LFG-DOP

Data-Oriented Parsing incorporating Lexical Functional Grammar

Ríona Finn
Supervisors: Dr Andy Way, Dr Mary Hearne

April 12th 2006

embarkinitiative
Investing in People and Ideas

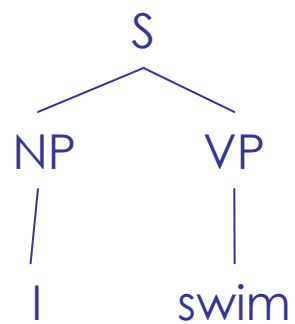
What is Parsing?

- To break (a sentence) down into its component parts of speech with an explanation of the form, function, and syntactical relationship of each part.*

*[www.dictionary.com]

- Input:
  "I swim"

- Output:
  (S (NP **I**) (VP **swim**))



Overview

- The DOP Model
- LFG Theory
- LFG-DOP
- Current experiments
- Future work

The DOP Model

- Proposed by Scha (1990)
- Formalised by Bod (1992)
- Defined by 4 elements:
  - Phrase-structure tree representations
  - Fragments
  - Composition operator
  - Probability model

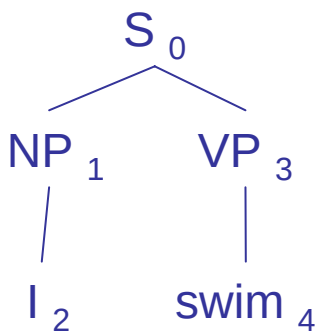
What makes DOP different?

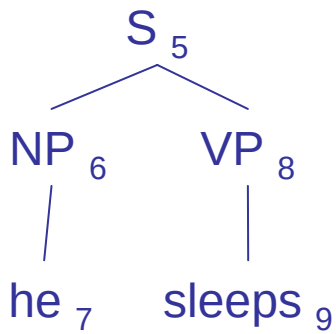
- extract the grammar directly from the corpus
- use arbitrarily large fragments
- use arbitrarily complex fragments
- exploit derivational redundency

[Bod, Scha & Sima'an, 2001]

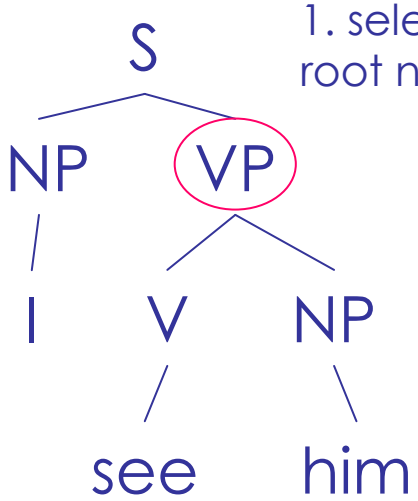
Corpus & Phrase Structure-Trees

- corpus:
  - I swim
  - he sleeps



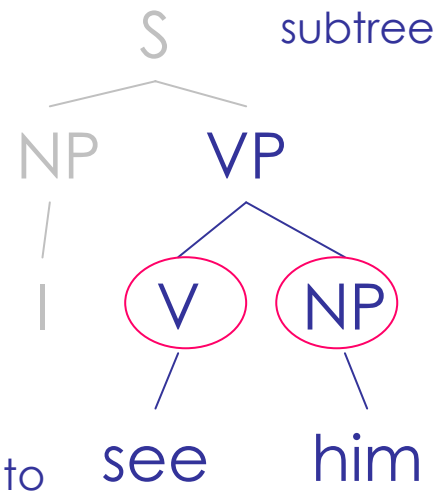


Root & Frontier

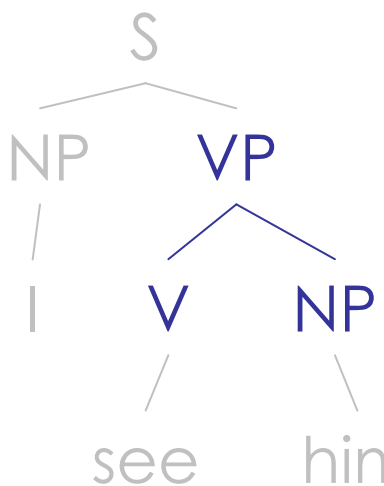


1. select VP as root node

2. delete all except the subtree VP dominates



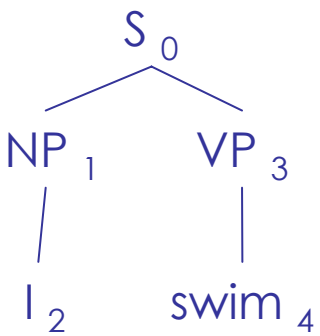
3. select V and NP to be the set of frontier nodes



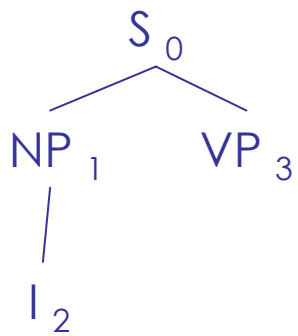
4. delete the subtrees V and NP dominate

Fragmentation

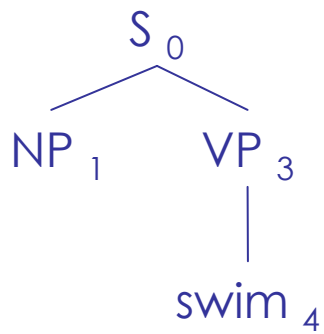
f1:



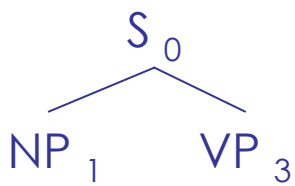
f2:



f3:



f4:

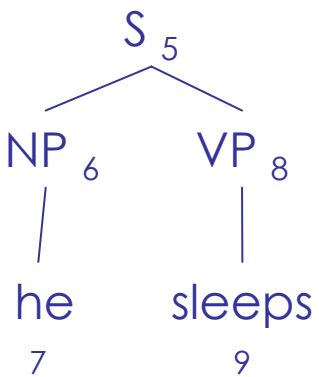


f5:

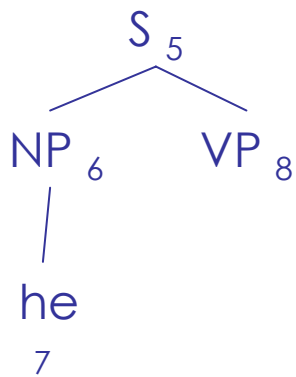
f6:

Fragmentation

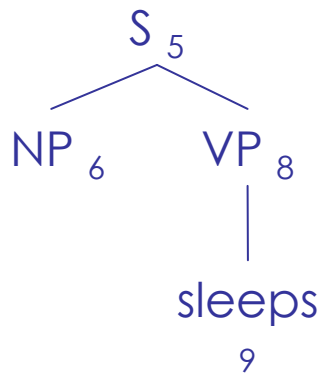
f7:



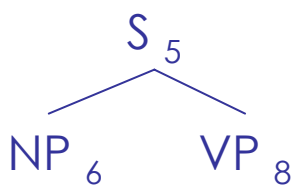
f8:



f9:



f10:

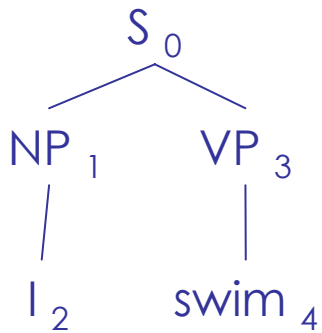


f11:

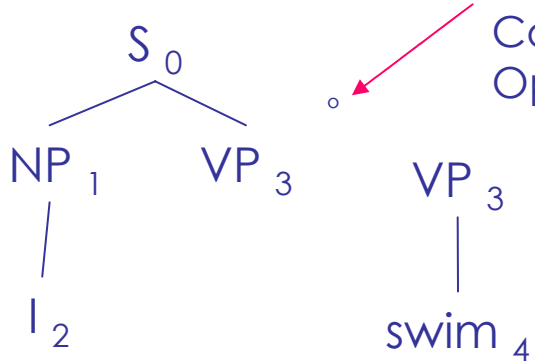
f12:

Composition

Derivation 1 = f1

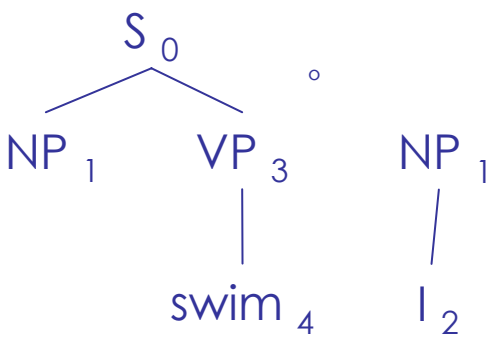


Derivation 2 = f2 ◦ f6

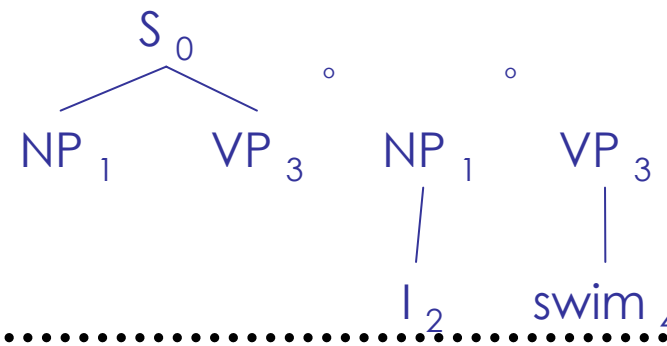


Left-Most Composition Operator

Derivation 3 = f3 ◦ f5



Derivation 4 = f4 ◦ f5 ◦ f6

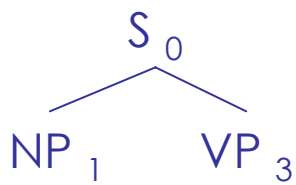


Fragment Probabilities

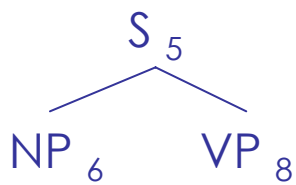
- P(f1): 1/12
- P(f2): 1/12
- P(f3): 1/12
- **P(f4): 2/12**
- P(f5): 1/12
- P(f6): 1/12
- P(f7): 1/12
- P(f8): 1/12
- P(f9): 1/12
- P(f10): 1/12
- P(f11): 1/12
- P(f12): 1/12

- P(fragment) = relative frequency in the fragment set

f4:



f10:



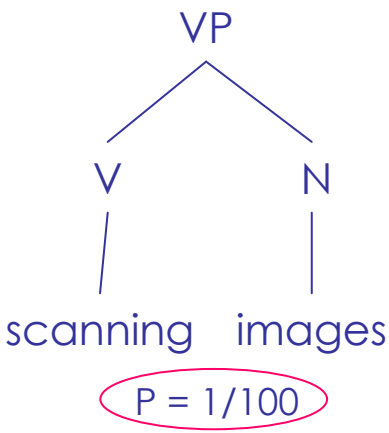
- But f(4) and f(10) are identical:

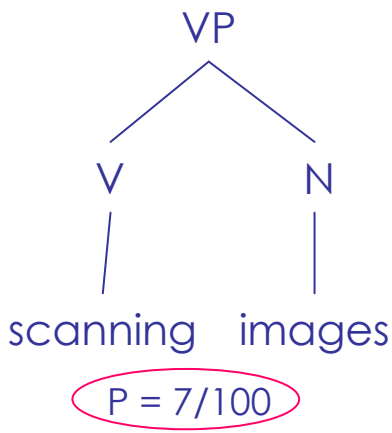
Probability Model

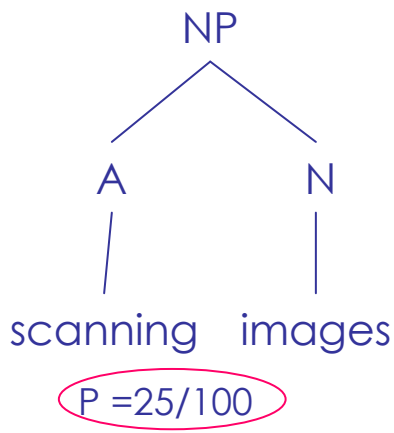
- P(derivation) = product of probabilities of fragments used to build that derivation
- P(parse) = sum of probabilities of derivations that yield that parse

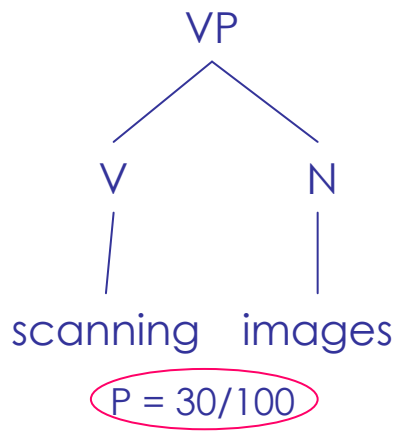
- MPD (Most Probable Derivation) does not distinguish which parse tree is the best parse
- MPP (Most Probable Parse) selects one most probable analysis

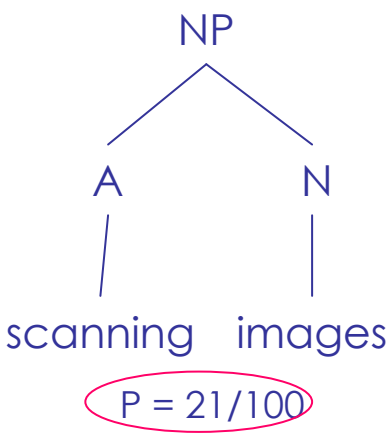
Derivation or Parse?

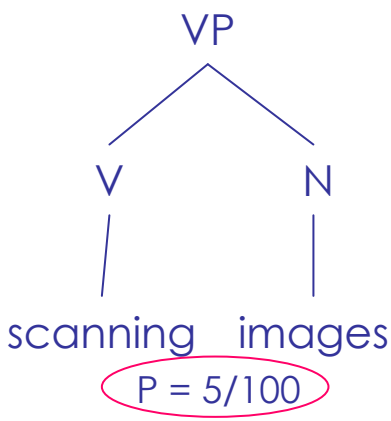












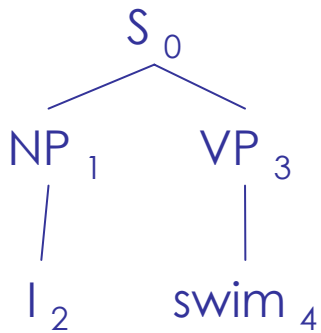
- 4 derivations (VP (V scanning) (N images))
  total probability 43/100
- 2 derivations (NP (A scanning) (N images))
  total probability 46/100
- MPD = (VP (V scanning) (N images))
- MPP = (NP (A scanning) (N images))

Derivation Probabilities

- P(Derivation 1)
  - P(f1) = 1/12 = *72/184*
- P(Derivation 2)
  - P(f2) * P(f6) = 1/12 * 1/12 = 1/144 = *6/184*
- P(Derivation 3)
  - P(f3) * P(f5) = 1/12 * 1/12 = 1/144 = *6/184*
- P(Derivation 4)
  - P(f4) * P(f5) * P(f6) = 2/12 * 1/12 * 1/12 = *1/184*

Most Probable Derivation

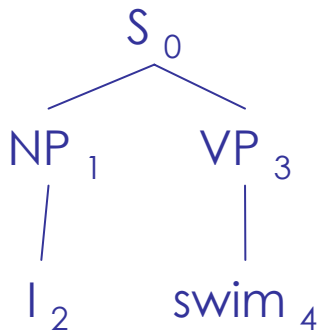
- Most probable derivation: f1
- P(f1) = (72/184)
- DOP Hypothesis:
  - parse accuracy increases with increasing fragment size



Most Probable Parse

- 4 identical derivations
- Sum probabilities

  72/184 + 6/184 + 6/184 + 1/184 = 85/184
- P(MPP) = 85/184

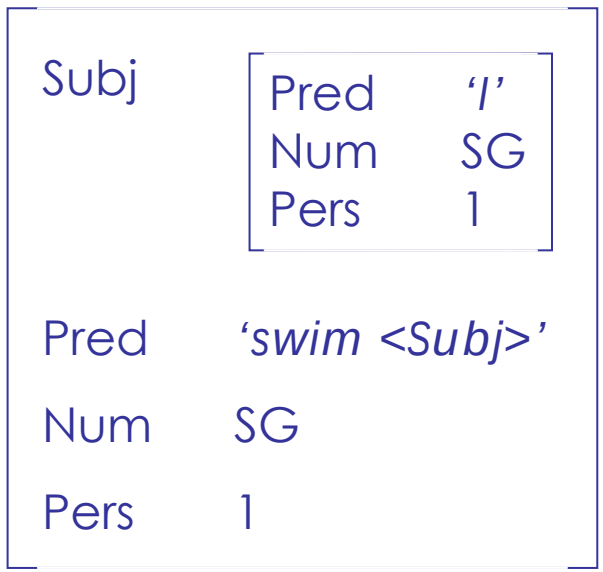


LFG

- Developed by Kaplan & Bresnan
- Beyond context-free
- Encodes grammatical features
  - number, case, tense
- Identifies grammatical functions of constituents
  - subj, obj, comp, adjunct

LFG

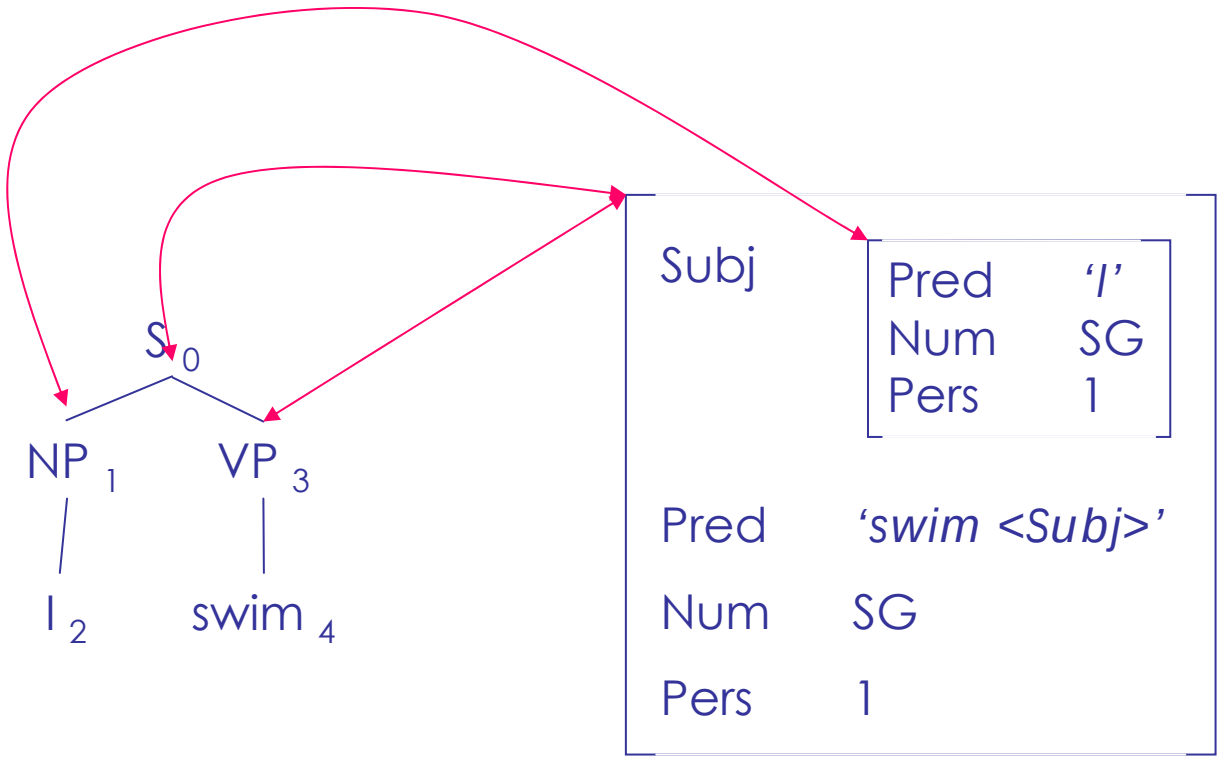
- 3 well-formedness conditions:
  - uniqueness
  - coherence
  - completeness



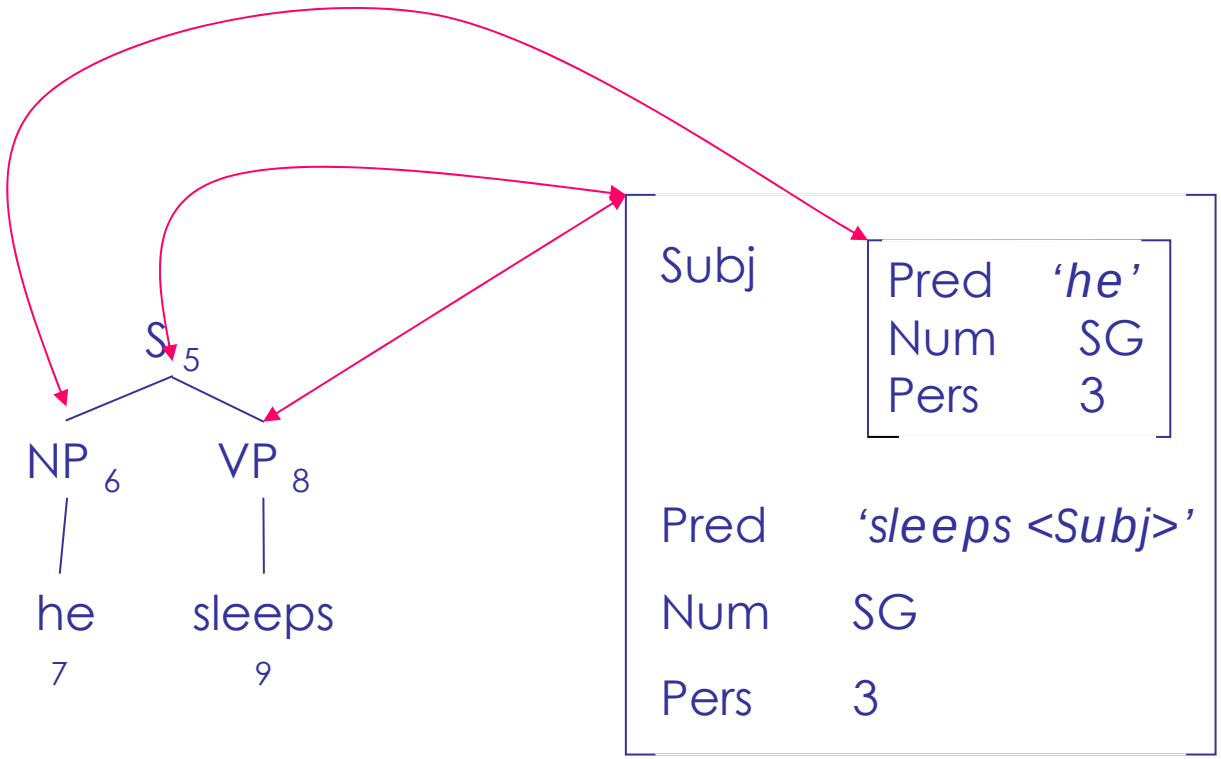
LFG-DOP

- Bod & Kaplan (1998)
- DOP Model based on syntactic representations of LFG
- Combines advantages of two approaches:
  - Linguistic adequacy of LFG
  - Robustness of DOP

LFG-DOP

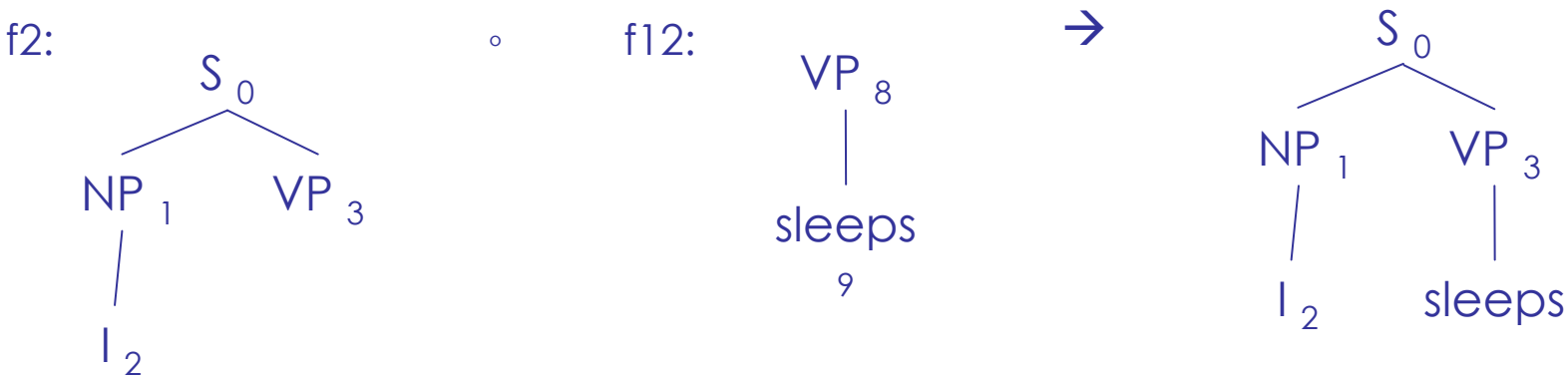


LFG-DOP



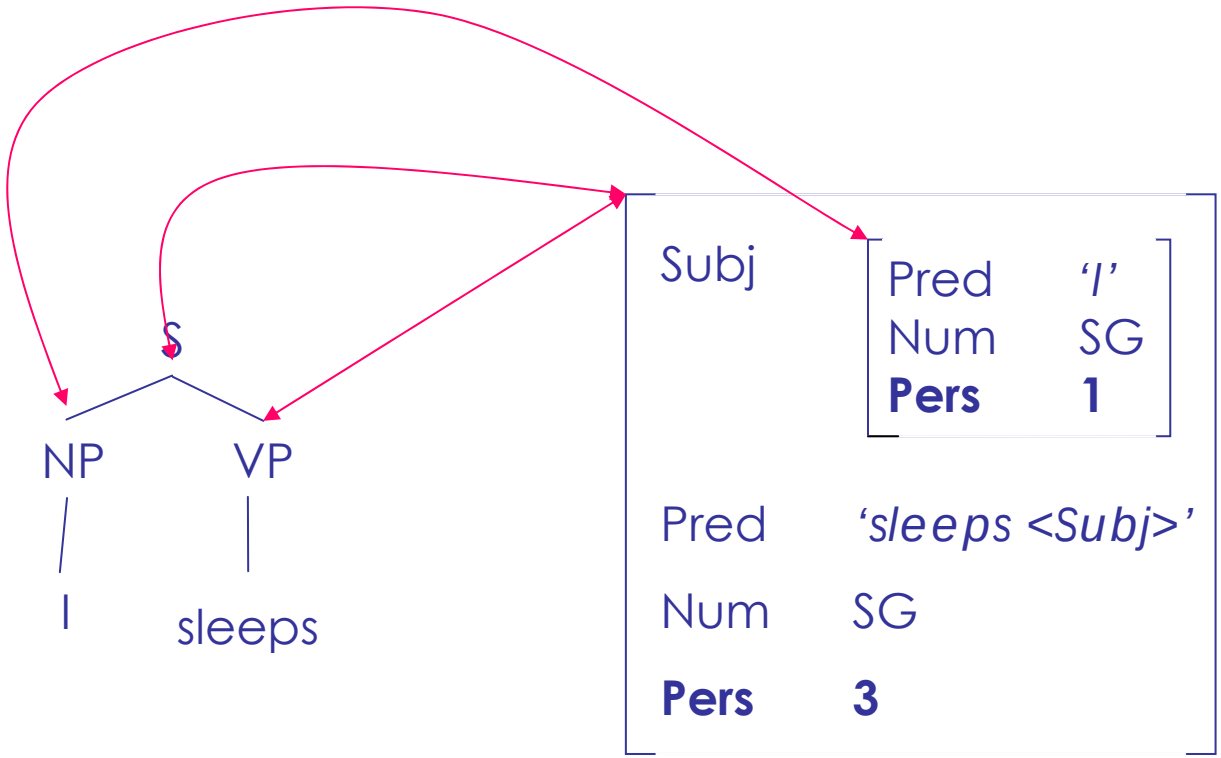
LFG
C-Structure

- "I sleeps" is possible in DOP



LFG
F-Structure

- But not in LFG-DOP



Current Experiments

- GF-DOP: Grammatical Function DOP
  - Subj, Obj(2), Obl(2), Comp, Xcomp, Poss, Adjunct
  - NP^Subj = "he"
  - NP^Obj = "him"
- English Xerox HomeCentre Corpus
- 8 training/test splits

Current Experiments

Version 1		f_inc	f_ex
baseline	[NPpro]	92.7425	92.7425
+Subj	[NPpro^SUBJ]	92.0252	92.8258
+Obj	[NPpro^OBJ]	92.1554	92.8683
+Subj+Obj	[NPpro^SUBJ, NPzero^Obj]	91.8324	92.7293
+AllTags	[NPpro^COMP]	90.7493	92.1227

Version 2		f_inc	f_ex
baseline-X	[NP]	93.3492	93.3492
+Subj-X	[NP^SUBJ]	92.0156	93.3808
+Obj-X	[NP^OBJ]	92.7961	92.9767
+Subj+Obj-X	[NP^SUBJ, NP^OBJ]	92.2569	93.1476
+AllTags-X	[NP^COMP]	91.9739	93.2397

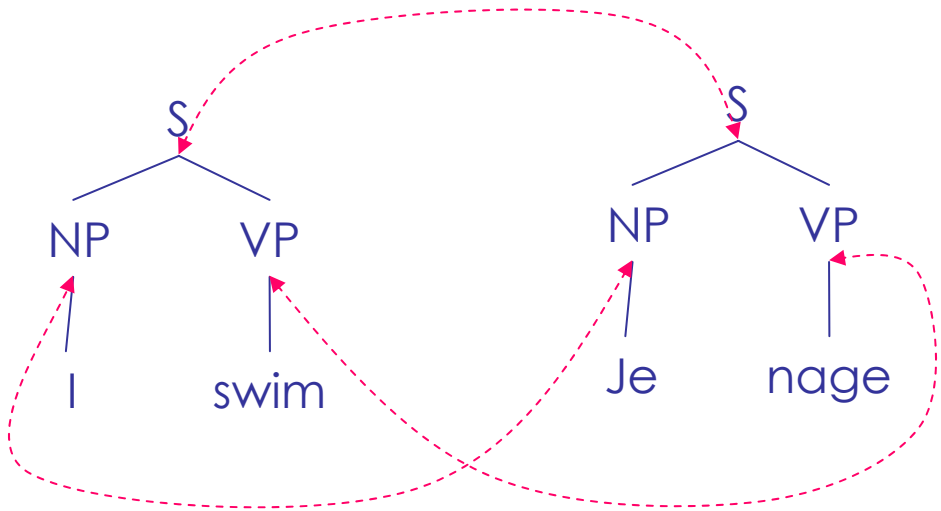
Future Work

- French Xerox HomeCentre Corpus
  - NUM, PERS, GENDER
    - "the man" → "l'homme"
    - "the men" → "les hommes"
- Freer word order languages
  - German: NOM, ACC, DAT, GEN
- Larger datasets
- Different functional annotations

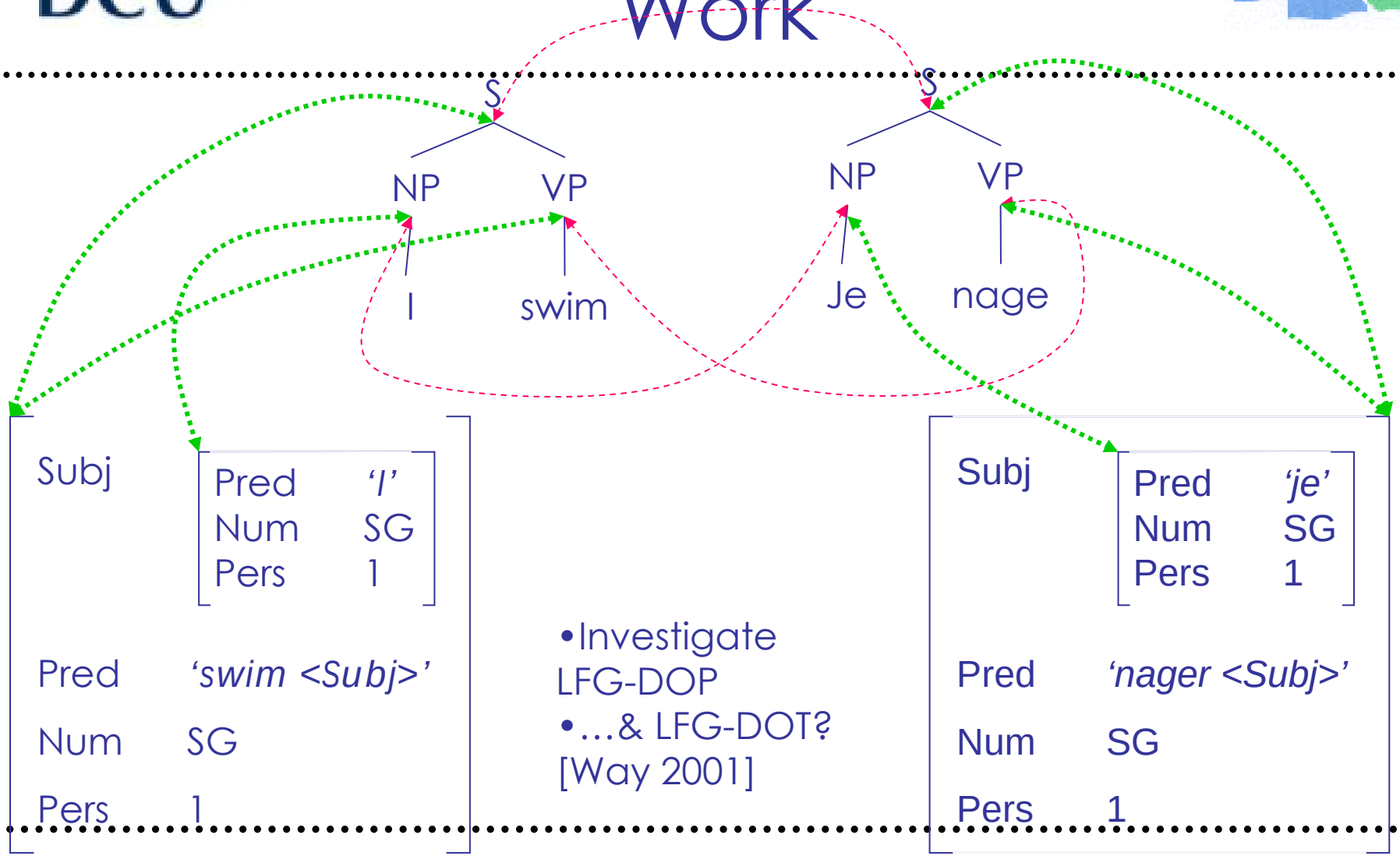
Future Work

- DOP Hypothesis
  - Holds true for DOP
  - Holds true for DOT

    (translation by dual parsing)



Future Work



- Investigate LFG-DOP
- …& LFG-DOT? [Way 2001]

References

- Bod, R. (1992): 'A Computational Model of Language Performance: Data-Oriented Parsing', in COLING: *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, **3**:855-859.
- Bod, R. and R.Kaplan (1998): 'A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis', in COLING: *Proceedings of the 17th International Conference on Computational Linguistics and 36th Conference of the Association for Computational Linguistics*, Montreal, Canada, **1**:145-151.
- Bod, R., Scha, R., and Sima'an, K., editors (2003). *Data-Oriented Parsing.* Stanford CA: CSLA Publications.
- Kaplan, R. and J. Bresnan, (1982): 'Lexical Functional Grammar: A Formal System for Grammatical Representation', in J.Bresnan (ed.) *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, Mass., pp. 173-281.
- Scha, R. (1990): 'Language Theory and Language Technology: Competence and Performance', in de Kort, Q. and G.Leerdam (eds) *Computertoepassingen in de Neerlandistiek* (in Dutch), Almere: LVNN-jaarboek.
- Way, A. (2001). LFG-DOT: A Hybrid Architecture for Robust MT. PhD thesis, University of Essex, Colchester, UK.