STS: The productivity of collocations

Anke Lüdeling
University of Osnabrück
aluedeli@uos.de

Thanks to Stefan Evert!

Collocations in computational linguistics

- identification/acquisition of collocations (for all kinds of applications from lexicography to parsing)
- (lexicographic) description of properties of collocations with focus on syntactic variability & usage (corpus evidence)
- underlying assumption: there is a finite number of collocations that can be distinguished from free syntactic combinations

Productive collocations?

- is the number of collocations in a given language L really finite?
- (is it possible to find and list all collocations in L?)
- or is there some element of productivity?
- if the formation of collocations is productive in some sense: what are the consequences for
  - the theoretical treatment of 'collocation'?
  - the treatment of collocations in computational linguistics?

Outline

- two families of definitions of 'collocation' and their implications with respect to productivity
- productivity in word formation
- case studies: productivity in collocations
- consequences for the definitions
- consequences for computational linguistics

Definitions of 'collocation' I: co-occurring words

"Collocations of a given word are statements of the habitual or customary places of that word." (Firth 1957, 181)

- finite number
- relation between collocations?
- collocations can only be learned from experience

Why do words co-occur?

- facts of life, 'corpus' domain, subcategorization
  Hund bellen, Kaffee trinken
  Glühwein trinken vs. *Glühsaft trinken*
  → compositional
- non-compositionality (in the widest sense)
  rote Rose, müde Mark, zum Streit kommen
- convention (clichés): *Zähne putzen, deine Freunde sind auch meine Freunde*

Definitions of 'collocation' II: semantic iregularity

"A collocation is defined as a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components."
(Choueka 1988)

semantic irregularity

- finite number
- relation between collocations?
- (in theory): collocations can be learned without linguistic experience
- (consecutive?)

Case study: *blau*

blaue Stunde	*41*	Firth, Choueka
blauer Himmel	*33*	Firth
blaue Blume	*29*	Firth, Choueka?
blaues Wunder	*5*	Firth?, Choueka
blaue Tarifzone	*1*	Choueka
blaues Stündchen	*1*	Choueka

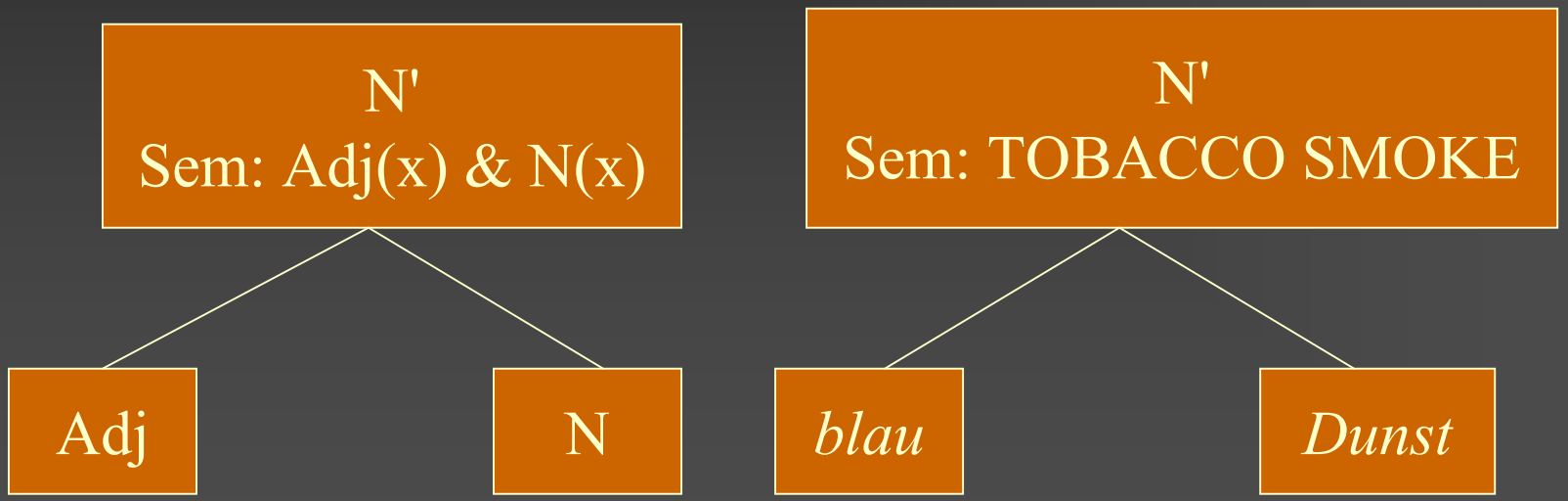
Definitions and co-occurrence measures

- co-occurrence measures find words that co-occur frequently (with some statistical significance): Firth
- manual correction of results according to linguistic criteria or 'intuition': Choueka

- again: both approaches assume that there is a **finite number** of collocations

finiteness

free combinations vs. collocations



productivity: complex words I

- productivity is a notion discussed in morphology
- to describe:
  - morphological processes are used to form new words according to given regularities
    -*bar* takes transitive verbs to form adjectives:
    lesbar 'readable', *interpretierbar* 'interpretable', ...
    -verb stems combine with nouns to form compounds:
    Weckruf 'wake-up call', *Esszimmer* 'dining room', ...
  - some processes seem to generate more new words than others
    -bar creates more new adjectives than *-sam*

productivity: complex words II

- qualitative aspect of productivity: detailed linguistic description of a morphological process
- quantitative aspect (Baayen 1992): some kind of measures for the productivity of a morphological process

cognitive perspective I

- how do speakers know that they can form new words of a given pattern?
- hypothesis:
  distribution of words formed by that pattern
- large portion of low frequency words indicates that new words can be formed
- small portion of low frequency words indicates that no new words can be formed

cognitive perspective II

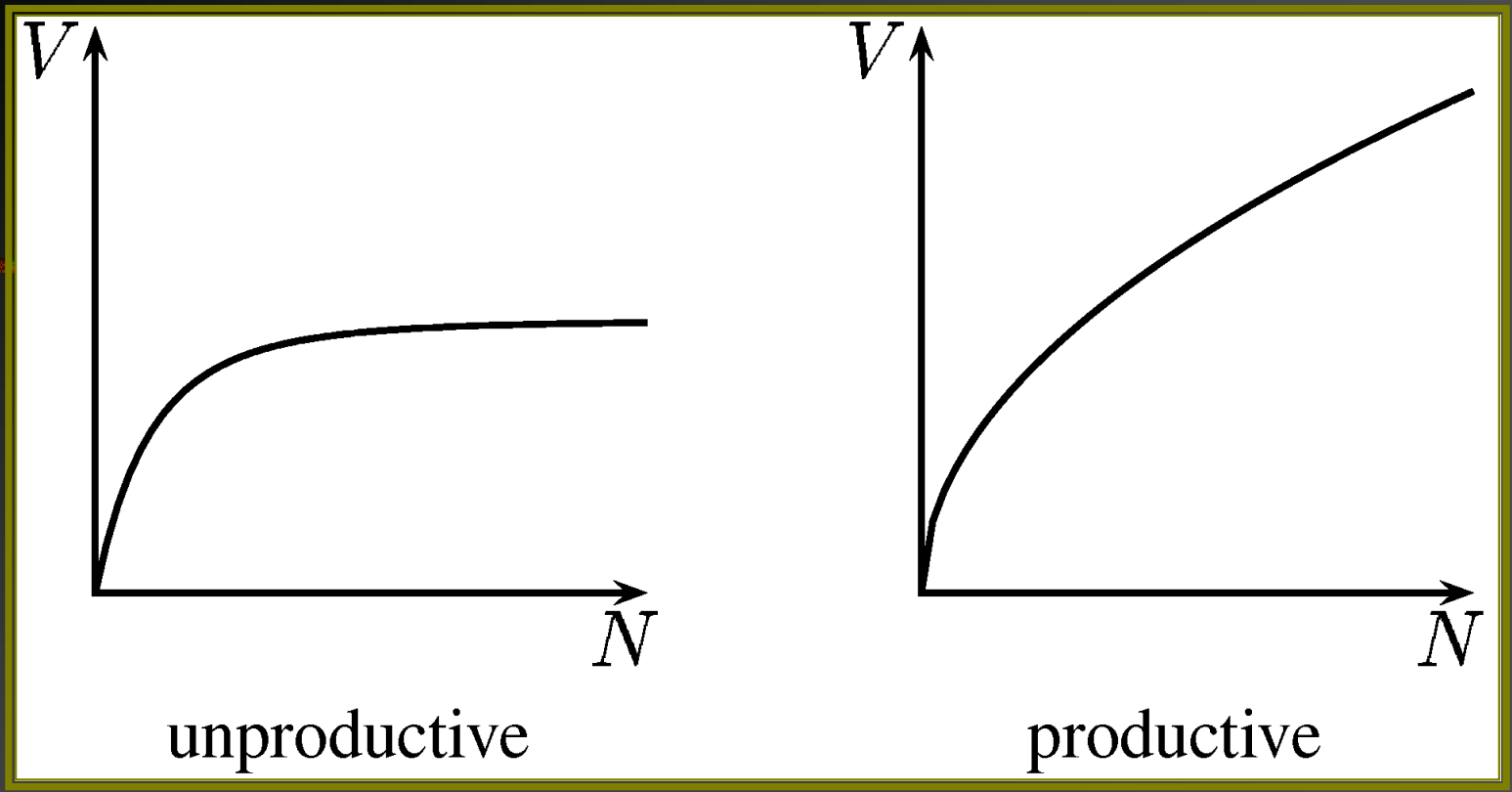
- *-bar* has 102 hapax legomena (words with occurrence frequency 1) in the STZ corpus
  → productive
- *-sam* has 7 hapax legomena in the STZ-corpus
  → less productive/unproductive

cognitive perspective III

- this presupposes that instances that are produced by the process are stored
- together with their frequencies
- the formation patterns are somehow 'found' through the analysis of the instances one has already seen
- the cognitive notion of productivity is based on linguistic experience!
- and as such is a diachronic process!

computational linguistic perspective I

- corpus-based model
- corpus models linguistic experience
- adding more text models diachronicity
- instances of a given morphological pattern must be found and stored together with their frequencies
- type-token curves:
  - the process is productive as long as new types appear if more tokens are added
  - this means that the proportion of hapax legomena is high



N = tokens, V = types:
in the distribution labeled 'productive' the number of types continues to grow as new tokens are sampled

productivity: quantitative aspects

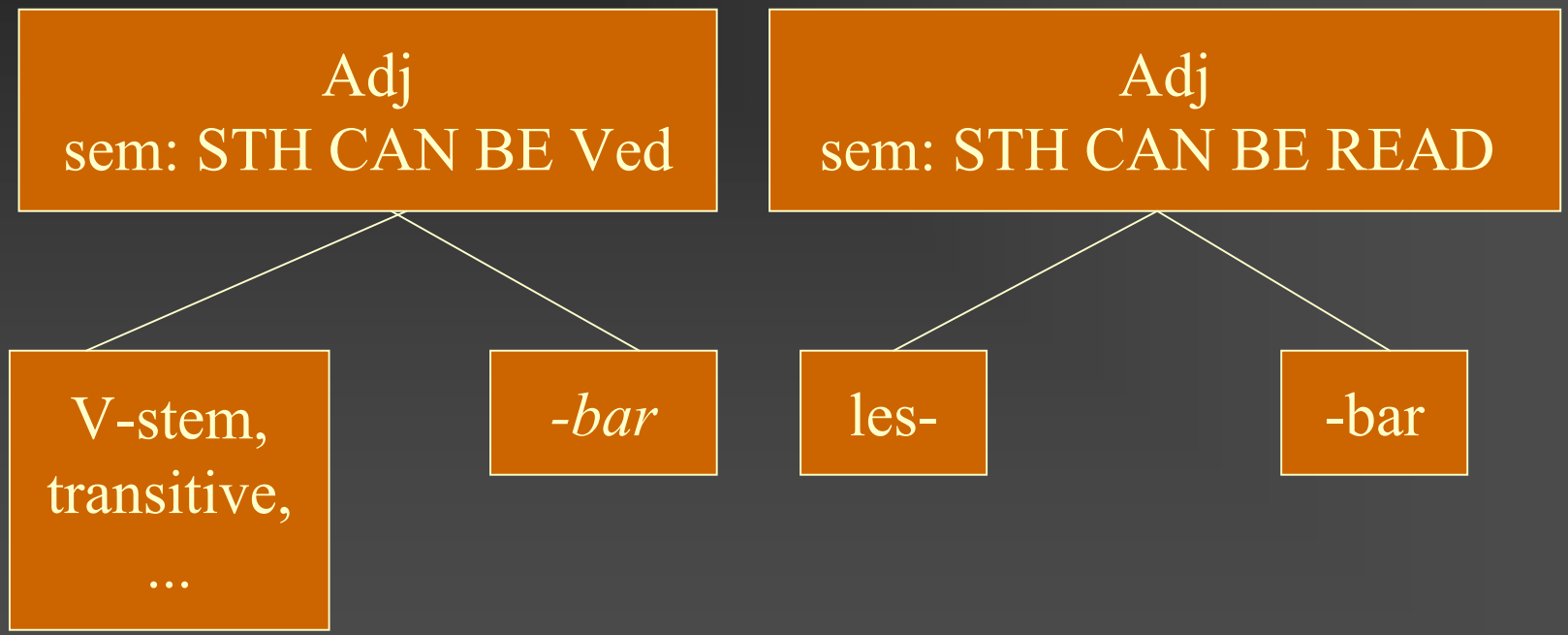
- calculate the probability that a new type formed by a given morphological process is found after a given amount of text is sampled
- approximation: slope of the type-token curve or V1/N
- (really: productive processes have an LNRE distribution (for Large Number of rare events, Baayen 2001) – we have to use LNRE-models)

productivity: a definition I

- Rules:
  - (X, Y) (V-trans,Adj) productive
  - (x_1, Y) (*V-trans, -bar*) productive
  - (x_1, y_1) (*les-, -bar*) not productive

- A rule is **productive** if
  not all terminal nodes are associated with phonological information
  (at least one variable is present).

productivity: a definition II



productivity: summary

- productive morphological processes
  - are regular
  - produce an in principle unlimited number of new formations
- collocations as we have seen them do not contain variables
- in what sense can they be productive?

variables within collocations: case study *blaues Auge* I

- readings:
  colour of iris,
  (other colour readings),
  bruised area around eye
- usage: in free combinations, in fixed combination:
  mit einem blauen Auge davonkommen
- indications for collocation:
  high frequency, blue not colour,
  translation "black eye"

blaues Auge II

- Vor ca. 2 Wochen habe ich bei einem Internationalen Meeting in Kassel zweimal in dem Hürdenrennen ganz stark an die Hürden angeschlagen und mein Knie blutete sehr stark. Es gab sofort ein richtig dickes und sofort blaues Knie.
- Heute möchte ich mich mal wieder melden. Seit 13 Wochen bin ich nun an der Dialyse. Zum Anfang lief auch einiges schief (Blauer Arm) aber das ist alles überstanden.
- Auch blieben mir die Prügel, die mir einen blauen geschwollenen Hintern bescherte, nicht aus.

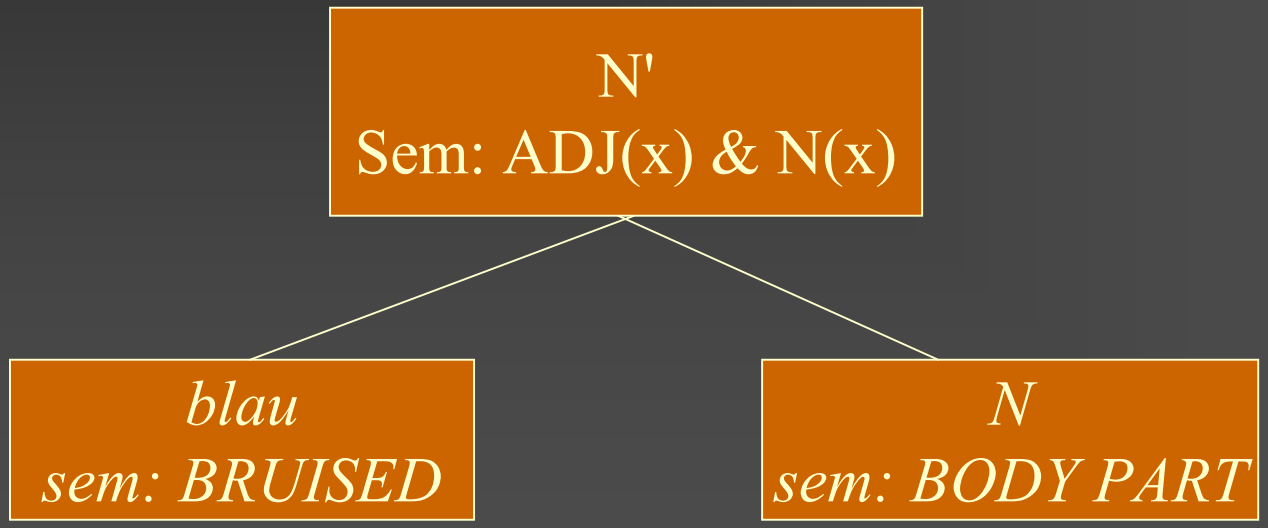
blaues Auge III

- Zu Verletzungen oder Unfällen kam es bei den Läufern daher nicht, wohl aber beim OL-Chef: ein blauer Ellenbogen beim Verlassen der unabgestreuten Brücke.

- low frequency combinations with body parts (HGC):
  Fuß, Bein, Fingernagel, Lippe, ...

blaues Auge IV

- regular
- in principle unlimited



blaues Auge V: consequences

- *blau* has to be listed in the lexicon as a polysemous entry with the reading 'bruised' in addition to the colour reading
- *blaues Auge* is a collocation only in the Firthian definition
- for acquisition: we have to look not for pairs (tuples) of words that occur together with high frequency but for semantic classes of words

blaues Auge VI: other examples

- *Zähne putzen* 'to brush one's teeth':
  N can be any noun that means 'teeth'
  Vorderzähne, Backenzähne, Schneidezähne, Milchzähne,
  Beisserchen, ...
- *zur Aufführung kommen* 'to be performed'
  N can be any event noun
  Streit, Eheschliessung, Gespräch, ...

- we can formulate a rule with a variable
- in principle unlimited number of formations

aside: approaches to variation

- syntactic variability (passivization, modification, ...)
- collocation potential (Kollokationspotenzial, Kollokationsradius)
- semi-compositionality (Heid)

aside: approaches to variation

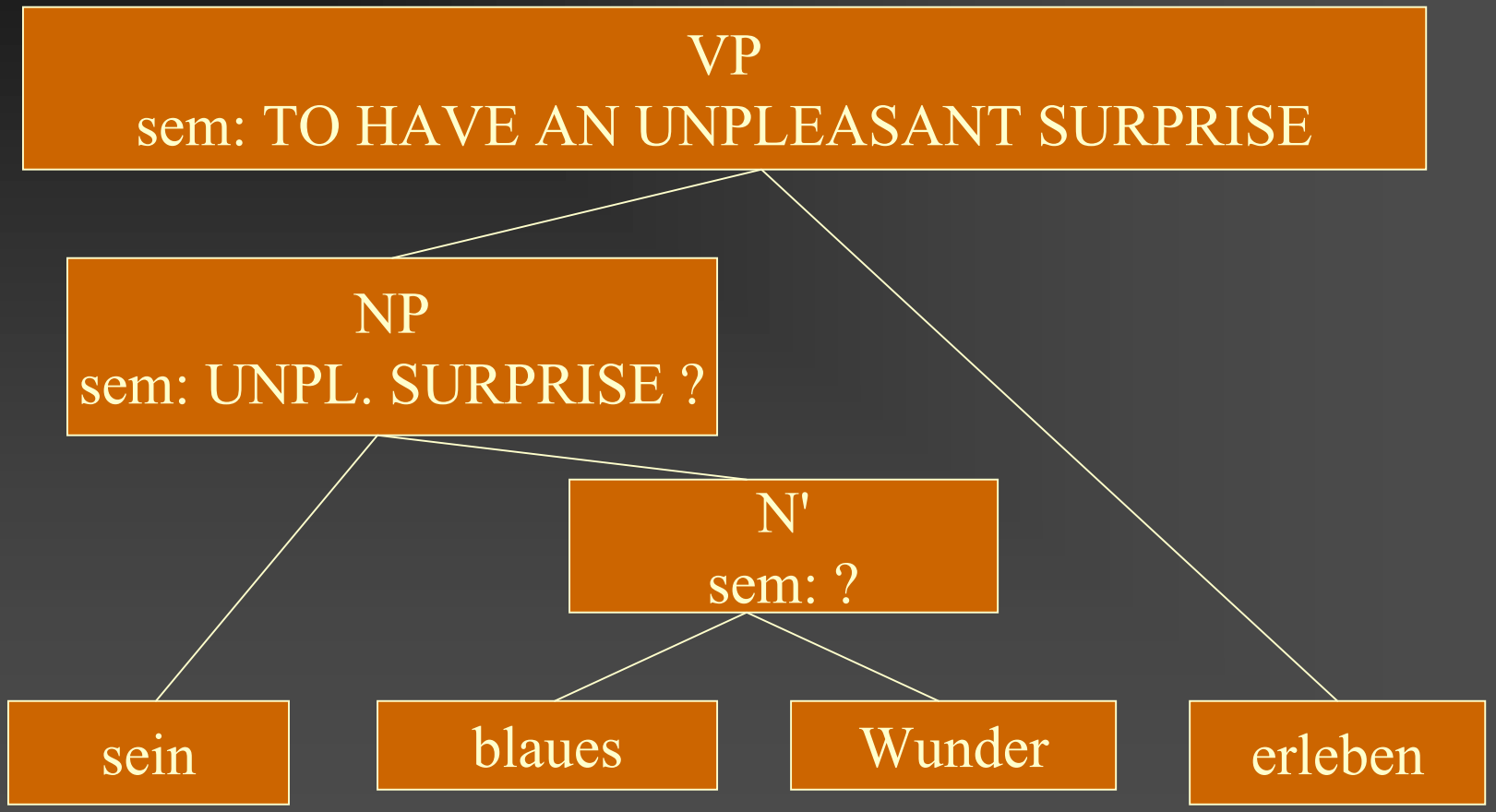
- lexical set (Kollokationsfeld)

  "Ein Kollokationsfeld setzt sich zusammen aus allen Synonymen mit dem gleichen Kollokationspotenzial. Zu einem solchen Kollokationsfeld gehören also etwa die Wörter *Steuern, Gebühren, Beiträge* und *Eintrittsgeld,* weil sie einen vergleichbaren Kollokationsradius (*erheben, zahlen, entrichten* usw.) aufweisen." (Hausmann, 127)

modification inside of collocations: case study *blaues Wunder* I

- *sein/ihr blaues Wunder erleben*
  "to have an unpleasant surprise"
- no lexical variation possible
- non-compositional semantics
- medium-low frequency (5)

blaues Wunder II

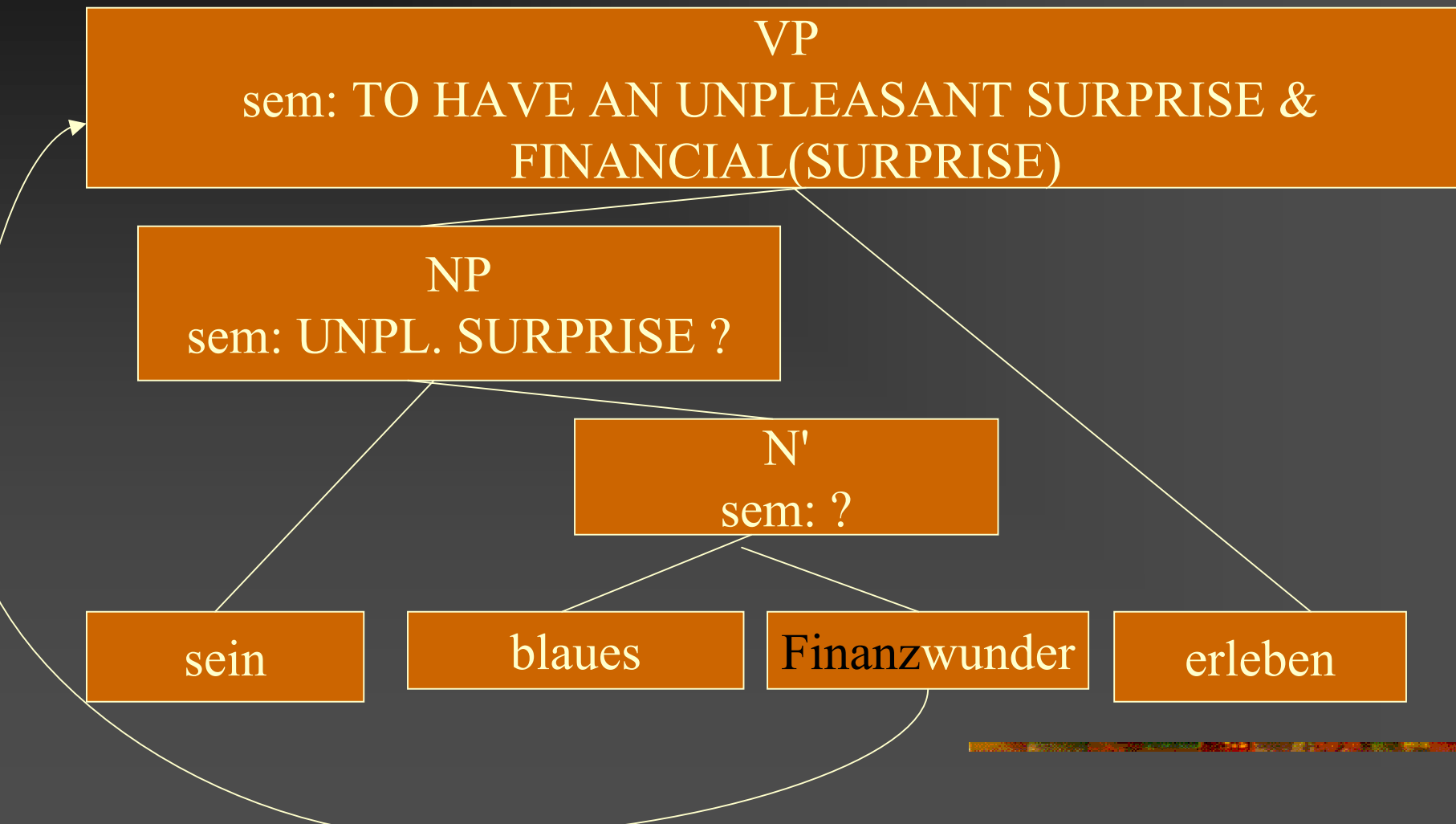


blaues Wunder III

- *sein blaues Finanzwunder erleben*
  'to have an unpleasant financial surprise'

- in principle unlimited: *Aktienwunder, Geschmackswunder, ...*
- head of compound must be *Wunder*

- semantics: modification outside of sollocation but regular!
- and in principle unlimited


VP
sem: TO HAVE AN UNPLEASANT SURPRISE & FINANCIAL(SURPRISE)
NP
sem: UNPL. SURPRISE ?
N'
sem: ?
sein
blaues
Finanzwunder
erleben

summary I: claims

- collocation formation can be productive – similar to word formation
- depending on the definition of 'collocation': there are there are two ways in which collocations can be productive
  - a variable in the collocation
    Firth
  - modification within the collocation
    Choueka, (Firth)

summary II: consequences

- we have to be even more careful when we acquire collocations:
  - collocations are not necessarily single instances
  - we might have to look for semantic similarity classes instead of single instances
  - the distrbution of the types in the similarity classes tells us something about the productivity