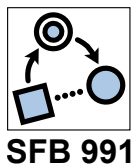
Mehrworteinheiten

MWE und NLP: Sag et al. (2002)

Timm Lichte

HHU Düsseldorf

2. Sitzung, 17.10.2017



Status von Sag et al. (2002)

- Übersichtsartikel = keine Darstellung neuer Daten und Theorien, aber Systematisierung und Pointierung der Forschungslage
- Fokus auf die Behandlung von “multiword expressions” (MWE) beim NLP
- Achtung: historischer Artikel von 2002 (vor Deep Learning, Distributioneller Semantik, …)

Aufbau

1. Introduction
2. Some kinds of MWE
    1. Fixed expressions
    2. Semi-fixed expressions
    3. Syntactically-flexible expressions
    4. Institutionalized phrases
3. Some analytic techniques
4. Conclusion

Abstract: Forschungshypothesen

The various kinds of multiword expressions should be analyzed in distinct ways, including listing “words with spaces”, hierarchically organized lexicons, restricted combinatoric rules, lexical selection, “idiomatic constructions” and simple statistical affinity. (p. 1)

An adequate comprehensive analysis of multiword expressions must employ both symbolic and statistical techniques. (p. 1)

1. Introduction: “Two key problems”

Disambiguation

- Domäne: “grammar development”
- “linguistic precision is inversely correlated with degree of sentence ambiguity”
- “Knowledge representation [...] has largely failed to provide completely satisfactory solutions.”
- stattdessen: stochastische Methoden

Multiword expressions

- “insufficient ongoing work investigating the nature of this problem or seeking computationally tractable techniques”

1. Introduction: MWE-Definition und -Status

Definition

idiosyncratic interpretations that cross word boundaries (or spaces)

Mengenverhältnis zu nicht-MWEs

- “same order of magnitude” (Jackendoff 1997)
- 41% in WordNet 1.7
- “almost certainly an underestimate”

1. Introduction: Probleme für NLP

I. Overgeneration problem

(1) a. telephone booth (American)
telephone box (British/Australian)
b. #telephone cabinet, telephone closet

II. Idiomaticity problem

(2) a. kick the bucket (‘die’)
b. in ~~the~~ line

III. Flexibility problem (of words-with-spaces approaches)

(3) a. look up the tower (‘glance up at the tower’, ‘consult a reference book about the tower’)
b. look the tower up (‘consult a reference book about the tower’)

IV. Lexical proliferation problem (of words-with-spaces approaches)

(4) take a walk, take a hike, take a trip, take a flight (LVC)

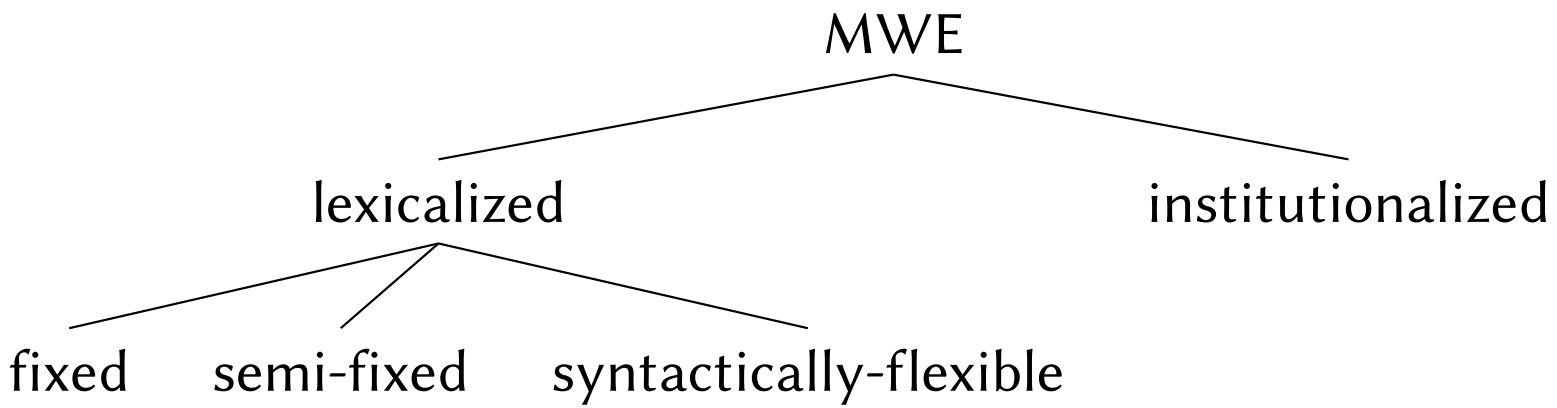
1. Introduction: State of the art

Large-scale, linguistically precise computational grammars:

- ParGram (LFG)
- XTAG (TAG)
- CCG
- LingGO (HPSG)
- FrameNet (CxG?)

"All of these projects are currently engaged (to varying degrees) in linguistically informed investigations of MWEs."

2. MWE-Arten



lexicalized

"partially **idiosyncratic** syntax or semantics, or contain 'words' which do not occur in isolation"

institutionalized

"syntactically and semantically **compositional**, but occur with markedly high frequency"

2.1 MWE-Arten: fixed

(5) a. by and large, in short, kingdom come, every which way
b. ad hoc, Palo Alto

fixed

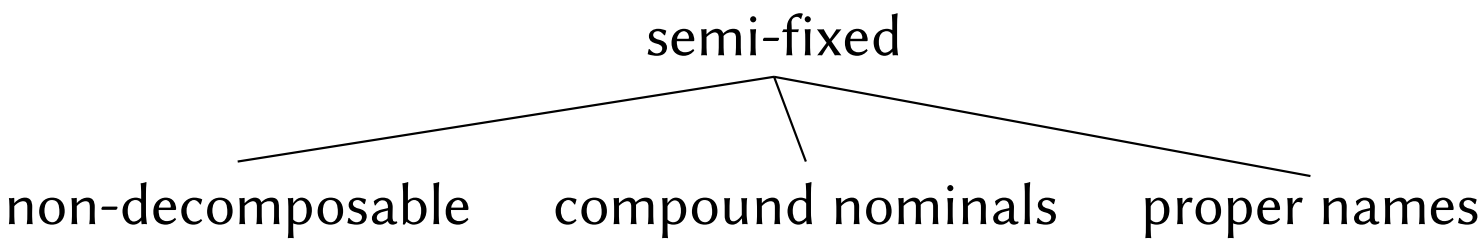
"immutable expressions in English that defy conventions of grammar and compositional interpretation"

"fully lexicalized and undergo neither morphosyntactic variation (cf. * *in shorter*) nor internal modification (cf. * *in very short*))"

2.2 MWE-Arten: semi-fixed

Semi-fixed

"Semi-fixed expressions adhere to **strict constraints on word order and composition**, but undergo **some degree of lexical variation**, e.g. in the form of inflection, variation in reflexive form, and determiner selection."



2.2 MWE-Arten: semi-fixed: semantic decomposability

Semantic decomposability

"describing how the overall sense of a given idiom is related to its parts" (formerly called semantic compositionality in Nunberg, Sag & Wasow 1994)

Decomposable

(6) a. spill the beans ('reveal secrets')
    b. let the cat out of the bag ('reveal secrets')

Non-decomposable

(7) a. kick the bucket ('die')
    b. trip the light fantastic ('Tanzbein schwingen')
    c. shoot the breeze ('Small-Talk halten')

2.2 MWE-Arten: semi-fixed: non-decomposable

(8) a. kick the bucket (‘die’)
b. trip the light fantastic (‘Tanzbein schwingen’)
c. shoot the breeze (‘Small-Talk halten’)

Morphosyntaktische Eigenschaften

- keine interne Modifikation (# *kick the great bucket in the sky*)
- keine Passivierung (* *the breeze was shot*)
- aber: Flexiblität bei der Flexion (*kicked the bucket*)
- aber: Flexiblität bei der Reflexivform (*wet oneself*)

Modellierung

- words-with-spaces ⟶ lexical proliferation
- fully compositional approach ⟶ idiomaticity and overgeneration problem

2.2 MWE-Arten: semi-fixed: compound nominals

(9) car park (‘Parkplatz’), attorney general (‘Justizminister’), part of speech (‘Wortart’)

Morphosyntaktische Eigenschaften

- “syntactically unalterable”
- aber: Flexiblität bei der Flexion (*attorneys general*)

Modellierung

- words-with-spaces ⟶ lexical proliferation
- fully compositional approach ⟶ idiomaticity and overgeneration problem

2.2 MWE-Arten: semi-fixed: proper names

(10) a. the San Francisco 49ers
b. the Orlando Raiders

Morphosyntaktische Eigenschaften

- "syntactically highly idiosyncratic"
- Teile sind elidierbar (*the (San Francisco) 49ers*)
- Definitartikel nicht immer obligatorisch (*the 49ers player*, *the 49ers and raiders*)

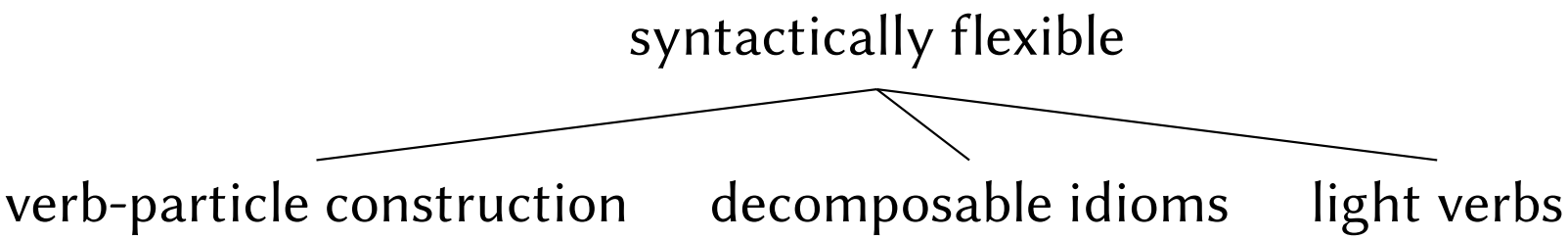
Modellierung

- words-with-spaces ⟶ lexical proliferation
- fully compositional approach ⟶ idiomaticity and overgeneration problem

2.3 MWE-Arten: syntactically flexible

syntactically flexible

"Whereas semi-fixed expressions retain the same basic word order throughout, syntactically-flexible expressions exhibit a much wider range of syntactic variability."



2.3 MWE-Arten: synt. flexible: verb-particle constructions

(11) write up, look up, break up, brush up on sth.

Semantische Eigenschaften

- "semantically idiosyncratic" oder "compositional"

Morphosyntaktische Eigenschaften

- transitiv → NP-Argument steht zwischen Verb und Partikel oder nach dem Partikel. (* *fall a truck off*)
- Adverbien können manchmal zwischen Verb und Partikel stehen. (*fight bravely on*)

Modellierung

- words-with-spaces ⟶ "impossible to capture" (lexical proliferation)
- fully compositional approach ⟶ idiomaticity and overgeneration problem

2.3 MWE-Arten: synt. flexible: decomposable idioms

(12) let the cat out of the bag, sweep under the rug

Morphosyntaktische Eigenschaften

- "tend to be syntactically flexible to some degree"
- "highly unpredictable"

Modellierung

- words-with-spaces ⟶ "incompatible"
- fully compositional approach ⟶ idiomaticity and overgeneration problem
- "semantic approach" (Nunberg, Sag & Wasow 1994)

2.3 MWE-Arten: synt. flexible: **light verbs**

(13) make a mistake, give a demo

Morphosyntaktische Eigenschaften

- "full syntactic variability"
- Passivierung (*a demo was given*)
- Extraktion (*How many demos did Kim give?*)
- interne Modifikation (*give a revealing demo*)

Modellierung

- words-with-spaces ⟶ nicht möglich
- fully compositional approach ⟶ overgeneration problem (* *make a demo*)

2.4 MWE-Arten: institutionalized phrases

Institutionalized phrases

"semantically and syntactically compositional, but statistically **idiomatic**"

(14) traffic light, telephone booth, fresh air, kindle exitement

"Statistisch idiomatisch"

- "conventionalized"
- "much higher relative frequency than any alternative lexicalization of the concept" (*traffic director*, anit-colloctation)
- "deminished decomposability" (*traffic light* 'Blinker')

Modellierung

- words-with-spaces ⟶ "incompatible"
- fully compositional approach ⟶ idiomaticity and overgeneration problem

Übung

fixed – semi-fixed – flexible – institutionalized

(15) jemanden im Stich lassen

(16) ins Gras beißen

(17) spill the beans

(18) ein Bad nehmen

(19) kingdom come

(20) Telefonhäuschen

(21) Gib's auf.

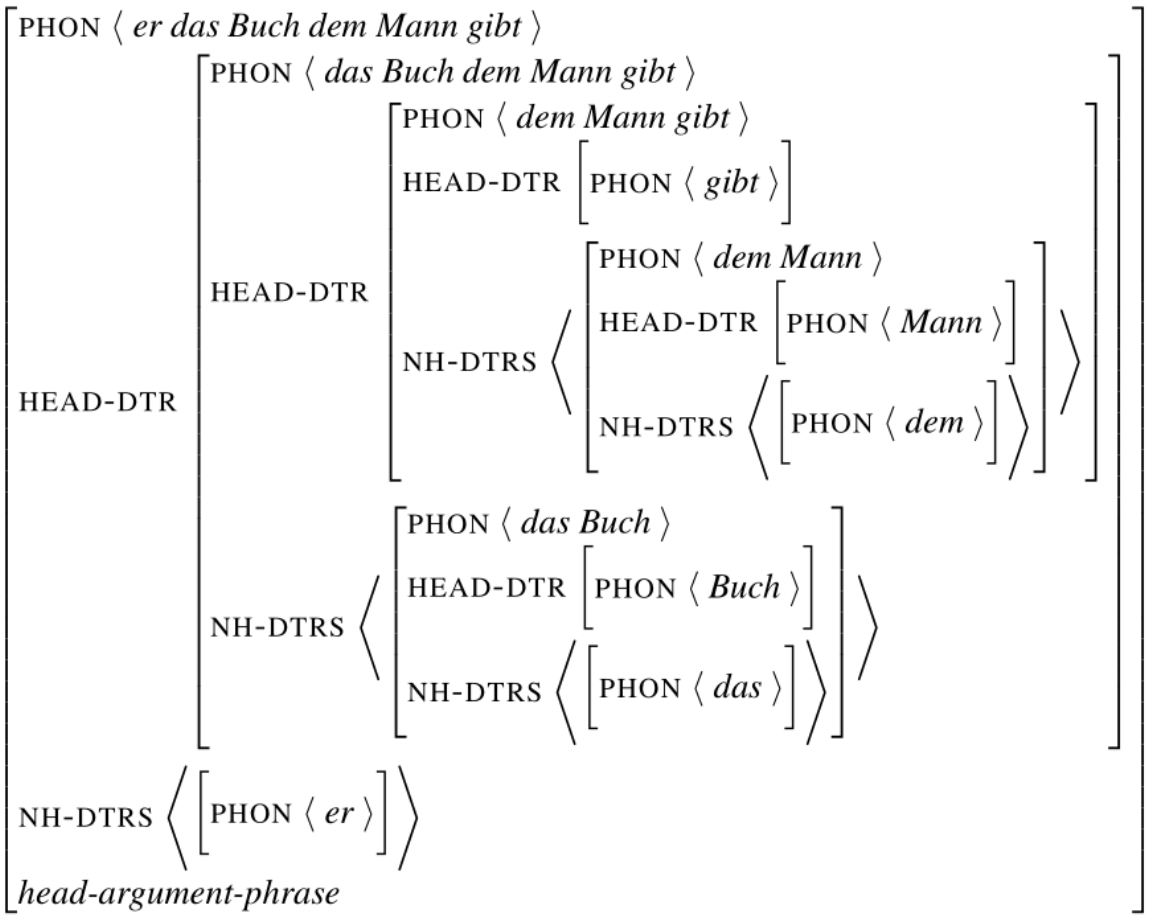
(22) bis der Arzt kommt
Ich glaube, mich tritt ein Pferd.

(23) John loaded the hay onto the wagon.
John loaded the wagon with the hay.

3. Analysetechniken: Allgemeines zur HPSG

HPSG (Head-driven Phrase Structure Grammar)[1,3]

- constraint-basierter Grammatikformalismus
- Modelle sind getypte Merkmalsstrukturen.



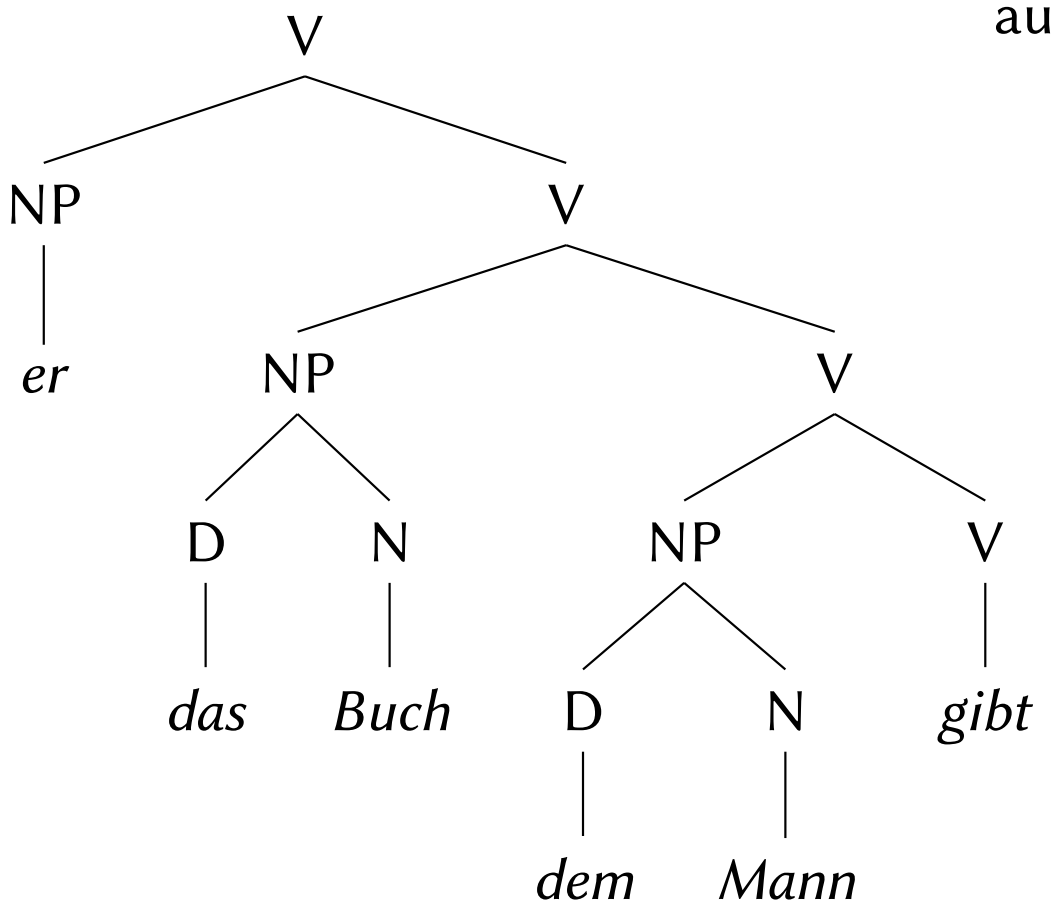
aus Müller [1: 55]

3. Analysetechniken: Allgemeines zur HPSG

HPSG (Head-driven Phrase Structure Grammar)[1,3]

- constraint-basierter Grammatikformalismus
- Modelle sind getypte Merkmalsstrukturen.

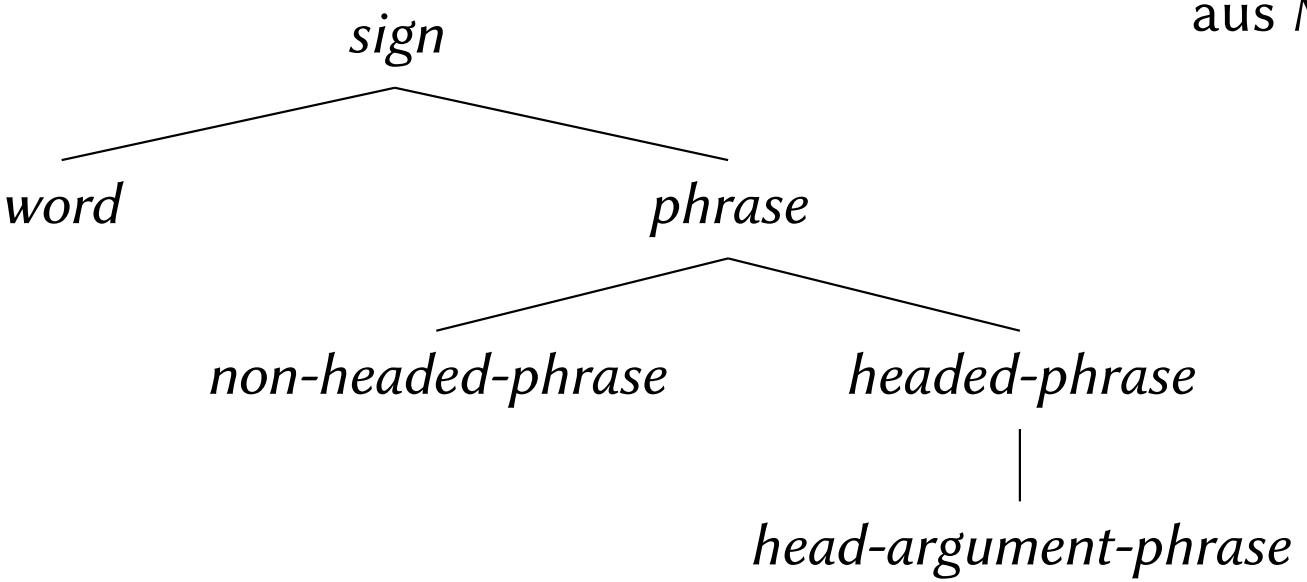
aus Müller [1: 55]



HPSG (Head-driven Phrase Structure Grammar)[1,3]

- Die Theorie steckt in Typenbeschränkungen und “Prinzipien”.

aus Müller [1]



$$\textit{headed phrase} \Rightarrow \begin{bmatrix} \textsc{head} & \boxed{1} \\ \textsc{head-dtr}|\textsc{head} & \boxed{1} \end{bmatrix}$$

HPSG (Head-driven Phrase Structure Grammar)[1,3]

- Das Lexikon liegt in der Typenbeschränkung von *word*:

$$word \Rightarrow gibt \vee dem \vee Buch \ldots$$

gibt:

$$\begin{bmatrix} \text{PHON} & \langle\, gibt \,\rangle \\ \text{HEAD} & \begin{bmatrix} \text{VFORM} & fin \\ verb & \end{bmatrix} \\ \text{SUBCAT} & \langle \text{NP}[nom], \text{NP}[acc], \text{NP}[dat] \rangle \\ word & \end{bmatrix}$$

aus Müller [1: 55]

3. Analysetechniken: fixed expressions

Words-with-spaces

```
ad_hoc_1  :=  intr_adj_l &
   [ STEM  <"ad", "hoc">,
     SEMANTICS  [KEY ad-hoc_rel] ].
```

3. Analysetechniken: semi-fixed expressions

Internal inflection

```
part_of_speech_1 := intr_noun_l &
    [ STEM < "part", "of", "speech" >,
      INFL-POS "1",
      SEMANTICS [KEY part_of_speech_rel ]].
```

Hierarchical lexicon with default constraint inheritance

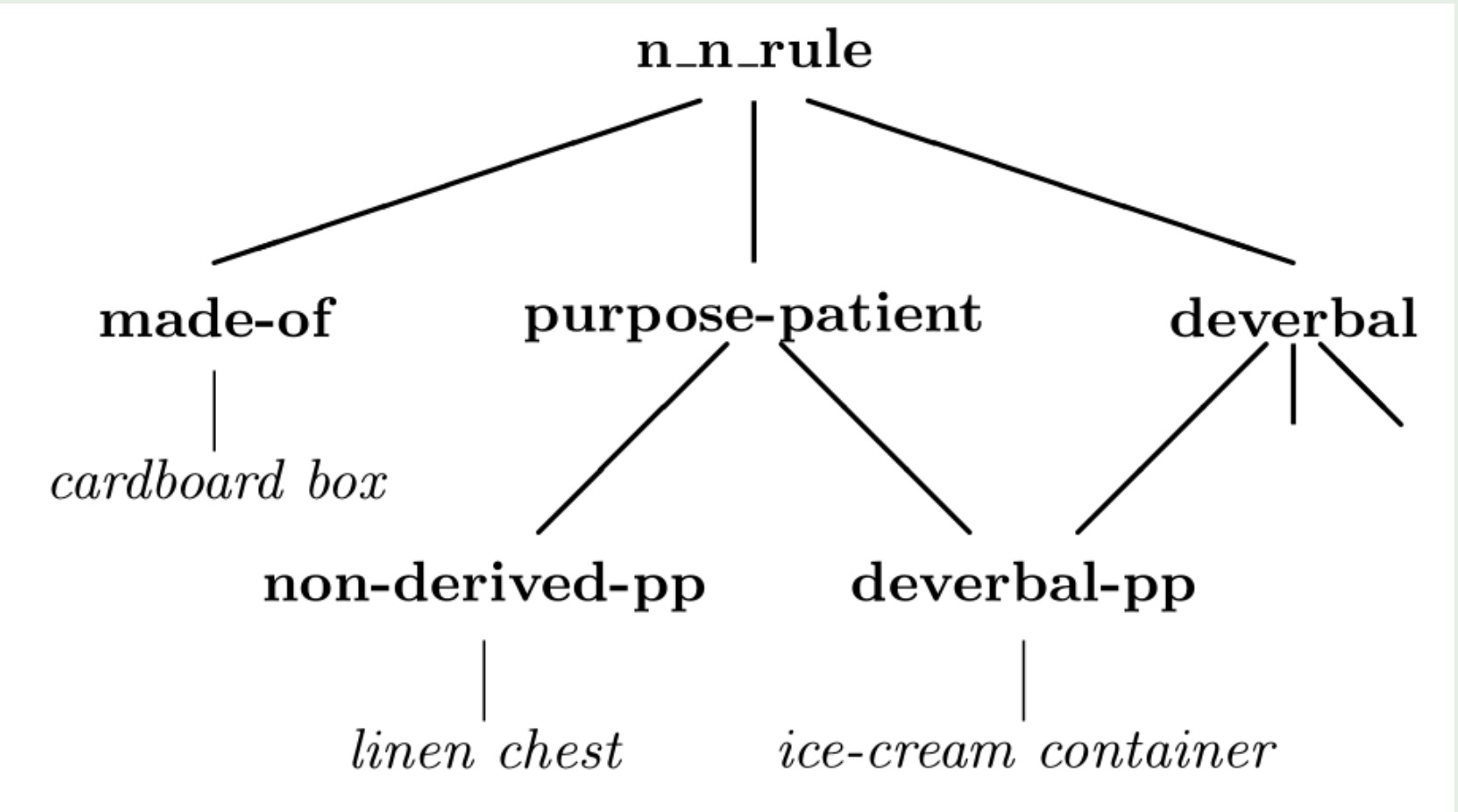
```
Name: [SPR / < > ]
USTeamName: [SPR < Det[definite] >, NUM / plural]
MassTeamName: [NUM singular]

oakland_raiders_1 := USTeamName &
    [ LEX-SIGNS / < oakland_1, raiders_1 >,
      SEMANTICS < oakland_raiders_rel > ].
```

(24) a. the (Oakland) Raiders
b. the (Miami) Heat

Circumscribed constructions

Vererbungshierarchien für die Modellierung von Semi-Produktivität:



3. Analysetechniken: syntactically-flexible expressions

Lexical Selection

"a sign associated with one word of the phrase selects for the other word(s)"

```
hand_out_v1 := mv_prep_particle_np_l &
    [ STEM < "hand" >,
      SEMANTICS [ KEY hand_out_rel,
                  --COMPKEY out_rel ] ].
```

Light verb: Das Verb (*make*) selegiert ein Nomen (*mistake*) mittels eines semantisch LVC-kompatiblen Typs (`make-arg-rel`).

Decomposable idioms: lexikalische Selektion oder Quasi-Inferenz

3. Analysetechniken: institutionalized expressions

- “treatment of frequency”?
- “semantic probabilities” … ???

4. Zusammenfassung

In this paper we hope to have shown that MWEs, which we have ***classified in terms of lexicalized phrases*** *(made up of fixed, semi-fixed and syntactically flexible expressions) and institutionalized phrases, are far more diverse and interesting than is standardly appreciated. Like the issue of disambiguation, MWEs constitute a key problem that must be resolved in order for linguistically precise NLP to succeed. Our goal here has been primarily to* ***illustrate the diversity of the problem****, but we have also* ***examined known techniques*** *— listing words with spaces, hierarchically organized lexicons, restricted combinatoric rules, lexical selection, idiomatic constructions, and simple statistical affinity.*

References

[1] Müller, Stefan. 2013. *Head-Driven Phrase Structure Grammar: Eine Einführung*. 3rd edn. (Stauffenburg Einführungen 17). Tübingen: Stauffenburg Verlag. `http://hpsg.fu-berlin.de/~stefan/Pub/hpsg-lehrbuch.html`.

[2] Nunberg, Geoffrey, Ivan A Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

[3] Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago & London: University of Chicago Press.

[4] 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.