Creating a semantically annotated corpus based on Discourse Representation Theory

Johan Bos Valerio Basile Kilian Evang Noortje Venhuizen

university of groningen

The Groningen Meaning Bank (GMB)

- Current state in data-driven computational semantics:
  - Several annotated corpora are available that include some semantic annotation (PropBank, Penn Discourse TreeBank, OntoNotes)
  - However, none of these resources contain annotations that are motivated by formal semantic theory
- The objectives of the **Groningen Meaning Bank** are:
  - Producing a corpus of texts annotated with quasi gold-standard Discourse Representation Structures (DRSs)
  - Making this resource available for research in a kerfuffle-free manner (only public-domain texts are included)

Discourse Representation Theory (DRT)

- DRT is a theory of analysing **meaning from text**, in principle language-neutral
- Many **linguistic phenomena** are studied in the framework provided by DRT (anaphora, scope, events, tense)
- DRT has a **model-theoretic backbone**, allowing applications to perform inferences on the basis of first-order logic

Discourse Representation Structures (DRSs)

- DRSs are visualised as a box with two parts:
  - Top part of the box: discourse referents
  - Bottom part of the box: properties of and relations between referents
  - DRSs (boxes) are recursive data structures
- Extensions to standard DRT:
  - neo-Davidsonian events (with VerbNet roles)
  - presuppositions (Van der Sandt, 1992)
  - rhetorical relations (Asher, 1993)

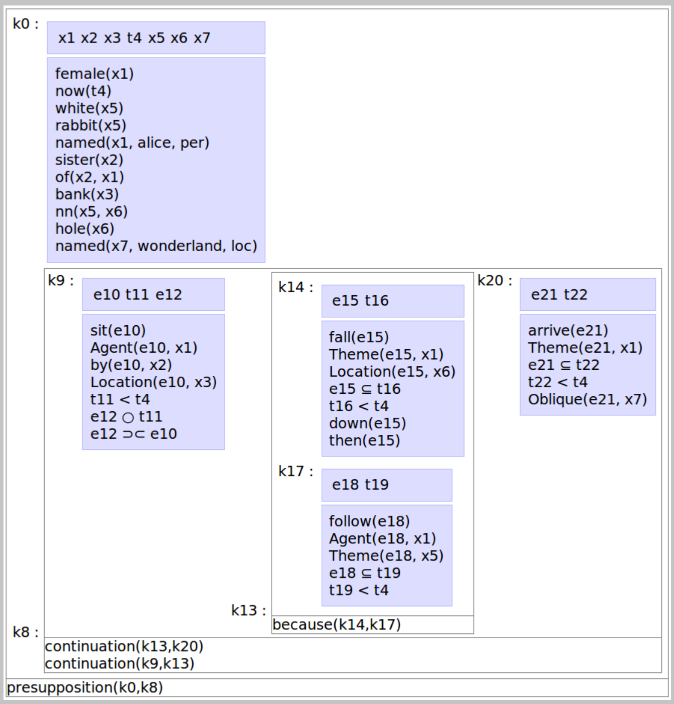
Annotation method

- Manually annotating a reasonably large corpus with gold-standard DRSs is obviously a hard and time-consuming task
- We use a bootstrapping approach that employs state-of-the-art NLP tools to get a reasonable approximation of the target annotations
- Human annotations are coming from two main sources: experts (linguists) and non-experts (players of a *game with a purpose*)
- The annotation of a text comprises several layers:
  - boundaries (for tokens and sentences)
  - tags (part of speech, named entities, word senses)
  - syntactic structure (based on combinatory categorial grammar)
  - semantic structure (including thematic roles and rhetorical relations)

Innovative features and possible impact of the GMB

- Comprises deep, rather than shallow semantics
  - This opens the way to empirical, data-driven approaches to computational semantics
- Integrates phenomena, instead of covering single phenomena in isolation
  - This will provide a better handle on explaining dependencies between various ambiguous linguistic phenomena
- Deals with text, not sentences.
  - This gives us means to deal with ambiguities on the sentence level that require the discourse context for resolving them

Example of a DRS for a small text



k0 :
x1 x2 x3 t4 x5 x6 x7
female(x1)
now(t4)
white(x5)
rabbit(x5)
named(x1, alice, per)
sister(x2)
of(x2, x1)
bank(x3)
nn(x5, x6)
hole(x6)
named(x7, wonderland, loc)

k9 :
e10 t11 e12
sit(e10)
Agent(e10, x1)
by(e10, x2)
Location(e10, x3)
t11 < t4
e12 ○ t11
e12 ⊃⊂ e10

k14 :
e15 t16
fall(e15)
Theme(e15, x1)
Location(e15, x6)
e15 ⊆ t16
t16 < t4
down(e15)
then(e15)

k17 :
e18 t19
follow(e18)
Agent(e18, x1)
Theme(e18, x5)
e18 ⊆ t19
t19 < t4

k13 :
because(k14,k17)

k20 :
e21 t22
arrive(e21)
Theme(e21, x1)
e21 ⊆ t22
t22 < t4
Oblique(e21, x7)

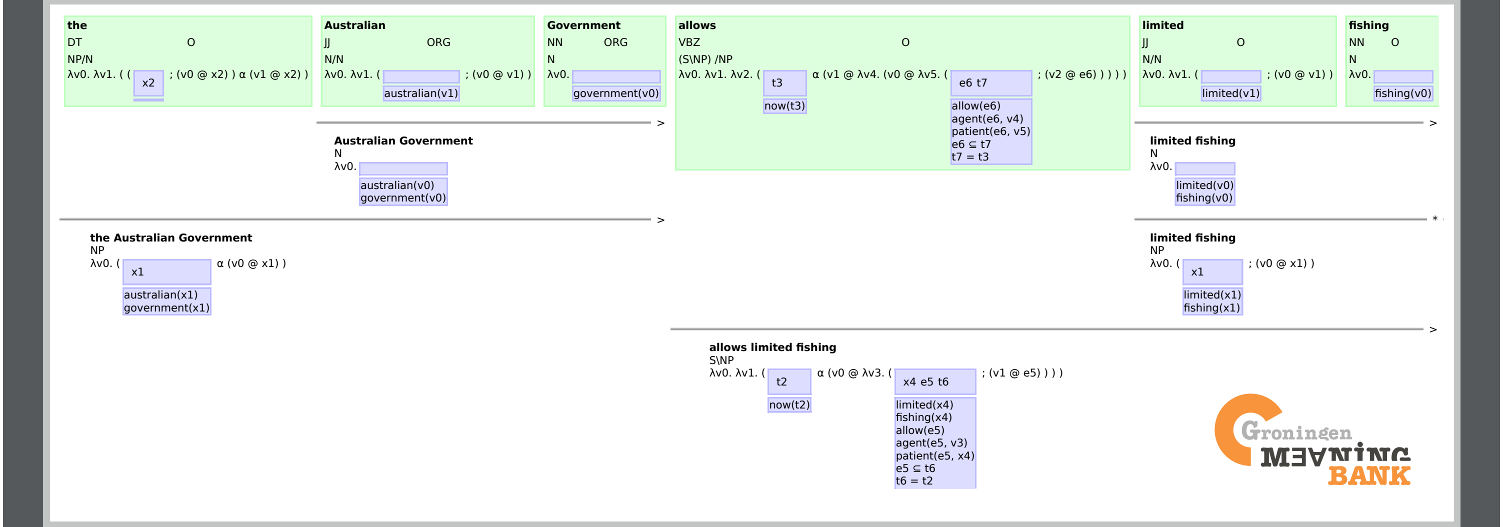
k8 :
continuation(k13,k20)
continuation(k9,k13)

presupposition(k0,k8)

Results

- Explorer: wiki-like interface for expert annotators
- Current corpus size (development version):
  - 70K documents, 1.3M sentences, 31M tokens
  - First stable release: 1,000 documents (GMB 1.0)

The GMB Explorer: visualisation tool for manipulating DRSs



the
DT O
NP/N
λv0. λv1. ((x2 ; (v0 @ x2)) α (v1 @ x2))

Australian
JJ ORG
N/N
λv0. λv1. (australian(v1) ; (v0 @ v1))

Government
NN ORG
N
λv0. government(v0)

allows
VBZ O
(S\NP) /NP
λv0. λv1. λv2. (t3 now(t3) α (v1 @ λv4. (v0 @ λv5. (e6 t7 allow(e6) agent(e6, v4) patient(e6, v5) e6 ⊆ t7 t7 = t3 ; (v2 @ e6)))))

limited
JJ O
N/N
λv0. λv1. (limited(v1) ; (v0 @ v1))

fishing
NN O
N
λv0. fishing(v0)

>

Australian Government
N
λv0. australian(v0) government(v0)

>

limited fishing
N
λv0. limited(v0) fishing(v0)

>

the Australian Government
NP
λv0. (x1 australian(x1) government(x1) α (v0 @ x1))

*

limited fishing
NP
λv0. (x1 limited(x1) fishing(x1) ; (v0 @ x1))

>

allows limited fishing
S\NP
λv0. λv1. (t2 now(t2) α (v0 @ λv3. (x4 e5 t6 limited(x4) fishing(x4) allow(e5) agent(e5, v3) patient(e5, x4) e5 ⊆ t6 t6 = t2 ; (v1 @ e5))))

Groningen MEANING BANK

http://gmb.let.rug.nl

johan.bos@rug.nl