-Spezialist*
  - *Computerspezialist* ***Default* bei Neubildung!**

Aber!


Merkel eröffnet
Klima Arena

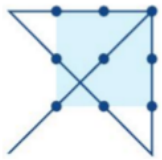
SINSHEIM. „Beim Klimaschutz mu
man die Menschen mitnehmen
Das hat Bundeskanzlerin Ange
Merkel bei der Eröffnung der Klim

2) Strukturelle Komplexität von MWEs IV



- **Zusammenfassung**
  - MWEs vom Typ Nominal-Kompositum können **typografisch** auch ***ein*** Wort sein
    - Englisch
      - Je nach Grad der Etablierung
      - Neubildungen: In der Regel getrennt (*open*)
    - Deutsch
      - Zusammen (*closed*), oder mit Bindestrich (*hyphenated*)
      - Trend zur Übernahme des englischen Schemas
- Aktuelle Methoden sind stark auf Englisch optimiert

2) Strukturelle Komplexität von MWEs



- Fragen?

Was sind *Multiword Expressions*?

- Keine allgemein akzeptierte Definition, z.B.

> - "a multiword unit or a collocation of words that co-occur together statistically more than chance" (Carpuat and Diab 2010)
> - "a sequence of words that acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002)
> - "idiosyncratic interpretations that cross word boundaries" (Sag et al. 2002)
> - "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim 2010)

Constant et al. (2017)

- Wichtige Aspekte

  1) **Idiosynkratisch** (d.h. ungewöhlich, unregelmäßig),

  2) **komplex** (= zusammengesetzt), d.h. (theoretisch) zerlegbar,

  3) statistisch **signifikant häufig**

3) MWEs und Kollokationen / *collocations* I

- **Kollokation** (von zwei oder mehr Elementen)
  - häufiges gemeinsames Auftreten in demselben ***Bereich***
  - *häufig* im Verhältnis zum Auftreten der Elemente alleine
- Bereich = Fenster einer best. Größe *N*
  - **Kontinuierlich** (nebeneinander), d.h. N-Gramme
    - zusammenhängende MWEs wie feststehende Ausdrücke, Nominalkomposita (wenn getrennt geschrieben!), Adjektiv-Nomen-Kombinationen
  - **Diskontinuierlich** (mit Zwischenräumen) / für größeres *N*
    - syntakt. flexiblere Ausdrücke (Verb-Partikel-Konstruktionen)

3) MWEs und Kollokationen / *collocations* II

- **Ursachen** von Kollokationen: **Syntaktische** Regularitäten

1) Für kleines *N* (*N*-Gram-Level)
   1) 'and the', 'of the', 'and it', ...
   2) 'the' / 'a(n)' + Nomen

3) MWEs und Kollokationen / *collocations* III

- **Ursachen** von Kollokationen: **Lexikalische** Regularitäten

1) Für kleines *N*
   1) Nominal-Komposita
   2) Adjektiv-Nomen-Kombinationen
2) Für größeres *N*
   1) Verb-Partikel-Konstruktionen, *phrasal verbs*

3) MWEs und Kollokationen / *collocations* IV

- **Ursachen** von Kollokationen: **Außersprachliche** Regularitäten
  - Für größeres *N*
    - Best. Wörter tauchen gemeinsam auf, weil die zugehörigen Konzepte “miteinander zu tun haben”.
    - Unterscheidung zwischen semant. Ähnlichkeit und semant. “Verbundenheit” (similarity vs. relatedness)

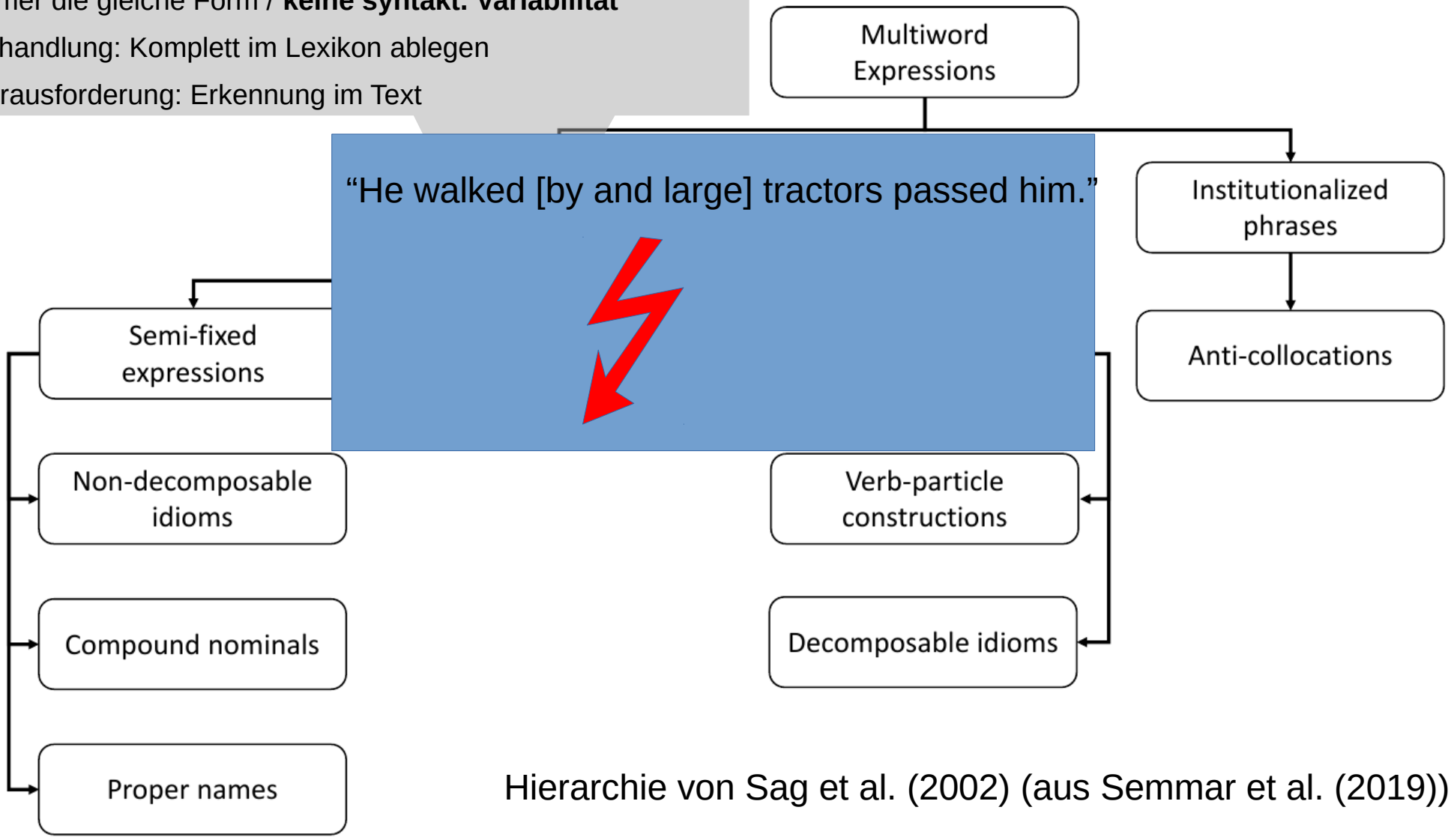
3) MWEs und Kollokationen / *collocations*

- Fragen?

Übersicht *Multiword Expressions*

Beispiel: 'by and large'; 'im Großen und Ganzen'

- Meist Funktionswörter (z.B. Adverbien), geschlossene Klasse
- Nicht sinnvoll zerlegbar / analysierbar
- Immer die gleiche Form / **keine syntakt. Variabilität**
- Behandlung: Komplett im Lexikon ablegen
- Herausforderung: Erkennung im Text

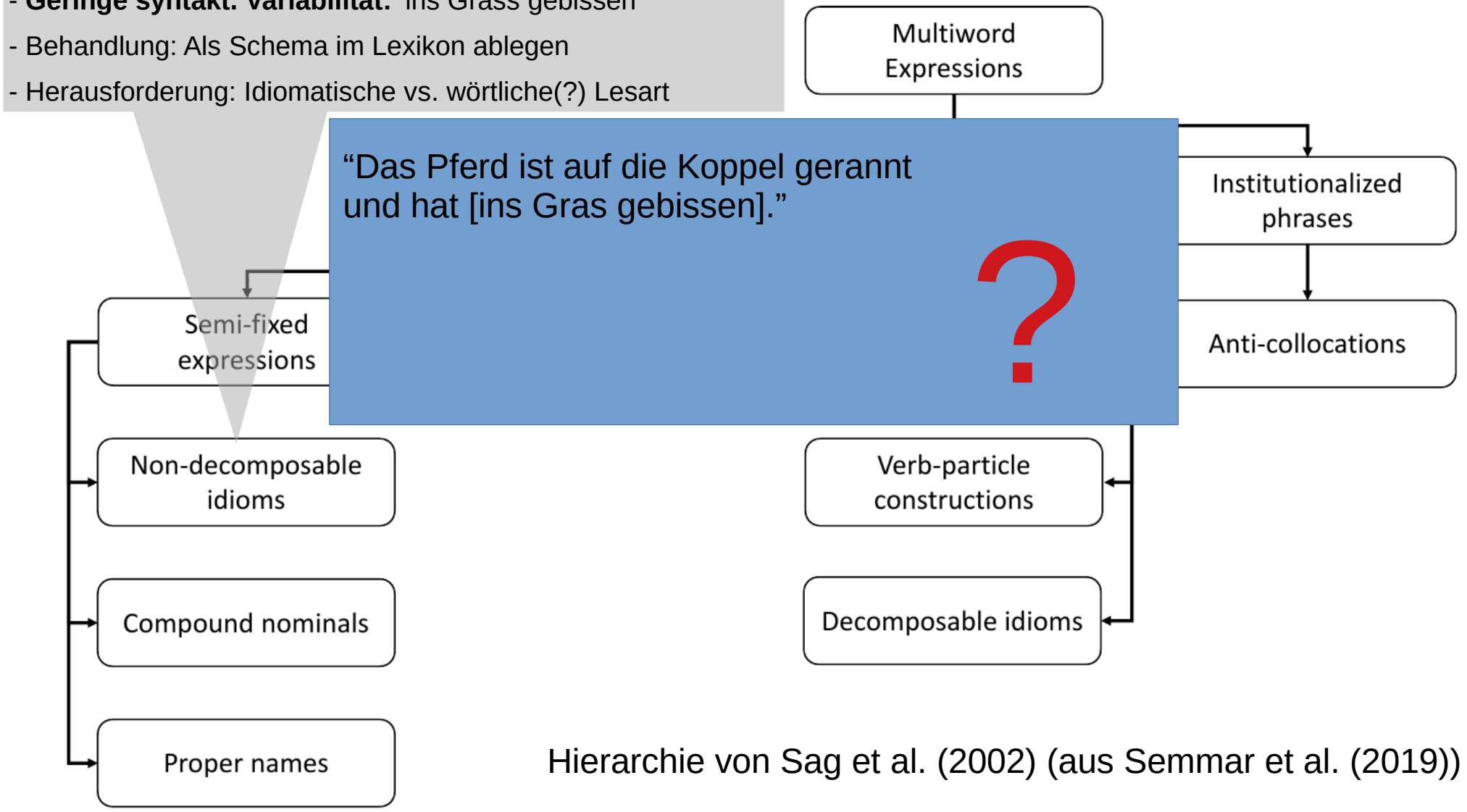


Hierarchie von Sag et al. (2002) (aus Semmar et al. (2019))

Übersicht *Multiword Expressions*

Beispiel: 'kick the bucket'; 'ins Gras beissen'

- Div. Wortarten, geschlossene Klasse
- Nicht sinnvoll zerlegbar / analysierbar
- **Geringe syntakt. Variabilität:** 'ins Grass gebissen'
- Behandlung: Als Schema im Lexikon ablegen
- Herausforderung: Idiomatische vs. wörtliche(?) Lesart



"Das Pferd ist auf die Koppel gerannt und hat [ins Gras gebissen]."

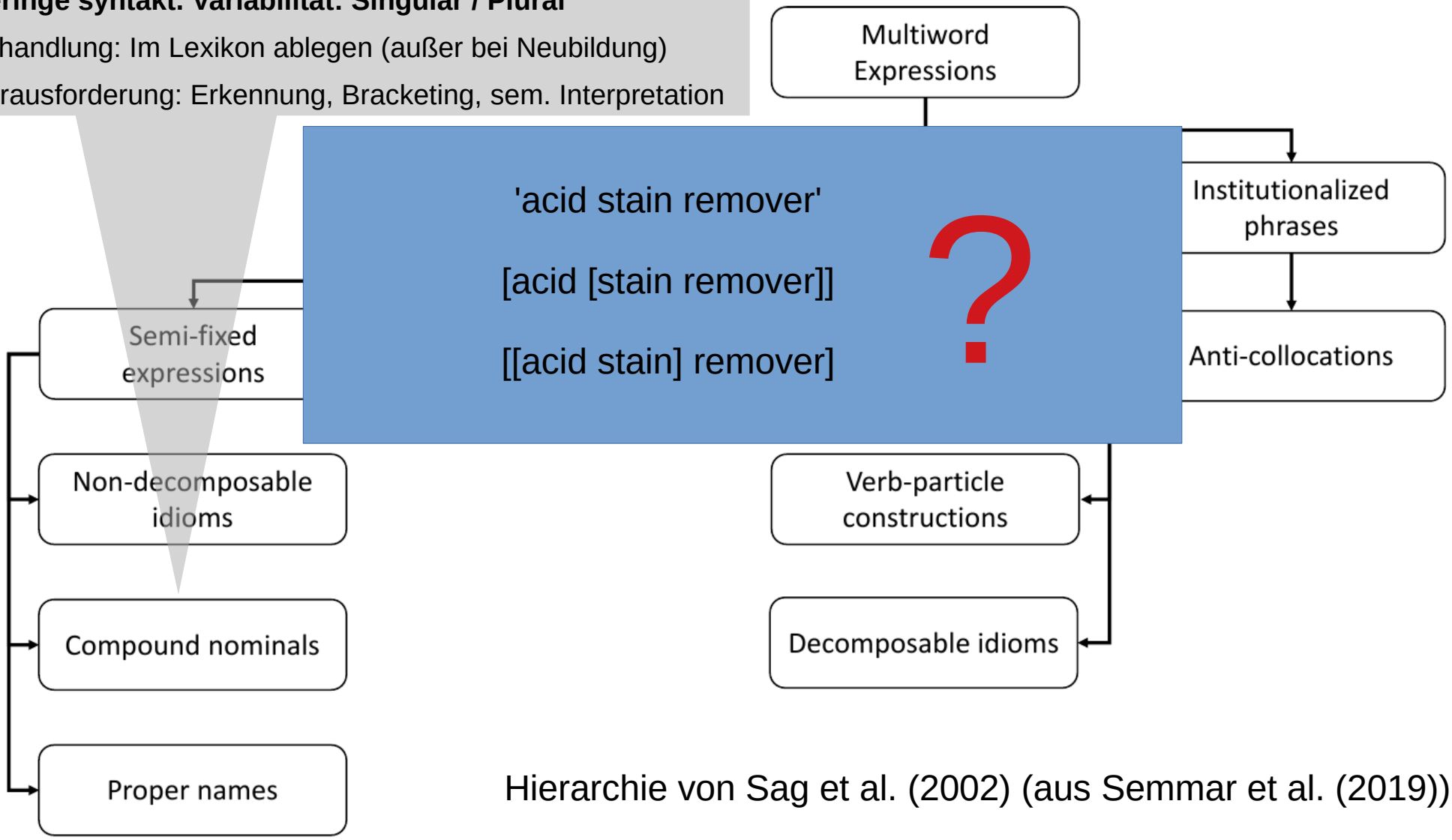
?

Hierarchie von Sag et al. (2002) (aus Semmar et al. (2019))

Übersicht *Multiword Expressions*

Beispiel: 'murder arrest', 'human stem cell research',

- Immer Nomen / Substantive, offene Klasse, sehr produktiv
- *Strukturell* zerlegbar / analysierbar; *semantisch* offen
- **Geringe syntakt. Variabilität: Singular / Plural**
- Behandlung: Im Lexikon ablegen (außer bei Neubildung)
- Herausforderung: Erkennung, Bracketing, sem. Interpretation

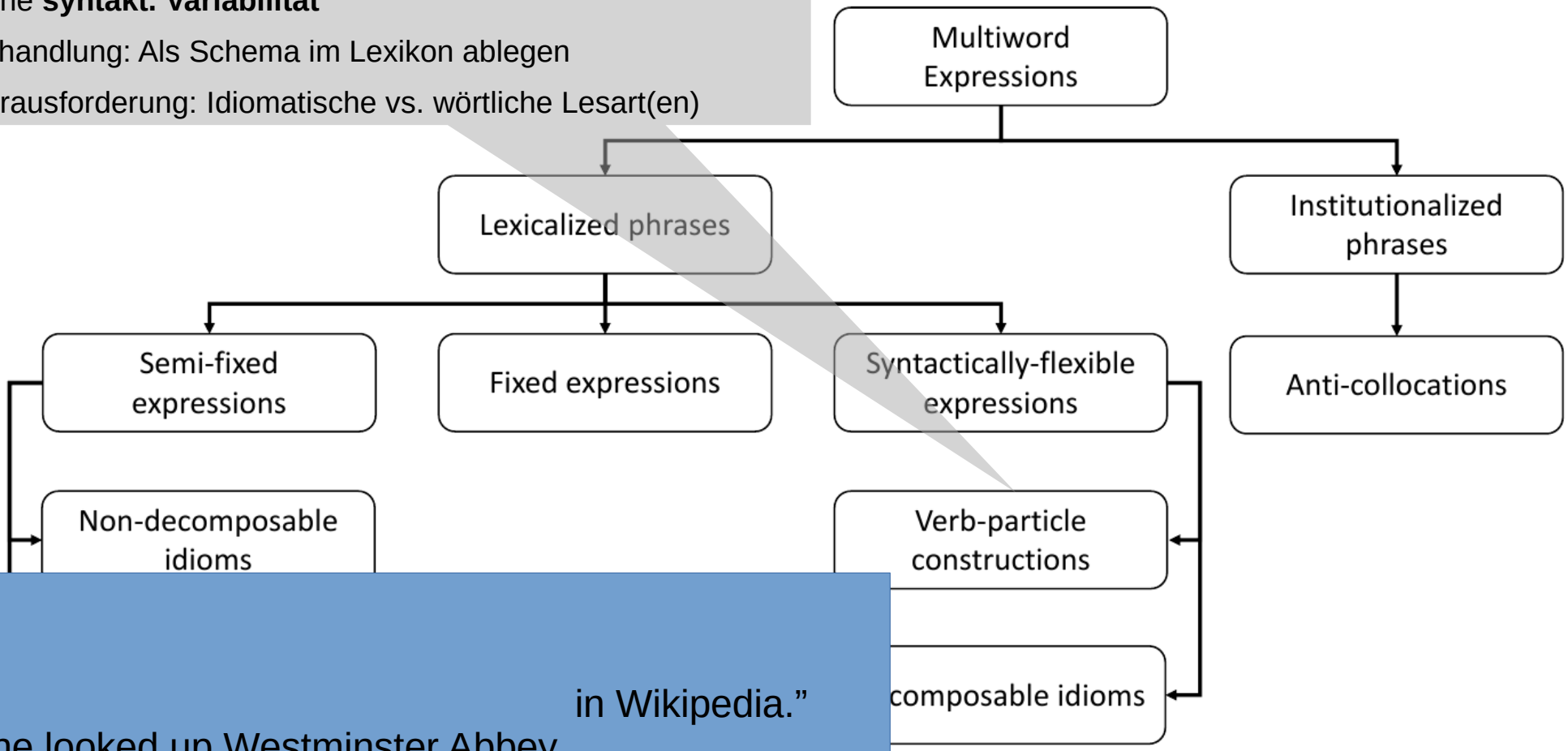


Hierarchie von Sag et al. (2002) (aus Semmar et al. (2019))

Übersicht *Multiword Expressions*

Beispiel: 'look up', 'call up', 'fight off'

- Immer Verb + Partikel
- *Strukturell* zerlegbar / analysierbar; *semantisch* offen
- Hohe **syntakt. Variabilität**
- Behandlung: Als Schema im Lexikon ablegen
- Herausforderung: Idiomatische vs. wörtliche Lesart(en)



“She looked up Westminster Abbey in Wikipedia.”

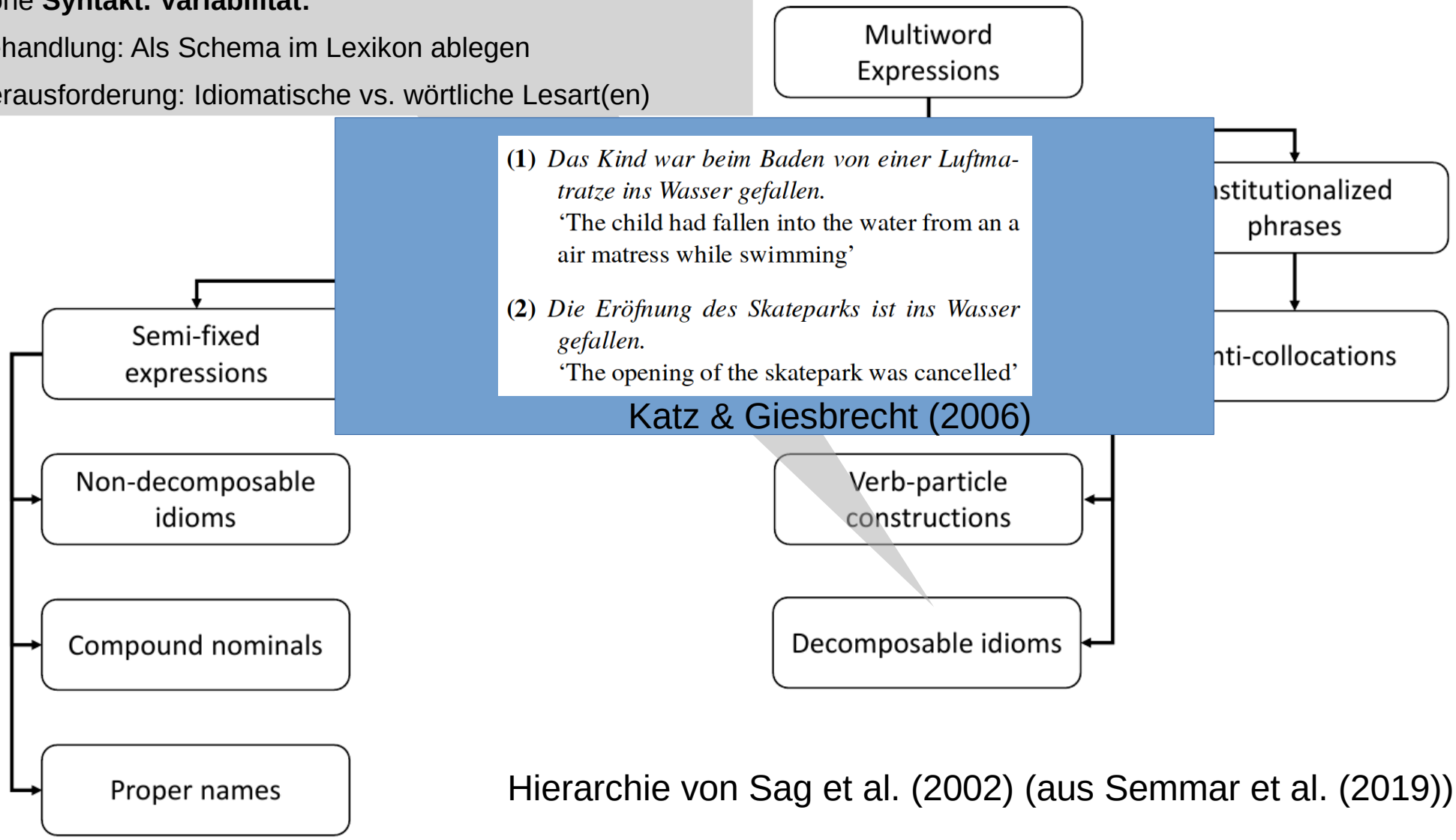
with admiration.”

et al. (2002) (aus Semmar et al. (2019))

Übersicht *Multiword Expressions*

Beispiel: 'sweep under the rug'; 'unter den Teppich kehren'

- Div. Wortarten, geschlossene Klasse
- *Strukturell* zerlegbar / analysierbar
- Hohe **Syntakt. Variabilität:**
- Behandlung: Als Schema im Lexikon ablegen
- Herausforderung: Idiomatische vs. wörtliche Lesart(en)



Hierarchie von Sag et al. (2002) (aus Semmar et al. (2019))

Übersicht *Multiword Expressions*

Bislang: **Wort** --> Bedeutung (=bezeichnetes **Konzept**)

Jetzt: **Konzept** --> mögliche **Wörter** / Versprachlichungen

Nur ein Bruchteil der sprachlich / semantisch möglichen Ausdrücke existiert / wird tatsächlich verwendet!

traffic light

?intersection regulator

?traffic director

.....

'Aufsehen'

'erregen'

'erzeugen'

'verursachen'

.....

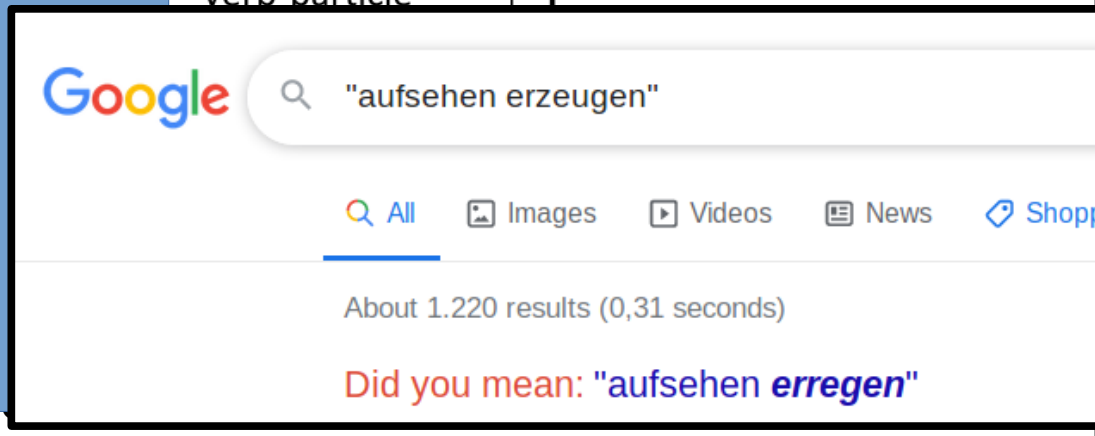
Multiword Expressions

Institutionalized phrases

Anti-collocations

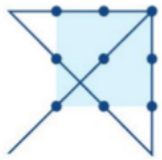
yntactically-flexible expressions

Verb-particle



Fragen?

Für die nächste Sitzung: 05.11.2019



- Church & Hanks (1990) lesen und eine Frage überlegen
- Dunning (1993) lesen und eine Frage überlegen
- Fragen per Mail
  - bis 04.11.2019, 12:00
  - mit Subject “MWE 05.11.2019”
  - an **mark-christoph.mueller@h-its.org**
- Vorbesprechung Referat vom 12.11.2019 (Menderes & Heide)
  - Di 05.11.2019