Die Parallel Meaning Bank

Kilian Evang, Christian Wurm

Düsseldorf, 22.12.2022

Einleitung

- Großes, tiefensemantisch annotiertes Korpus
- Teilweise parallel: Englisch, Deutsch, Italienisch, Niederländisch, neuerdings auch Chinesisch, Japanisch, Französisch
- Im Aufbau seit 2015, aktuelles Release 4.0.0
- Vorgängerin: Groningen Meaning Bank (nur Englisch, 2010–2015)
- Zweck: Semantic Parsing, empirische Untersuchung des Syntax-Semantik-Interface verschiedener Sprachen...
- Überwiegend freie Texte verschiedener Genres: Nachrichtenmeldungen, Fabeln, Tatoeba...

Methodik

- Semiautomatische Annotation
  - eine NLP-Pipeline kreiert Vorannotationen
  - Annotator*innen korrigieren diese
- Dokumente in Releases haben verschiedene Status:
  - bronze = nur automatische Vorannotation
  - silver = ein Mensch hat mindestens eine Sache korrigiert (und wenn es eklatante Fehler gäbe, hätte er*sie sie vermutlich direkt mitkorrigiert)
  - gold = ein Mensch hat markiert, dass alle Annotationsschichten korrekt sind

Annotationsschichten (1/2)

- Baumbanken wie Penn Treebank oder Universal Dependencies annotieren „nur“ Syntax
- PropBank annotiert „nur“ semantische Frames und Rollen
- PMB annotiert sehr viele Schichten
- Teilweise sind sie ein „Nebenprodukt“ der NLP-Pipeline, die für die Vorannotation genutzt wird
- Letztlich wird alle Information in die DRS integriert, die am Ende herauskommt

Annotationsschichten (2/2)

1. Segmentierung (in Sätze und Tokens)
2. Semantische Tags (ähnlich wie POS-Tags, aber ohne morphosyntaktische Information, sondern semantische Funktion, insbesondere für Funktionswörter wichtig)
3. Synset (nach WordNet, feinkörnige Unterscheidung der Wortbedeutung von Inhaltswörtern)
4. Semantische Rollen
5. Skopus
6. Anaphora
7. CCG-Kategorien

Integration der Annotationsshichten

1. Basierend auf den CCG-Kategorien baut EasyCCG (Lewis and Steedman, 2014) eine Derivation für jeden Satz
2. Basierend auf allem baut Boxer (Bos, 2015) eine DRS für jedes Dokument
   - 2.1 weist Tokens λ-DRSs zu (basierend auf Kategorien, semantischen Tags, Synsets, semantischen Rollen)
   - 2.2 berechnet Semantik für Satz (deterministisch anhand Derivation)
   - 2.3 wendet leeren Modifizierer an (Resultat: DRS ohne λ-Abstraktion)
   - 2.4 löst Anaphern auf (basierend auf Anaphern-Annotation)
   - 2.5 integriert DRSen in eine große DRS mit Diskursrelationen

Ausgabeformate

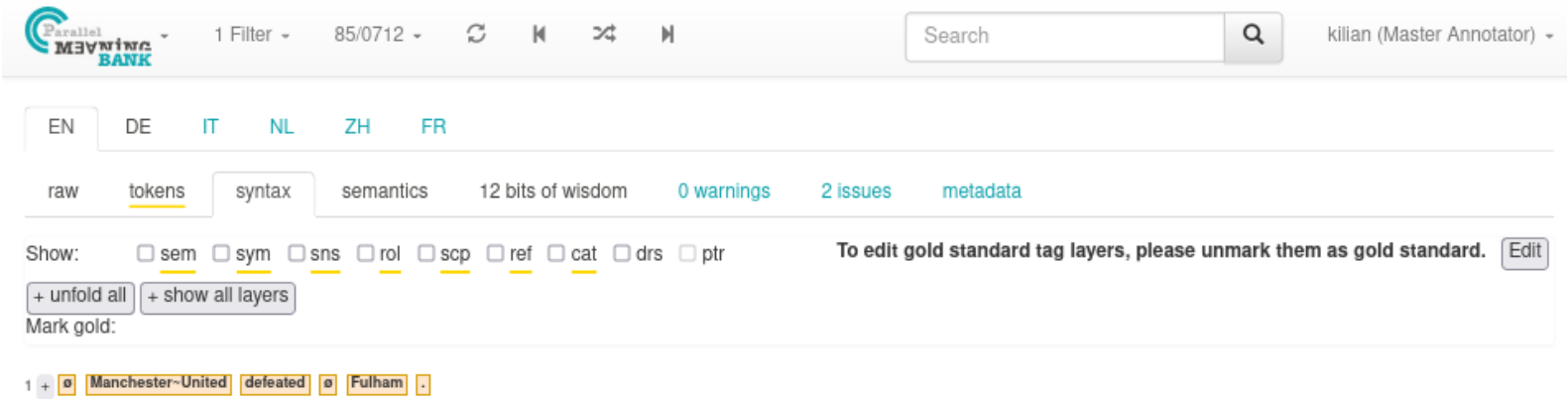
- Discourse Representation Structure (box notation)
- Discourse Representation Structure (clause notation)
- Discourse Representation Graph (graph notation)
- Discourse Representation Graph (sequence notation)

PMB-basierte Forschung

- Theoretische Semantik
- Semantic Parsing
- Textgenerierung
- Semantic Tagging und Vektorrepräsentationen

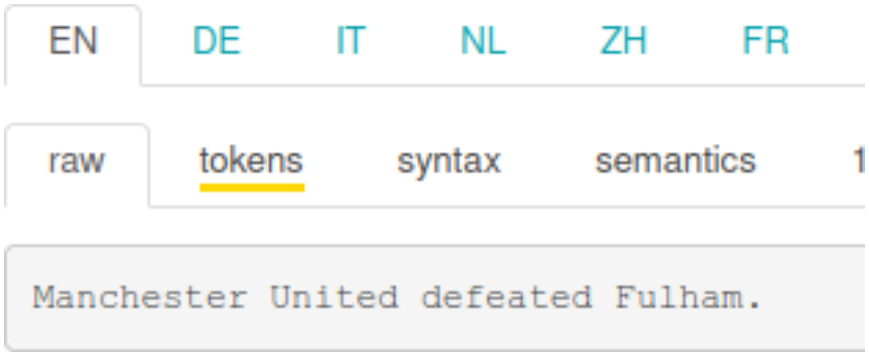
`https://pmb.let.rug.nl/publications.php`

PMB Explorer



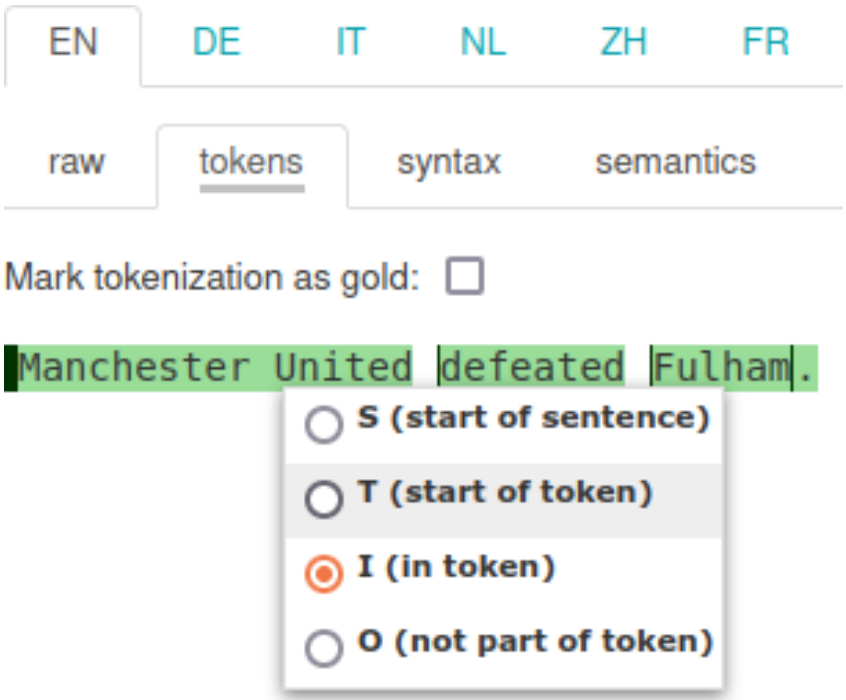
https://pmb.let.rug.nl/explorer

Rohdaten



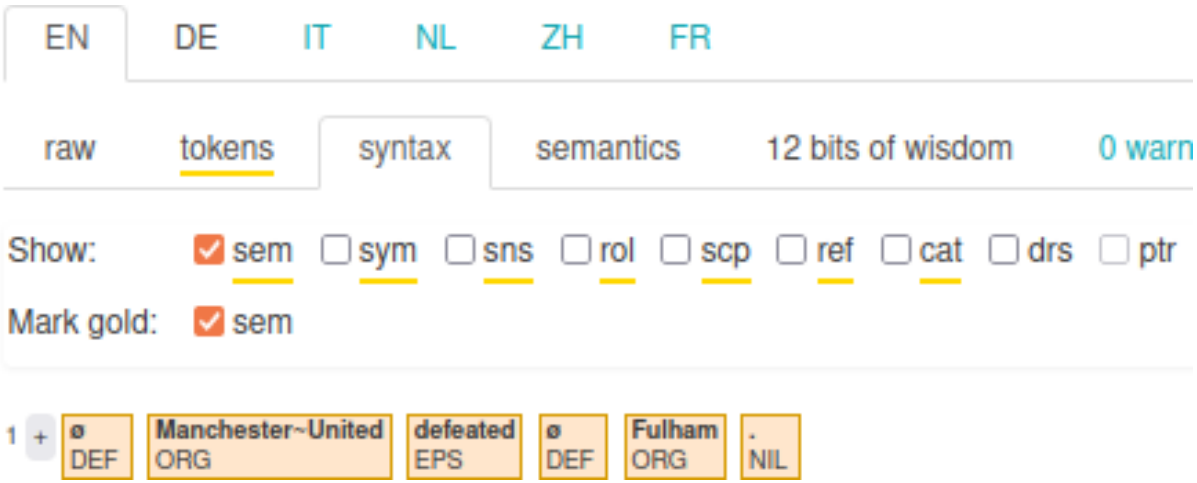
https://pmb.let.rug.nl/explorer/explore.php?part=85&doc_id=0712&type=raw

Segmentierung



Eigennamen werden als ein Token behandelt, auch, wenn sie Leerzeichen enthalten.

Semantische Tags

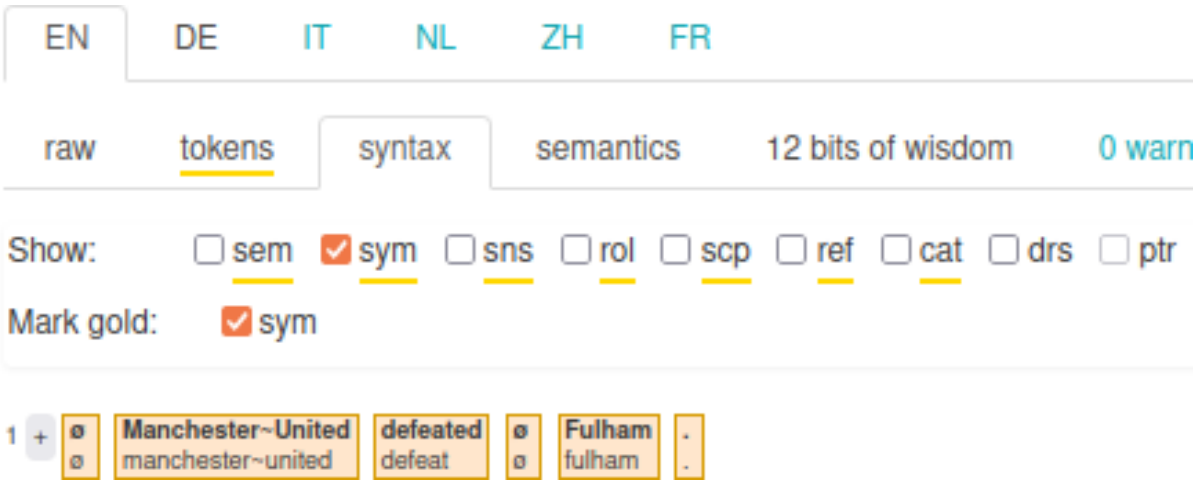


Semantische Tags (Abzianidze and Bos, 2017)

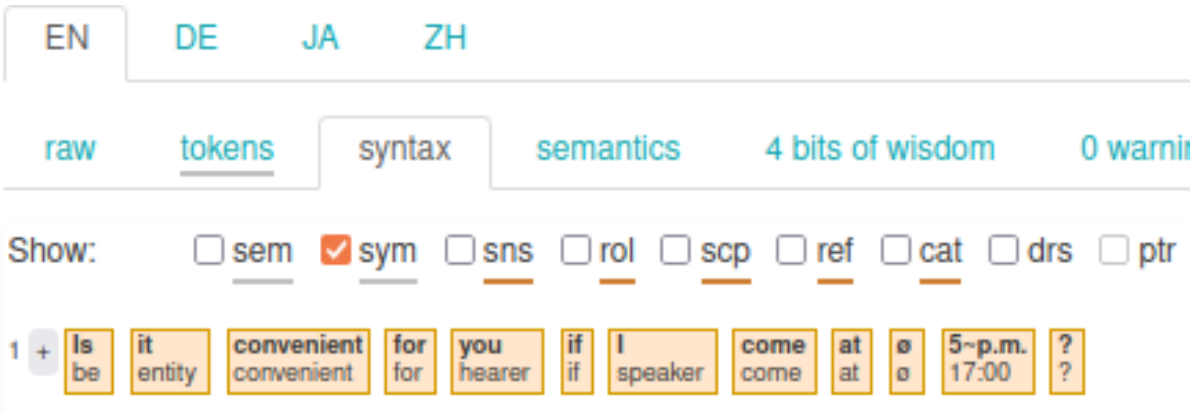
ANA anaphoric	PRO	anaphoric & deictic pronouns: *he, she, I, him*
	DEF	definite: *the, lo*^{IT}*, der*^{DE}
	HAS	possessive pronoun: *my, her*
	REF	reflexive & reciprocal pron.: *herself, each_other*
	EMP	emphasizing pronouns: *himself*
ACT speech act	GRE	greeting & parting: *hi, bye*
	ITJ	interjections, exclamations: *alas, ah*
	HES	hesitation: *err*
	QUE	interrogative: *who, which, ?*
ATT attribute	QUC*	concrete quantity: *two, six_million, twice*
	QUV*	vague quantity: *millions, many, enough*
	COL*	colour: *red, crimson, light_blue, chestnut_brown*
	IST	intersective: *open, vegetarian, quickly*
	SST	subsective: *skillful* *surgeon*, *tall* *kid*
	PRI	privative: *former, fake*
	DEG*	degree: *2 meters* *tall*, *20 years* *old*
	INT	intensifier: *very, much, too, rather*
	REL	relation: *in, on, 's, of, after*
	SCO	score: *3-0*, *grade* *A*
COM comparative	EQU	equative: *as* *tall as John*, *whales* *are* *mammals*
	MOR	comparative positive: *better, more*
	LES	comparative negative: *less, worse*
	TOP	superlative positive: *most, mostly*
	BOT	superlative negative: *worst, least*
	ORD	ordinal: *1st, 3rd, third*
UNE unnamed entity	CON	concept: *dog, person*
	ROL	role: *student, brother, prof., victim*
	GRP*	group: *John* {,} *Mary* *and* *Sam gathered*, *a* *group* *of people*
DXS deixis	DXP*	place deixis: *here, this, above*
	DXT*	temporal deixis: *just, later, tomorrow*
	DXD*	discourse deixis: *latter, former, above*
LOG logical	ALT	alternative & repetitions: *another, different, again*
	XCL	exclusive: *only, just*
	NIL	empty semantics: {.}, *to, of*
	DIS	disjunction & exist. quantif.: *a, some, any, or*
	IMP	implication: *if, when, unless*
	AND	conjunction & univ. quantif.: *every, and, who, any*

MOD modality	NOT	negation: *not, no, neither, without*
	NEC	necessity: *must, should, have* *to*
	POS	possibility: *might, could, perhaps, alleged, can*
DSC discourse	SUB	subordinate relations: *that, while, because*
	COO	coordinate relations: *so,* {,}, {;}, *and*
	APP	appositional relations: {,}, *which,* {()}, {—}
	BUT	contrast: *but, yet*
NAM named entity	PER	person: *Axl_Rose, Sherlock_Holmes*
	GPE	geo-political entity: *Paris, Japan*
	GPO*	geo-political origin: *Parisian, French*
	GEO	geographical location: *Alps, Nile*
	ORG	organization: *IKEA, EU*
	ART	artifact: *iOS_7*
	HAP	happening: *Eurovision_2017*
	UOM	unit of measurement: *meter, $, %, degree_Celsius*
	CTC*	contact information: *112, info@mail.com*
	URL	URL: `http://pmb.let.rug.nl`
	LIT*	literal use of names: *his name is* *John*
	NTH*	other names: *table* *1a*, *equation* *(1)*
EVE events	EXS	untensed simple: *to* *walk*, *is* *eaten, destruction*
	ENS	present simple: *we* *walk*, *he* *walks*
	EPS	past simple: *ate, went*
	EXG	untensed progressive: *is* *running*
	EXT	untensed perfect: *has* *eaten*
TNS tense & aspect	NOW	present tense: *is* *skiing*, *do* *ski*, *has* *skied*, *now*
	PST	past tense: *was* *baked*, *had* *gone*, *did* *go*
	FUT	future tense: *will, shall*
	PRG*	progressive: *has been* *being* *treated*, *aan_het*^{NL}
	PFT*	perfect: *has* *been* *going/done*
TIM temporal entity	DAT*	full date: *27.04.2017, 27/04/17*
	DOM	day of month: *27th* *December*
	YOC	year of century: *2017*
	DOW	day of week: *Thursday*
	MOY	month of year: *April*
	DEC	decade: *80s, 1990s*
	CLO	clocktime: *8:45_pm, 10_o'clock, noon*

Symbole (1/2)

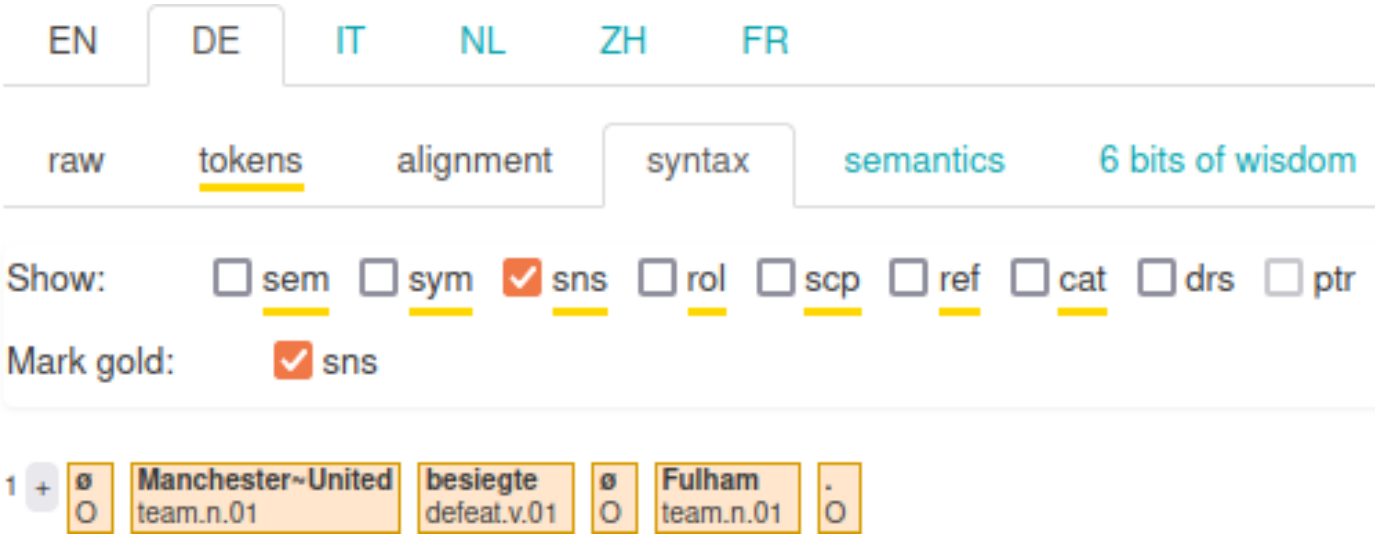

EN
DE
IT
NL
ZH
FR
raw
tokens
syntax
semantics
12 bits of wisdom
0 warn
Show:
sem
sym
sns
rol
scp
ref
cat
drs
ptr
Mark gold:
sym
1
ø
ø
Manchester~United
manchester~united
defeated
defeat
ø
ø
Fulham
fulham
.
.

Symbole (2/2)

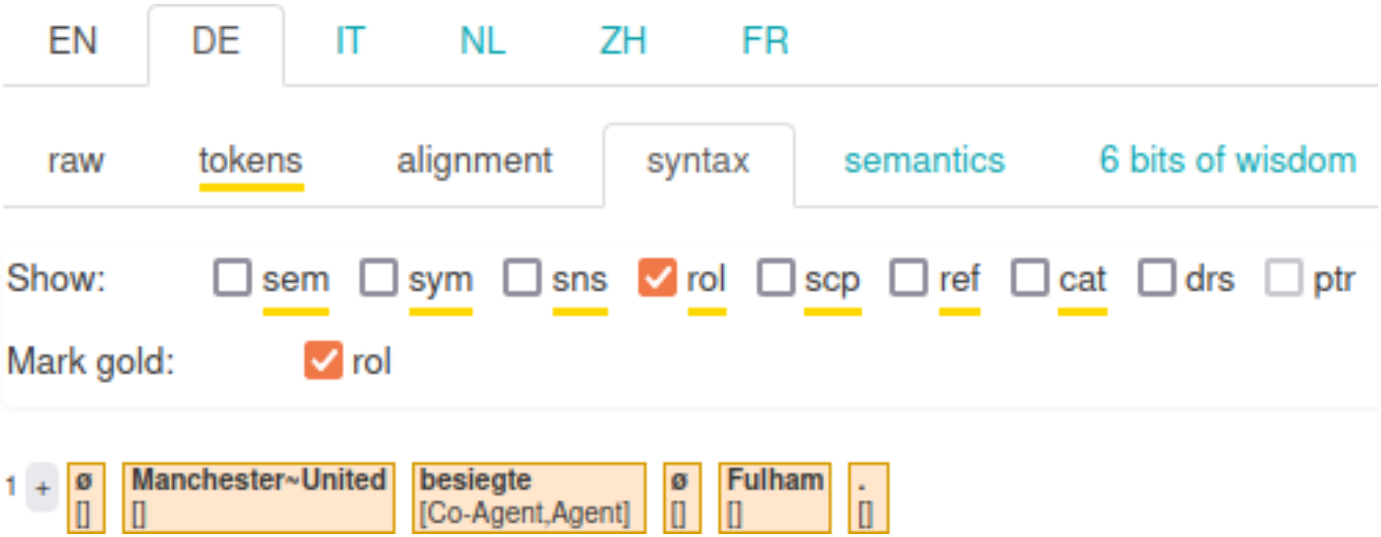


https://pmb.let.rug.nl/explorer/explore.php?part=28&doc_id=2273&type=der.xml

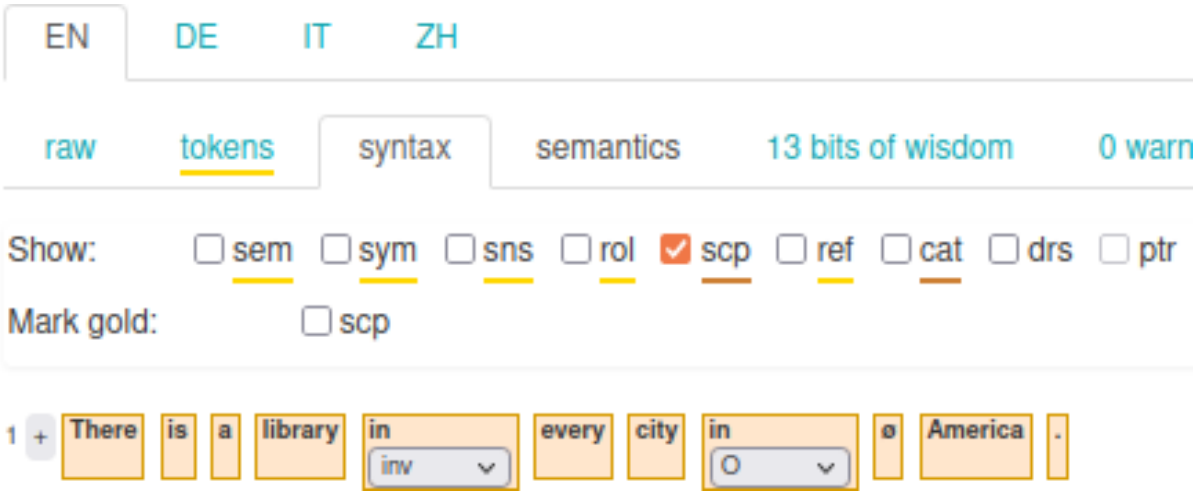
Synsets


EN
DE
IT
NL
ZH
FR
raw
tokens
alignment
syntax
semantics
6 bits of wisdom
Show:
sem
sym
sns
rol
scp
ref
cat
drs
ptr
Mark gold:
sns
1
+
ø
O
Manchester~United
team.n.01
besiegte
defeat.v.01
ø
O
Fulham
team.n.01
.
O

Semantische Rollen

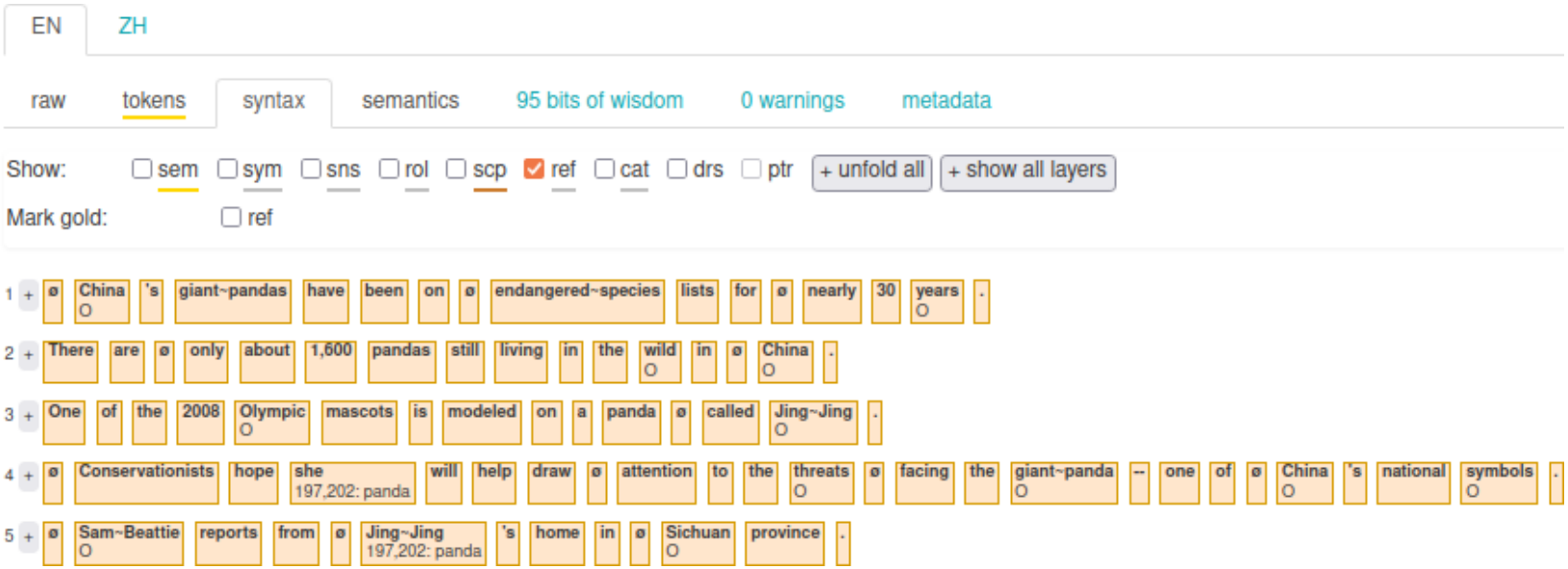


Skopus



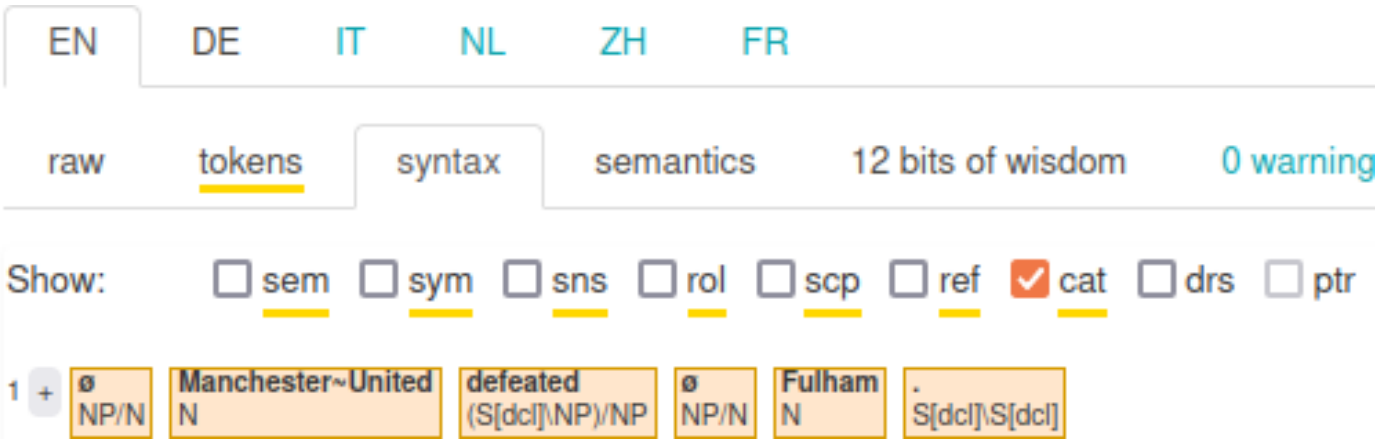
https://pmb.let.rug.nl/explorer/explore.php?part=00&doc_id=1428&type=der.xml

Koreferenz

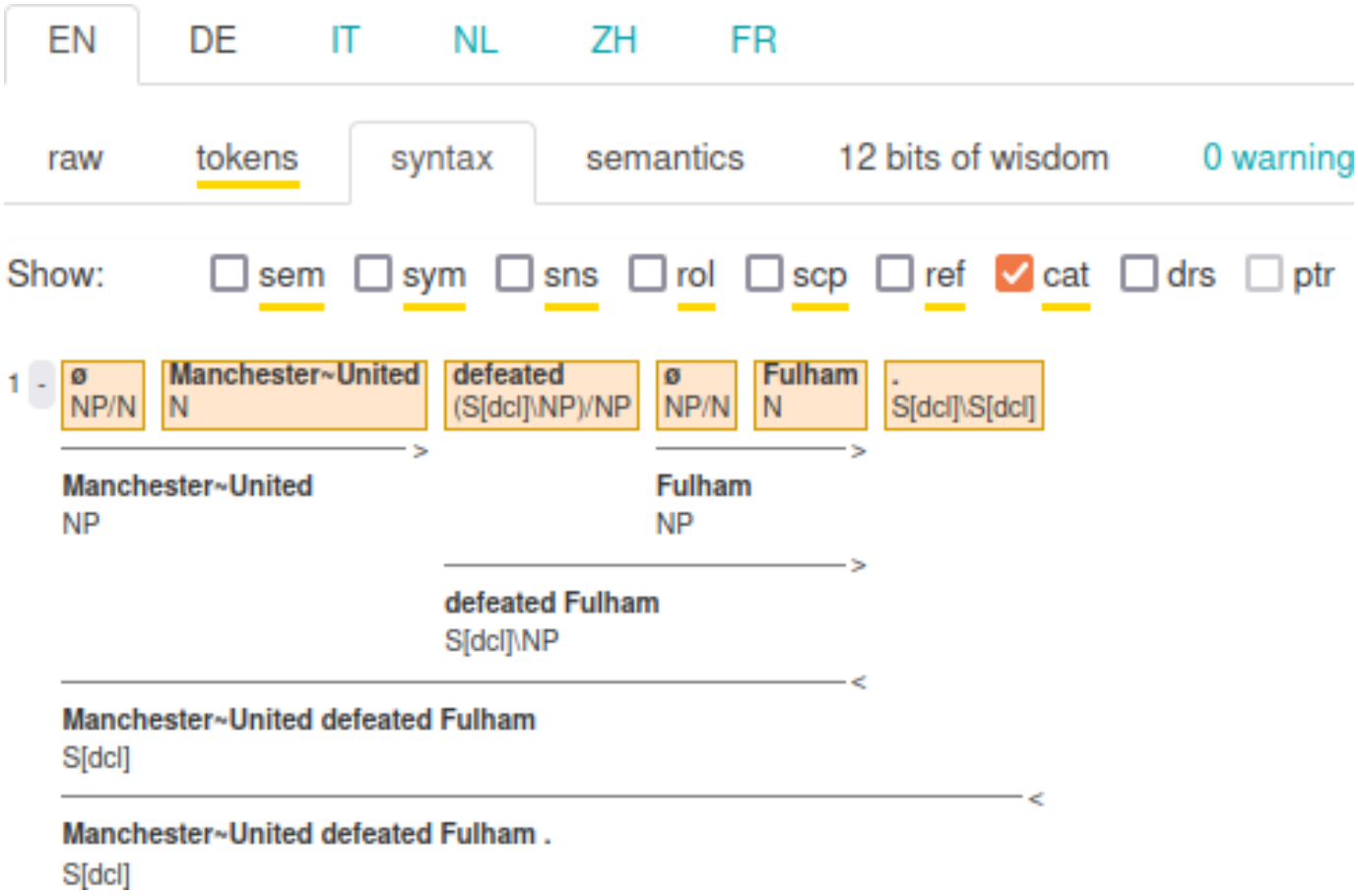


https://pmb.let.rug.nl/explorer/explore.php?part=00&doc_id=0252&type=der.xml

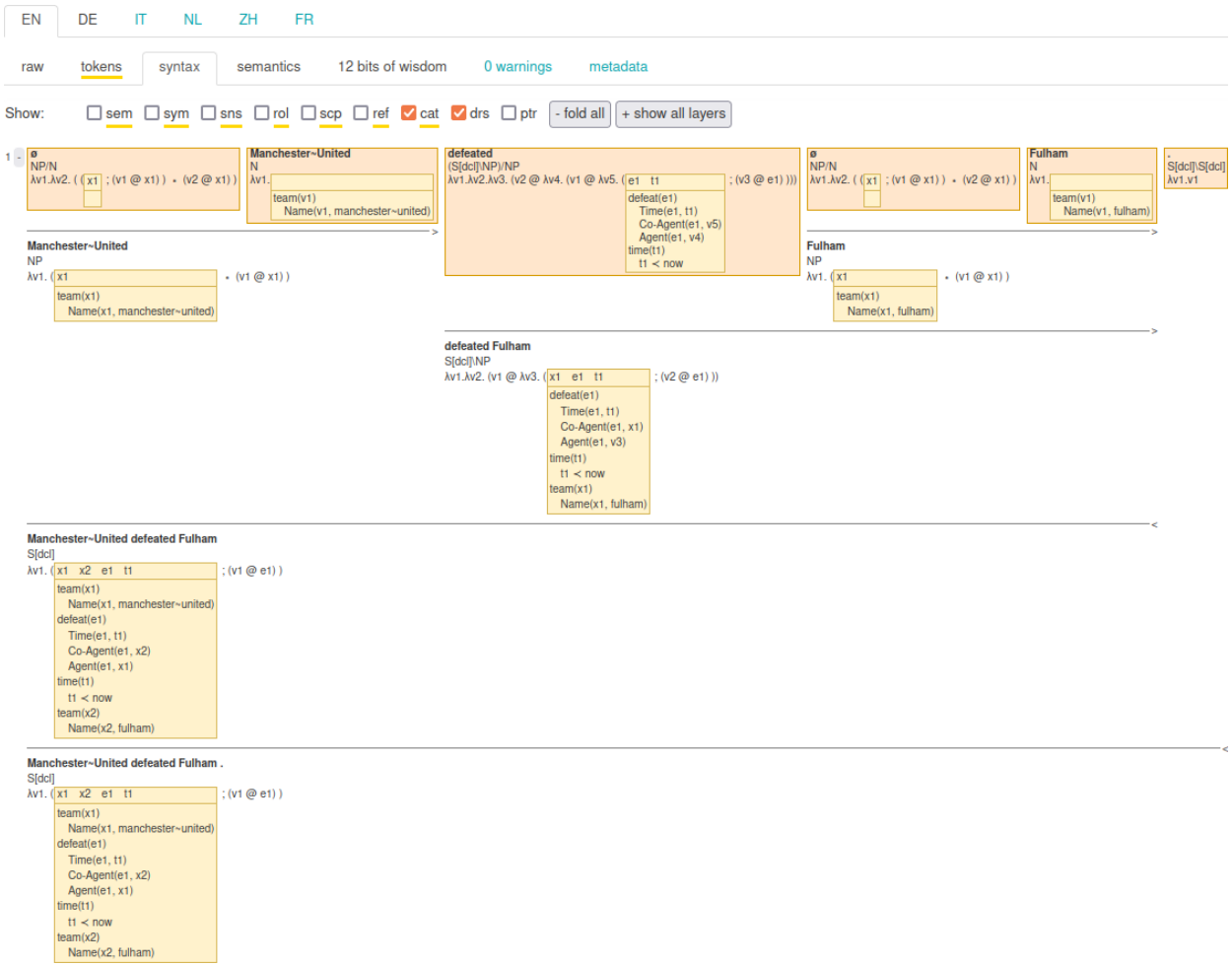
CCG-Kategorien



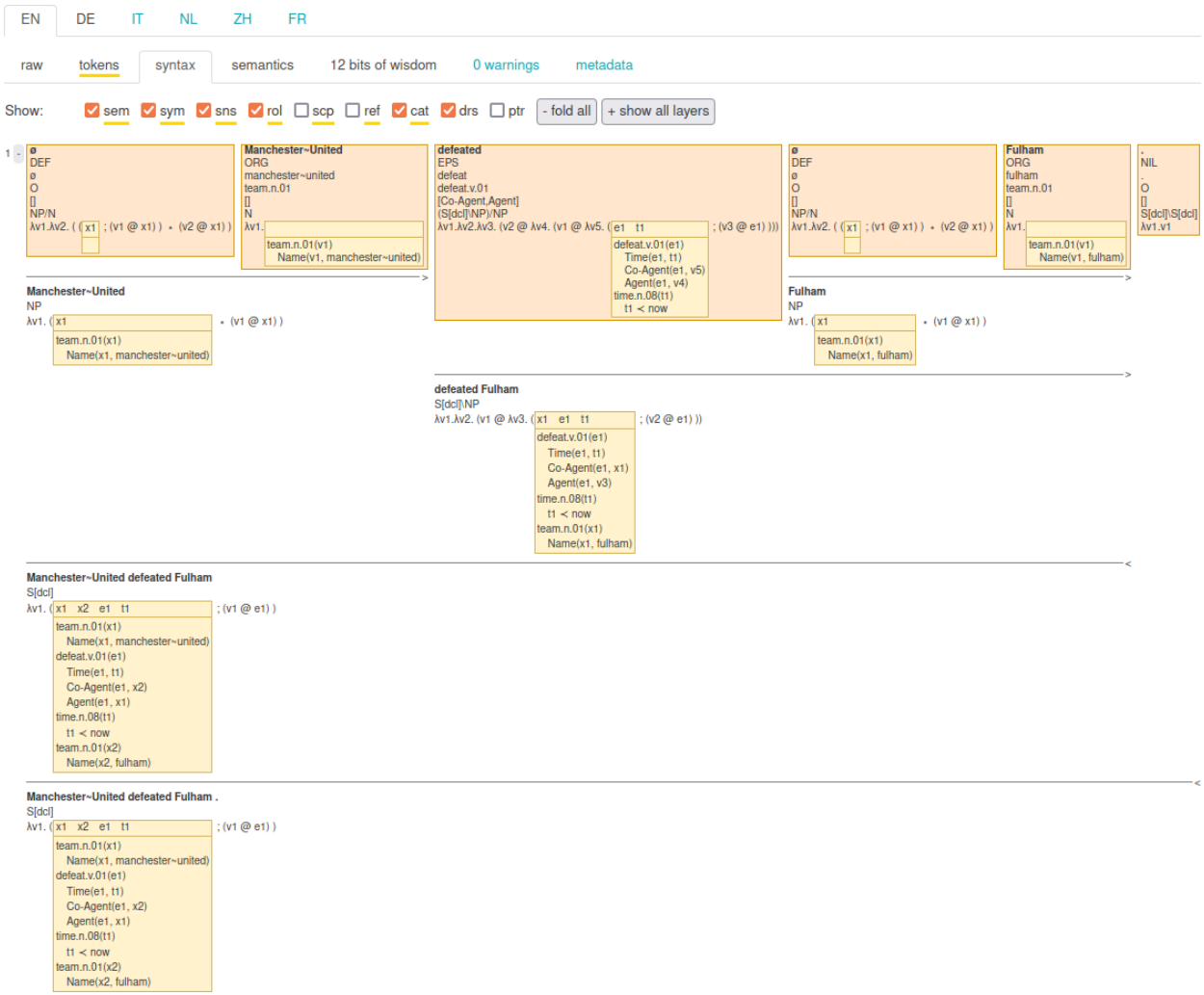
CCG-Derivationen



CCG-Derivationen mit Semantik



Alles


EN DE IT NL ZH FR
raw tokens syntax semantics 12 bits of wisdom 0 warnings metadata
Show: sem sym sns rol scp ref cat drs ptr - fold all + show all layers
Manchester~United defeated Fulham
Manchester~United defeated Fulham .

Discourse Representation Graph

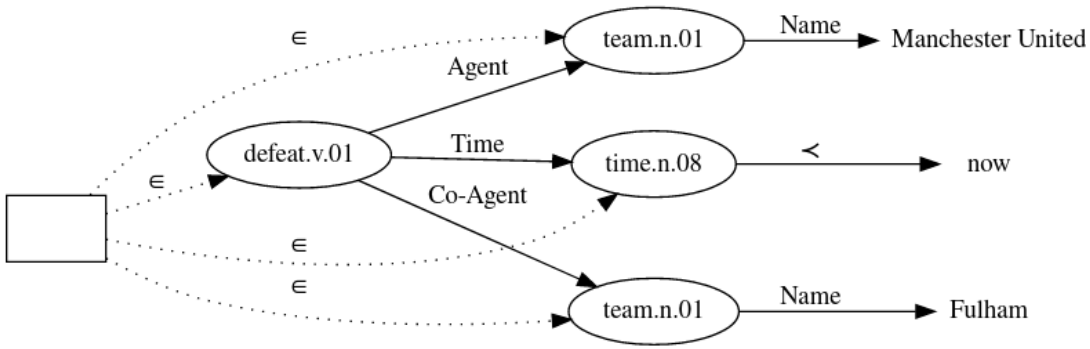
EN DE IT NL ZH FR

raw tokens syntax semantics 12 bits of wisdom 0 warnings metadata

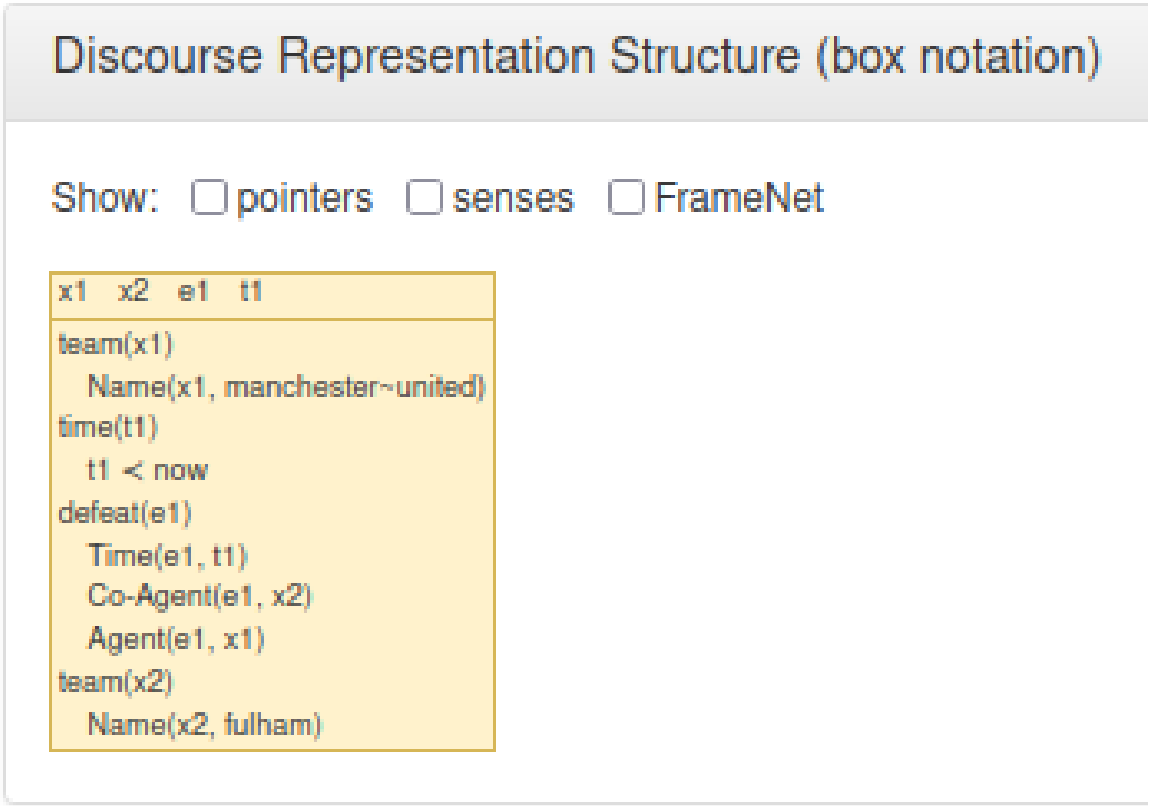
Discourse Representation Graph (sequence notation)

```
team.n.01   Name "Manchester United"     % Manchester United [0-17]
defeat.v.01 Agent -1 Time +1 Co-Agent +2 % defeated          [18-26]
time.n.08   TPR now                      %
team.n.01   Name "Fulham"                % Fulham.           [27-34]
```

Discourse Representation Graph (graph notation)



Discourse Representation Structure (Box Notation)

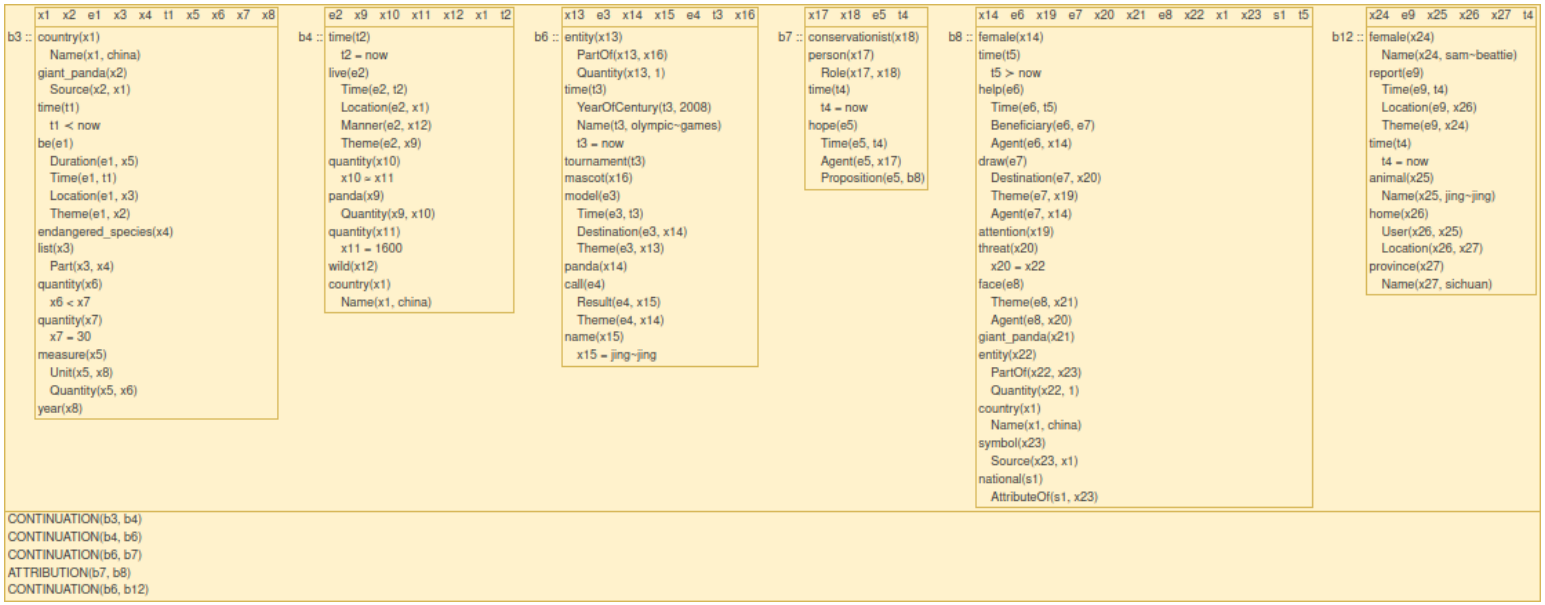


Discourse Representation Structure (Clause Notation)

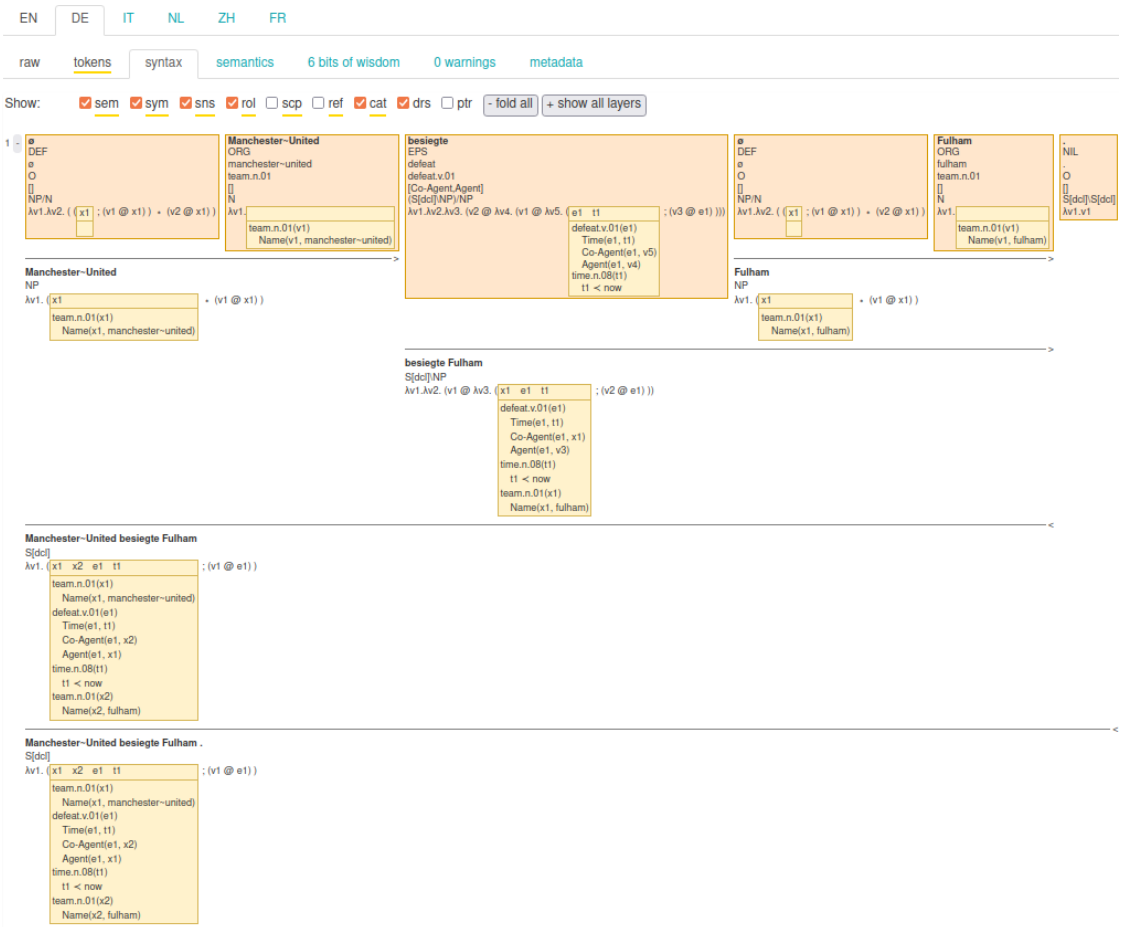
Discourse Representation Structure (clause notation)

```
b1 REF x1                          % Manchester~United [0...17]
b1 Name x1 "manchester~united"     % Manchester~United [0...17]
b1 PRESUPPOSITION b3               % Manchester~United [0...17]
b1 team "n.01" x1                  % Manchester~United [0...17]
b3 REF e1                          % defeated [18...26]
b3 REF t1                          % defeated [18...26]
b3 Agent e1 x1                     % defeated [18...26]
b3 Co-Agent e1 x2                  % defeated [18...26]
b3 TPR t1 "now"                    % defeated [18...26]
b3 Time e1 t1                      % defeated [18...26]
b3 defeat "v.01" e1                % defeated [18...26]
b3 time "n.08" t1                  % defeated [18...26]
b2 REF x2                          % Fulham [27...33]
b2 Name x2 "fulham"                % Fulham [27...33]
b2 PRESUPPOSITION b3               % Fulham [27...33]
b2 team "n.01" x2                  % Fulham [27...33]
                                   % . [33...34]
```

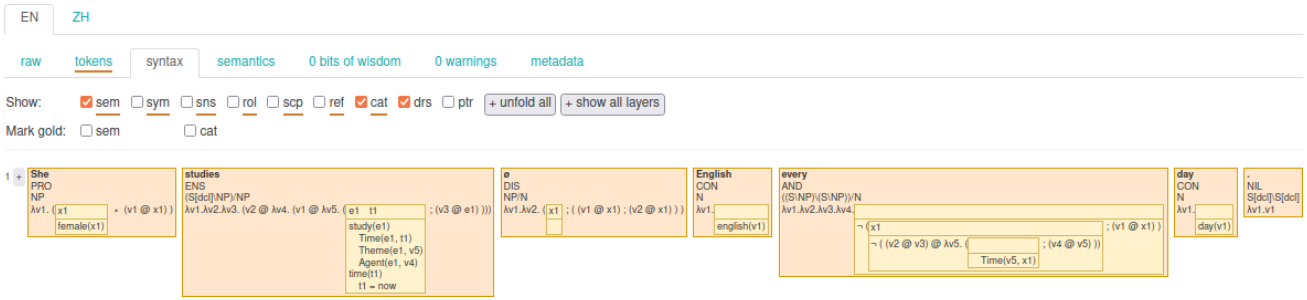
DRS eines Dokuments mit mehreren Sätzen



Parallele Daten

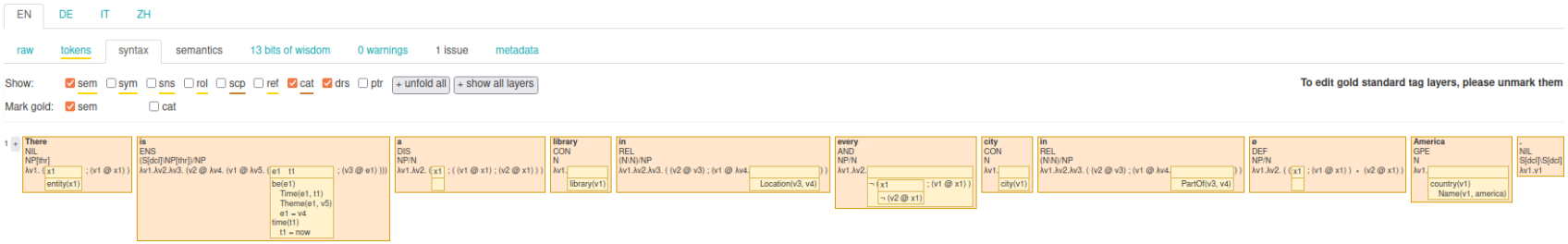


every als temporaler VP-Modifizierer (1/2)



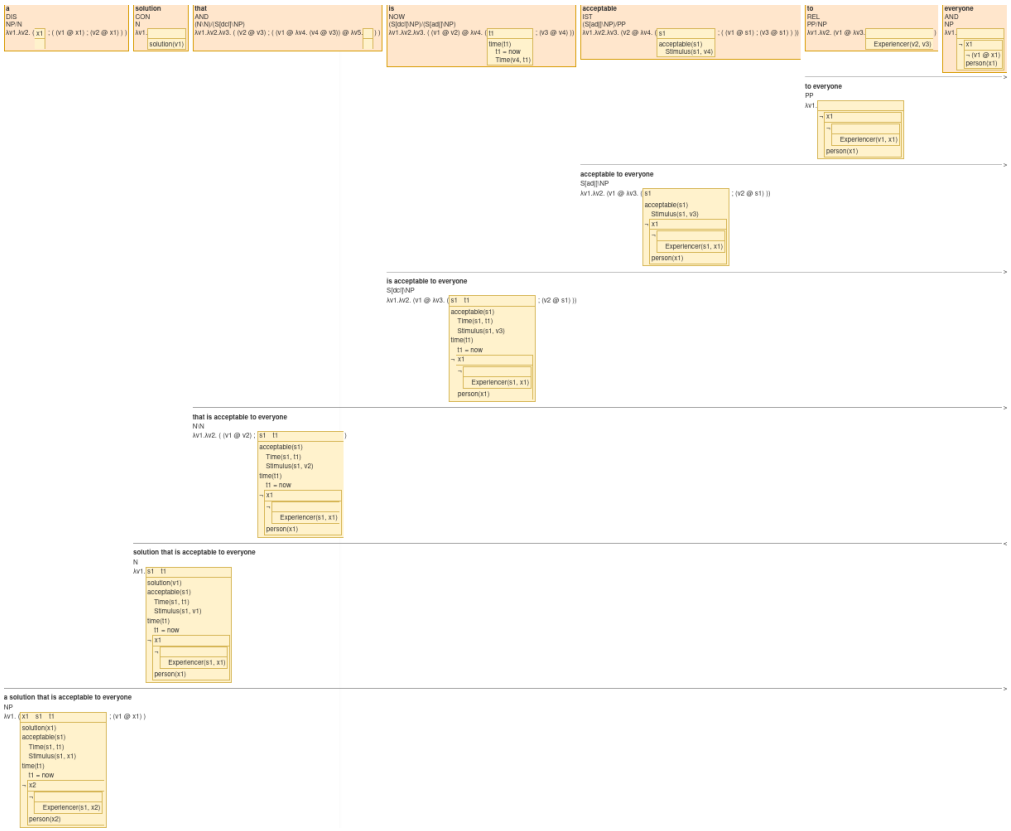
`https://pmb.let.rug.nl/explorer/explore.php?part=00&doc_id=1428&type=der.xml`

every als Determinierer (1/2)



https://pmb.let.rug.nl/explorer/explore.php?part=00&doc_id=1428&type=der.xml

that als Relativpronomen (1/2)



https://pmb.let.rug.nl/explorer/explore.php?part=00&doc_id=1428&type=der.xml

Zu beachten

- Universelle Determinierer werden AND getaggt und existenzielle DIS
- Nicht nur das semantische Tag bestimmt die λ-DRS eines Funktionsworts, sondern auch seine CCG-Kategorie (und Symbole, Synsets, Skopus- und Koreferenztags...)
- Universelle Implikation wird als doppelte Negation ausgedrückt
- Intersektive Relativpronomen werden AND getaggt (Information aus dem Relativsatz wird einfach mit der Information aus dem Bezugsnomen gemerget)

Zusammenfassung

Kategorialgrammatiken sind superspannend!

Literatur I

Abzianidze, L. and Bos, J. (2017). Towards universal semantic tagging. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Bos, J. (2015). Open-domain semantic parsing with boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 301–304, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Lewis, M. and Steedman, M. (2014). A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar. Association for Computational Linguistics.