Multiword expressions - a pain in the neck for NLP

Sag, Baldwin, Bond, Copestake, Flickinger (2002)

presented by
William Blacoe

Saarland University, Saarbrücken

Outline

Outline

Current Situation

What is needed for Natural Language Understanding?

- world knowledge?
- disambiguation?
- symbolic or statistical information?
- domain knowledge?

MWE complexity and omnipresence is underappreciated in NLP

Motivation

Problems with understanding MWEs

- **overgeneration**
  e.g. "telephone booth" $\rightarrow$ "telephone closet"
- **idiomaticity**
  e.g. how predict that "kick the bucket" is not literal?
- **flexibility**
  a words-with-spaces approach is often too rigid
- **lexical proliferation**
  listing all possible valid cases for lexical, syntactic or semantic selection
- **tractability**

State of the Art

Current formal approaches

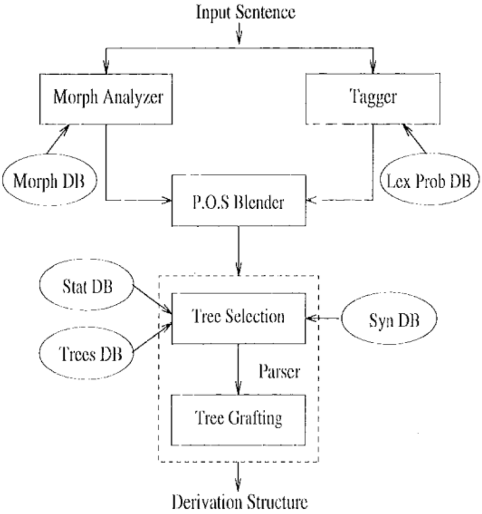
- ParGram
- XTAG
- CCG
- LinGO
- FrameNet

ParallelGrammar (ParGram)

$$
\begin{bmatrix}
\text{PRED} & \text{'see/voir/sehen<}(\uparrow \text{SUBJ}),(\uparrow \text{OBJ}>\text{'} \\
\text{TENSE} & \text{FUT} \\
\text{SUBJ} & \begin{bmatrix}
\text{PRED} & \text{'Maria'} \\
\text{NTYPE} & \begin{bmatrix} \text{PROPER} & \text{NAME} \end{bmatrix} \\
\text{PERS} & 3 \\
\text{GEND} & \text{FEM} \\
\text{NUM} & \text{SG} \\
\text{CASE} & \text{NOM}
\end{bmatrix} \\
\text{OBJ} & \begin{bmatrix}
\text{PRED} & \text{'Hans'} \\
\text{NTYPE} & \begin{bmatrix} \text{PROPER} & \text{NAME} \end{bmatrix} \\
\text{PERS} & 3 \\
\text{GEND} & \text{MASC} \\
\text{NUM} & \text{SG} \\
\text{CASE} & \text{ACC}
\end{bmatrix} \\
\text{PASSIVE} & - \\
\text{STMT-TYPE} & \text{DECLARATIVE} \\
\text{VTYPE} & \text{MAIN}
\end{bmatrix}
$$

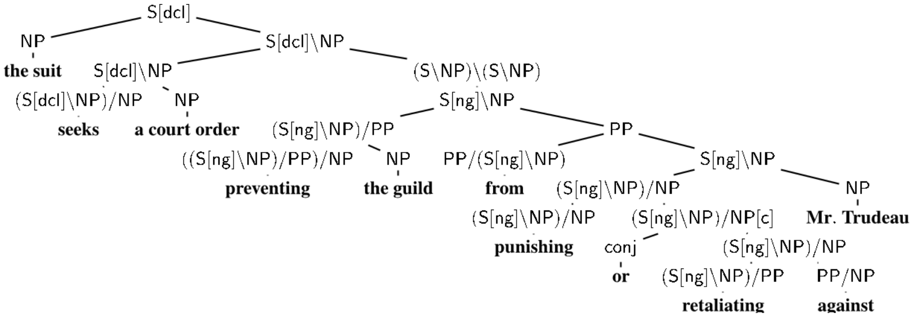
(Butt et al., 1999)

X-Tree Adjoining Grammar (XTAG)



(Doran et al., 1994)

Combinatory Categorial Grammar (CCG)



Categorisation

4 Types of MWEs

- lexicalised phrases
  - fixed expressions
  - semi-fixed expressions
  - syntactically flexible expressions
- institutionalised phrases

Outline

Type 1

fixed expressions

- examples: *ad hoc*, *Palo Alto*
- no morpho-syntactic variation
- not compositional

Type 2

semi-fixed expressions

- non-decomposeable idioms:
  kick the bucket, *shoot the breeze*, but not *spill the beans*
- compound nominals: *attorney general*, *part of speech*
  pluralisation is more complex than just adding an *s* at the end
- proper names: *(the/those) (San Francisco) 49ers*, but not *the Oakland 49ers*
- no syntactic variation, e.g. passivisation
- limited morphological variation, e.g. number
- words-with-spaces is no real solution

Type 3

syntactically flexible expressions

- verb-particle constructions: *write up*, *look up*, *brush up on*
$\rightarrow$semi-compositional, but the semantics of up depends highly on the verb
- semi-compositional: *eat up* $\rightarrow$ *gobble up*
is particle-initial possible with transitive VPCs?
fall off a truck but not *fall a truck off*,
however *call Kim up* and *call up Kim*

Type 3

syntactically flexible expressions

- decomposeable idioms: *spill the beans*, *let the cat out of the bag*
  somewhat compositional
- light verbs: *make a mistake*, *give a demo*
  highly idiosyncratic
  hard to predict which noun can be selected
  syntactically very variable, but not fully compositional
- not representable as words-with-spaces
- only partially compositional due to problems with overgeneration and idiomaticity

Type 4

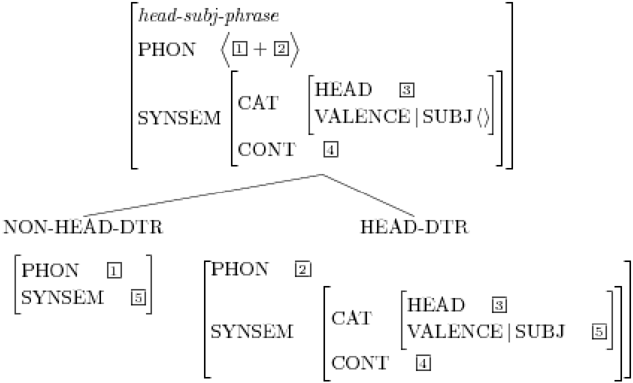
institutionalised phrases

- examples: *traffic light*, *telephone booth*
- semantically and syntactically fully compositional but statistically idionsyncratic
- very high frequency, lexical variants have particularly low frequency

Outline

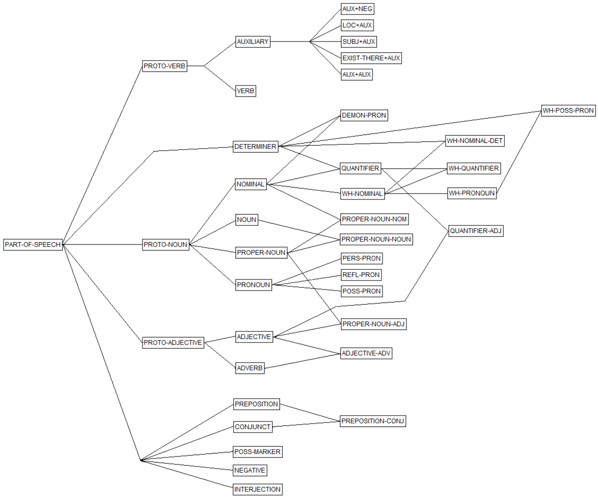
Formal Frameworks

constraint-based HPSG (head-driven phase structure grammar)



Formal Frameworks

constraint-based HPSG (head-driven phase structure grammar)



Formal Frameworks

constraint-based HPSG (head-driven phase structure grammar)
implemented as LKB (lexical knowledge base)

```
scissor := pair-noun-lxm &
[ ORTH.LIST.FIRST "scissor",
  SEM.RELS.LIST.FIRST.PRED "scissor_rel" ].
```

Formal Frameworks

ERG (English resource grammar) is compatible with LKB
and uses MRS (minimal recursion semantics) for semantic entries

every dog probably chased some white cat

TOP **h1**

LZT < [**prpstn_rel** HNL **h1** SOA **h21**], [**_every_rel** HNL **h3** BV **x4** RESTR **h5** BODY **h6**], [**_dog_rel** HNL **h8** INST **x4**], [**_probably_rel** HNL **h9** ARG **h10**], [**_chase_v_rel** HNL **h12** EVENT **e2** [TENSE **past** MOOD **indic**] ARG1 **x4** ARG2 **x13**], [**_some_rel** HNL **h14** BV **x13** RESTR **h15** BODY **h16**], [**_white_rel** HNL **h18** ARG **x13**], [**_cat_rel** HNL **h18** INST **x13**] >

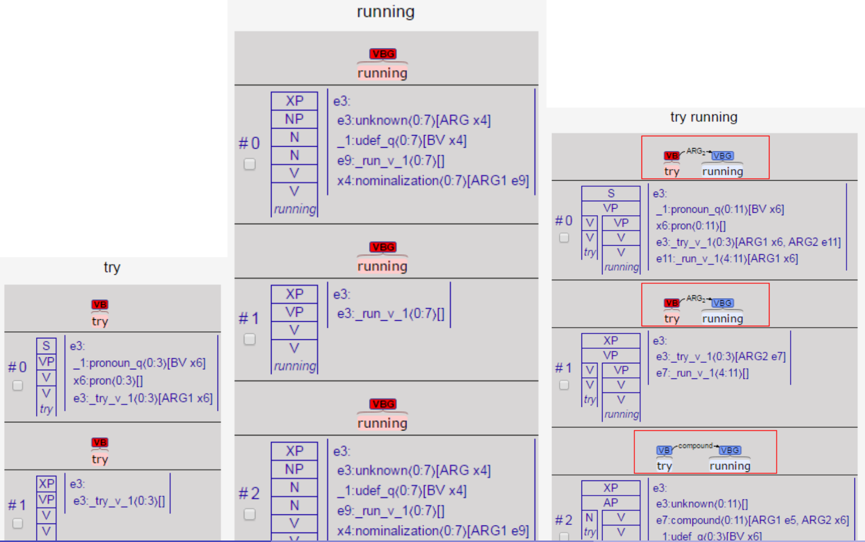
H-CONS < **h5 qeq h8**, **h10 qeq h12**, **h15 qeq h18**, **h21 qeq h9** >

$$\mathrm{prpstn}(\mathrm{probably}(\mathrm{every}(x, \mathrm{dog}(x), \mathrm{some}(y, \mathrm{white}(y) \wedge \mathrm{cat}(y), \mathrm{chase}(x, y)))))$$
$$\mathrm{prpstn}(\mathrm{every}(x, \mathrm{dog}(x), \mathrm{probably}(\mathrm{some}(y, \mathrm{white}(y) \wedge \mathrm{cat}(y), \mathrm{chase}(x, y)))))$$
$$\mathrm{prpstn}(\mathrm{every}(x, \mathrm{dog}(x), \mathrm{some}(y, \mathrm{white}(y) \wedge \mathrm{cat}(y), \mathrm{probably}(\mathrm{chase}(x, y)))))$$
$$\mathrm{prpstn}(\mathrm{probably}(\mathrm{some}(y, \mathrm{white}(y) \wedge \mathrm{cat}(y), \mathrm{every}(x, \mathrm{dog}(x), \mathrm{chase}(x, y)))))$$
$$\mathrm{prpstn}(\mathrm{some}(y, \mathrm{white}(y) \wedge \mathrm{cat}(y), \mathrm{probably}(\mathrm{every}(x, \mathrm{dog}(x), \mathrm{chase}(x, y)))))$$
$$\mathrm{prpstn}(\mathrm{some}(y, \mathrm{white}(y) \wedge \mathrm{cat}(y), \mathrm{every}(x, \mathrm{dog}(x), \mathrm{probably}(\mathrm{chase}(x, y)))))$$

(Copestake and Flickinger, 2000)

Formal Frameworks

ERG + MRS



Encoding MWEs

for fixed epxressions use words-with-spaces

```
ad_hoc_1 := intr_adj_l &
  [ STEM < "ad", "hoc" >,
    SEMANTICS [KEY ad-hoc_rel ]].
```

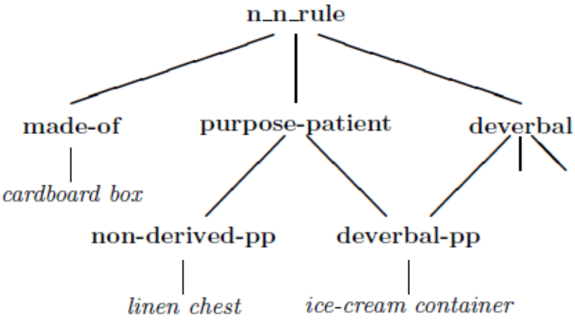
for semi-fixed expressions

- encode internal inflection
- hierarchical lexicon

```
part_of_speech_1 := intr_noun_l &
  [ STEM < "part", "of", "speech" >,
    INFL-POS "1",
    SEMANTICS [KEY part_of_speech_rel ]].
```

Encoding MWEs

for syntactically flexible expressions encode semi-productivity in a type graph (here: noun compounds)



Encoding MWEs

for syntactically flexible expressions

- encode lexical selection in co-occuring word's representation
- use semantic relations for light verbs
- for semantics of decomposeable idioms combine predicates with idiomatic interpretation of involved words

for institutionalised phrases use frequency information

Outline

Conclusions

- useful classification of MWEs
- disambiguation is key (not treated in this paper)
- existing approaches help, but there is much work to be done
- symbolic treatment of grammar is not enough
  statistical information is necessary

References

Miriam Butt, Stefanie Dipper, Anette Frank, and Tracy Holloway King. Writing large-scale parallel grammars for english, french, and german. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the Lexical Functional Grammar Conference*, Manchester, UK, 1999.

Ann Copestake and Dan Flickinger. An open-source grammar development environment and broad-coverage english grammar using hpsg. In *Proceedings of the Second conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000.

Christy Doran, Dania Egedi, Beth Ann Hockey, B Srinivas, and Martin Zaidel. Xtag system - a wide coverage grammar for english. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 922–928, 1994.