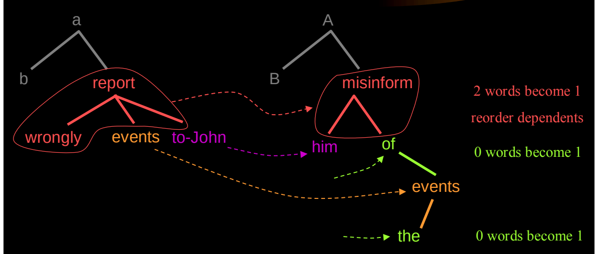
Learning Non-Isomorphic Tree Mappings for Machine Translation

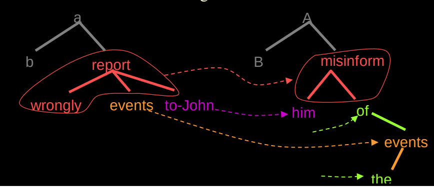
Jason Eisner - Johns Hopkins Univ.



"wrongly report events to-John" → "him misinform of the events"

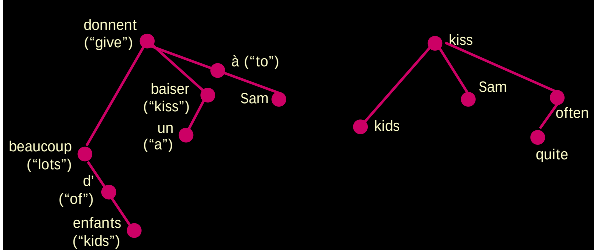
Syntax-Based Machine Translation

- Previous work assumes essentially isomorphic trees
  - Wu 1995, Alshawi et al. 2000, Yamada & Knight 2000
- But trees are *not* isomorphic!
  - Discrepancies between the languages
  - Free translation in the training data



Synchronous Tree Substitution Grammar

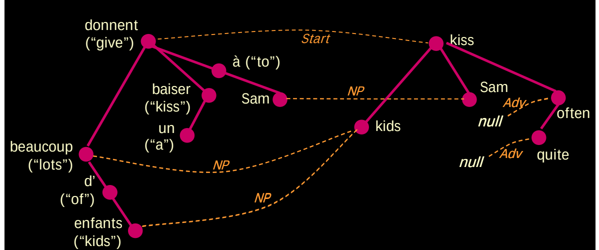
Two training trees, showing a **free translation** from French to English.



"beaucoup d'enfants donnent un baiser à Sam" → "kids kiss Sam quite often"

Synchronous Tree Substitution Grammar

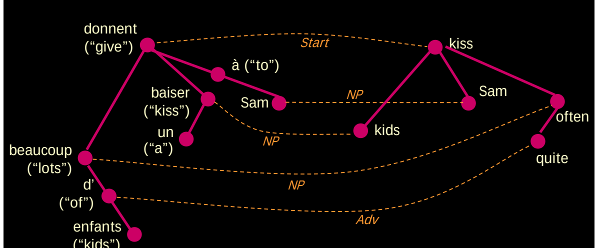
Two training trees, showing a **free translation** from French to English.
A possible alignment is shown in orange.



"beaucoup d'enfants donnent un baiser à Sam" → "kids kiss Sam quite often"

Synchronous Tree Substitution Grammar

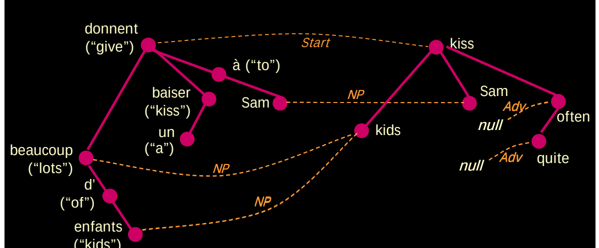
Two training trees, showing a **free translation** from French to English.
A possible alignment is shown in orange.
A much worse alignment ...



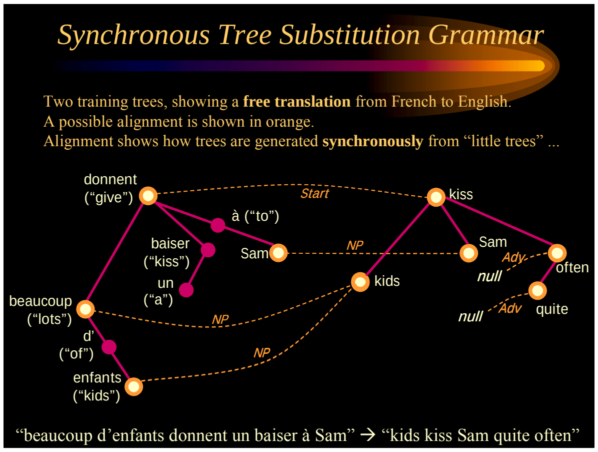
"beaucoup d'enfants donnent un baiser à Sam" → "kids kiss Sam quite often"

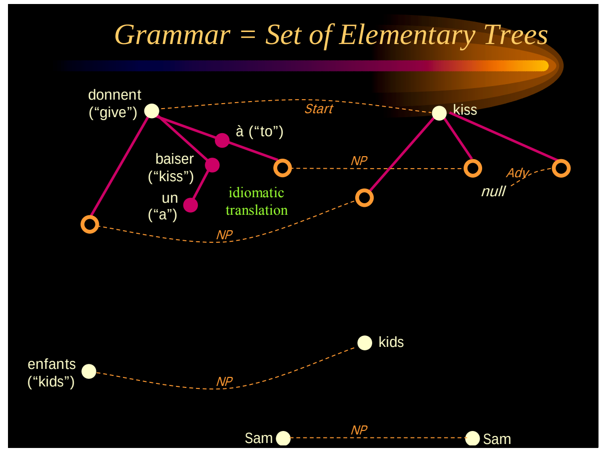
Synchronous Tree Substitution Grammar

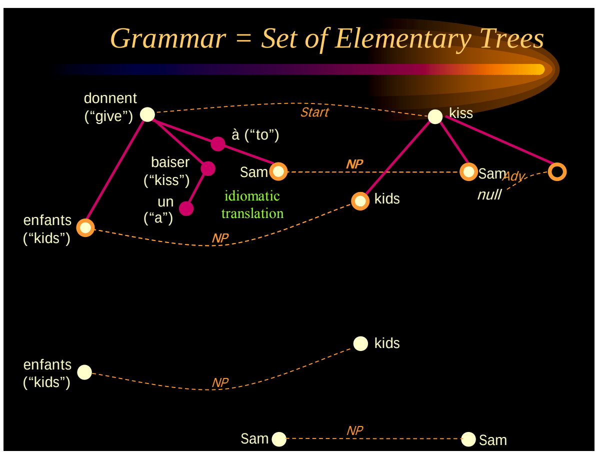
Two training trees, showing a **free translation** from French to English.
A possible alignment is shown in orange.

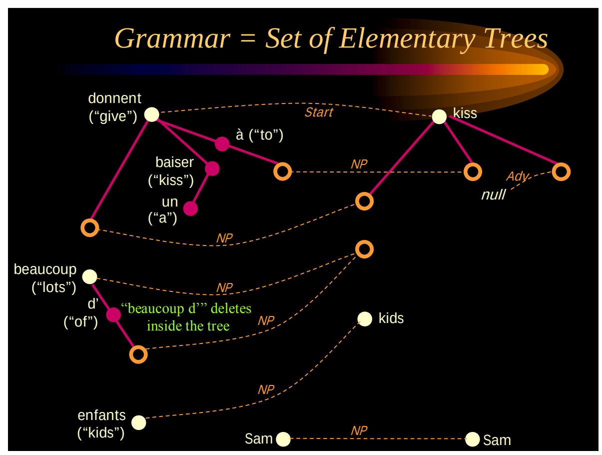


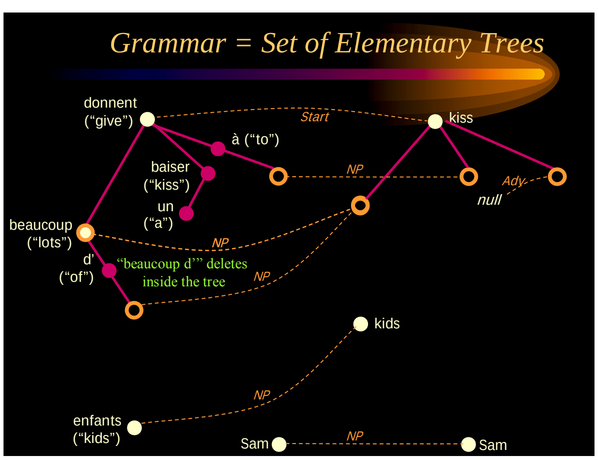
"beaucoup d'enfants donnent un baiser à Sam" → "kids kiss Sam quite often"

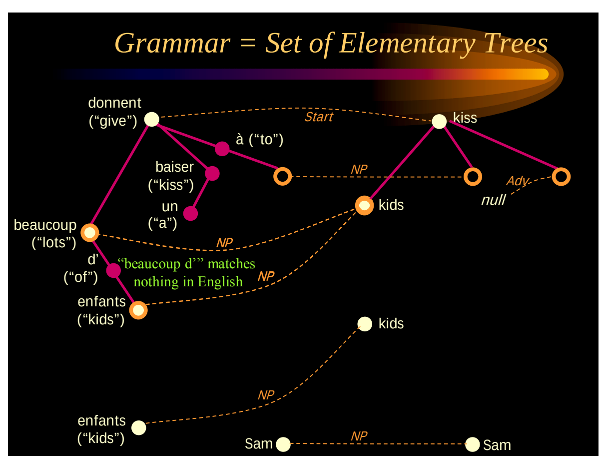

Synchronous Tree Substitution Grammar
Two training trees, showing a **free translation** from French to English.
A possible alignment is shown in orange.
Alignment shows how trees are generated **synchronously** from "little trees" ...
donnent ("give")
à ("to")
baiser ("kiss")
un ("a")
Sam
beaucoup ("lots")
d' ("of")
enfants ("kids")
Start
NP
NP
NP
kiss
Sam
kids
Adv
often
null
Adv
quite
null
"beaucoup d'enfants donnent un baiser à Sam" → "kids kiss Sam quite often"


Grammar = Set of Elementary Trees
donnent ("give")
à ("to")
baiser ("kiss")
un ("a")
idiomatic translation
Start
NP
NP
kiss
Adv
null
enfants ("kids")
NP
kids
Sam
NP
Sam

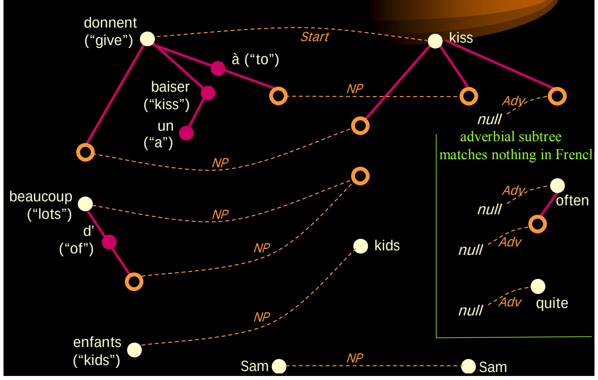

Grammar = Set of Elementary Trees
donnent ("give")
à ("to")
baiser ("kiss")
un ("a")
Sam
idiomatic translation
enfants ("kids")
Start
NP
NP
kiss
Sam
kids
Adv
null
enfants ("kids")
NP
kids
Sam
NP
Sam


Grammar = Set of Elementary Trees
donnent ("give")
à ("to")
baiser ("kiss")
un ("a")
Start
NP
NP
kiss
Adv
null
beaucoup ("lots")
d' ("of")
"beaucoup d'" deletes inside the tree
NP
NP
kids
enfants ("kids")
NP
Sam
NP
Sam


Grammar = Set of Elementary Trees
donnent ("give")
à ("to")
baiser ("kiss")
un ("a")
beaucoup ("lots")
d' ("of")
"beaucoup d'" deletes inside the tree
Start
NP
NP
NP
kiss
Adv
null
kids
NP
enfants ("kids")
Sam
NP
Sam


Grammar = Set of Elementary Trees
donnent ("give")
à ("to")
baiser ("kiss")
un ("a")
beaucoup ("lots")
d' ("of")
"beaucoup d'" matches nothing in English
enfants ("kids")
Start
NP
NP
NP
kiss
kids
Adv
null
kids
NP
enfants ("kids")
Sam
NP
Sam

Grammar = Set of Elementary Trees

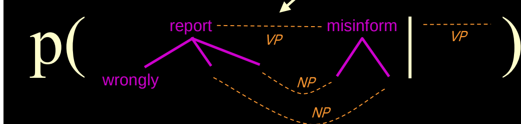


Probability model similar to PCFG

Probability of generating training trees T1, T2 with alignment A

$$P(T1, T2, A) = \Pi\ p(t1,t2,a \mid n)$$

probabilities of the "little" trees that are used



is given by a maximum entropy model

*Form of model of **big tree pairs***

Joint model $P_\theta(T1,T2)$.

Wise to use noisy-channel form: $P_\theta(T1 \mid T2)$ * $P_\theta(T2)$

But any joint model will do.

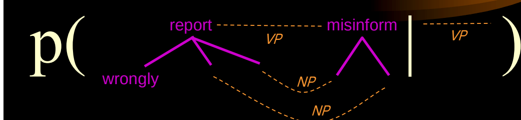
train on paired trees (hard to get)

could be trained on zillions of target-language trees

In synchronous TSG, aligned big tree pair is generated by choosing a sequence of little tree pairs:

$$P(T1, T2, A) = \Pi\ p(t1,t2,a \mid n)$$

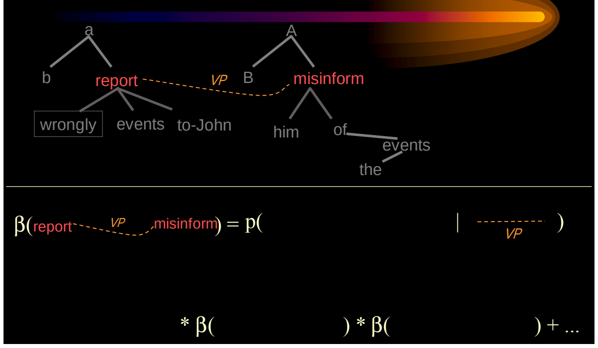
*Maxent model of **little tree pairs***



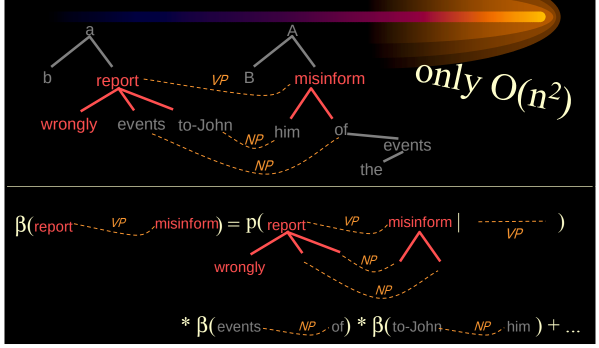
FEATURES

- report+wrongly ↔ misinform? (use dictionary)
- report ↔ misinform? (at root)
- wrongly ↔ misinform?
- verb incorporates adverb child?
- verb incorporates child 1 of 3?
- children 2, 3 switch positions?
- common tree sizes & shapes?
- ... etc. ....

Inside Probabilities



Inside Probabilities



$P(T1, T2, A) = \prod p(t1,t2,a \mid n)$

- Alignment: find A to max $P_\theta(T1,T2,A)$
- Decoding: find T2, A to max $P_\theta(T1,T2,A)$
- Training: find θ to max $\sum_A P_\theta(T1,T2,A)$

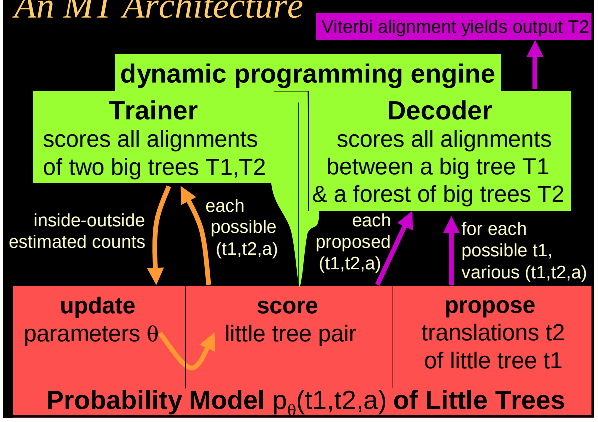
- **Do everything on little trees instead!**
- Only need to train & decode a model of $p_\theta(t1,t2,a)$
- But not sure how to break up big tree correctly
  - So try all possible little trees & all ways of combining them, by dynamic prog.

Alignment Pseudocode

for each node **c1** of **T1** (bottom-up)
  for each possible little tree **t1** rooted at **c1**
    for each node **c2** of **T2** (bottom-up)
      for each possible little tree **t2** rooted at **c2**
        for each matching **a** between frontier nodes of **t1** and **t2**
          p = p(t1,t2,a)
          for each pair **(d1,d2)** of frontier nodes matched by **a**
            p = p * β(d1,d2) // inside probability of kids
          β(c1,c2) = β(c1,c2) + p // our inside probability

Nonterminal states are used in practice but not shown here
For EM training, also find outside probabilities

An MT Architecture



Related Work

- **Synchronous grammars** (Shieber & Schabes 1990)
  - Statistical work has allowed only 1:1 (isomorphic trees)
    - Stochastic inversion transduction grammars (Wu 1995)
    - Head transducer grammars (Alshawi et al. 2000)
- **Statistical tree translation**
  - Noisy channel model (Yamada & Knight 2000)
    - Infers tree: trains on (string, tree) pair, not (tree, tree) pair
    - But again, allows only 1:1, plus 1:0 at leaves
- **Data-oriented translation** (Poutsma 2000)
  - Synchronous DOP model trained on already *aligned* trees
- **Statistical tree generation**
  - Similar to our decoding: construct forest of appropriate trees, pick by highest prob
  - Dynamic prog. search in packed forest (Langkilde 2000)
  - Stack decoder (Ratnaparkhi 2000)

What Is New Here?

- Learning full elementary tree pairs, not rule pairs or subcat pairs
  - Previous statistical formalisms have basically assumed isomorphic trees
- Maximum-entropy modeling of elementary tree pairs
- New, flexible formalization of synchronous Tree Subst. Grammar
  - Allows either dependency trees or phrase-structure trees
  - "Empty" trees permit insertion and deletion during translation
  - Concrete enough for implementation (cf. informal previous descriptions)
  - TSG is more powerful than CFG for modeling trees, but faster than TAG
- Observation that dynamic programming is surprisingly fast
  - Find all possible decompositions into aligned elementary tree pairs
  - $O(n^2)$ if both input trees are fully known and elem. tree size is bounded

Status & Thanks

- Developed and implemented during JHU CLSP summer workshop 2002 (funded by NSF)
- Other team members: Jan Hajič, Bonnie Dorr, Dan Gildea, Gerald Penn, Drago Radev, Owen Rambow, and students: Martin Cmejrek, Yuan Ding, Terry Koo, Kristen Parton

- Also being used for other kinds of tree mappings:
  - between deep structure and surface structure, or semantics and syntax
  - between original text and summarized/paraphrased/plagiarized version
- Results forthcoming (that's why I didn't submit a full paper ☺)

Summary

- Most MT systems work on strings
- We want to translate trees – want to respect syntactic structure
- But don't assume that translated trees are structurally isomorphic!

- ➔ TSG formalism: Translation locally replaces tree structure and content.
- ➔ Parameters: Probabilities of local substitutions (use maxent model)
- ➔ Algorithms: Dynamic programming (local substitutions can't overlap)

- EM training on <English tree, Czech tree> pairs can be fast:
  - Align O(n) tree nodes with O(n) tree nodes, respecting subconstituency
  - Dynamic programming – find all alignments and retrain using EM
  - Faster than aligning O(n) words with O(n) words
  - If correct training tree is unknown, a well-pruned parse forest still has O(n) nodes