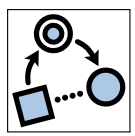
Mehrworteinheiten

MWE revisited: Baldwin & Kim (2010)

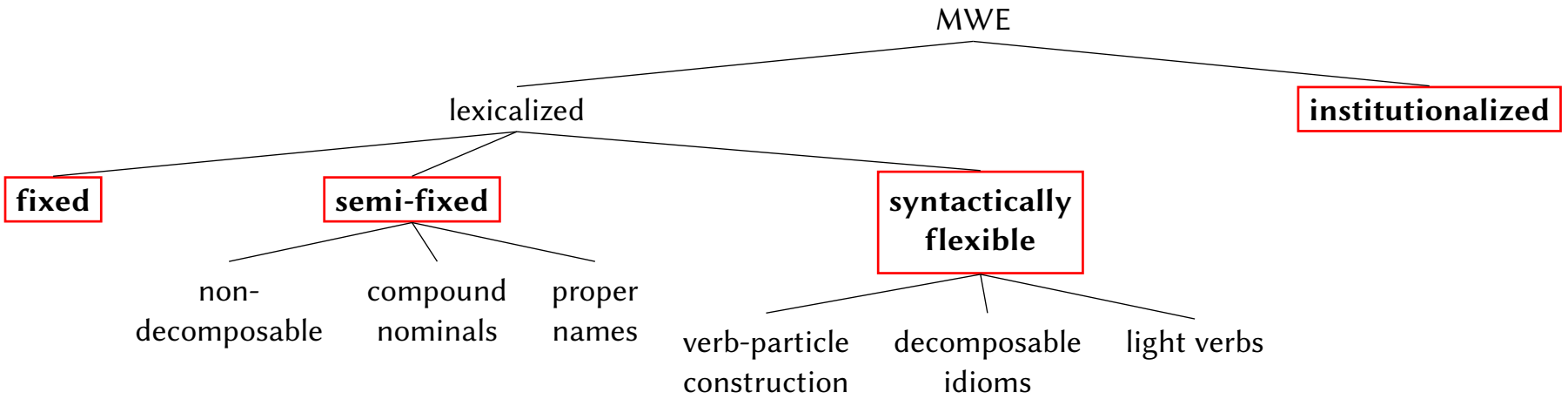
Timm Lichte

HHU Düsseldorf

3. Sitzung, 24.10.2017

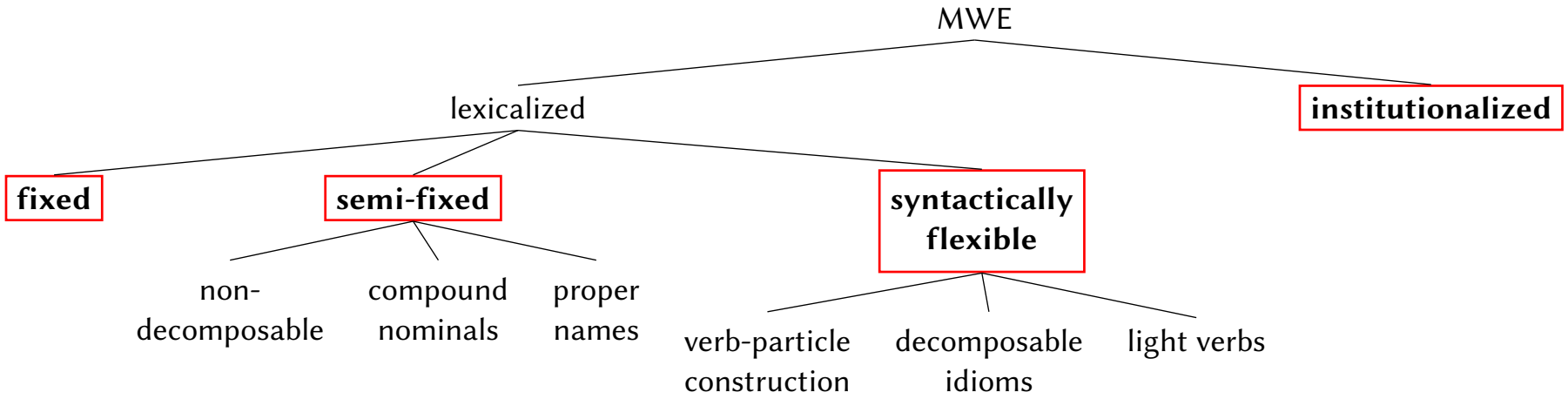


Rückblick: Sag et al. (2002)



MWE: idiosyncratic interpretations that cross word boundaries (or spaces)
fixed: defy conventions of grammar; no morphosyntactic variation; no internal modification
semi-fixed: strict constraints on word order; non-decomposable; some degree of lexical variation (inflection, relfexive form, determiner selection)
syntactically flexible: much wider range of syntactic variability (passivization, extraction), decomposable
institutionalized: statistically idiomatic; semantically and syntactically compositional

Rückblick: Sag et al. (2002)



	fixed	semi-fixed	flexible	instutional
morphological variation	−	+	+	+
internal modification	−	−?	+	+
passivization/extraction	−	−	+	+
decomposable	−	−	+	+
syntactically idiomatic	+	−	−	−
semantically idiomatic	+	+	+	−
statistically idiomatic	?	?	?	+

Status von Baldwin & Kim (2010)

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damera (eds.), Handbook of Natural Language Processing, 267–292. 2nd edn. Boca Raton, FL: CRC Press.

- Handbuchartikel = keine Darstellung neuer Daten und Theorien, aber Systematisierung und Pointierung der Forschungslage
- Fokus auf empirische Seite von “multiword expressions” (MWE)

Aufbau

1. Introduction: Core properties

Multiword expressions (informal definition)

expressions [...] which have surprising properties not predicted by their component words

(1) a. top dog ('person who is in charge')
b. dog days ('period of inactivity')

(2) a. In a nutshell, the administrator can take advantage of the database's many features through a single interface.
b. You should also jot down the serial number of your television video.

(3) a. She likes to take a long bath for relaxation after exams.
b. Kim hates to put her friends out.

Morphosyntax: nominal, verbial, adverbial
Semantics: components preserve or loose their original semantics
Syntax: internal modification, non-continguous

1. Introduction: Häufigkeit, Produktivität, Modellierung

Häufigkeit und Produktivität

- "same order of magnitude as the number of simplex words in a speaker's lexicon (Jackendoff 1997; Tschichold 1998; Pauwels 2000)"
- "MWE are continuously created as languages evolve (e.g. *shock and awe*, *carbon footprint*, *credit crunch*)"
- Vermutung: Jede Sprache hat MWEs.

Modellierung

- "a modest body of research on modelling MWEs which has been integrated into NLP applications"
- "used heavily" in "machine translation"
- "Explicit lexicalised MWE data helps [...]"
  - Vereinfachung der syntaktischen Strukturen
  - Vermeidung von Parsefehlern
  - Verbesserung beim semantischen Taggen
  - Verbesserung der Wort-Alinierung bei MT

2. Linguistic properties of MWEs: Definition

Multiword expressions (final definition)

Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity.

"decomposed into multiple lexems" ≠ "made up of multiple whitespace-delimited words"

marketing manage versus *lighthouse*

⇒ "we tend to relax this restriction and allow for **single-word MWEs**"

2.1 Linguistic properties of MWEs: Idiomaticity

Idiomaticity

markedness or deviation from the basic properties of the component lexemes

Compositionality

the degree to which the features of the parts of a MWE combine to predict the features of the whole

"While compositionality is often construed as applying exclusively to semantic idiomatic (hence by 'non-compositional MWE', researchers tend to mean a semantically-idiomatic MWE), in practice it can apply across all the same levels as idiomaticity."

2.1 Linguistic properties of MWEs: Types of idiomaticity

Lexical idiomaticity

Lexical idiomaticity occurs when one or more components of an MWE are not part of the conventional English lexicon.

(4) ad hoc

Syntactic idiomaticity

Syntactic idiomaticity occurs when the syntax of the MWE is not derived directly from that of its components.

(5) by and large (anomalous coordination ⇒ adverbial)

Can also occur at a "constructional" level: verb-particle constructions, determinerless PPs

2.1 Linguistic properties of MWEs: Types of idiomaticity

Semantic idiomaticity

Semantic idiomaticity is the property of the meaning of a MWE not being explicitly derivable [predictable] from its parts.

Not predictable:

(6) middle of the road ('non-extremism, especially in political views')

Partial predictability:

(7) blow hot and cold ('constantly change opinion')

Predictable with additions:

(8) bus driver ('one who drives ~~like~~ a bus')

Decomposable (with varying degree):

(9) spill the beans (`reveal'(secret')`)

2.1 Linguistic properties of MWEs: Types of idiomaticity

Pragmatic idiomaticity

Pragmatic idiomaticity is the condition of a MWE being associated with a fixed set of situations or a particular context.

(10) good morning, all aboard

2.1 Linguistic properties of MWEs: Types of idiomaticity

Statistical idiomaticity

Statistical idiomaticity occurs when a particular combination of words occurs with markedly high frequency, relative to the component words or alternative phrasings of the same expression.

	flawless	immaculate	impeccable	spotless	
condition	+	−	+	+	
credentials	−	−	+	−	
hair	−	+	?	−	
house	?	+	?	+	
logic	+	−	+	−	(from Cruse 1986)

(11) a. black and white television
b. #white and black television (from Benor and Levy 2006)

2.2 Other properties of MWEs

MWE haben oft, **aber nicht immer**, die folgenden Eigenschaften:

- Crosslingual variation
- Single-word paraphrasability
- Proverbiality (indicators of more informal situations)
- Distinct prosody

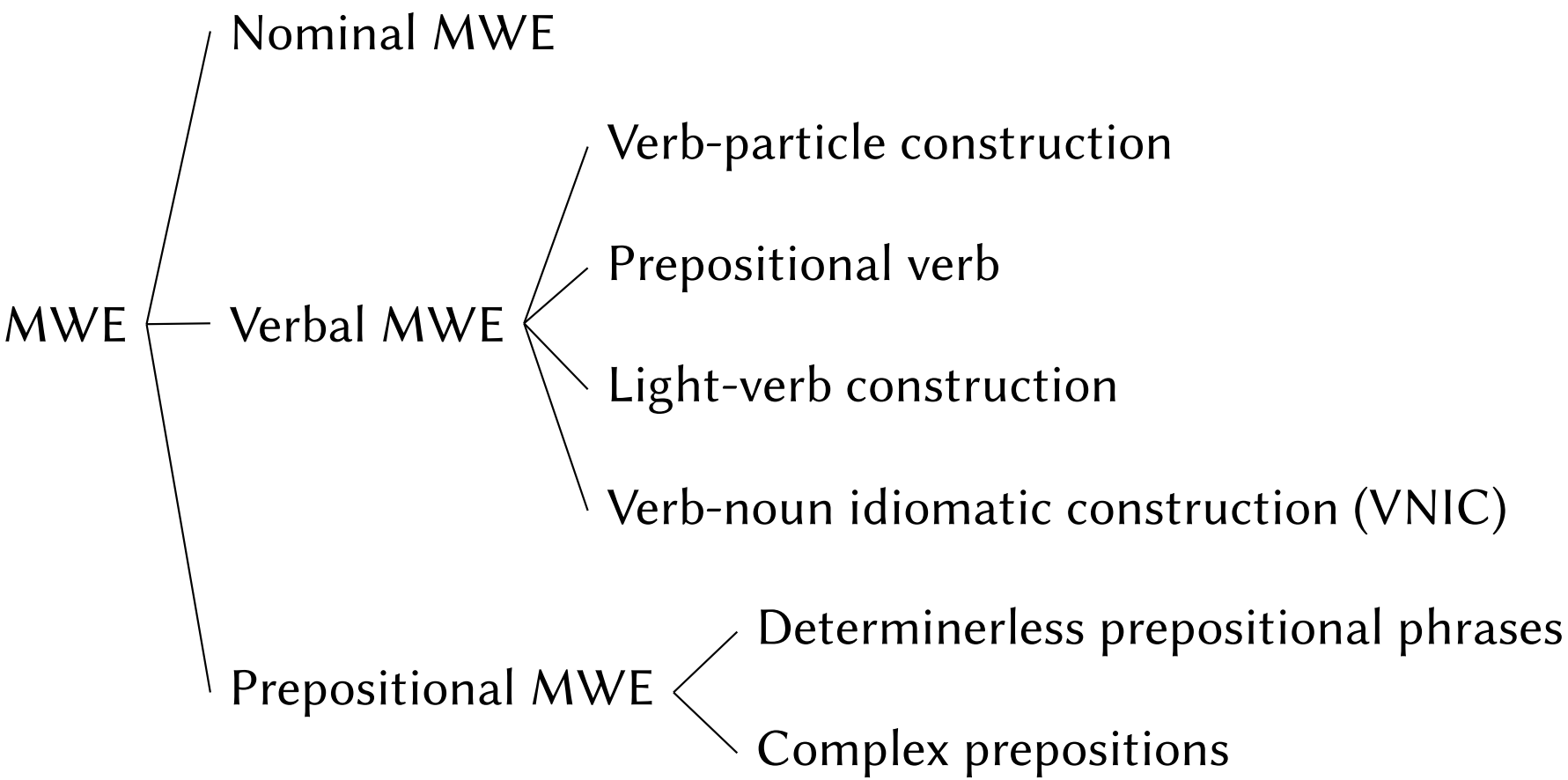
2.3 Testing an expression for MWEhood

Zusammenfassung

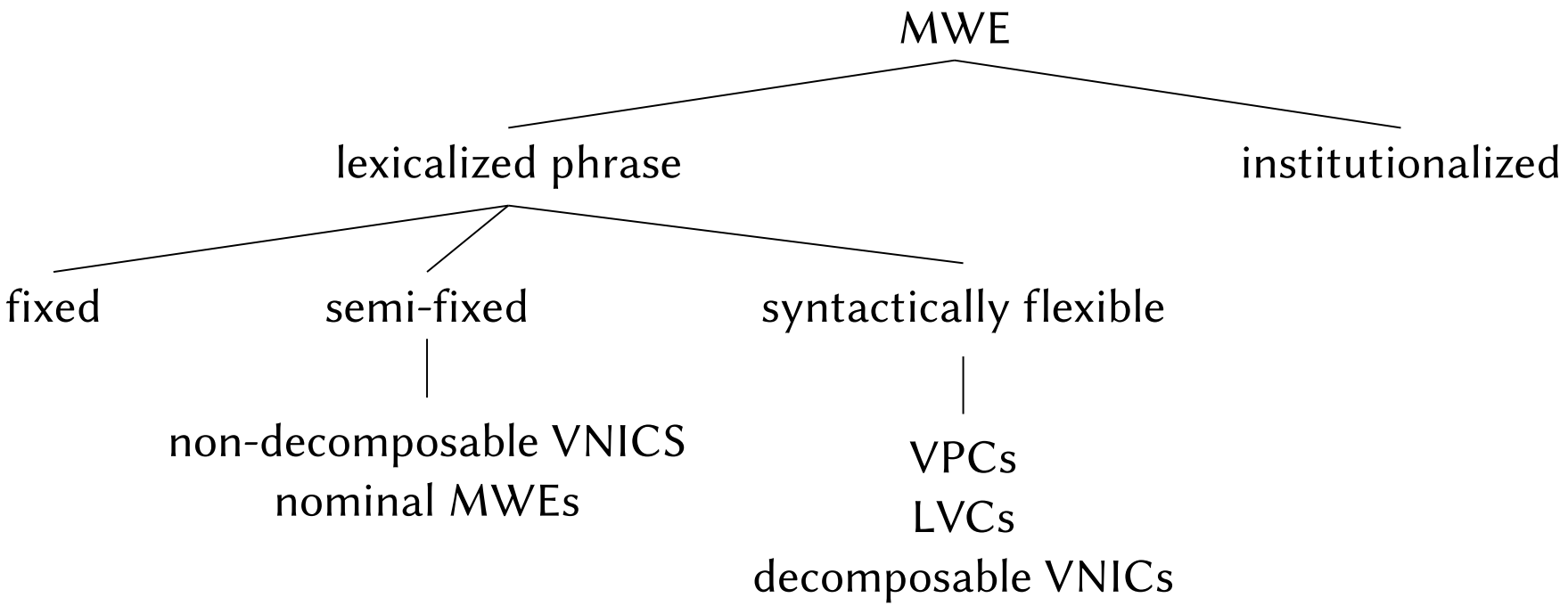
	Lexical	Syntactic	Semantic	Pragmatic	Statistical
all aboard	–	–	–	+	+
bus driver	–	–	+	–	+
by and large	–	+	+	–	+
kick the bucket	–	–	+	–	+
look up	–	–	+	–	+
shock and awe	–	–	–	+	+
social butterfly	–	–	+	–	+
take a walk	–	–	+	–	?
to and fro	?	+	–	–	+
traffic light	–	–	+	–	+
eat chocolate	–	–	–	–	–

collocation = statistically idiomatic MWE

3. Types of MWE



4. MWE classification

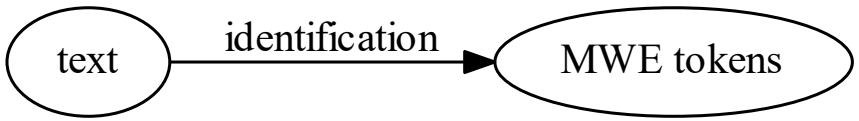


(aus Sag et al. (2002))

5.1 Research issues: identification

Identification

the task of determining individual occurrences of MWEs in running text



(12) One fine evening a young princess put on her bonnet and clogs, and went out to take a walk by herself in a wood; ... she ran to pick it up; ...

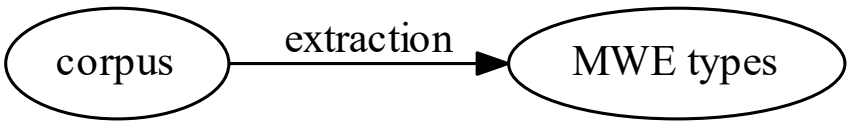
Disambiguation via:

- tagger, chunker, parser
- word sense disambiguation
- canonical form

5.2 Research issues: extraction

Extraction

the MWE lexical items attested in a predetermined corpus are extracted out into a lexicon



Methoden:

- collocation extraction via association measures (pattern-based)
- relative similarity between MWE components and the context (WSD-like)
- syntactic fixedness

5.3–5.4 Further research issues

Internal syntactic disambiguation

- bracketing: [[*glass window*] *cleaner*] vs. [*glass* [*window cleaner*]]

MWE interpretation

- semantic relations between components of nominal compunds
- paraphrases
- relative to a generalised semantic inventory (using, e.g., WordNet)

Fazit

Baldwin & Kim (2010) entwickeln eine Taxonmie der Idiomatizität (mit einem generalisierten Kompositionalitätsbegriff), die die Taxonomie von Sag et al. (2002) (fixed, semi-fixed, syntactically flexible) nicht ersetzen, sondern ergänzen soll.

References

[1] Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damera (eds.), *Handbook of natural language processing*, 2nd, 267–292. Boca Raton, FL: CRC Press.

[2] Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing* (Lecture Notes in Computer Science 2276), 1–15. Berlin: Springer.