ADVENTURES IN MEANING BANKING

Johan Bos

ADVENTURES IN MEANING BANKING

① A new adventure
② Choice of currency
③ Preparation
④ Off we go!
⑤ Help from strangers
⑥ Well on our way
⑦ Sweet memories
⑧ Bumps on the road
⑨ Lost in translation

Groningen MEMING BANK

Chapter I

A NEW ADVENTURE

WHERE ARE WE NOW?

Computational semantics is now at a point where computational syntax was twenty (thirty?) years ago:

- robust semantic parsing only for small domains
- semantic tasks often carried out in isolation
- lack of large semantically annotated linguistic resources

OUR GOAL

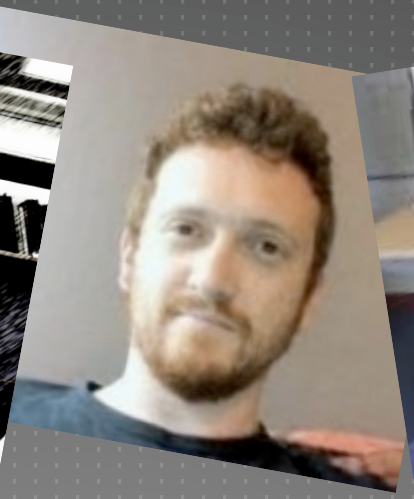
- Syntactic parsing made a breakthrough by developing treebanks in the 1990s
- We want to develop a large semantically annotated resource and ...
- ... cause a similar breakthrough in computational semantics!

THE MOTLEY CREW

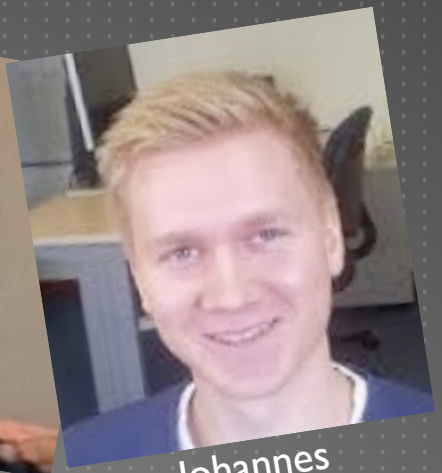
Kilian Evang

Noortje Venhuizen

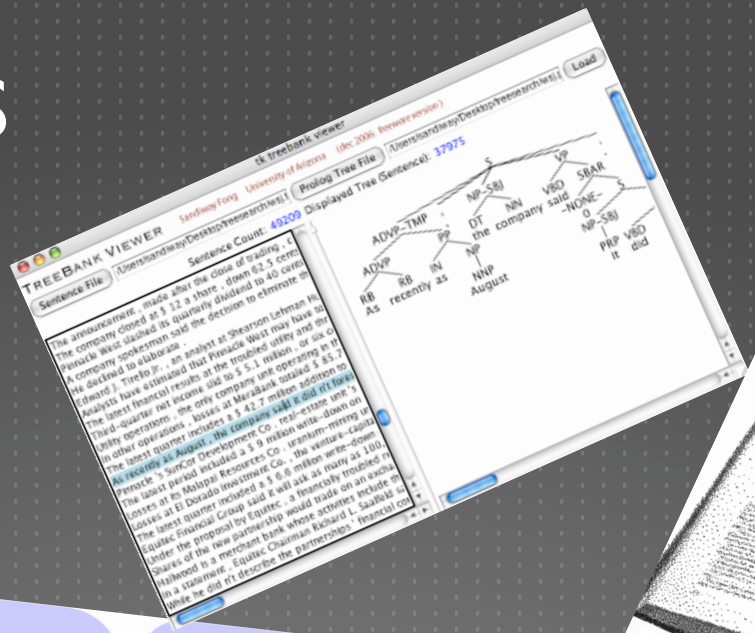
Valerio Basile

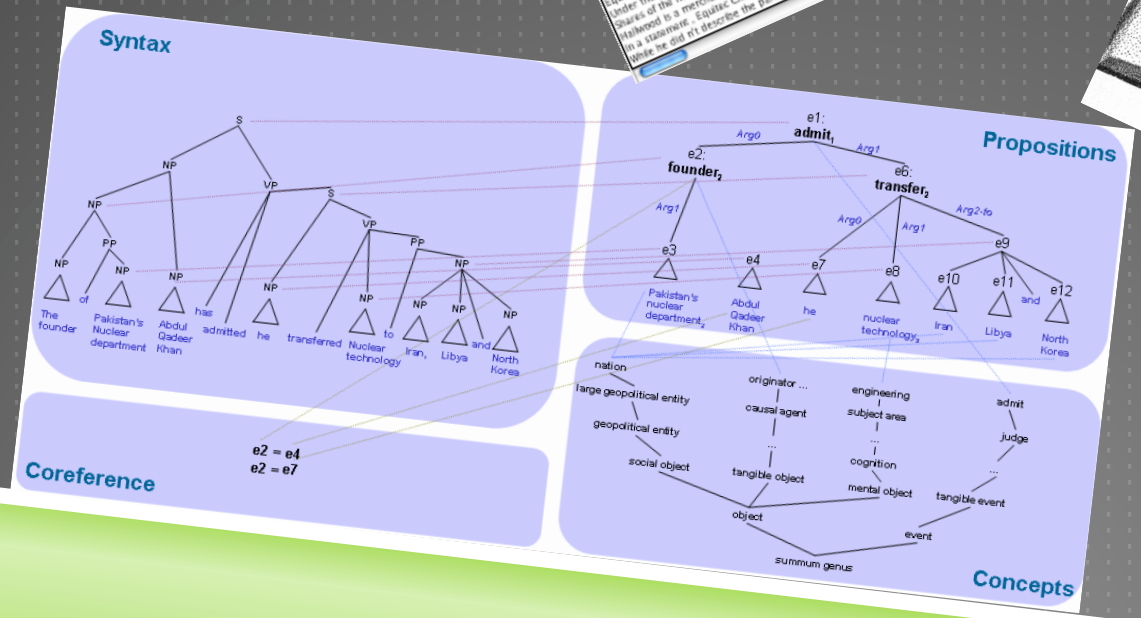


Johannes Bjerva



PREVIOUS ADVENTURES





THE GRONINGEN MEANING BANK

- large (English) corpus of public domain texts
- annotated with deep (formal), not shallow semantics
- largest annotated units are texts (not sentences)
- integrates various phenomena in one formalism
- automatically produced, manually corrected

Groningen

Groningen MEANING BANK

OBJECTIVES OF THE GMB

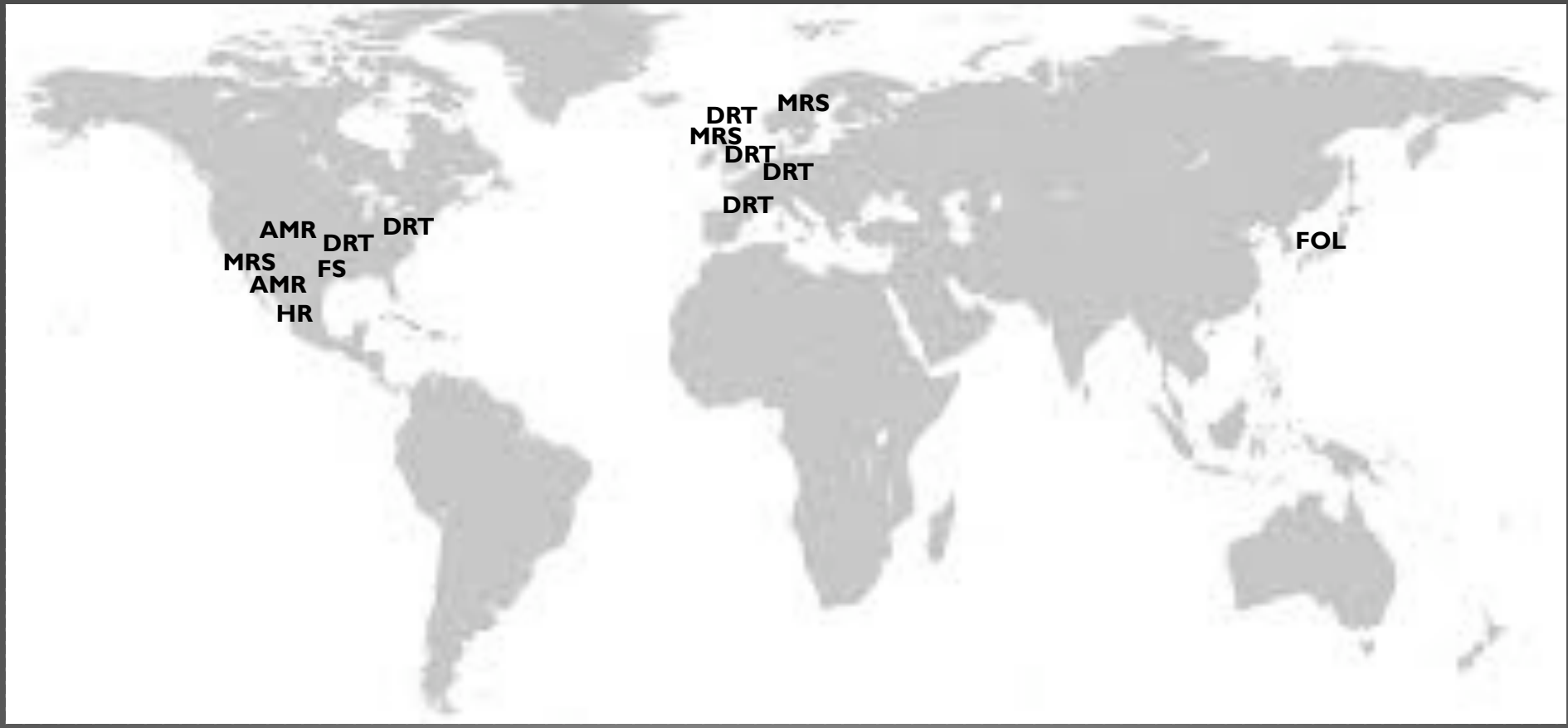
- develop, refine, and evaluate wide-coverage semantic parsers
- develop NLG components from formal meaning representations
- revise semantic theory where needed

Groningen MEANING BANK

Chapter 2

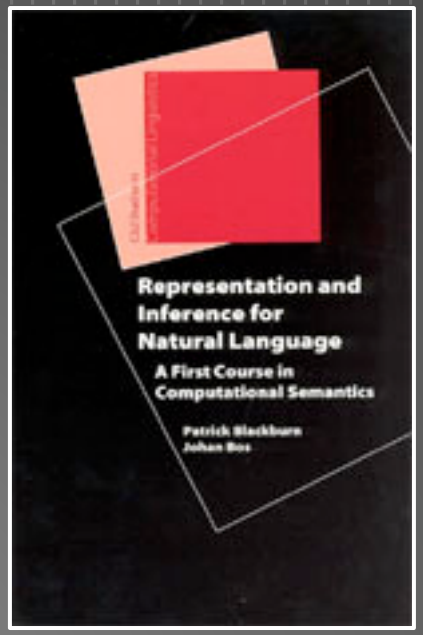
CHOICE OF CURRENCY

DISTRIBUTION OF REPRESENTATION EMPIRES



REPRESENTATION AND INFERENCE

- syntactic representation serve to build meaning representations
- meaning representations serve to perform inferences



Blackburn & Bos 2005

EXAMPLE: VENICE

Turn left ~~and~~ or right to reach San Marco.

EXAMPLE: BOLT

Bolt is faster than everyone else. **YES**

Bolt is in last position. **NO**

EXAMPLE: PETS

No pets are allowed in this area.

All pets must be on the leash in this area.

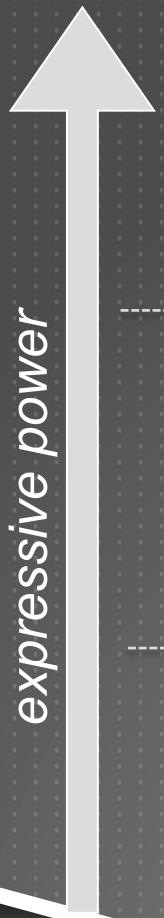
EXAMPLE: BUSH

"... when there's more trade, there's more commerce."

not informative

George W. Bush, at the Summit of the Americas in Quebec City, April 21, 2001 (source: Language Log 24/10/2004)

CONTROLLING INFERENCE



λx λP higher-order logic

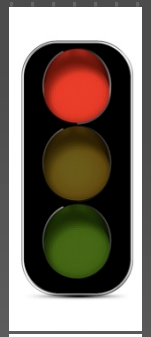
∀P ∃P second-order logic

∀x
∃x first-order logic (predicate logic)

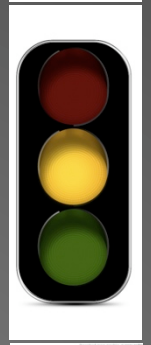
discourse
representation
structure

[] <> modal logics
description logics

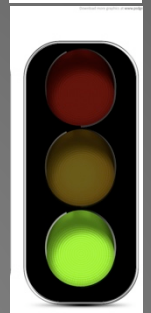
¬ ∧
→ ∨ propositional logic



undecidable



semi-decidable



decidable

Butch stole a chopper.
It was parked in a garage.

no events

x y u v
Butch(x) chopper(y) stole(x,y) u=y garage(v) parked-in(u,v)

Davidsonian

x y e u v e’
Butch(x) chopper(y) stole(e,x,y) u=y garage(v) parked(e’,u) in(e’,v)

neo-Davidsonian

x y e u v e’
Butch(x) chopper(y) stole(e) agent(e,x) theme(e,y) u=y garage(v) parked(e’) theme(e’,u) location(e’,v)

Hobbsian

x y e u v e’ a b c d
Butch(x) chopper(y) stole(e,x,a,b) agent(e,x) u=y garage(v) parked(e’,u,c,d) in(e’,v)

Chapter 3

PREPARATION

SURVIVAL TOOLS

- Good tools are needed to reach wide coverage
- Tokenisation, Tagging, Parsing
- Compositional semantics, semantic resolution: Boxing

BOXER

- Input: syntactic tree (categorial grammar derivation)
  CCG: see Steedman
- Output: resolved semantic representation
  (box, i.e. a Discourse Representation Structure)

Boxer works with
the C&C parser and easyCCG

Online demo:
http://www.let.rug.nl/basile/gsb/webdemo/demo.php

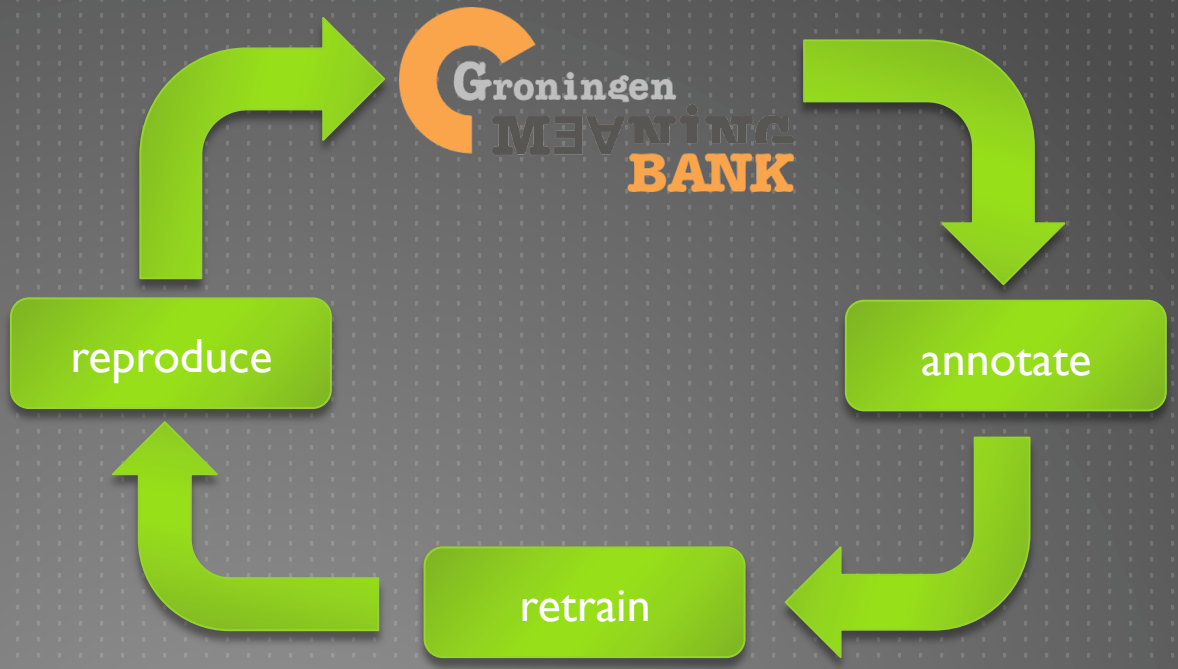
Groningen MEANING BANK

Chapter 4

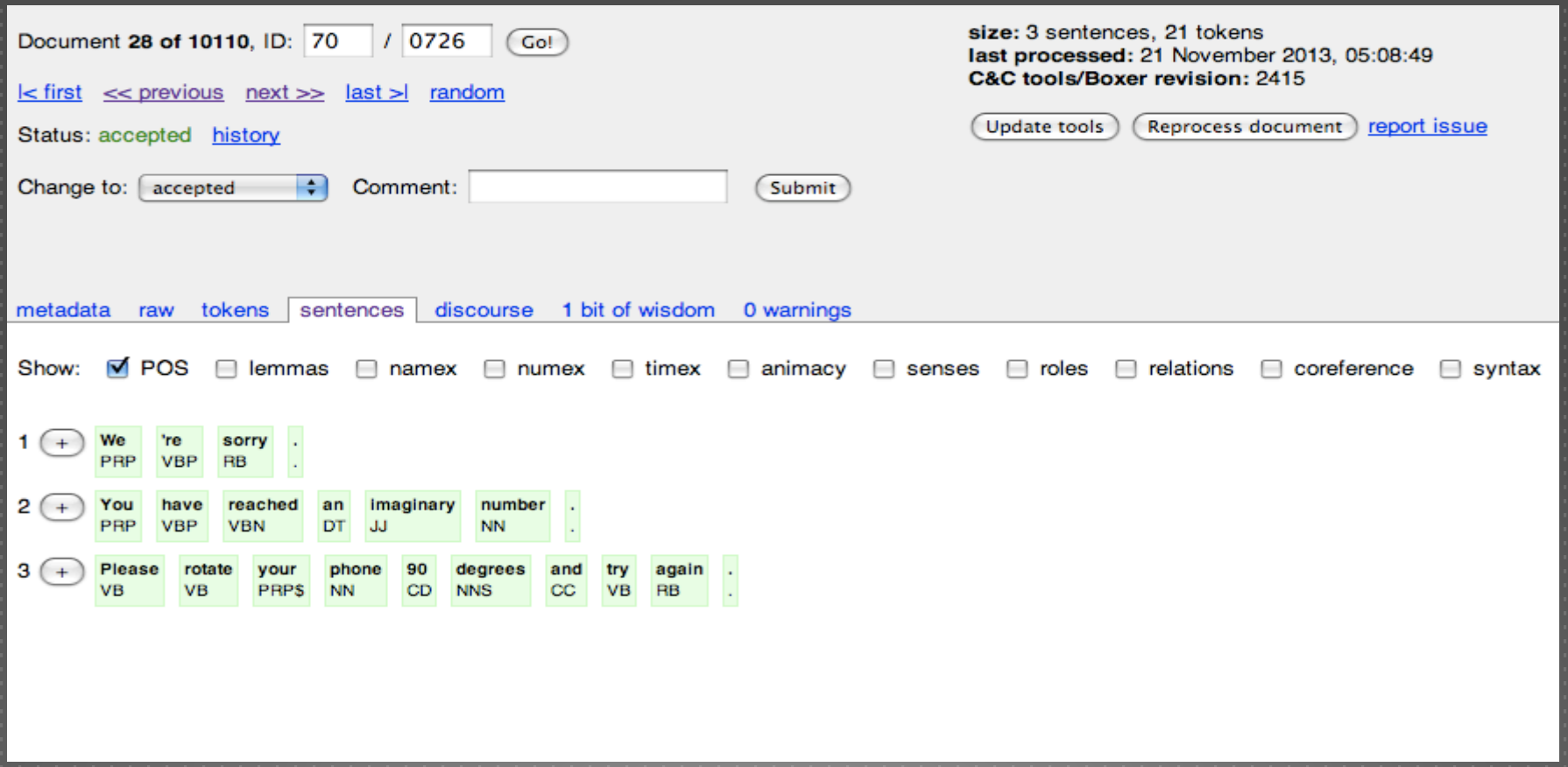
OFF WE GO!

ITERATIVE DEVELOPMENT

- Annotations are produced by machines, corrected by humans (crowdsourcing + experts)
- A lot of it is done by (statistical) NLP tools: C&C tools + **Boxer**

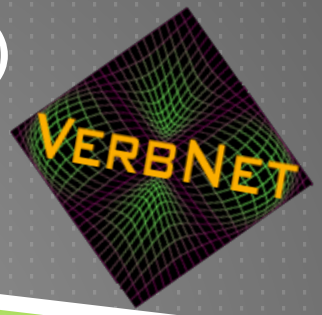


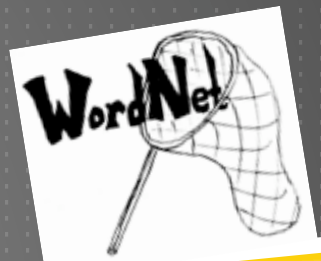
MEANING BANK DEVELOPMENT



LAYERS IN THE MEANING BANK

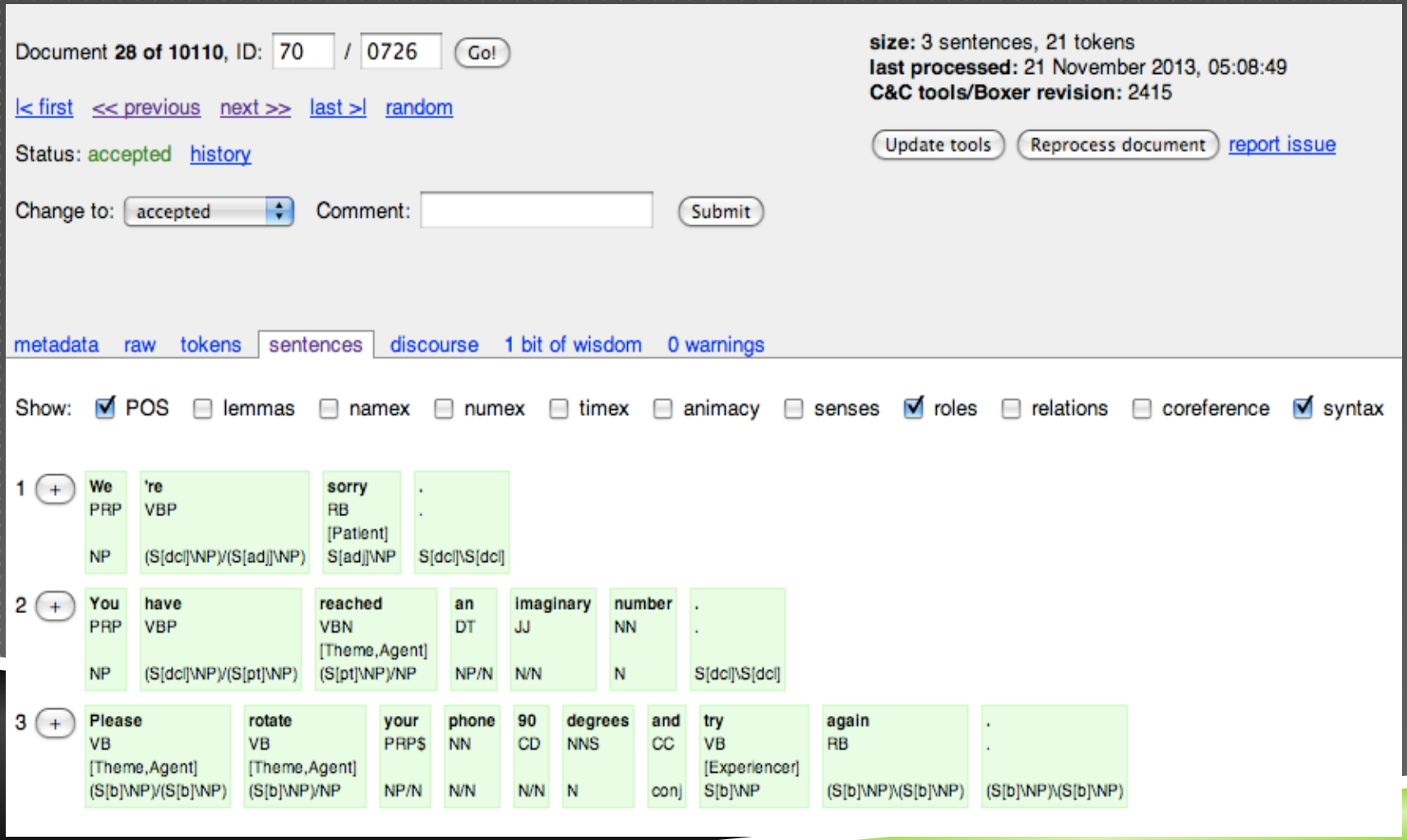
- PTB part of speech tags
- Animacy classes
- Named entity types
- WordNet senses
- VerbNet roles
- CCG derivations (syntax)
- DRT boxes (semantics)
- SDRT rhetorical relations (discourse)



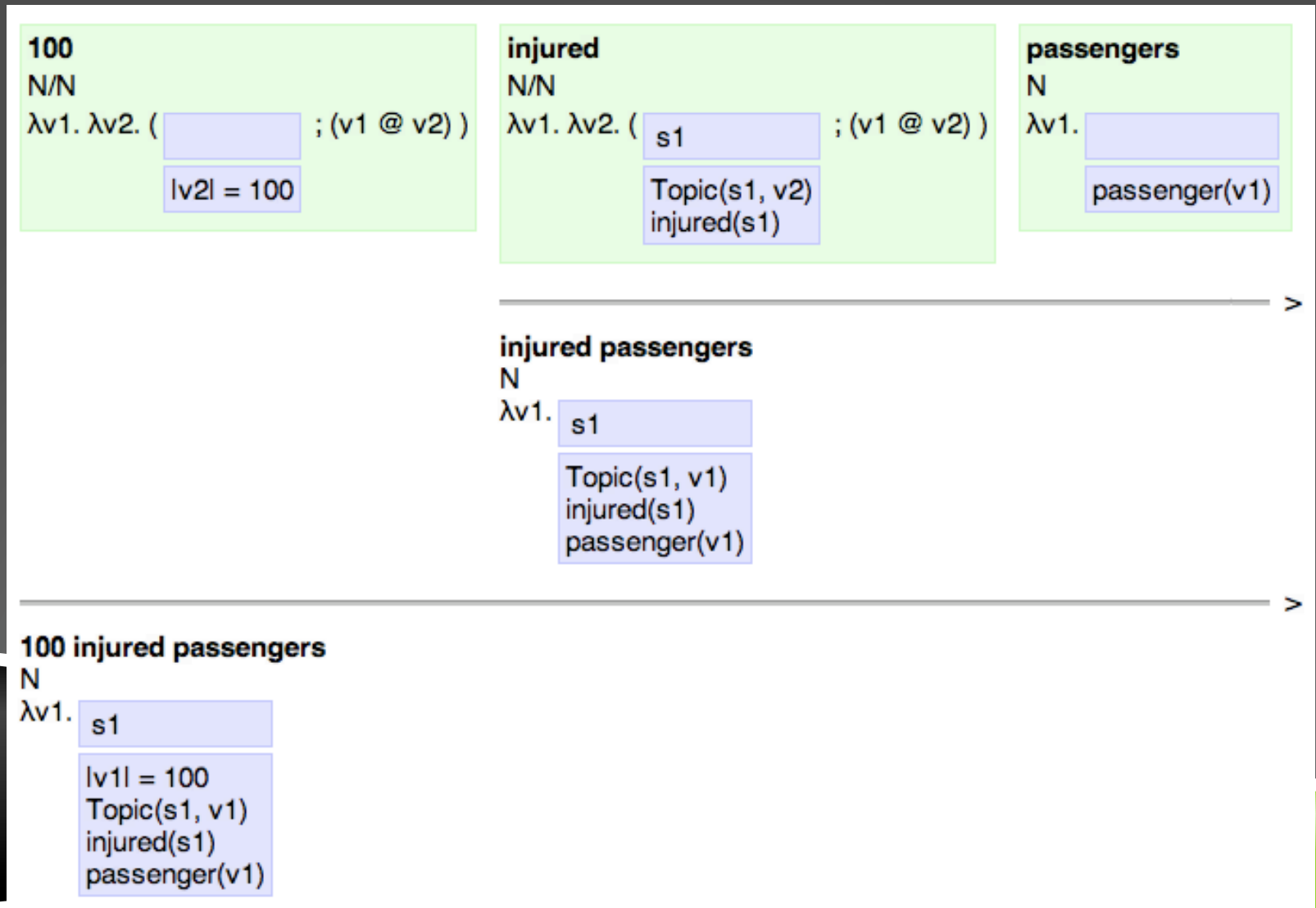


Groningen MEANING BANK

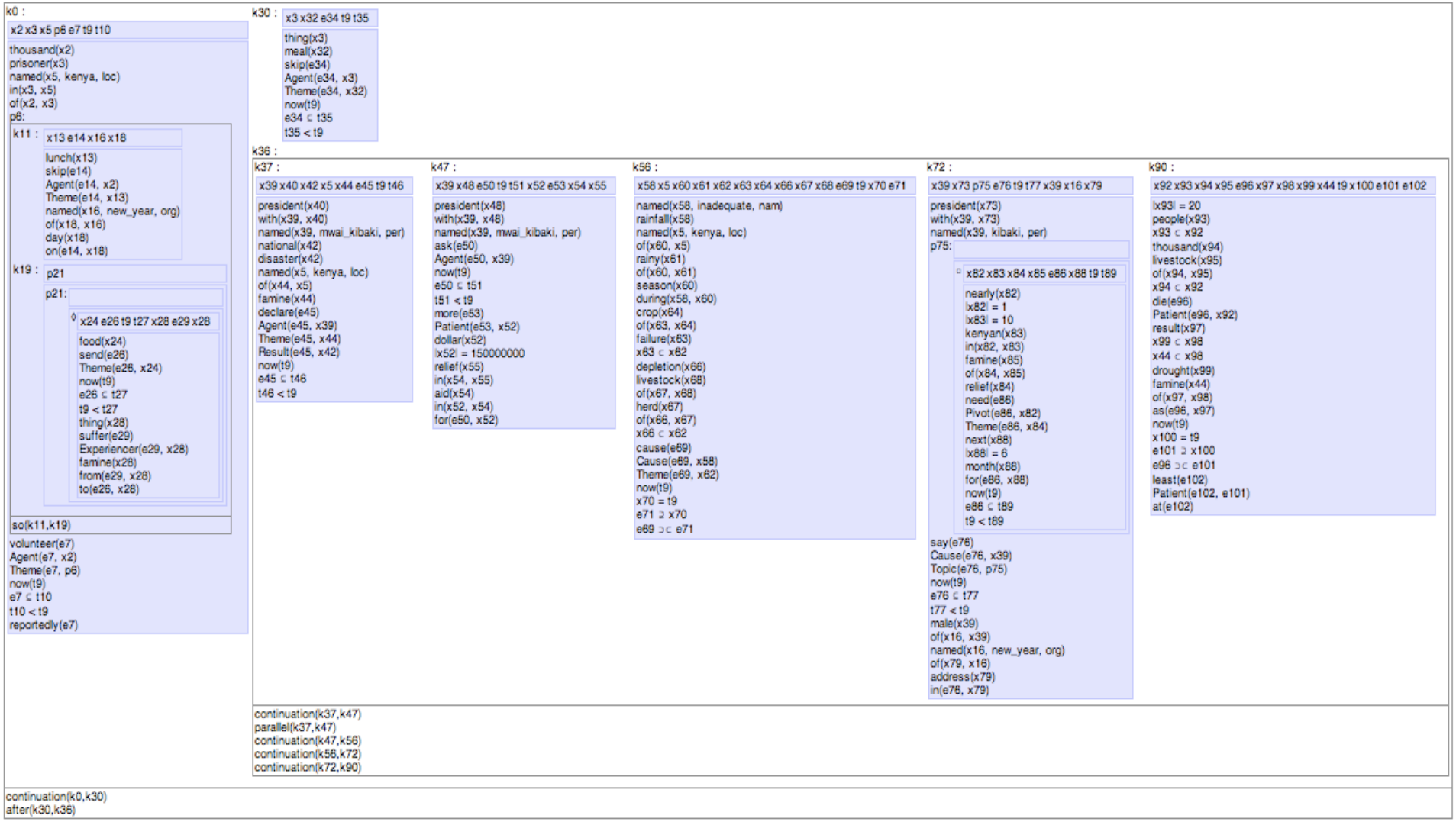
ANNOTATION LAYERS



THE COMPOSITIONAL SEMANTICS LAYER



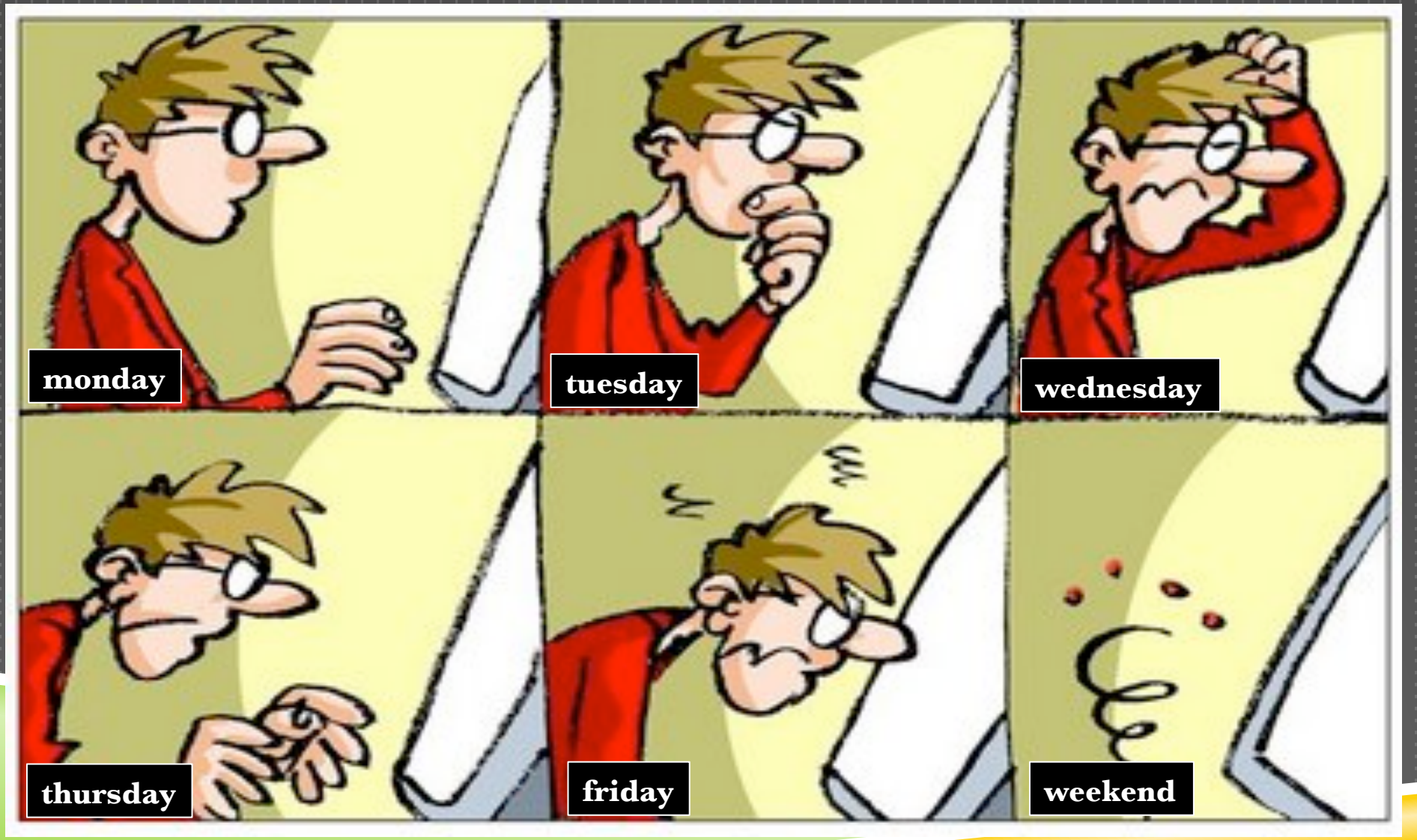
SEGMENTED DISCOURSE REPRESENTATION STRUCTURE (SDRS)



Chapter 5

HELP FROM STRANGERS

TRADITIONAL ANNOTATION



AMAZON'S MECHANICAL TURK

CROWDSOURCING

- Outsourcing tasks to a distributed group of people
- The internet provides infrastructure
- The GMB has two crowds:

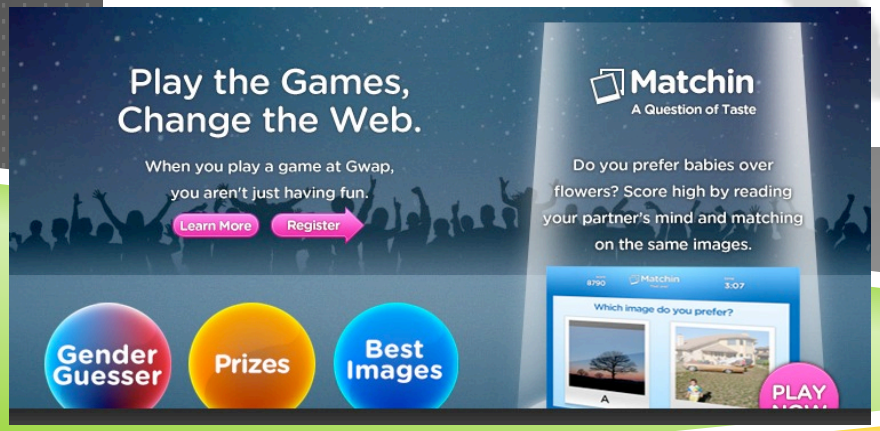
Expert Annotators

GWAP players

GWAP (GAME WITH A PURPOSE)

OPEN MIND word expert

#les jeux de mots



With **Duolingo** you learn a language for free while helping to translate the web

wordrobe

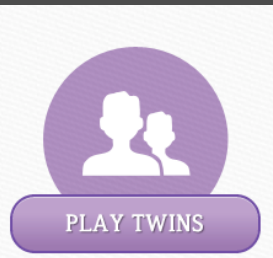
play what you mean

www.wordrobe.org

WORDROBE PHILOSOPHY

- not a single game, but a series of games that share same structure and scoring strategies
- each linguistic phenomenon that requires annotation corresponds to a different game
- every game consists of multiple-choice questions
- each question is presented by a text snippet plus a (small) number of possible answers
- these questions (and answers) are automatically generated from the corpus

"TWINS" (HOMOGRAPHS)



Twins Questions left until drawer is completed: 2

U.S.-led forces are hunting down remnants of Afghanistan's ousted Taleban regime who frequently carry out hit and run attacks on coalition and Afghan government **forces**, mainly in the country's eastern and southern regions.

- noun
- verb

Place your bet: low high

answer skip

"SENSES" (WORDNET)

Senses Questions left until drawer is completed: 1

gameimage

Russian **officials** say at least five more people have died from a wave of extremely cold weather gripping the nation, bringing the death toll to 43 in the past week.

- a worker who holds or is invested with an office
- someone who administers the rules of a game or sport

Place your bet: low high

answer skip

"POINTERS" (PRONOUNS)

Pointers Questions left until drawer is completed: 3

Venezuela's *president* is urging *President Bush* to use **his** second term in office to strengthen ties with Latin American nations.

- 1) Venezuela
- 2) president
- 3) President Bush

Place your bet: low high

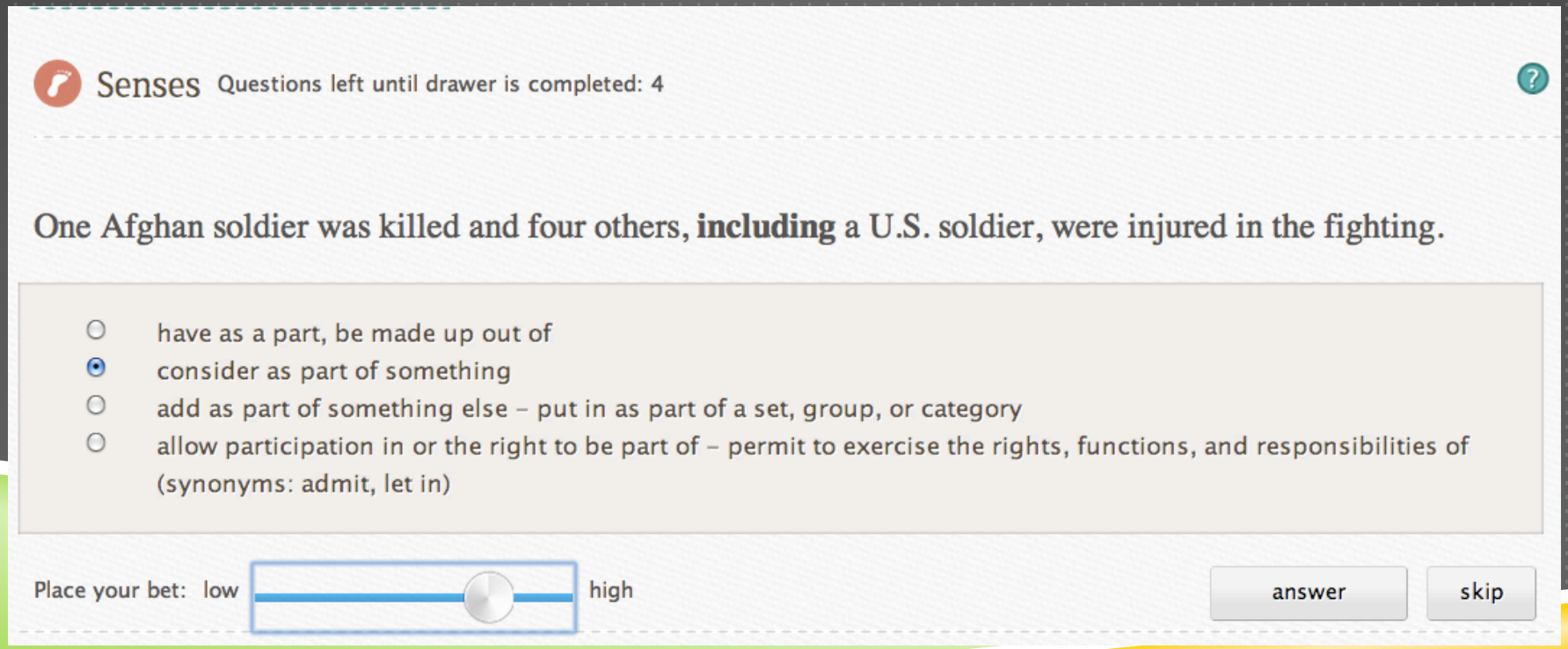
answer skip

SCORING IN WORDROBE

- Every answer increases the score of a player
- The more overlap of a player's answer with other players, the higher the score
- Total scores for a game are calculated over answers given in the last N days (N=50)

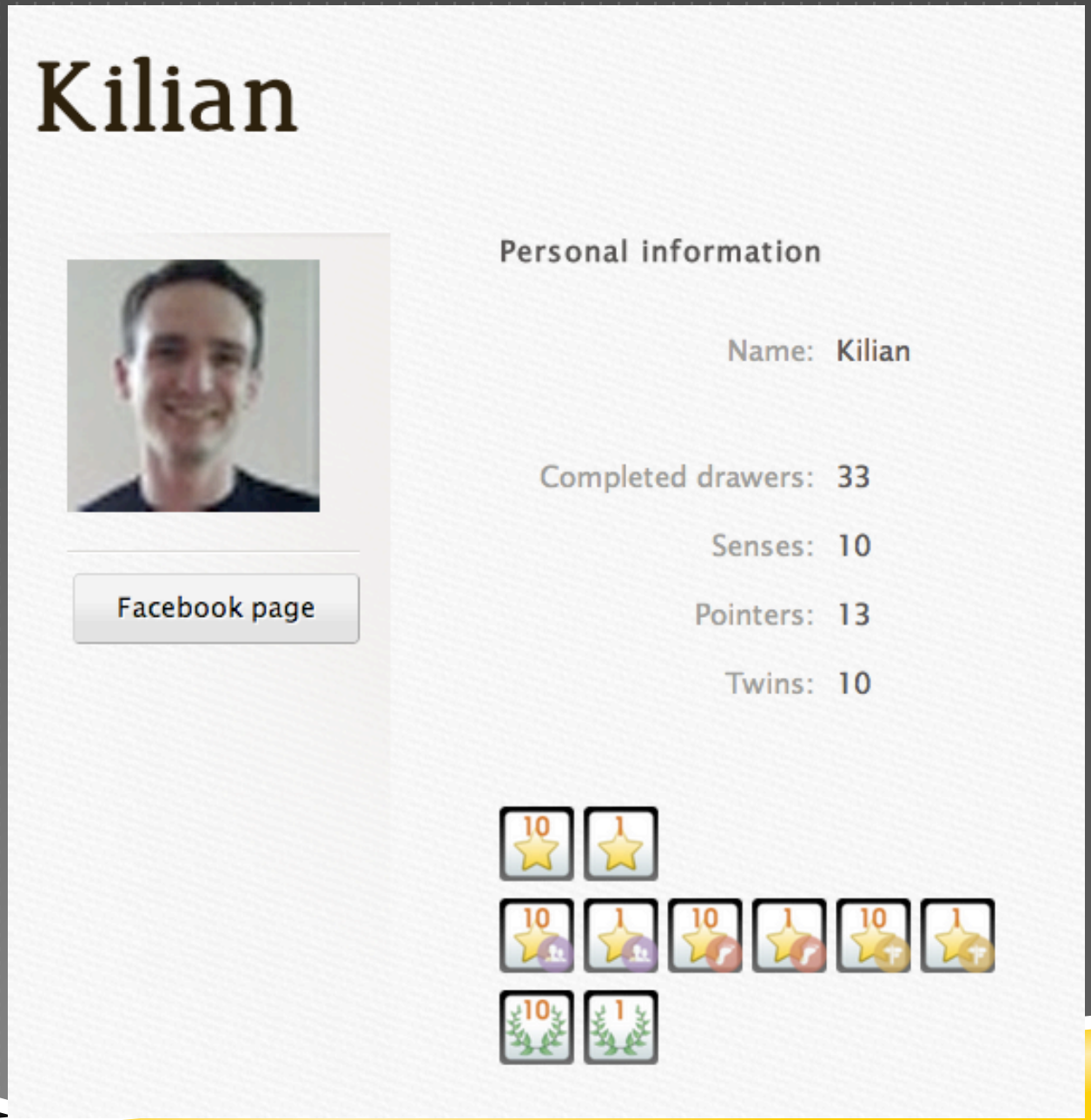
A SCORE WITH A TWIST

- In addition, Wordrobe players can take risks and bet on the correctness of an answer
- The higher the bet, the more points you can win (or loose)



WHY DO PEOPLE PLAY WORDROBE?

- unlocking achievements
- outperforming other players
- learning about language
- help computational linguists



Chapter 6

WELL ON OUR WAY

SEMANTIC PHENOMENA IN THE GMB

Included

- Scope (negation, quantifiers)
- Anaphoric pronouns
- Word senses (WordNet)
- Thematic Roles (VerbNet)
- Presuppositions
- Tense and aspect
- neo-Davidsonian events
- Rhetorical relations

Planned

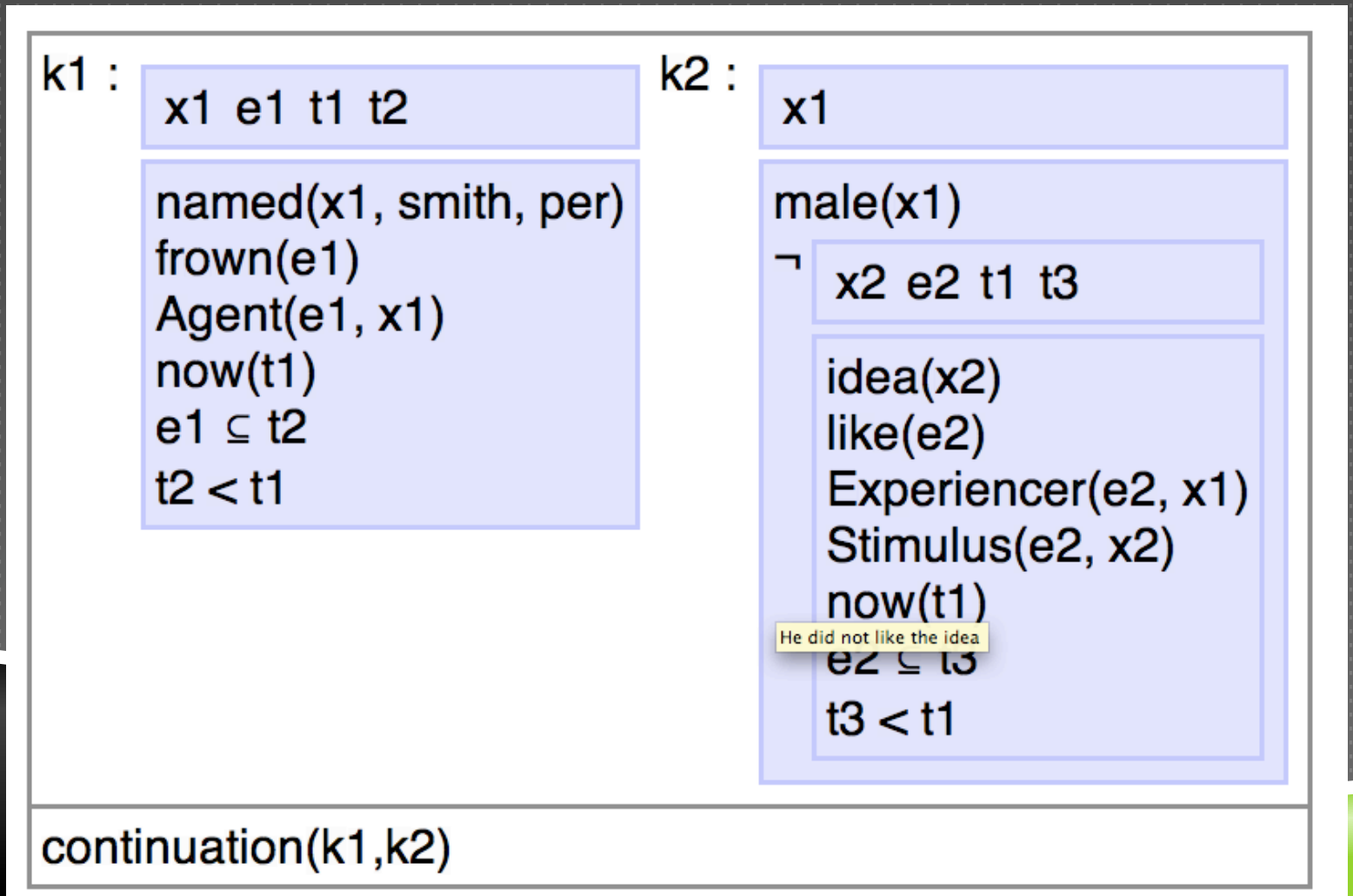
- Collective vs distributive NPs
- Reflexive pronouns
- Verb phrase ellipsis
- Background knowledge

Groningen MEANING BANK

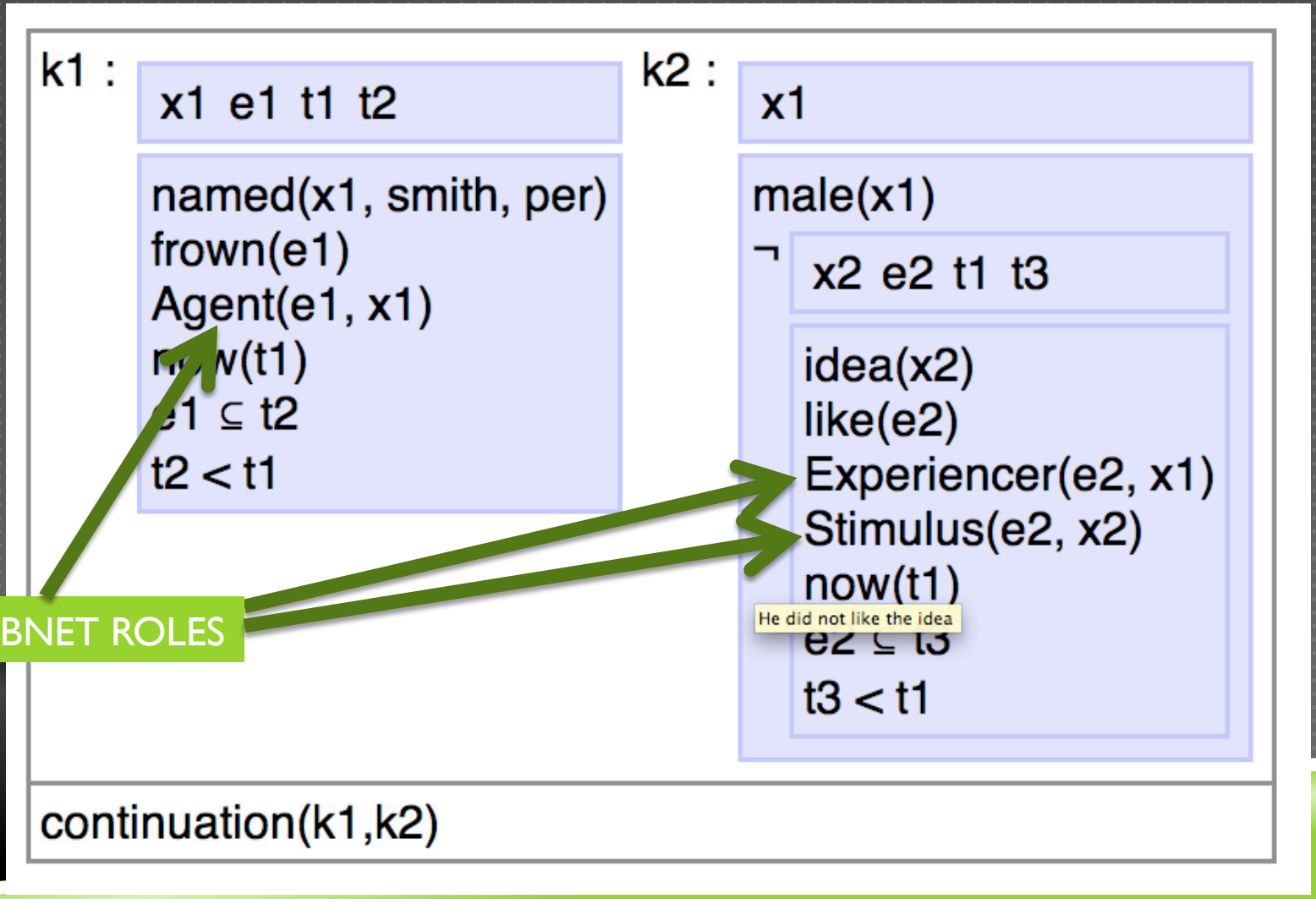
GMB: CORPUS SIZE

	genre	texts	sentences	words	s/t	w/s
Voice of America	newswire	9,207	57,174	1,238,576	6.2	21.7
CIA world factbook	almanac	514	4,436	112,516	8.6	25.4
Aesop's Fables	narrative	224	949	23,105	4.2	24.3
jokes	humor	122	443	7,531	3.6	17.0
MASC		35	291	6,985	8.3	24.0
RTE		1,338	1,537	29,854	1.1	19.4
		11,440	**64,830**	**1,418,567**	**5.7**	**21.9**

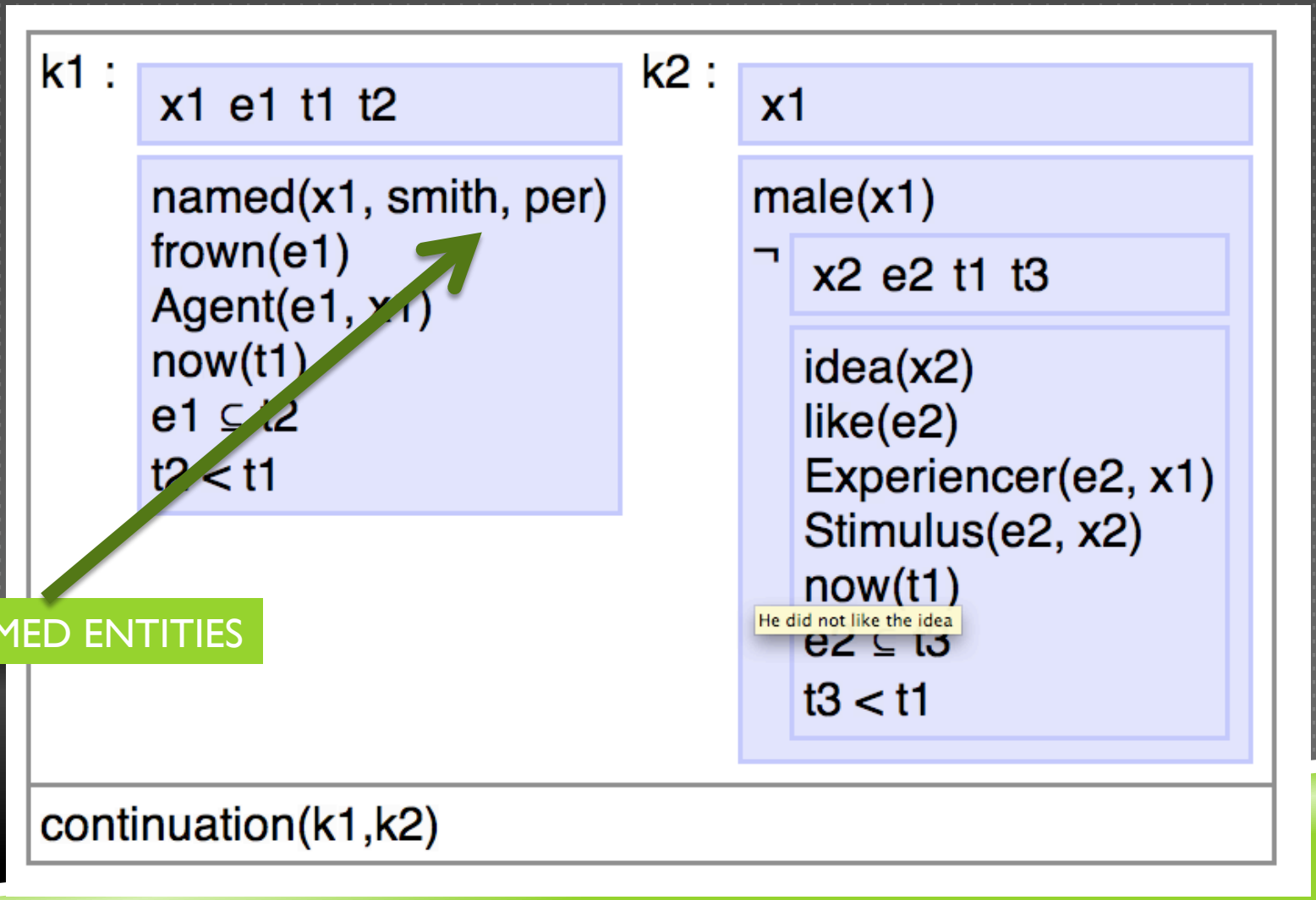
INTEGRATION OF SEMANTICS



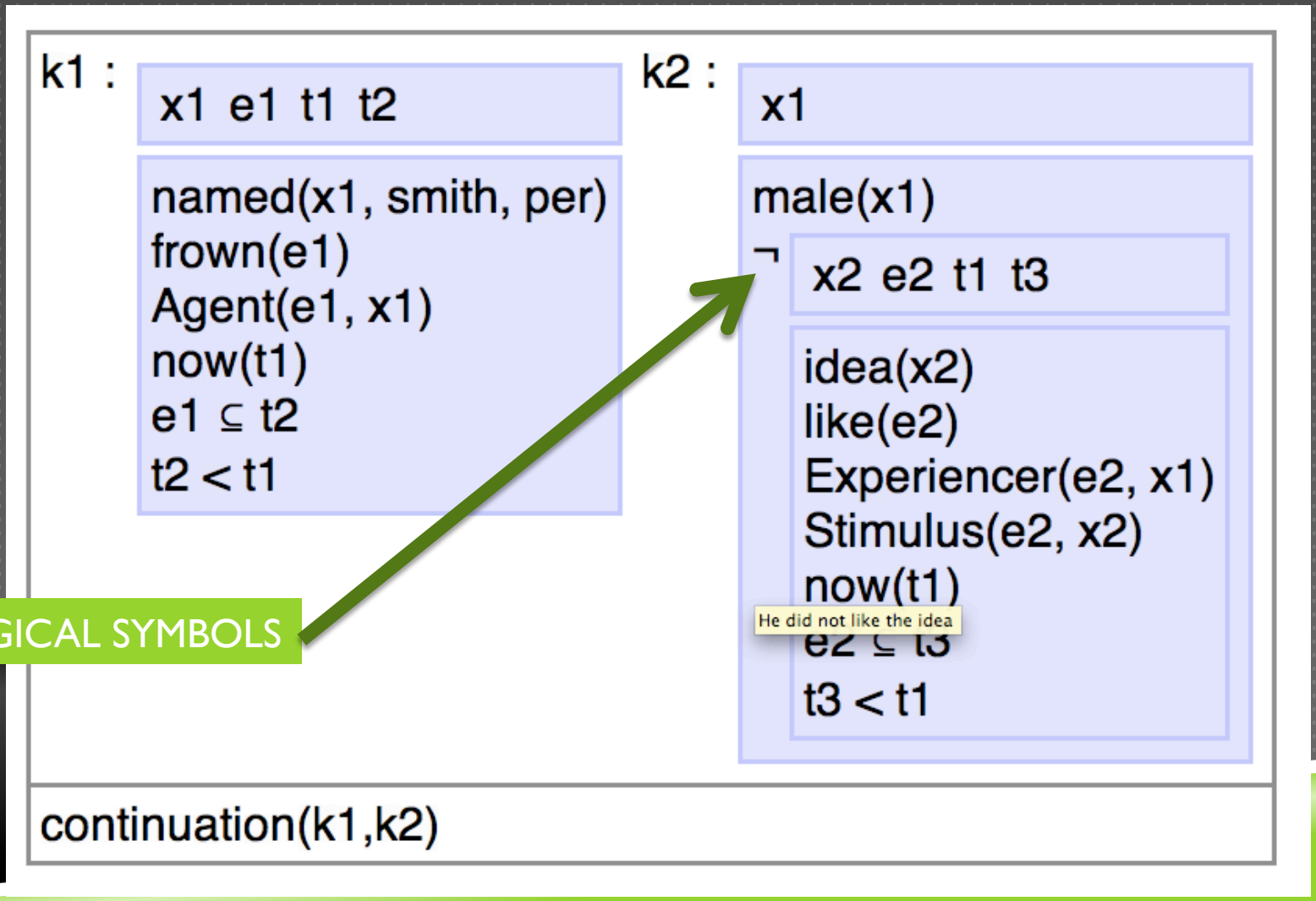
INTEGRATION OF SEMANTICS



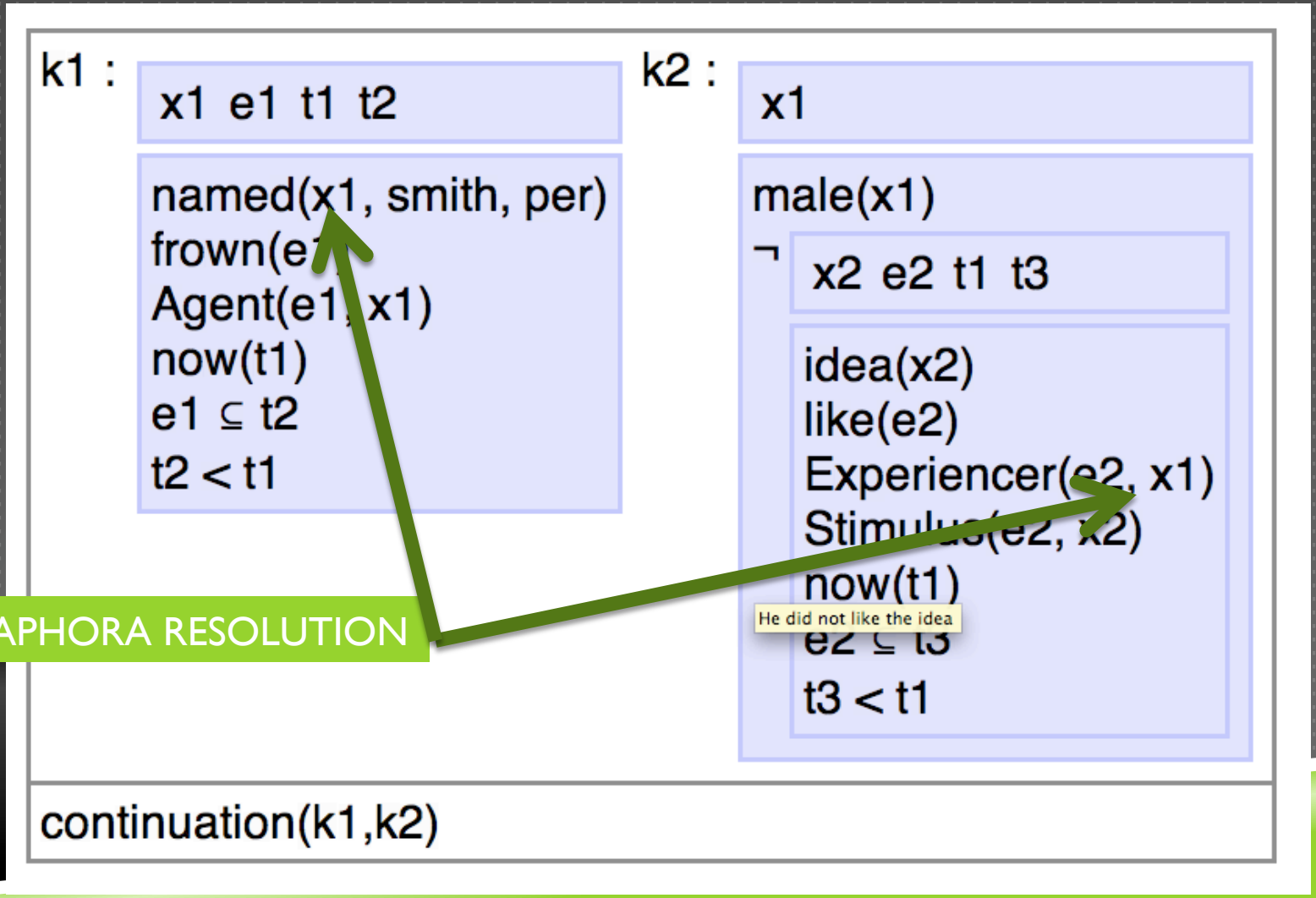
INTEGRATION OF SEMANTICS



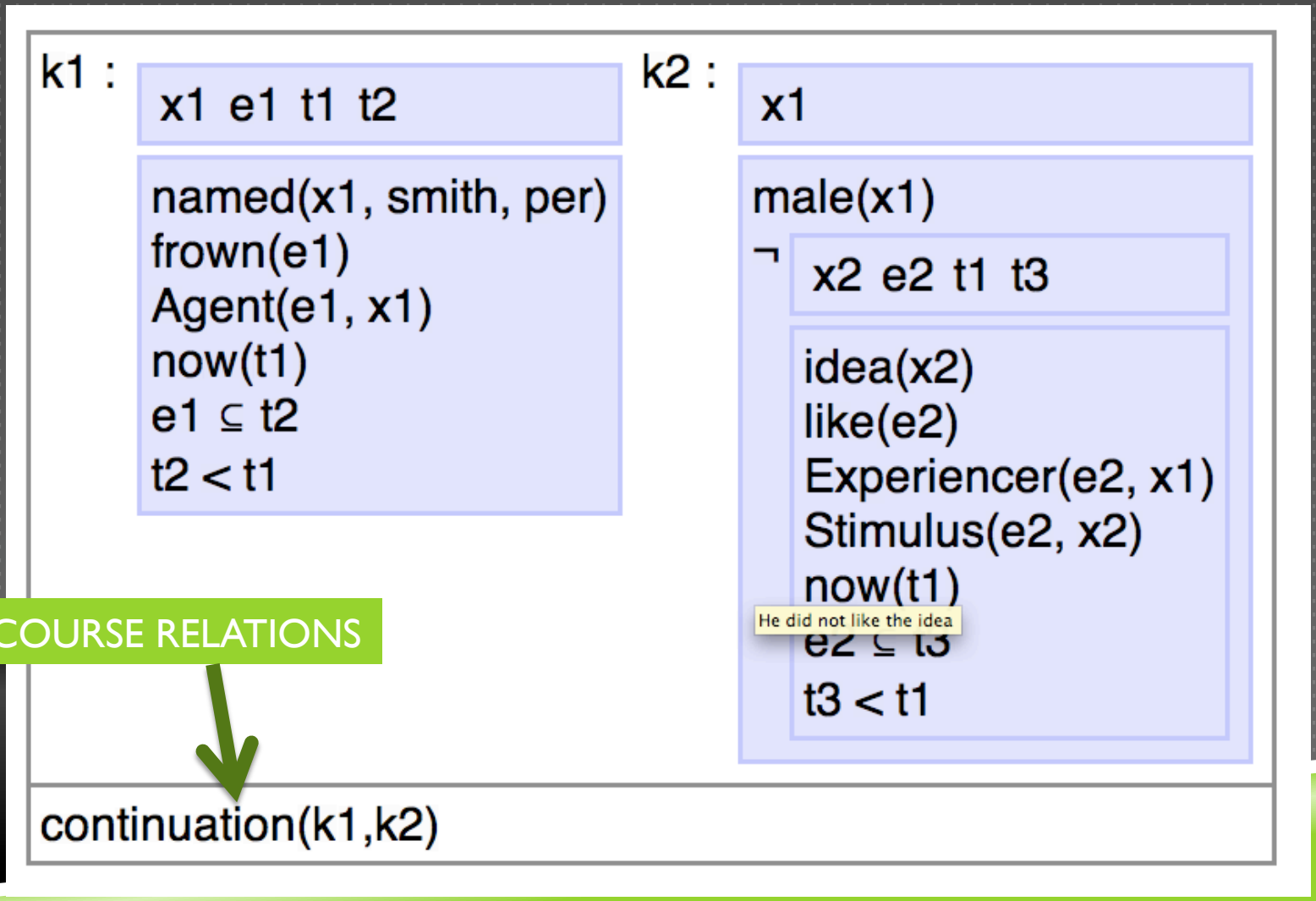
INTEGRATION OF SEMANTICS



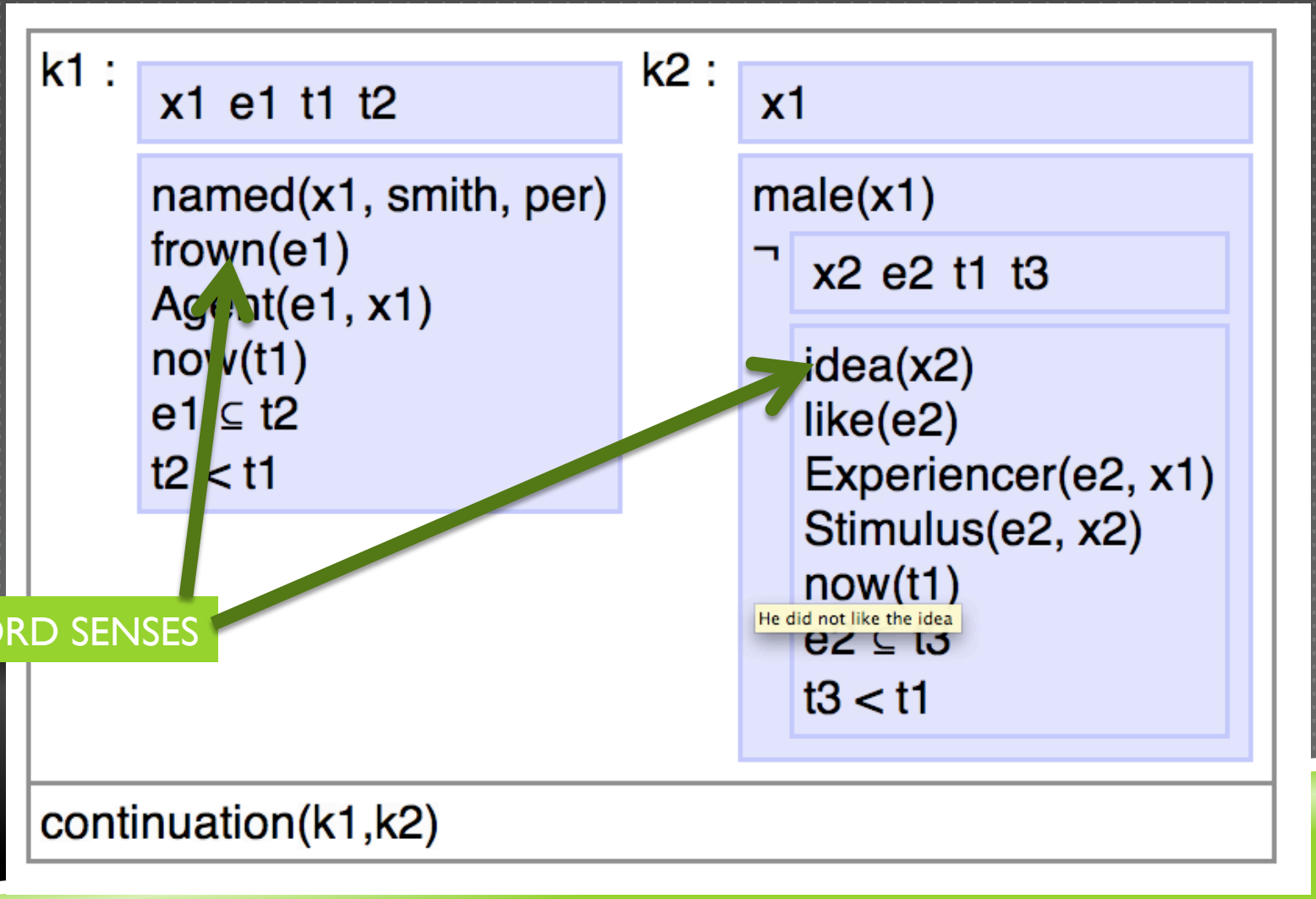
INTEGRATION OF SEMANTICS



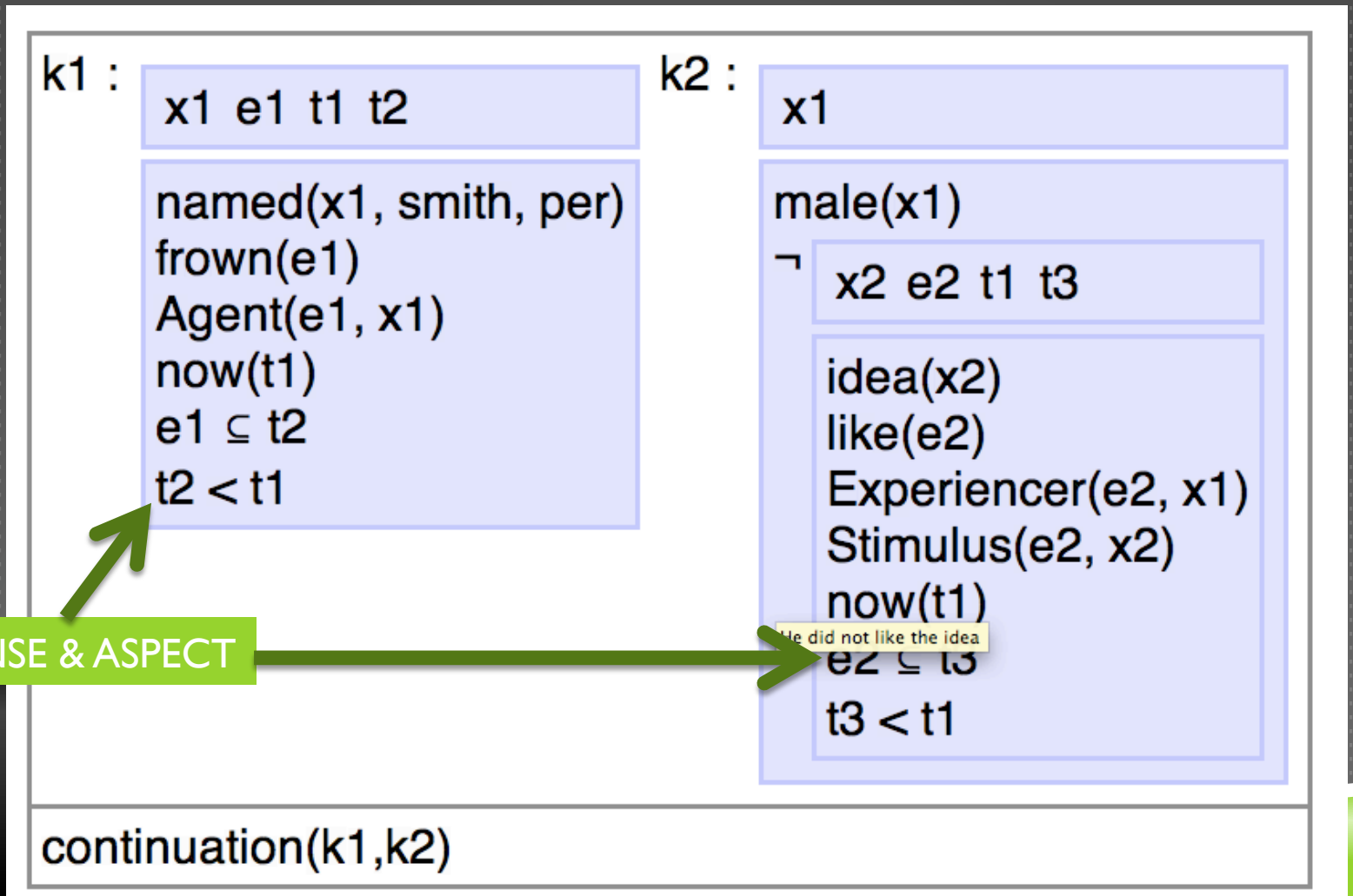
INTEGRATION OF SEMANTICS



INTEGRATION OF SEMANTICS



INTEGRATION OF SEMANTICS



THE GMB IN NUMBERS

- 1,285 wordrobe players
- 62,102 answers (6,982 "bits of wisdom")

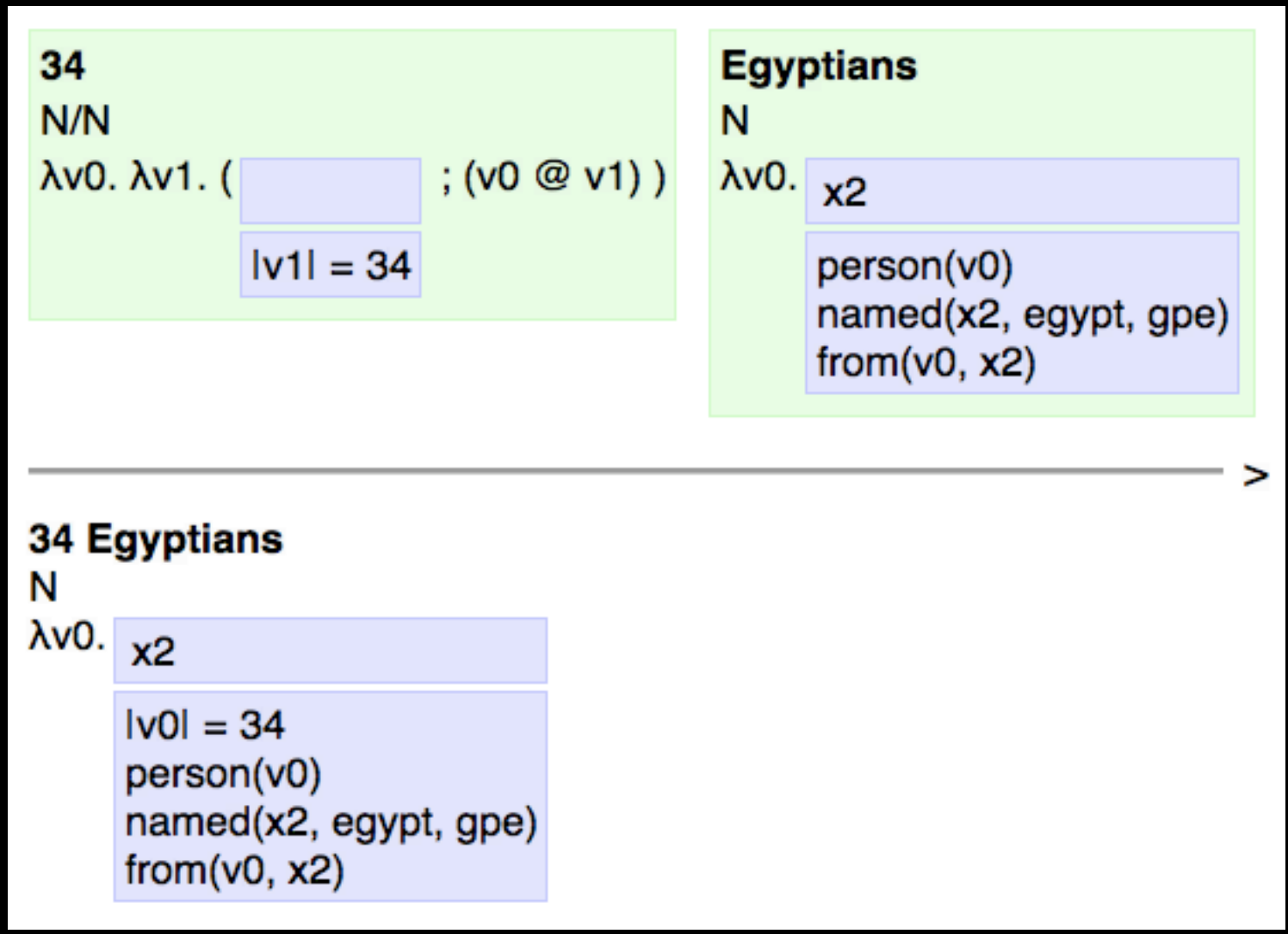
- 30 expert annotators
- 50,013 "bits of wisdom"

- 5 stable releases
- 233 downloads

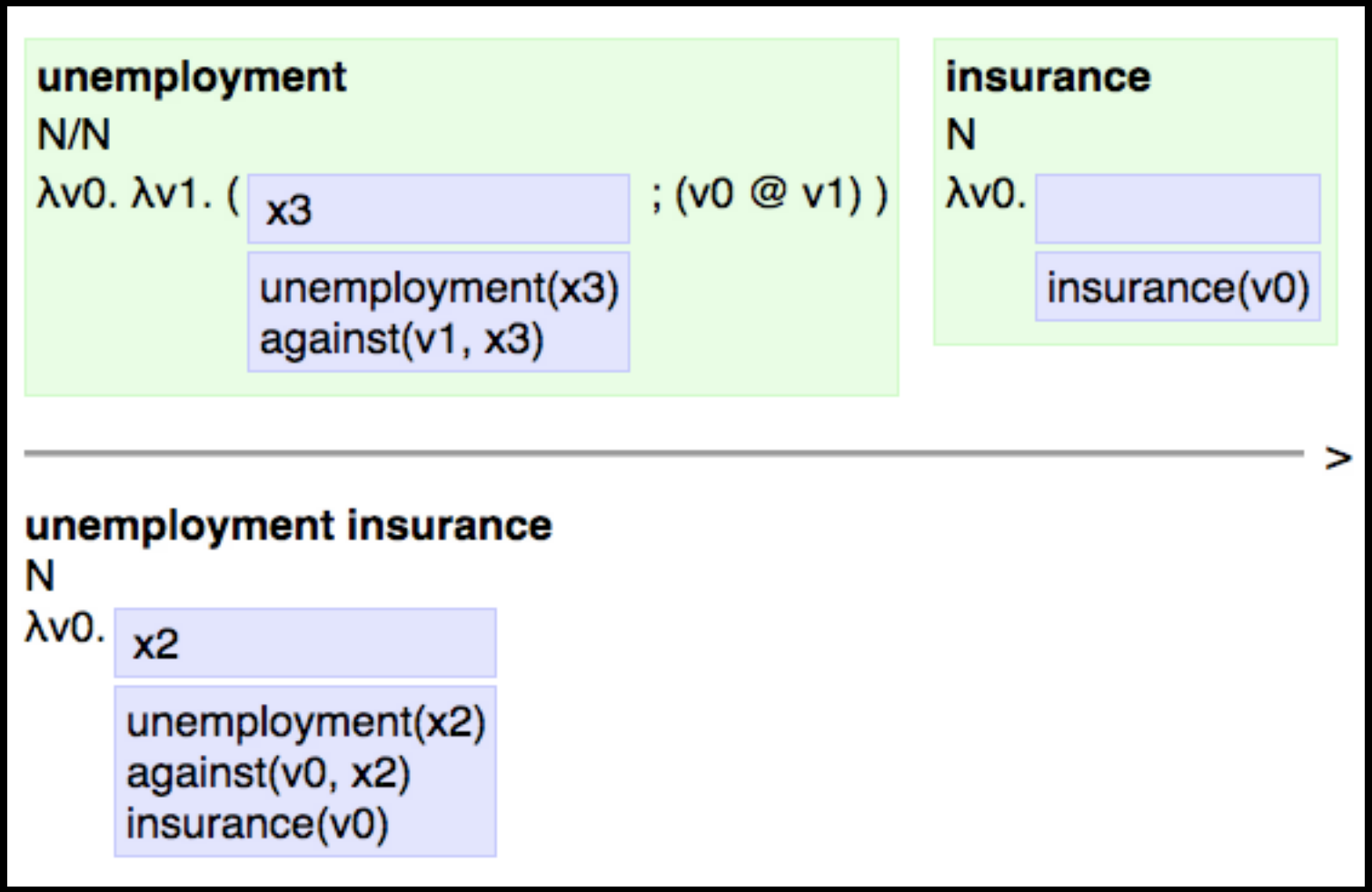
Chapter 7

SWEET MEMORIES

FEATURE #1: DECOMPOSITION



FEATURE #2: NOUN-NOUN COMPOUNDS



FEATURE #3: LONG-DISTANCE DEPENDENCIES

Two	documents	that	a	former	FBI	agent	said	he	took
N/N	N	(NP\NP)/(S[dcl]/NP)	NP/N	N/N	N/N	N	(S[dcl]\NP)/S[dcl]	NP	(S[dcl]\NP)/NP

p4: x9 x10 s11 x12 e13 p14 t16 t17 x18 s19 p20 e21 t23 x24 e25

|x9| = 2
document(x9)
Topic(s11, x10)
former(s11)
named(x12, fbi, org)
of(x10, x12)
agent(x10)
say(e13)
Cause(e13, x10)
Topic(e13, p14)

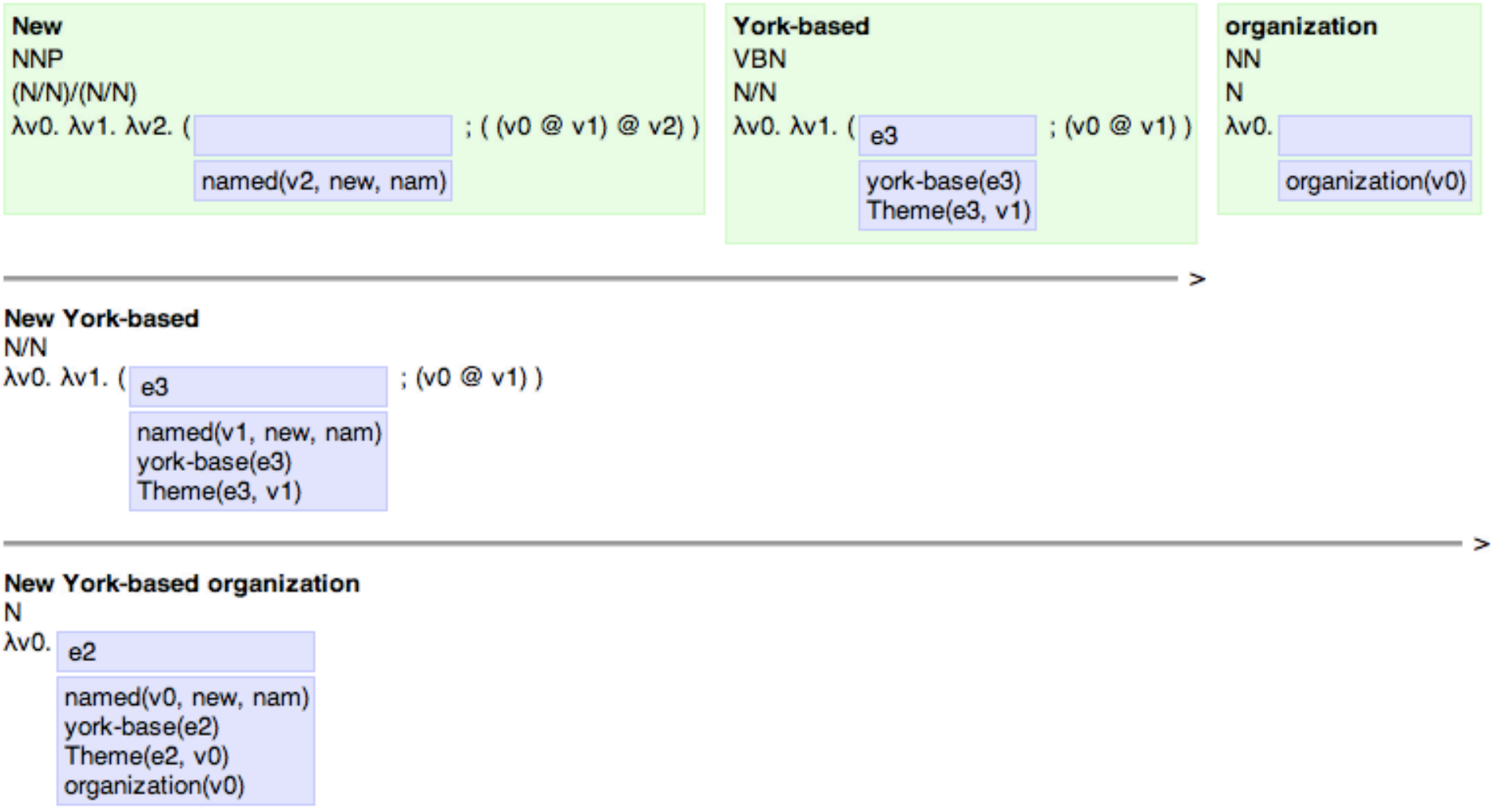
p14: x10 x27 e28 t30 t31 x33 x34 x35 x37 x38 x40 x42 x43 x44 x45 x46

male(x10)
x9 = x27
take(e28)
Agent(e28, x10)
Theme(e28, x27)

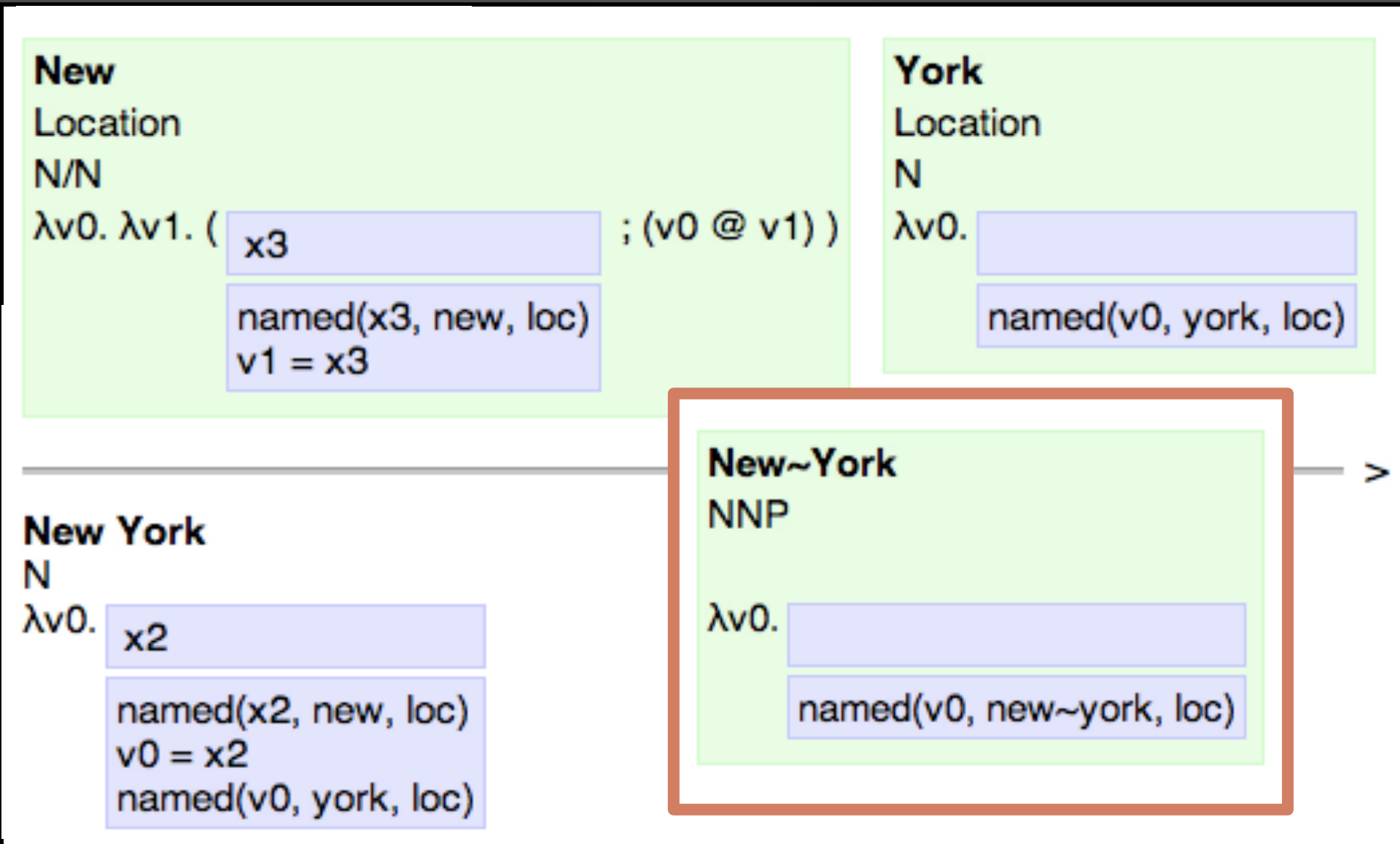
Chapter 8

BUMPS ON THE ROAD

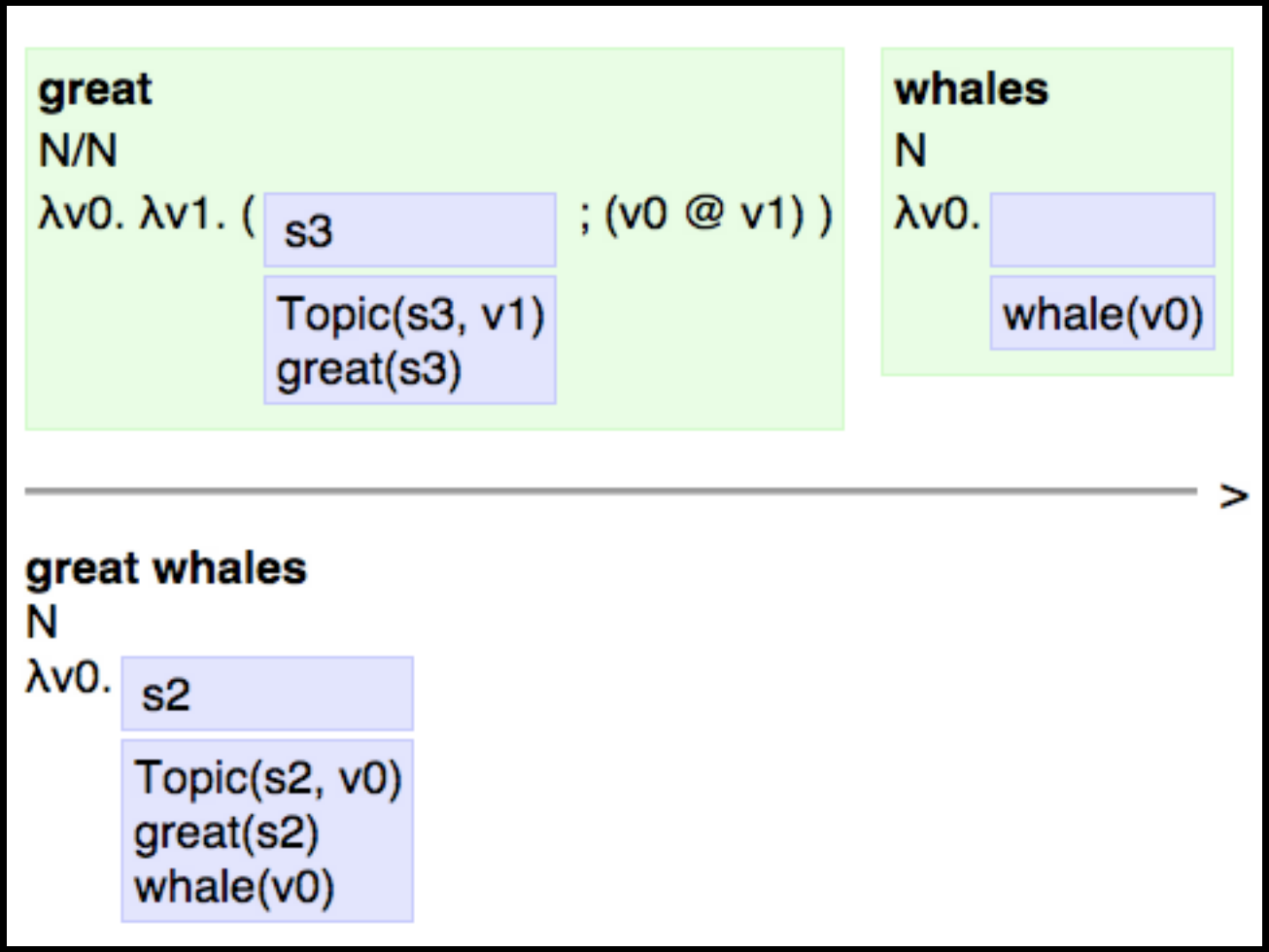
BUMP #1: SEGMENTATION ISSUES



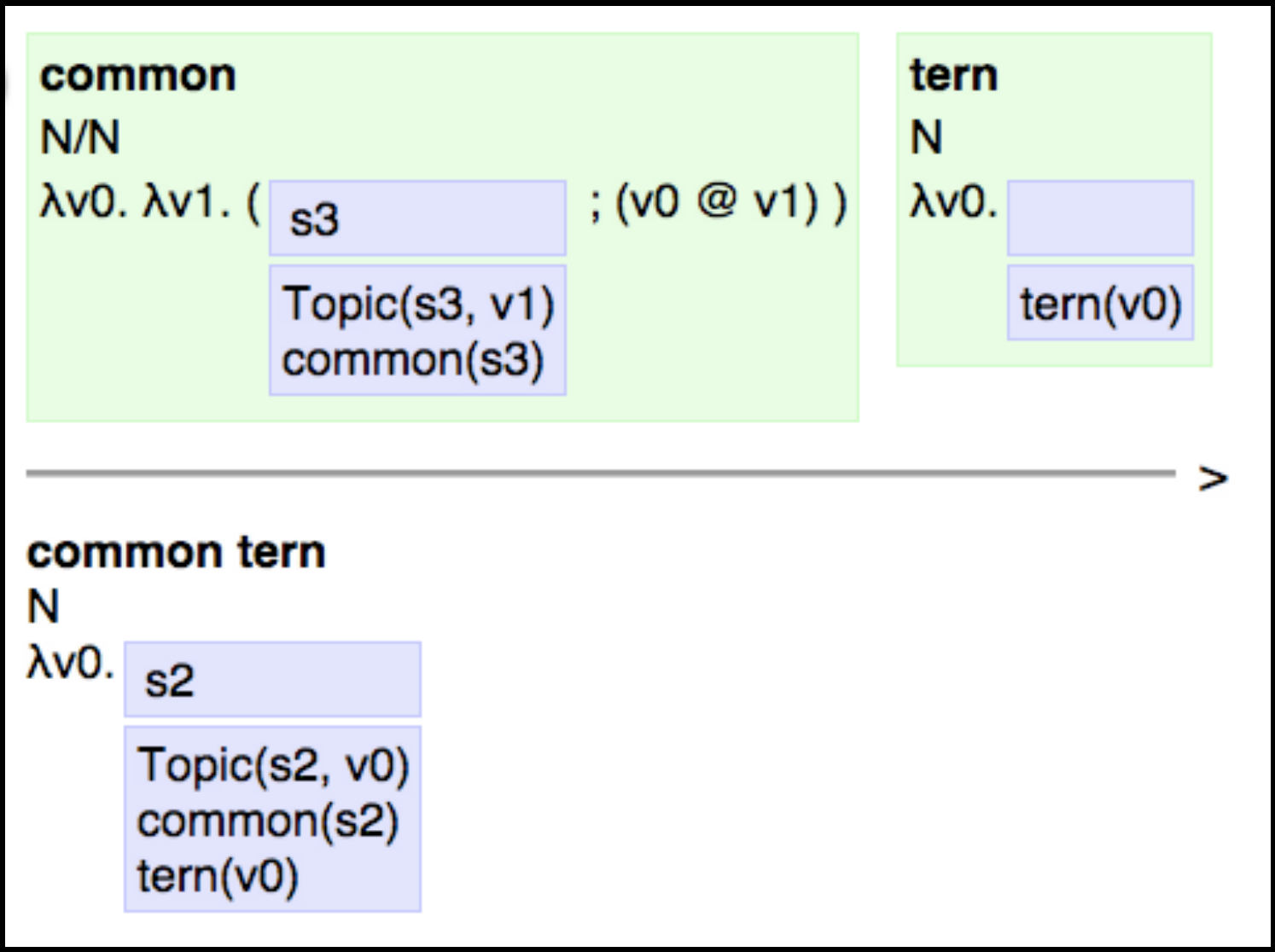
BUMP #1: SEGMENTATION ISSUES



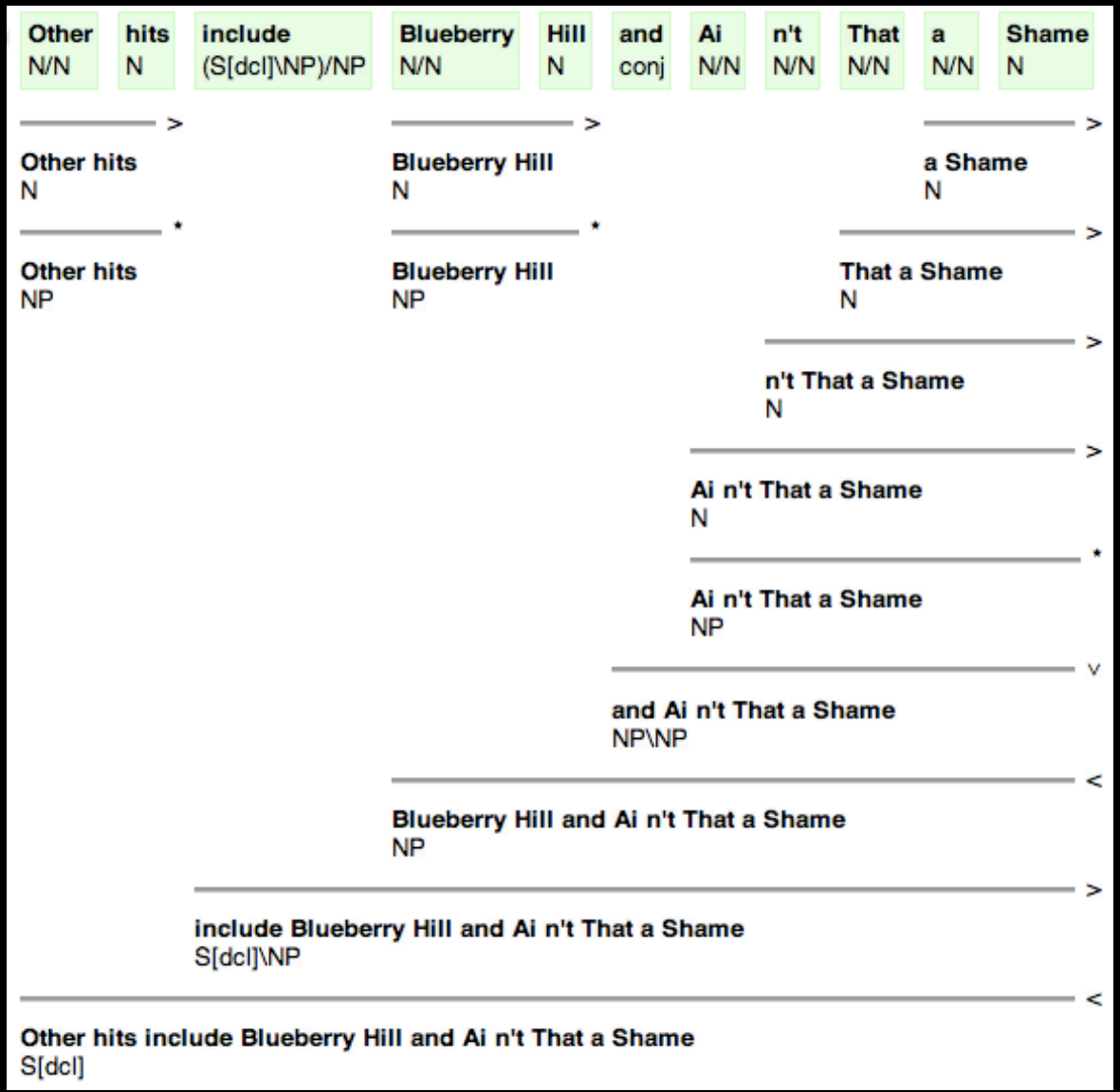
BUMP #2: COMPOUND CLASS NAMES



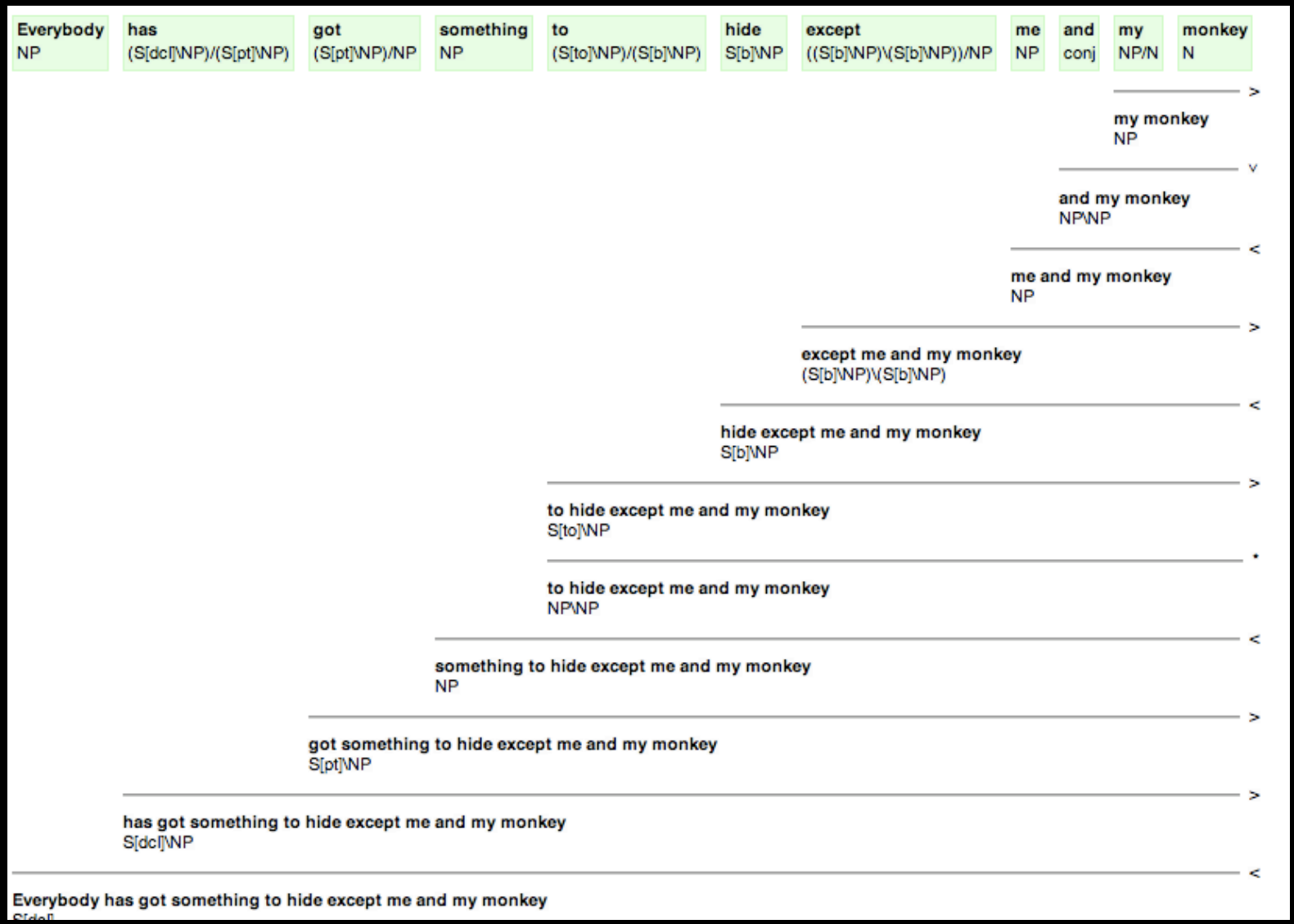
BUMP #2: COMPOUND CLASS NAMES



BUMP #3: ANALYSIS OF NAMES OF CREATIVE WORKS



BUMP #3: ANALYSIS OF NAMES OF CREATIVE WORKS

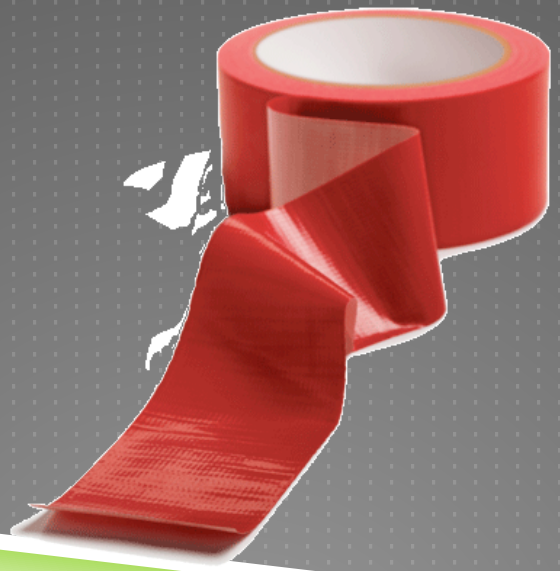


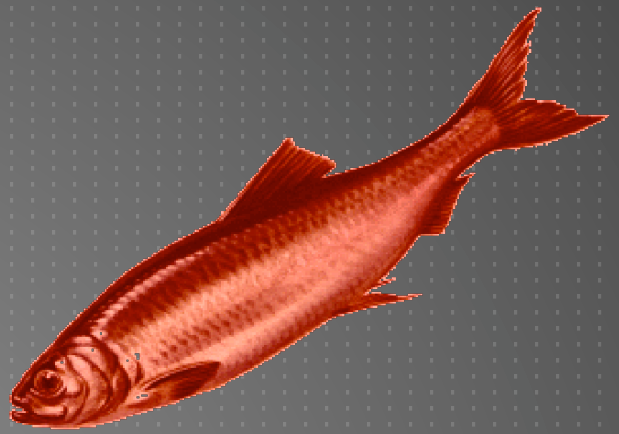
BUMP #4: INTERNAL STRUCTURE OF NAMED ENTITIES

- Los Angeles Mayor Antonio Villaraigosa
- U.S. Secretary of State Condoleezza Rice
- Cobourg Glasgow Rangers Supporters Club
- Air Force Brigadier General Thomas Hemingway
- Burmese Information Minister Brigadier General Kyaw Hsann

BUMP #5: NON-LITERAL INTERPRETATION

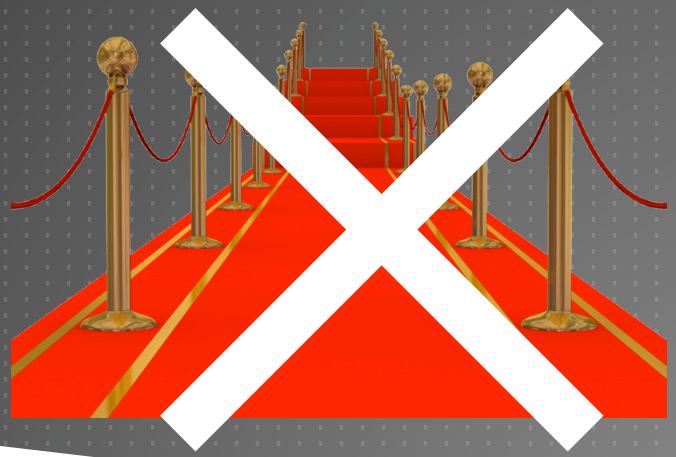
- Today Argentina gets out the red carpet for the UK Duke of York.
- The investment climate suffers from red tape and corruption.
- The plot of the mystery was full of red herrings.

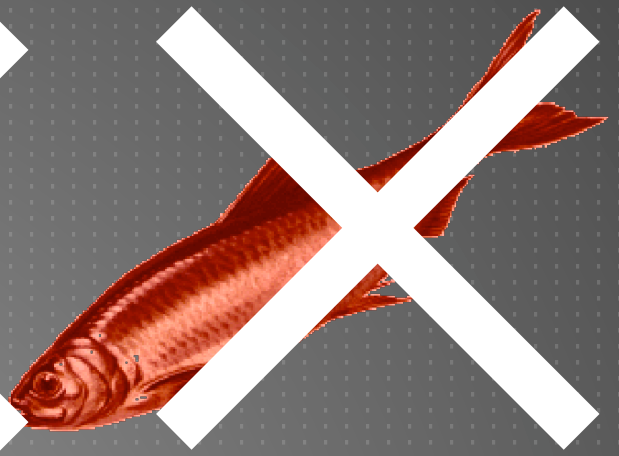




BUMP #5: NON-LITERAL INTERPRETATION

- Today Argentina gets out the red carpet for the UK Duke of York.
- The investment climate suffers from red tape and corruption.
- The plot of the mystery was full of red herrings.





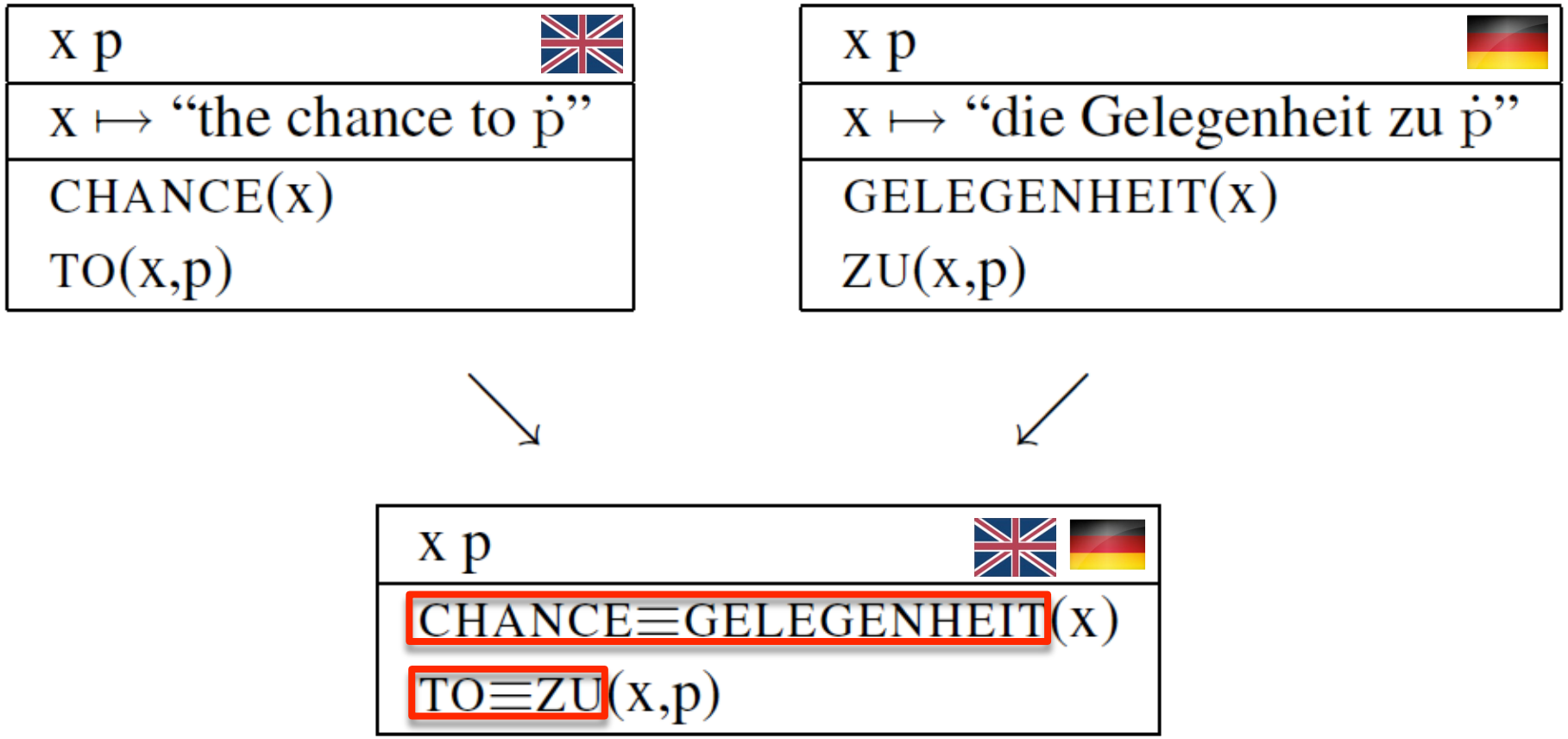
Chapter 9

LOST IN TRANSLATION

PARALLEL MEANING BANKING

- Input: parallel corpus (text aligned)
- Idea: translation ought to be meaning-preserving
- Capture the meaning of a translation by a cross-lingual semantic representation that aligns non-logical symbols of both/all languages
- Introduce new machinery for dealing with faithful (=), loose (≈), and informative translations (C)

EXAMPLE 1


x p
x ↦ “the chance to ṗ”
CHANCE(x)
TO(x,p)
x p
x ↦ “die Gelegenheit zu ṗ”
GELEGENHEIT(x)
ZU(x,p)
x p
CHANCE≡GELEGENHEIT(x)
TO≡ZU(x,p)

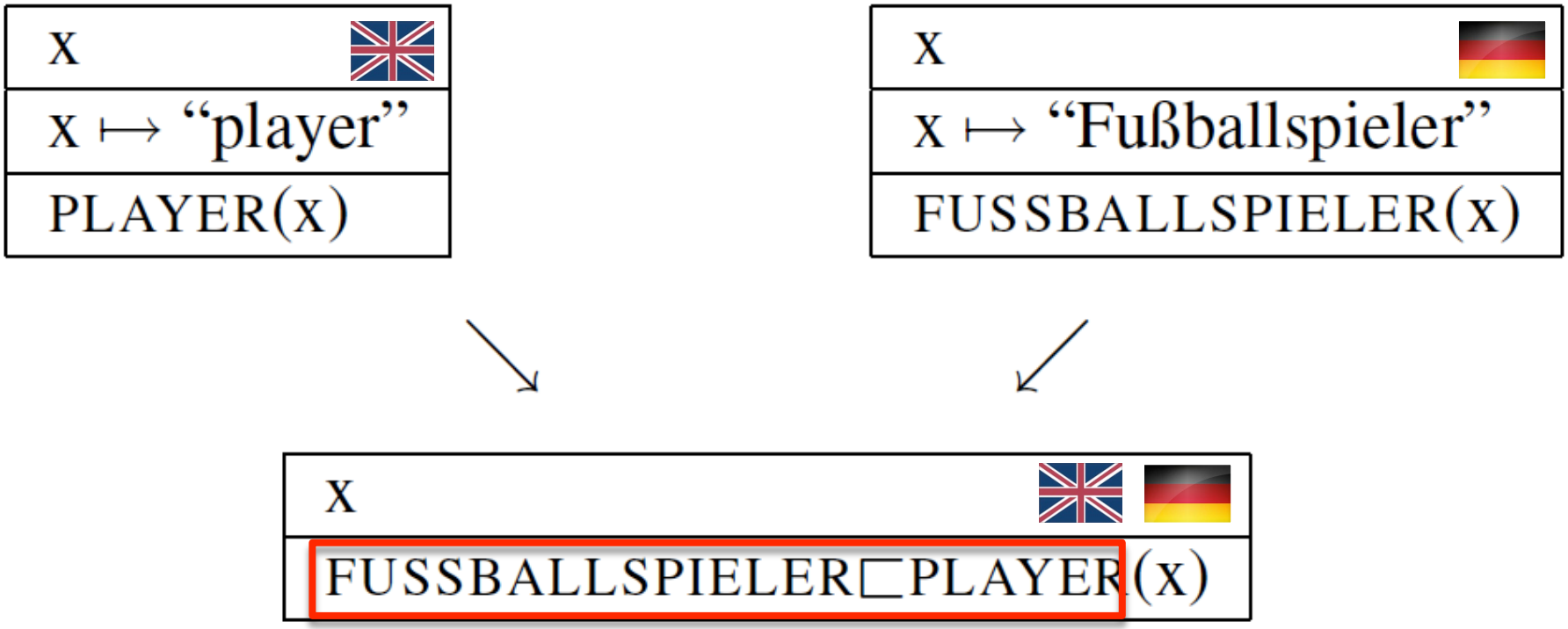
EXAMPLE 2

x e y
x ↦ “The Magpies” e ↦ “ẋ have produced ẏ”
MAGPIES(x) AGENT(e,x) PRODUCE(e) THEME(e,y)

x e y
x ↦ “Die Elstern” e ↦ “ẋ brachten ẏ hervor”
ELSTERN(x) AGENT(e,x) HERVORBRINGEN(e) THEME(e,y)

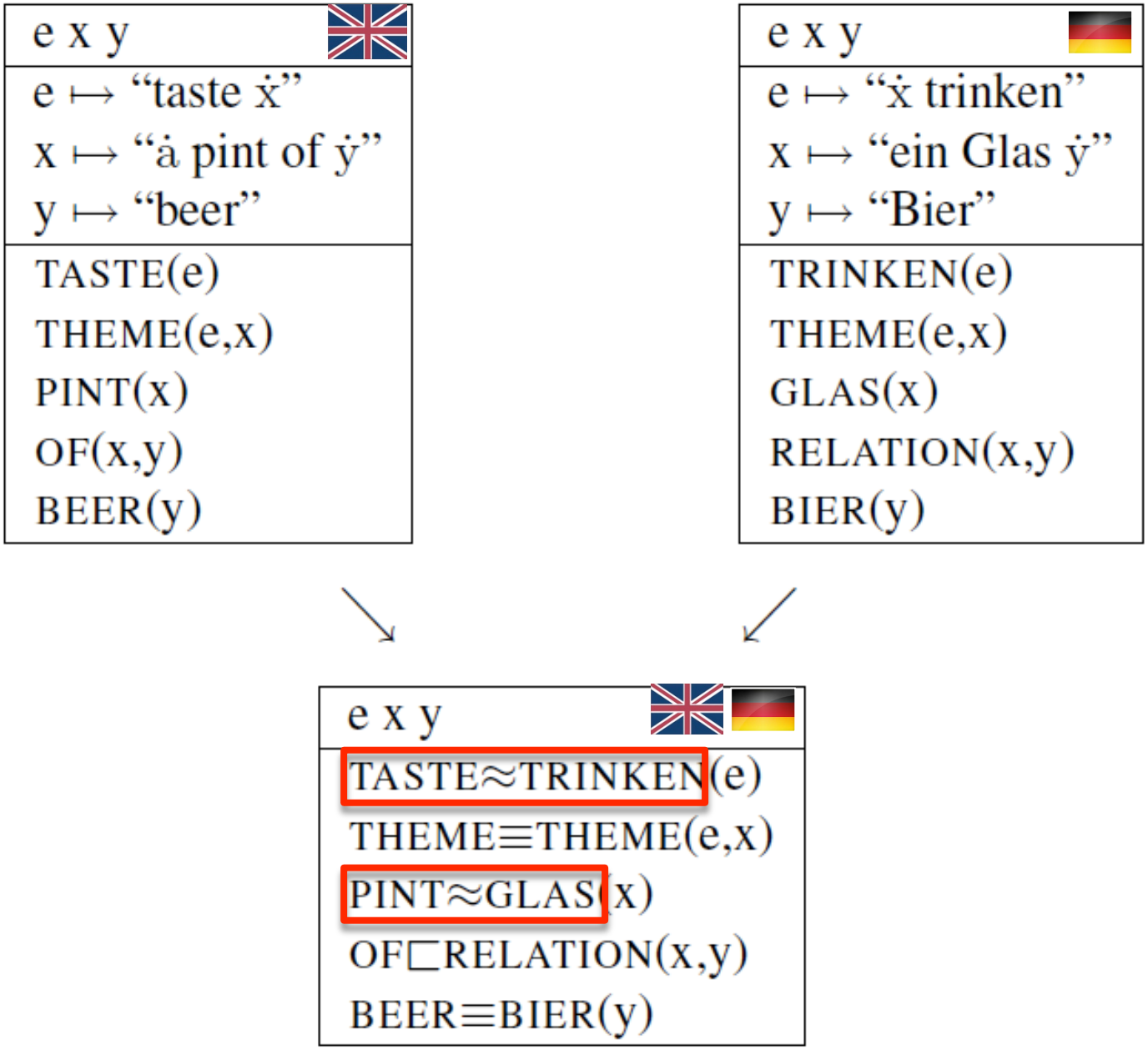
x e y
MAGPIES≡ELSTERN(x) AGENT(e,x) PRODUCE≡HERVORBRINGEN(e) THEME(e,y)

EXAMPLE 3



The “Magpies” have produced some of Britain’s finest **players**.
Die “Elstern” brachten einige der besten **Fussballspieler** Grossbritanniens hervor.

EXAMPLE 4


e x y
e ↦ “taste ẋ”
x ↦ “a pint of ẏ”
y ↦ “beer”
TASTE(e)
THEME(e,x)
PINT(x)
OF(x,y)
BEER(y)
e x y
e ↦ “ẋ trinken”
x ↦ “ein Glas ẏ”
y ↦ “Bier”
TRINKEN(e)
THEME(e,x)
GLAS(x)
RELATION(x,y)
BIER(y)
e x y
TASTE≈TRINKEN(e)
THEME≡THEME(e,x)
PINT≈GLAS(x)
OF⊏RELATION(x,y)
BEER≡BIER(y)

Chapter 10

THE END

CONCLUSION

- The GMB is a public domain corpus with formal semantic annotations for texts (not sentences)
- Anyone can contribute! (experts and street folk)
- Integration of semantic information into one formalism (DRS)
- Everyone can download it (free for research purposes)

http://gmb.let.rug.nl

Groningen MEANING BANK

REFERENCES

- Valerio Basile, Johan Bos, Kilian Evang, Noortje Venhuizen (2012): *Developing a large semantically annotated corpus*. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp 3196–3200, Istanbul, Turkey.

- Noortje Venhuizen, Valerio Basile, Kilian Evang, Johan Bos (2013): *Gamification for Word Sense Labeling*. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013), pp 397–403, Potsdam, Germany

- J. Bos (2014): Semantic Annotation Issues in Parallel Meaning Banking. Proceedings of the Tenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-10), pp 17–20, Reykjavik, Iceland.

http://www.let.rug.nl/bos/publications.html

Groningen MEANING BANK