# MWE Annotation in Monolingual and Parallel Treebanks in INESS

Gyri Smørdal Losnegaard, Victoria Rosén

University of Bergen
(Norway)

PARSEME WG4
IPIPAN, Warsaw, Poland
September 17, 2013

# INESS

INfrastructure for the Exploration of Syntax and Semantics

INESS is a specialized infrastructure for linguistic research

Two main goals:

1. To host and provide services to treebanks from other projects, providing powerful visualization and search facilities through an ordinary web browser
2. To develop a large LFG parsebank for Norwegian (deep parsing with c-structures and f-structures)

# NorGram

NorGram is a computational LFG grammar for Norwegian

Development began in 1999 in the NorGram project, led by Helge Dyvik

Part of the Parallel Grammar Project (ParGram)

Further developed in: LOGON, TREPIL, XPAR and INESS

Large lexicon based on NorKompLeks

Hand-written lexicon, especially for function words

# Types of MWEs in NorGram

1. Words with spaces
2. Particle verbs
3. Words taking selected prepositions
4. Idioms

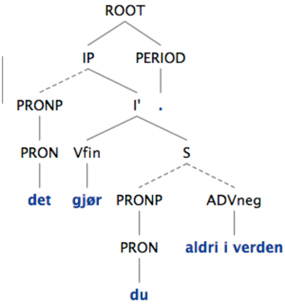
# Words with spaces

(1) a. aldri i verden
never in world.DEF
“never in my life”

b. som regel
as rule
“usually”

c. til syvende og sist
to seventh and last
“ultimately”

## C-structure



## F-structure

```
   PRED        'gjøre<[33:du], [23:den]>'
   TNS-ASP   36| TENSE pres, MOOD indicative |
              | PRED      'den'                                   |
              | NTYPE   27| NSYN pronoun |                         |
   TOPIC      | GEND    26| NEUT +, MASC -, FEM - |                |
              | PRON-FORM den, PERS 3, NUM sg, DEF +,             |
            23| CASE obl, REF +, PRON-TYPE pers                   |
   ADJUNCT   | { | PRED 'aldri-i-verden' | } |
            1| {2| ADV-TYPE neg          | } |
   OBJ        [23]
              | PRED      'du'                                    |
              | NTYPE   35| NSYN pronoun |                        |
   SUBJ       | GEND    34| NEUT - |                              |
              | REF +, PRON-TYPE pers, PRON-FORM du, PERS 2,      |
            33| NUM sg, DEF +, CASE nom                           |
0| VTYPE main, VFORM fin, STMT-TYPE decl
```

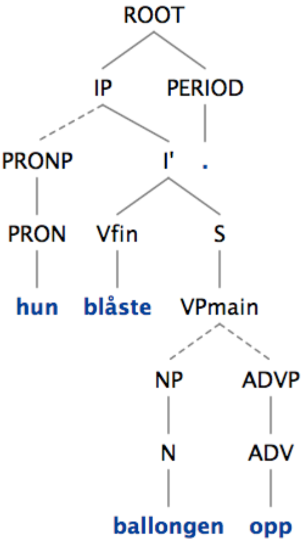
# Particle verbs

(2) a. Hun blåste ballongen opp.
She blew balloon.DEF up.
"She blew the balloon upwards."

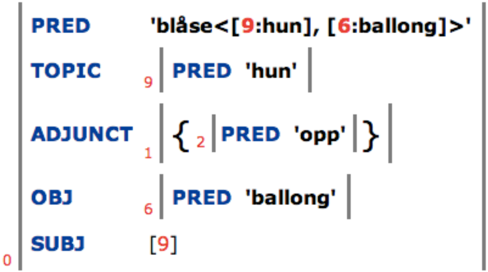
b. Hun blåste ballongen opp.
She blew balloon.DEF up.
"She inflated the balloon."

c. Hun blåste opp ballongen.
She blew up balloon.DEF.
"She inflated the balloon."

## C-structure



## F-structure


PRED 'blåse<[9:hun], [6:ballong]>'

TOPIC 9 PRED 'hun'

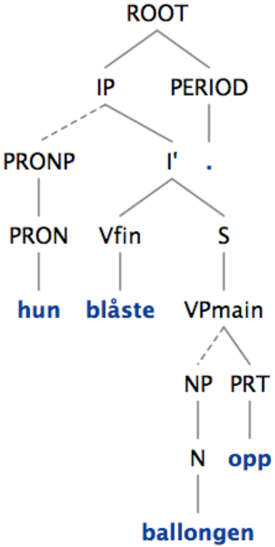
ADJUNCT 1 { 2 PRED 'opp' }

OBJ 6 PRED 'ballong'

SUBJ [9]

0

## C-structure



## F-structure

PRED 'blåse*opp<[9:hun], [6:ballong]>'

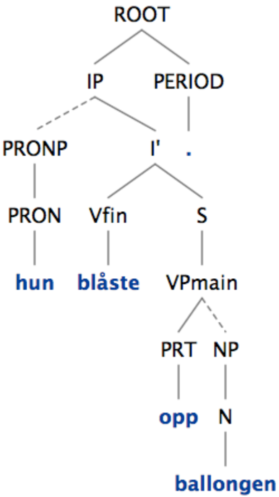
TOPIC $_9$ [ PRED 'hun' ]

OBJ $_6$ [ PRED 'ballong' ]

SUBJ [9]

0

## C-structure



## F-structure

```
  PRED    'blåse*opp<[13:hun], [11:ballong]>'
  TOPIC 13 [ PRED  'hun' ]
  OBJ   11 [ PRED  'ballong' ]
0 SUBJ    [13]
```

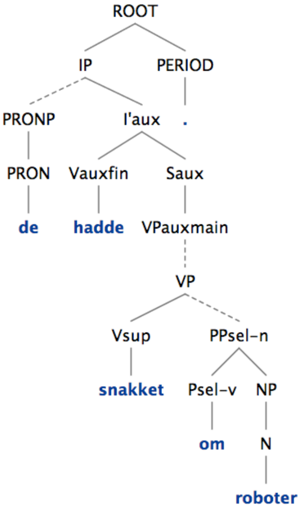
# Words (verbs, adjectives, nouns) taking selected prepositions

(3) a. De hadde **snakket om** roboter.
They had talked of robots.
“They had discussed robots.”

b. Jeg er **redd for** mørket.
I am afraid of dark.DEF.
“I’m afraid of the dark.”

c. Leksikonet gir oss ikke noe **svar på** hvordan vi
Lexicon.DEF gives us not any answer on how we
bør leve heller.
should live either.
“The lexicon doesn’t tell us how to live either.”

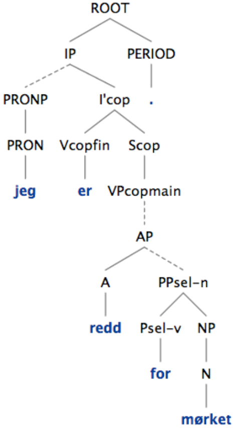
## C-structure



## F-structure

```
  PRED    'ha-perf<[42:snakke*om]>[26:de]'
  TOPIC  26[ PRED 'de' ]
            [ PRED    'snakke*om<[26:de], [46:robot]>' ]
  XCOMP     [ OBL-TH 46[ PRED 'robot' ]               ]
         42 [ SUBJ    [26]                            ]
0 SUBJ    [26]
```

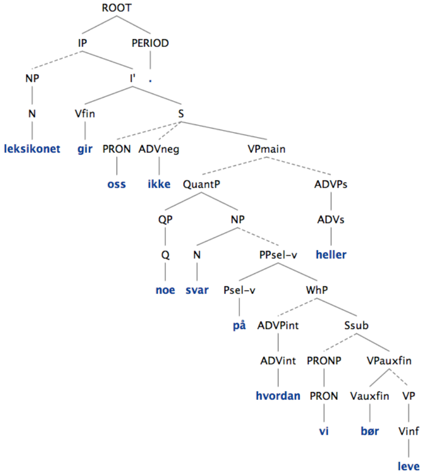
## C-structure



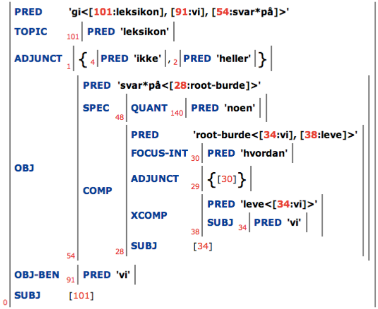
## F-structure

```
  PRED      'være<[45:jeg], [49:redd*for]>'
  TNS-ASP   62[TENSE pres, MOOD indicative]
              [PRED     'jeg'
              [NTYPE    47[NSYN pronoun]
  TOPIC       [GEND     46[NEUT -]
              [REF +, PRON-TYPE pers, PRON-FORM jeg, PERS 1,
            45[NUM sg, DEF +, CASE nom
              [PRED     'redd*for<[26:mørke]>'
              [GEND     [46]
              [               [PRED    'mørke'
              [               [NTYPE      [NSEM 12[COMMON count]]
  PREDLINK    [OBL-TH         [        11[NSYN common          ]
              [               [GEND    10[NEUT +, MASC -, FEM -]
              [               [PTYPE nosem, PFORM for, PERS 3,
              [             26[CASE obl, NUM sg, DEF +
            49[NUM sg, DEF -, ATYPE predicative
  SUBJ      [45]
 0VFORM fin, STMT-TYPE decl, VTYPE main
```

## C-structure



## F-structure



```
  PRED      'gi<[101:leksikon], [91:vi], [54:svar*på]>'
  TOPIC  101[PRED 'leksikon']
  ADJUNCT  1{ 4[PRED 'ikke'], 2[PRED 'heller'] }
  OBJ         PRED  'svar*på<[28:root-burde]>'
              SPEC  48[QUANT 140[PRED 'noen']]
                     PRED       'root-burde<[34:vi], [38:leve]>'
                     FOCUS-INT 30[PRED 'hvordan']
              COMP   ADJUNCT   29{[30]}
                     XCOMP      PRED 'leve<[34:vi]>'
                               38 SUBJ 34[PRED 'vi']
         54         28 SUBJ     [34]
  OBJ-BEN 91[PRED 'vi']
0 SUBJ      [101]
```

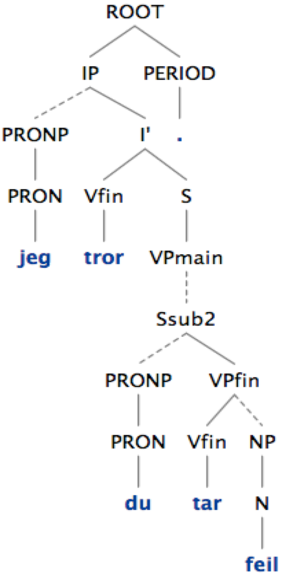
# Idioms

(4) a. Jeg tror du **tar feil**.
I think you take error.
“I think you’re wrong.”

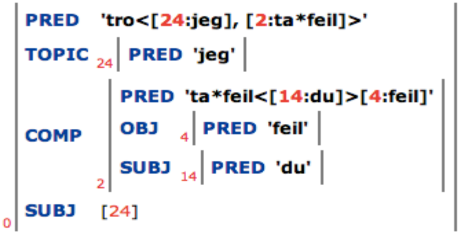
b. Tannlegen **gjorde et nummer av** at hun hadde overbitt.
Denist.DEF made a number of that she had overbite.
“The dentist made a big deal out of her having an overbite.”

c. Alt **kom for en dag**.
All came for a day.
“Everything was revealed.”

## C-structure

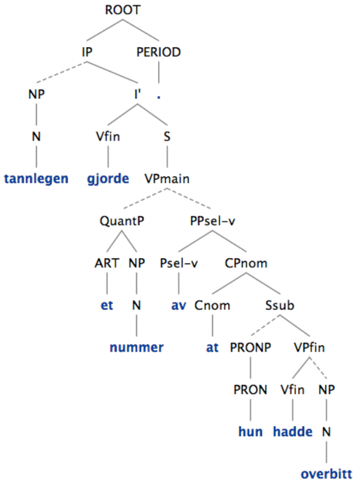


## F-structure

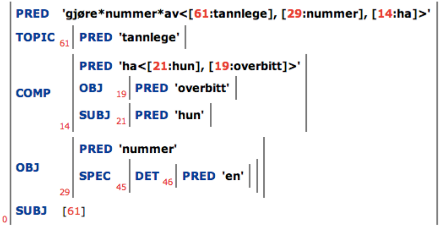


```
PRED     'tro<[24:jeg], [2:ta*feil]>'
TOPIC 24 [ PRED 'jeg' ]
         [ PRED  'ta*feil<[14:du]>[4:feil]' ]
COMP     [ OBJ  4 [ PRED 'feil' ]           ]
         [ SUBJ 14 [ PRED 'du' ]            ]
       2
0 SUBJ   [24]
```

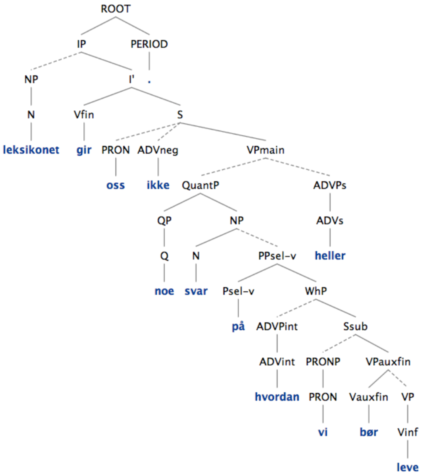
## C-structure



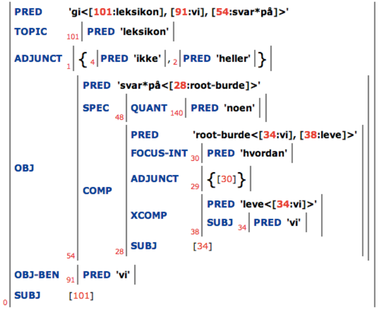
## F-structure


```
PRED    'gjøre*nummer*av<[61:tannlege], [29:nummer], [14:ha]>'
TOPIC 61 [ PRED 'tannlege' ]
         [ PRED 'ha<[21:hun], [19:overbitt]>'   ]
COMP     [ OBJ  19 [ PRED 'overbitt' ]          ]
         [ SUBJ 21 [ PRED 'hun' ]               ]
      14
         [ PRED 'nummer'                        ]
OBJ      [ SPEC 45 [ DET 46 [ PRED 'en' ] ]     ]
      29
SUBJ    [61]
0
```

## C-structure



## F-structure


PRED 'gi<[101:leksikon], [91:vi], [54:svar*på]>'
TOPIC 101 PRED 'leksikon'
ADJUNCT 1 { 4 PRED 'ikke' , 2 PRED 'heller' }
OBJ
PRED 'svar*på<[28:root-burde]>'
SPEC 48 QUANT 140 PRED 'noen'
COMP
PRED 'root-burde<[34:vi], [38:leve]>'
FOCUS-INT 30 PRED 'hvordan'
ADJUNCT 29 {[30]}
XCOMP 38 PRED 'leve<[34:vi]>'
SUBJ 34 PRED 'vi'
28 SUBJ [34]
54
OBJ-BEN 91 PRED 'vi'
0 SUBJ [101]

# Overview: MWEs in the NorGram lexicon

NorKompLeks:

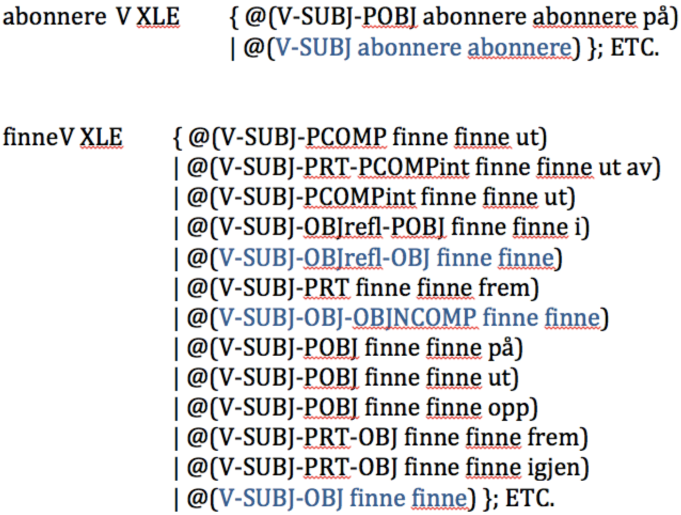
- Verbs with selected prepositions
  - 2523 verbs
  - 6 different argument frames

  - 1773 verbs
  - 9 different argument frames
- Words with spaces
  - 388 expressions
  - 210 ADV, 164 N, 14 A

# Overview: MWEs in the NorKompLeks

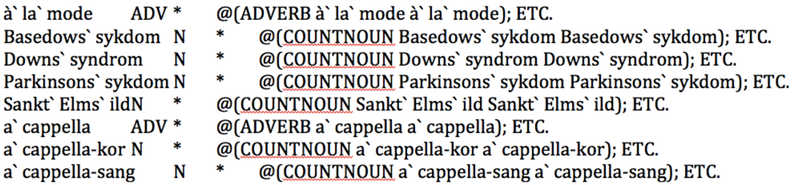
- NKL verb frames with selected preposition
  - V-SUBJ-POBJ: 1474
  - V-SUBJ-OBJ-POBJ: 576
  - V-SUBJ-OBJrefl-POBJ: 366
  - V-SUBJ-OBJrefl-PRT-POBJ: 23
  - V-SUBJ-PRT-POBJ: 81
  - V-SUBJ-PRT-OBJ-POBJ: 3
- NKL verb frames with particle verbs
  - V-SUBJ-PRT-OBJ: 1076
  - V-SUBJ-OBJrefl-PRT-POBJ:23
  - V-SUBJ-PRT-PCOMP:2
  - V-SUBJ-PRT: 418
  - V-SUBJ-OBJrefl-PRT: 166
  - V-SUBJ-PRT-POBJ: 81
  - V-SUBJ-PRT-OBJ-POBJ: 3
  - V-SUBJ-PRT-PCOMPint: 1
  - V-SUBJ-PRT-NCOMPsom: 3

# NorKompLeks verb entries



```
abonnere V XLE     { @(V-SUBJ-POBJ abonnere abonnere på)
                   | @(V-SUBJ abonnere abonnere) }; ETC.


finneV XLE     { @(V-SUBJ-PCOMP finne finne ut)
               | @(V-SUBJ-PRT-PCOMPint finne finne ut av)
               | @(V-SUBJ-PCOMPint finne finne ut)
               | @(V-SUBJ-OBJrefl-POBJ finne finne i)
               | @(V-SUBJ-OBJrefl-OBJ finne finne)
               | @(V-SUBJ-PRT finne finne frem)
               | @(V-SUBJ-OBJ-OBJNCOMP finne finne)
               | @(V-SUBJ-POBJ finne finne på)
               | @(V-SUBJ-POBJ finne finne ut)
               | @(V-SUBJ-POBJ finne finne opp)
               | @(V-SUBJ-PRT-OBJ finne finne frem)
               | @(V-SUBJ-PRT-OBJ finne finne igjen)
               | @(V-SUBJ-OBJ finne finne) }; ETC.
```

# NorKompLeks entries for words with spaces



```
à` la` mode       ADV *      @(ADVERB à` la` mode à` la` mode); ETC.
Basedows` sykdom   N     *       @(COUNTNOUN Basedows` sykdom Basedows` sykdom); ETC.
Downs` syndrom     N     *       @(COUNTNOUN Downs` syndrom Downs` syndrom); ETC.
Parkinsons` sykdom N     *       @(COUNTNOUN Parkinsons` sykdom Parkinsons` sykdom); ETC.
Sankt` Elms` ildN    *       @(COUNTNOUN Sankt` Elms` ild Sankt` Elms` ild); ETC.
a` cappella        ADV *     @(ADVERB a` cappella a` cappella); ETC.
a` cappella-kor N    *       @(COUNTNOUN a` cappella-kor a` cappella-kor); ETC.
a` cappella-sang     N     *       @(COUNTNOUN a` cappella-sang a` cappella-sang); ETC.
```

# Overview: MWEs in the NorGram lexicon

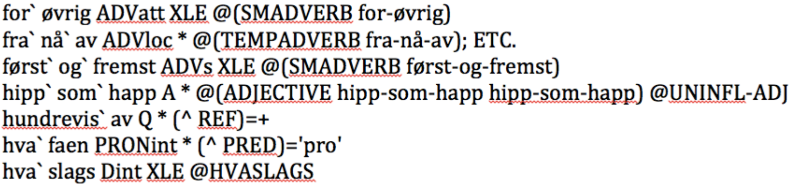
MWEs added to the hand-coded lexicon during grammar development:

- Words with spaces
  - 309 expressions, 30 lexical categories
- Particle verbs
  - 703 verbs
  - 18 different argument frames
- Verbs with selected prepositions
  - 513 verbs
  - 15 different argument frames
- Idioms
  - 38 idioms

# Overview: Words with spaces in the NorGram lexicon

| MWEs in NorGram: lexical categories | | | | |
|---|---|---|---|---|
| ADVloc: 63 | PRON: 16 | Qgen: 4 | Cpur: 1 | CPidiom: 1 |
| P: 43 | Cadv: 16 | CONJ: 4 | CPadv-ell: 1 | Dint: 1 |
| ADVs: 42 | ADVdeg: 13 | POSSint: 4 | TAG: 1 | Psel-v: 1 |
| N: 38 | A: 13 | ADVcmt: 4 | ADVint: 1 | ADVpar: 1 |
| ADV: 26 | PRT: 12 | ADVatt: 3 | PRTidiom: 1 | DA: 1 |
| Q: 25 | ADVneg: 5 | Pvbobj: 3 | ADVwhmod: 1 | Ppost: 1 |

# Hand-coded entries for words with spaces



```
for` øvrig ADVatt XLE @(SMADVERB for-øvrig)
fra` nå` av ADVloc * @(TEMPADVERB fra-nå-av); ETC.
først` og` fremst ADVs XLE @(SMADVERB først-og-fremst)
hipp` som` happ A * @(ADJECTIVE hipp-som-happ hipp-som-happ) @UNINFL-ADJ
hundrevis` av Q * (^ REF)=+
hva` faen PRONint * (^ PRED)='pro'
hva` slags Dint XLE @HVASLAGS
```

# MWEs in the Sofie Parallel Treebank

- Small parallel treebank, 8 languages
- Danish, English, Estonian, Georgian, German, Icelandic, Norwegian, Swedish
- Sentence aligned

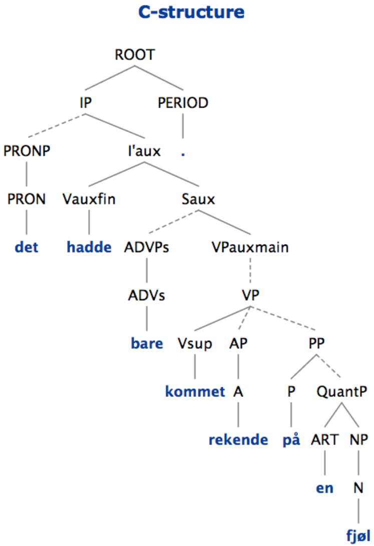
# Future work: challenges

- Collocations, decomposable idioms
  - MWE-analysis or compositional?
- Semi-fixed expressions
  - Groups or “families” of expressions
  - Non-sequential expressions
  - Patterns with semi-open slots

# Decomposable idioms

(5) a. Det hadde bare kommet rekende på en fjøl.
It had just come drifting on a board.
“It had just been dumped on her.”

# Decomposable idioms

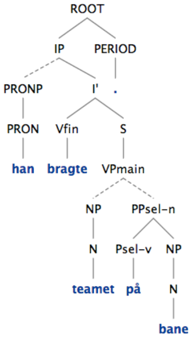


# Decomposable idioms

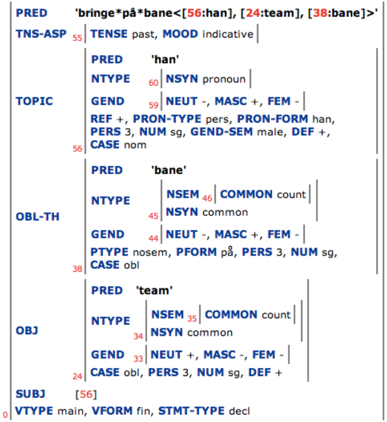
(6) a. Han bragte teamet på bane.
He brought the subject on court.
"He brought up the subject."

# Decomposable idioms

## C-structure

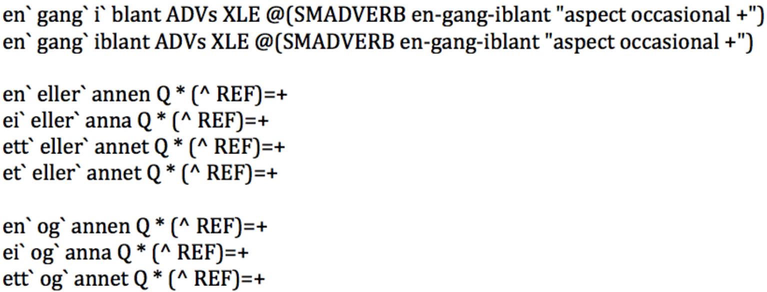


## F-structure



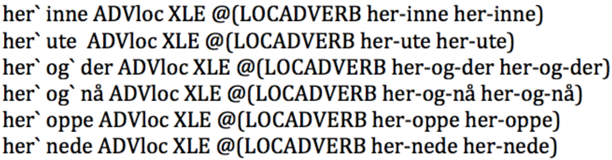
```
   PRED     'bringe*på*bane<[56:han], [24:team], [38:bane]>'
   TNS-ASP  55 TENSE past, MOOD indicative
                PRED    'han'
                NTYPE   60 NSYN pronoun
   TOPIC        GEND    59 NEUT -, MASC +, FEM -
                REF +, PRON-TYPE pers, PRON-FORM han,
                PERS 3, NUM sg, GEND-SEM male, DEF +,
             56 CASE nom
                PRED    'bane'
                              NSEM 46 COMMON count
                NTYPE      45 NSYN common
   OBL-TH       GEND    44 NEUT -, MASC +, FEM -
                PTYPE nosem, PFORM på, PERS 3, NUM sg,
             38 CASE obl
                PRED  'team'
                           NSEM 35 COMMON count
                NTYPE   34 NSYN common
   OBJ          GEND  33 NEUT +, MASC -, FEM -
             24 CASE obl, PERS 3, NUM sg, DEF +
   SUBJ      [56]
 0 VTYPE main, VFORM fin, STMT-TYPE decl
```

# Groups



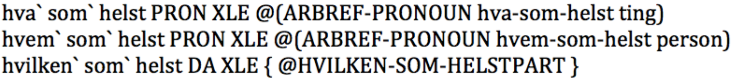
```
en` gang` i` blant ADVs XLE @(SMADVERB en-gang-iblant "aspect occasional +")
en` gang` iblant ADVs XLE @(SMADVERB en-gang-iblant "aspect occasional +")

en` eller` annen Q * (^ REF)=+
ei` eller` anna Q * (^ REF)=+
ett` eller` annet Q * (^ REF)=+
et` eller` annet Q * (^ REF)=+

en` og` annen Q * (^ REF)=+
ei` og` anna Q * (^ REF)=+
ett` og` annet Q * (^ REF)=+
```

# Groups



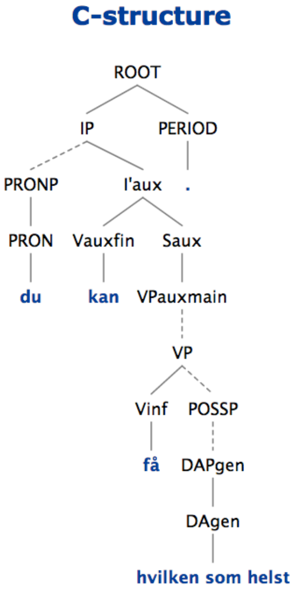
```
her` inne ADVloc XLE @(LOCADVERB her-inne her-inne)
her` ute  ADVloc XLE @(LOCADVERB her-ute her-ute)
her` og` der ADVloc XLE @(LOCADVERB her-og-der her-og-der)
her` og` nå ADVloc XLE @(LOCADVERB her-og-nå her-og-nå)
her` oppe ADVloc XLE @(LOCADVERB her-oppe her-oppe)
her` nede ADVloc XLE @(LOCADVERB her-nede her-nede)
```

# Non-sequential (discontinuous) expressions



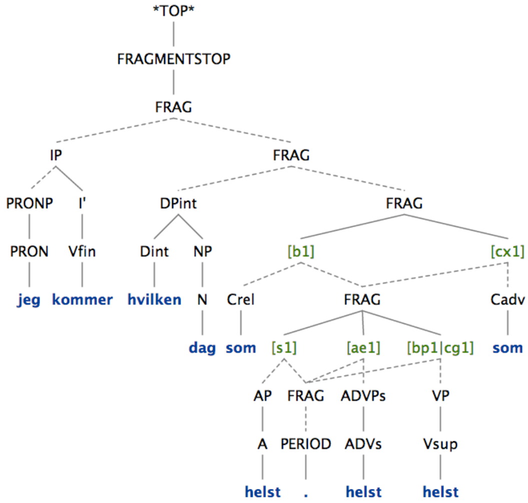
```
hva` som` helst PRON XLE @(ARBREF-PRONOUN hva-som-helst ting)
hvem` som` helst PRON XLE @(ARBREF-PRONOUN hvem-som-helst person)
hvilken` som` helst DA XLE { @HVILKEN-SOM-HELSTPART }
```

# Non-sequential expressions



# Non-sequential expressions

## C-structure



# Conclusion

Code MWEs before or after parsing?

- In a parsebank, before parsing
- Missing MWEs are discovered during annotation and are iteratively added to the lexicon

Code MWEs as one or several units?

- In the c-structure, words-with-spaces or several units (which may be discontinuous), depending on the MWE type
- In the f-structure, as one unit (single predicate)

This dual representation in LFG reflects the nature of MWEs very well.

# Thank you!

`http://clarino.uib.no/iness`