university of groningen

Towards Surface Realisation from Discourse Representation Structures

Valerio Basile

Computational Linguistics In the Netherlands

20 January 2012

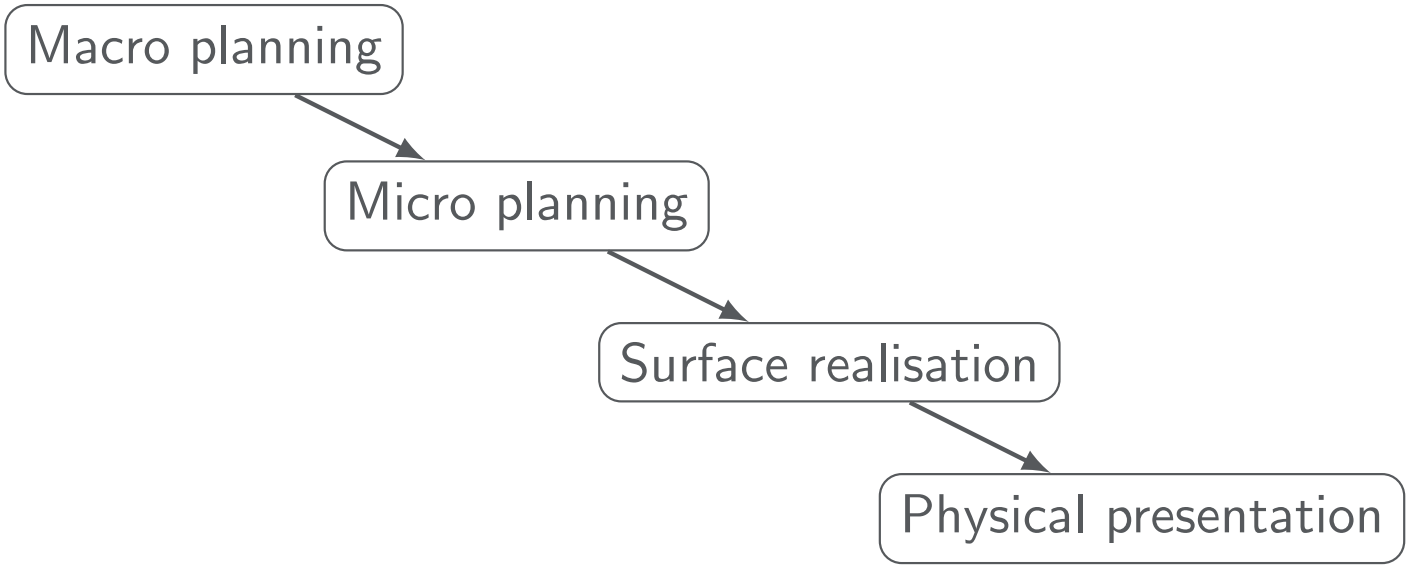
Overview

- Problem definition
  - Open-domain, robust Surface Realisation
- Proposed solution
  - Discourse Representation Theory
  - Stochastical semantic-based Surface Realisation
  - The Groningen Meaning Bank
- A baseline system
- Conclusion

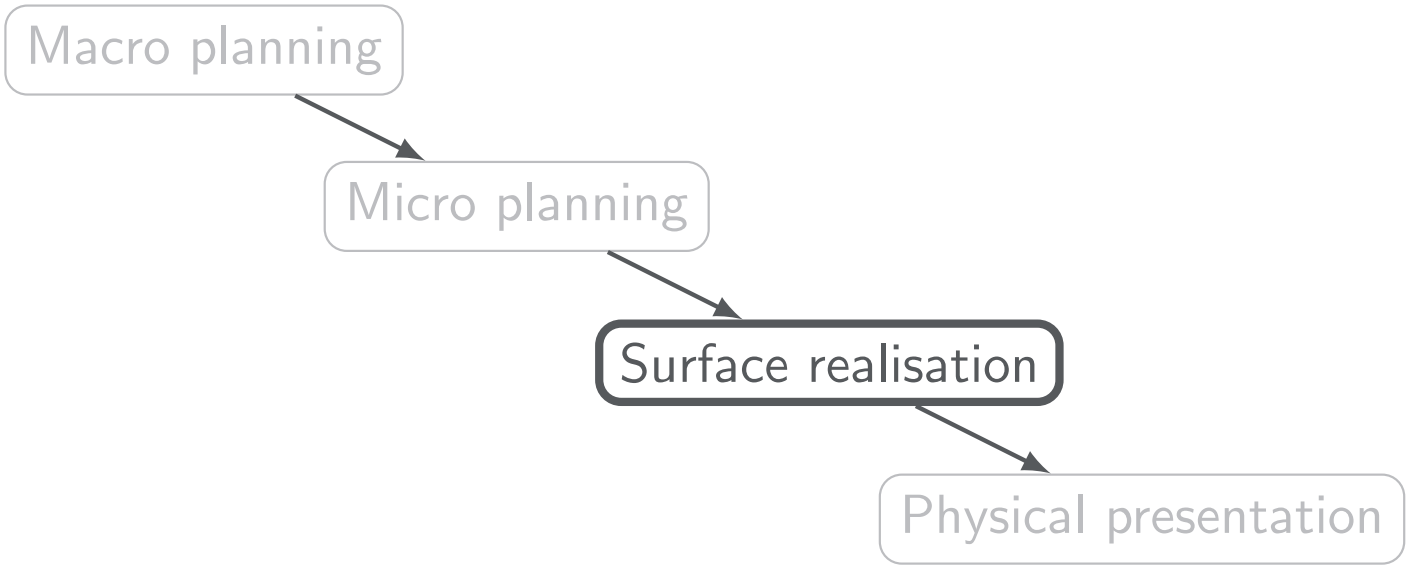
Natural Language Generation

What is Natural Language Generation?

- automatically producing a text
- input = conceptual description
- written, descriptive, in English



Surface Realisation



Mapping the output of the content planning component into a text (words, punctuation symbols, structure, ...).

- Depends on the input representation format
- Syntax-based vs. Semantics-based

Goal

A Surface Realisation system that...

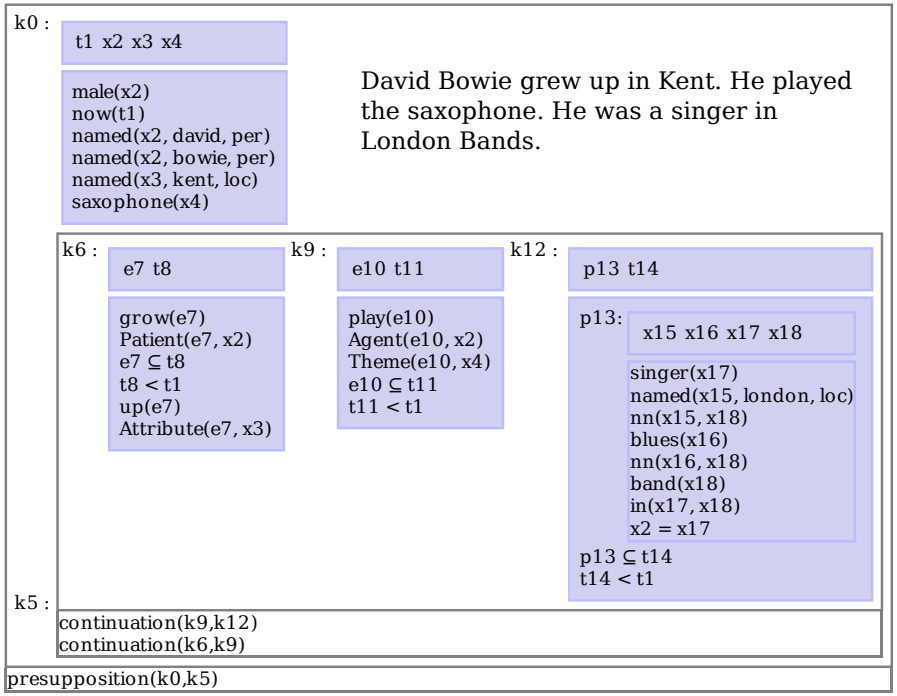
- uses only semantics, rather than syntax
- generates full texts, rather than sentences
- generates open-domain text, rather than domain-specific
- is in principle language-independent

What does the input for such system look like?

Discourse Representation Structures

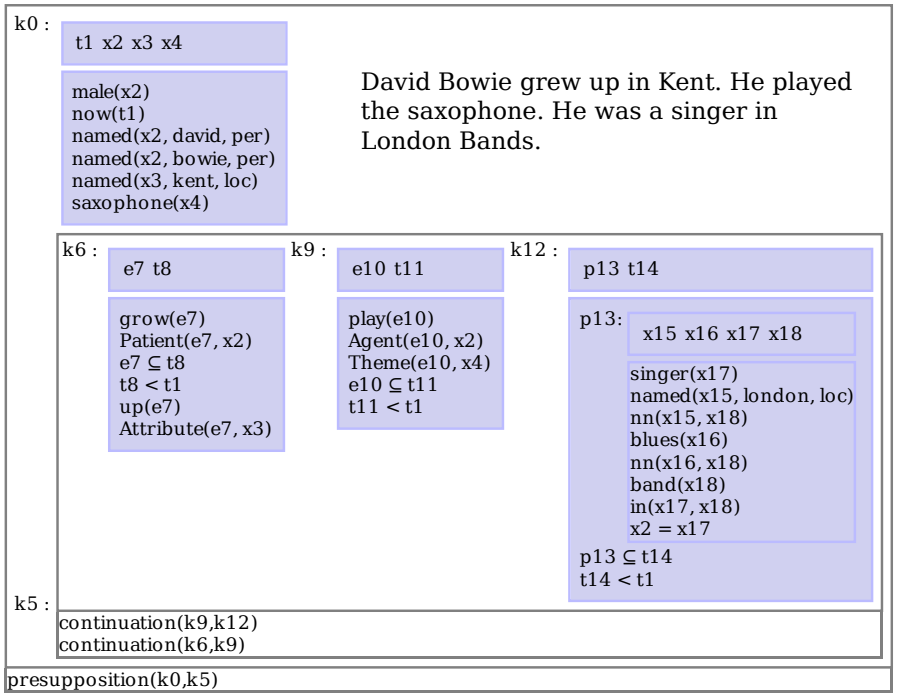
Discourse Representation Structures from Discourse Representation Theory are a suitable input for Surface Realisation.

- Formal theory of meaning
- Language-neutral
- Many linguistic phenomena



Discourse Representation Structures

- Recursive structure
  - Representing subordination, negation, coordination, ...
  - Scope of referring expressions
- Extensions
  - Presuppositions
  - Rhetorical relations



Stochastic Semantics-based SR

Is there enough information in semantic representations to produce grammatically correct surface forms?

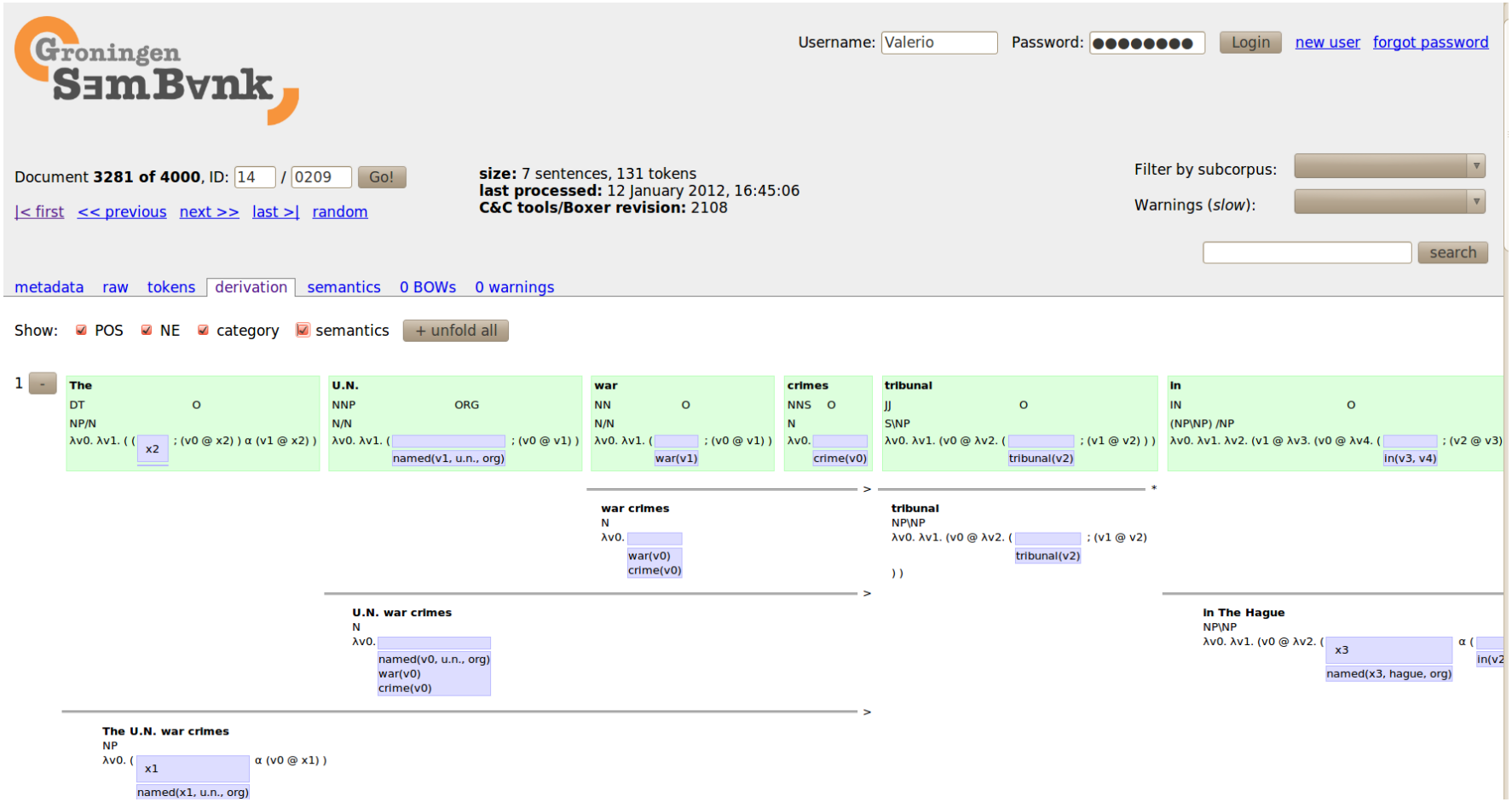
The Groningen Meaning Bank

- Multi-layer annotated linguistic resource.
  - tokenization, CCG derivations, thematic roles, word senses, named entities, semantics (DRSs), rhetorical relations
- Automatically annotated, manual corrections
- 70K documents, 1.3M sentences, 31M tokens, first stable release: 1,000 documents (GMB 1.0)

The Groningen Meaning Bank allows us to automatically evaluate Surface Realisation:

1. Use DRSs from GMB as input
2. Compare the output with the original text, e.g. with BLEU score

The Groningen Meaning Bank

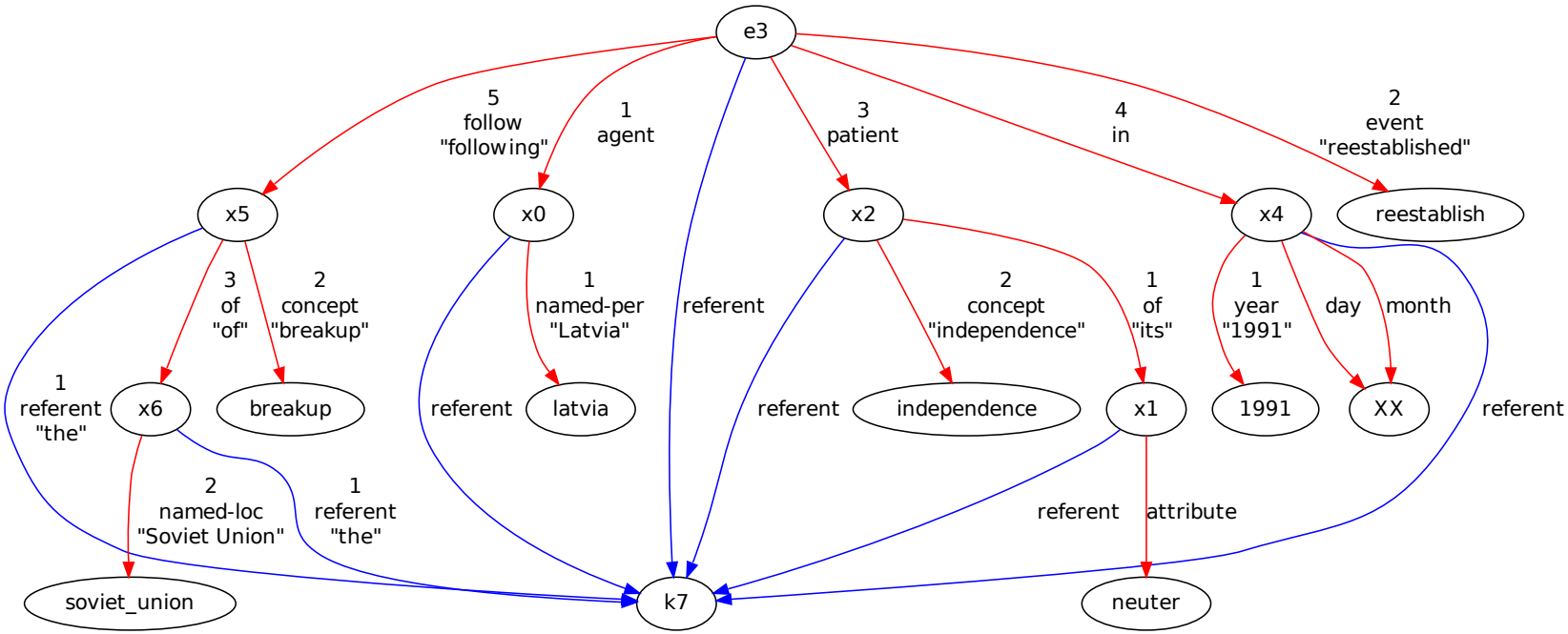


From DRS to graph

- DRSs are recursive
- Machine learning systems don't work well with recursive structures
- We need to transform DRSs into flat structures for ML
- Straightforward translation from DRS to graph with two distincs sets of arcs, for argument structure and discourse structure

From DRS to graph

"Latvia reestablished its independence in 1991 following the breakup of the Soviet Union."



Input representation format - tuples

"Latvia reestablished its independence in 1991 following the breakup of the Soviet Union ."

```
x5   referent-0   k7           s0   1   [ the ]
x5   concept-0    breakup      k7   2   [ breakup ]
x5   of-0         x6           k7   3   [ of ]
e3   referent-0   k7           s0   0   [ ]
e3   agent-0      x0           k7   1   [ ]
e3   event-0      reestablish  k7   2   [ reestablished ]
e3   patient-0    x2           k7   3   [ ]
e3   in-0         x4           k7   4   [ in ]
e3   follow-0     x5           k7   5   [ following ]
x6   referent-0   k7           s0   1   [ the ]
x6   named-loc    soviet_union k7   2   [ Soviet Union ]
x4   day-1        XX           k7   0   [ ]
x4   month-1      XX           k7   0   [ ]
x4   referent-0   k7           s0   0   [ ]
x4   year-1       1991         k7   1   [ 1991 ]
x2   referent-0   k7           s0   0   [ ]
x2   of-0         x1           k7   1   [ its ]
x2   concept-0    independence k7   2   [ independence ]
x1   attribute-0  neuter       k7   0   [ ]
x1   referent-0   k7           s0   0   [ ]
x0   referent-0   k7           s0   0   [ ]
x0   named-per    latvia       k7   1   [ Latvia ]
```

A baseline system

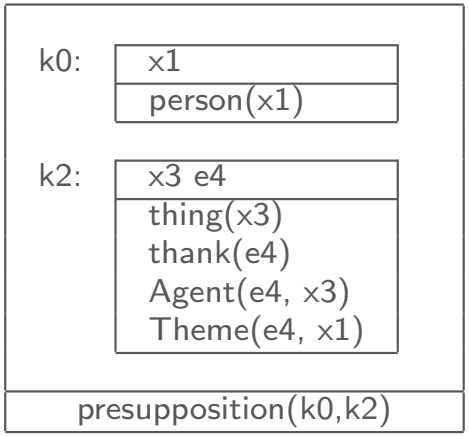
- No discourse structure, only argument structure
- No morphology (use symbols)
- Based on SVM
- Problem: partial order / total order of arcs

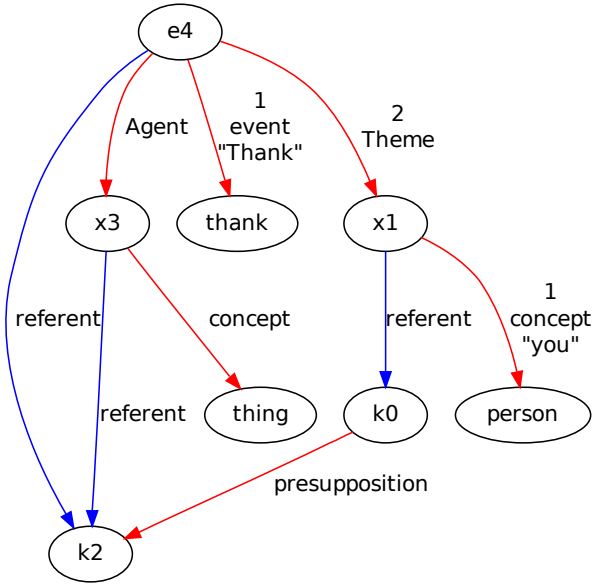
Future works

This is preliminary work, lots of challenges are ahead.

- Which ML technique is the best for this task?
- Generating rhetorical relations and complete texts with negations, coordinations, implications (AKA taking discourse structure into account)
- Morphology and word order

Thanks!


k0:
x1
person(x1)
k2:
x3 e4
thing(x3)
thank(e4)
Agent(e4, x3)
Theme(e4, x1)
presupposition(k0,k2)


e4
Agent
1
event
"Thank"
2
Theme
x3
thank
x1
referent
concept
referent
1
concept
"you"
referent
thing
k0
person
presupposition
k2