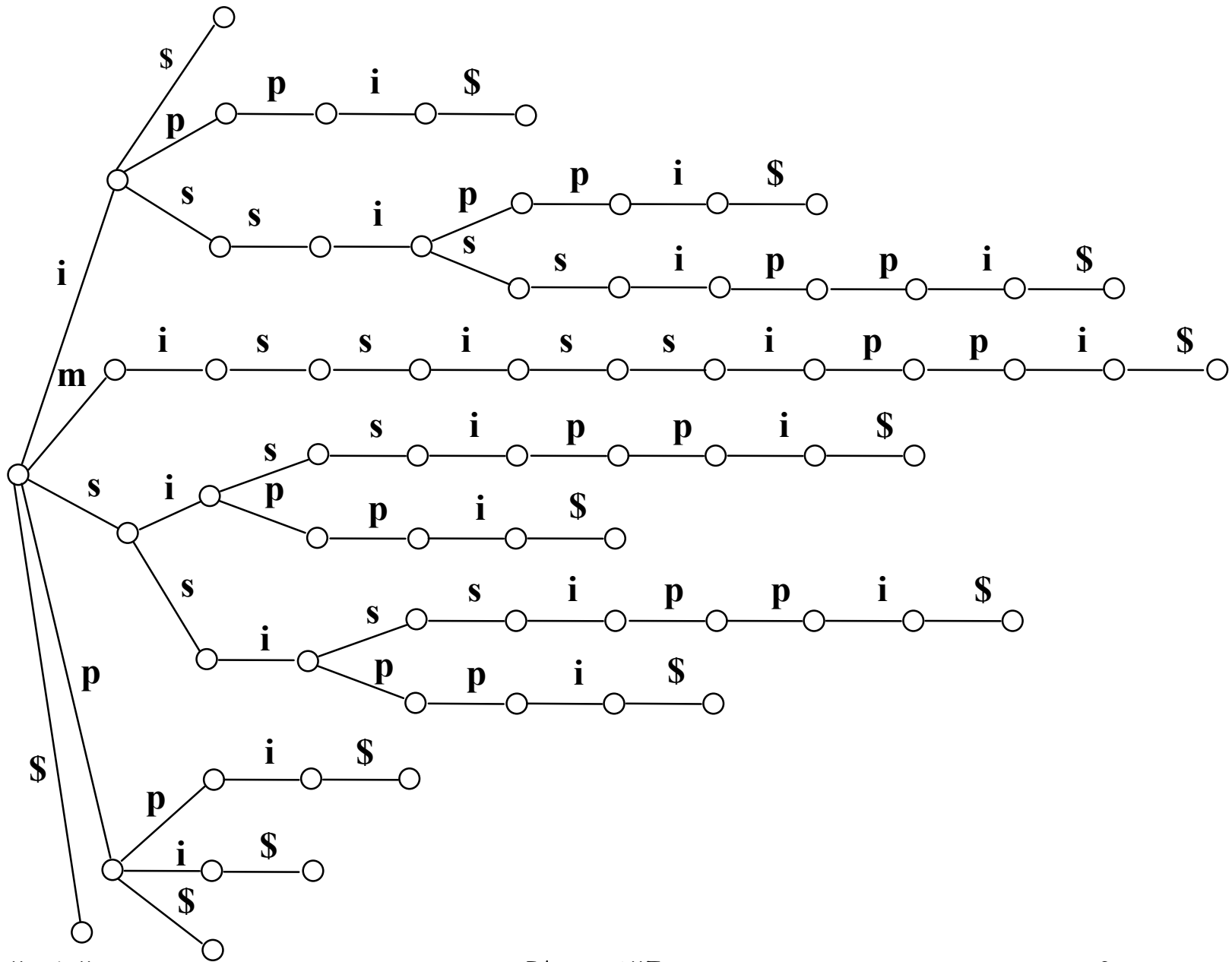


# 5 Ideas on Statistical Machine Translation

**Martin Kay**

*Stanford University and  
The University of the Saarland*

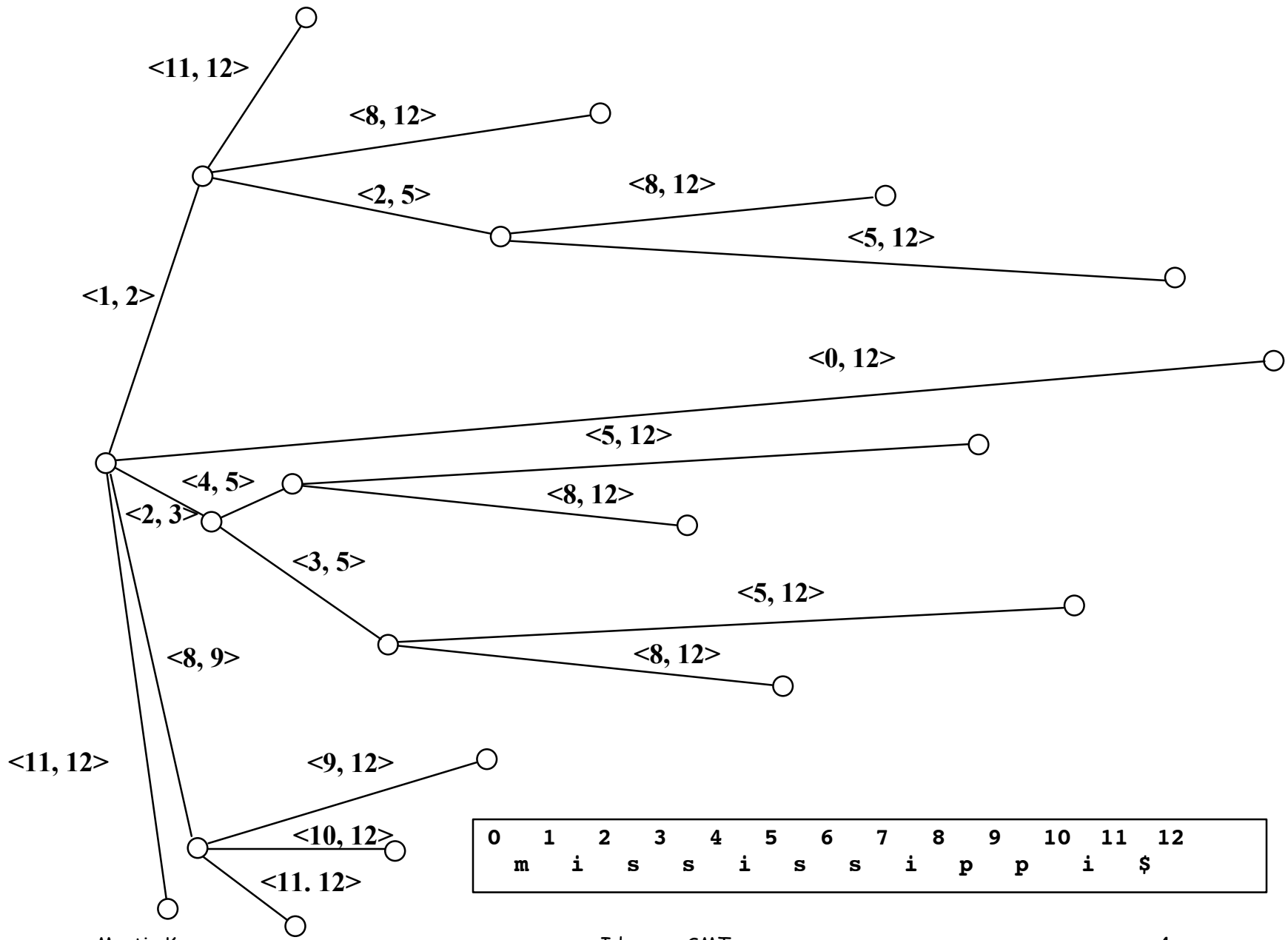
# Suffix Trees



Martin Kay

Ideas on SMT

3



Martin Kay

Ideas on SMT

4

0	1	2	3	4	5	6	7	8	9	10	11	12
m	i	s	s	i	s	s	i	p	p	i	i	\$

# Suffix Trees

## Enable one to discover

- whether the text from which it is constructed contains a given substring (in time proportional to the length of the substring and unrelated to the length of the text.)
- where all the occurrences of a substring are located (in time proportional to the number of those occurrences.)
- All substrings of length  $n$  (or greater) that occur  $k$  (or more) times.

Has size and construction time proportional to the size of the text

# Suffix Trees and Translation Models

A phrase is a sequence of characters that is (1) long enough, and (2) occurs frequently enough to be interesting

# ① Flexible Phrase Boundaries

Decompose compounds and morphologically complex words and word sequences.

**Lebensversicherungsgesellschaftsangestellter  
Zweihundertsiebenundfünfzig**

**without a trace  
without a shadow of a doubt  
in the final analysis  
all things considered**

**mensa  
mensam  
mensae  
mensarum  
mensis**

## **Klonmeister aus Korea**

Auf dem Jahrestreffen der "American Association for the Advancement of Science" konnte das Wissenschaftsland Südkorea seinen spektakulären Forschungserfolg zum therapeutischen Klonen präsentieren Von Gero von Randow für ZEIT.de

**darselo  
mostrártela  
estudiándola**

## Find sequences that

1. consist of at least  $L$  characters
2. are repeated at least  $R$  times
3. are not always preceded by the same character
4. are not always followed by the same character
5. do not cross a sentence boundary

# A Few Alignments

-ly

-ment

-s

les

anti-

contraire

clockwise

dans le sense des aiguilles

une montre

conditioning

climatisation

lighter

cigar

# A Few More

anticlockwise

gegen den Uhrzeigersinn

Hand break

Feststellbremse

break

-bremse

turned to "AUX"

in die "AUX" Stellung  
gedreht

all-wheel drive

der Allradantrieb

ignition key

Zündschlüssel

the engine and

abstellen und

safety

Sicherheits-

# Discontinuous items ???

Phrasal verbs  
as far as ... is concerned  
...

## ② Outside/Inside Phrases

Identity in the phrase table depends on identity of their insides. Coverage of the input or output string depends on their outsides.

kleinen

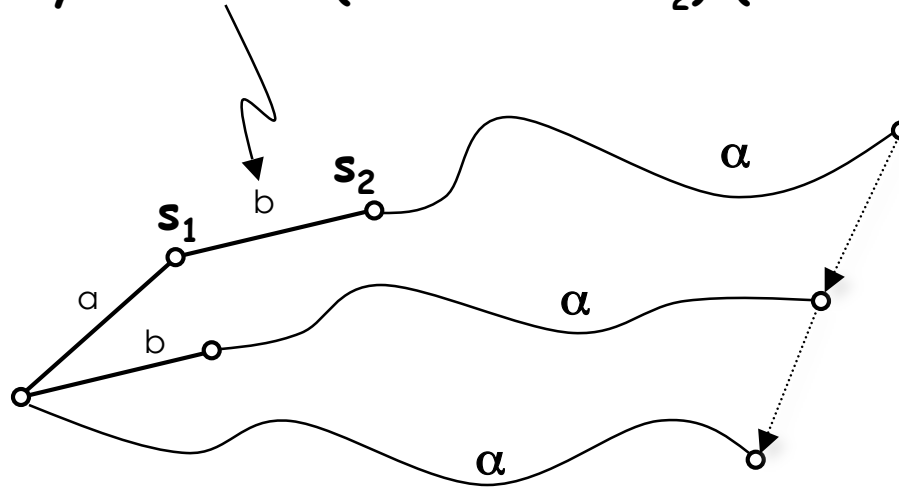
What about

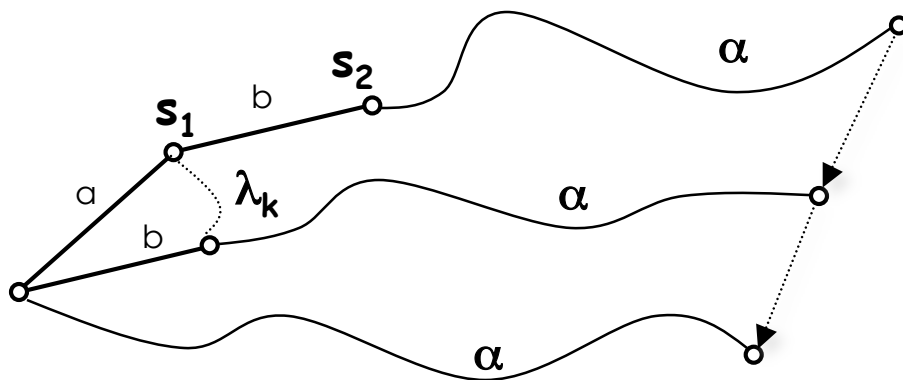
den kleinen Kindern

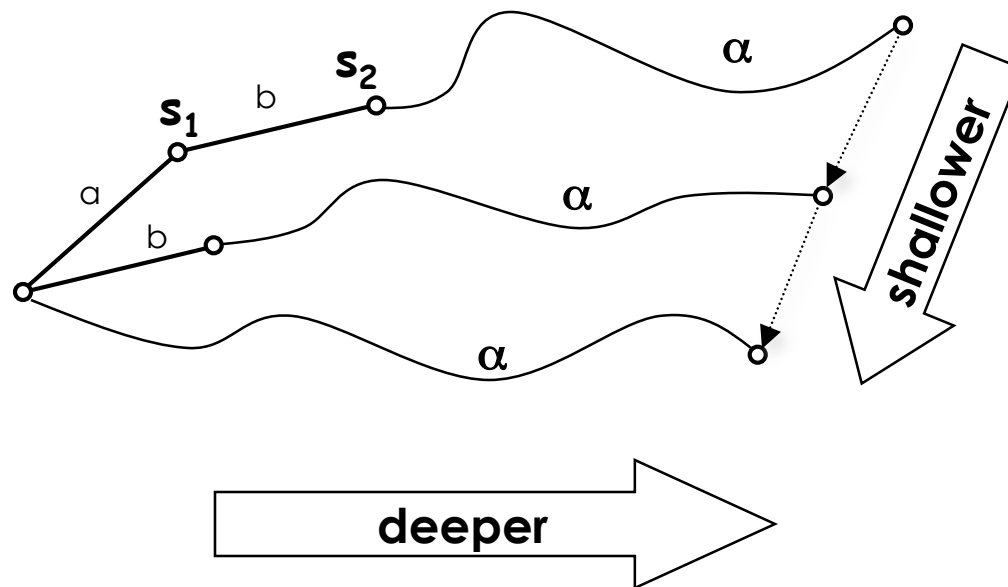
# ③ Suffix Trees and Language Models

- For long strings, two occurrences is a lot more than one, and so on for three, four ...
- 5-tuples are a lot better than 3-tuples (Google)
- So what about variable tuples?

Probability of this is  $(\text{branches at } s_2)/(\text{branches at } s_1)$





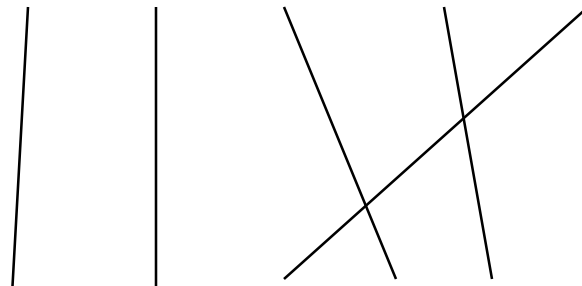


# Distortion

# Distortion

**A really bad idea: line crossings**

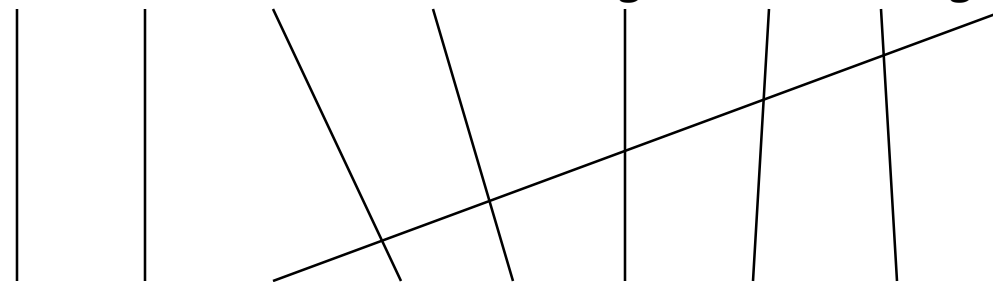
das Buch, das Maria Dem Jungen gab



2

the book that Maria gave the boy

das Buch, das Maria dem anderen Jungen aus Polen gab



5

the book that Maria gave the other boy from Poland

# Distortion

A fairly bad idea: Levenstein distance

So etwas interesstantes hätte ich sehen wollen

I would\_have wanted to\_see something that interesting

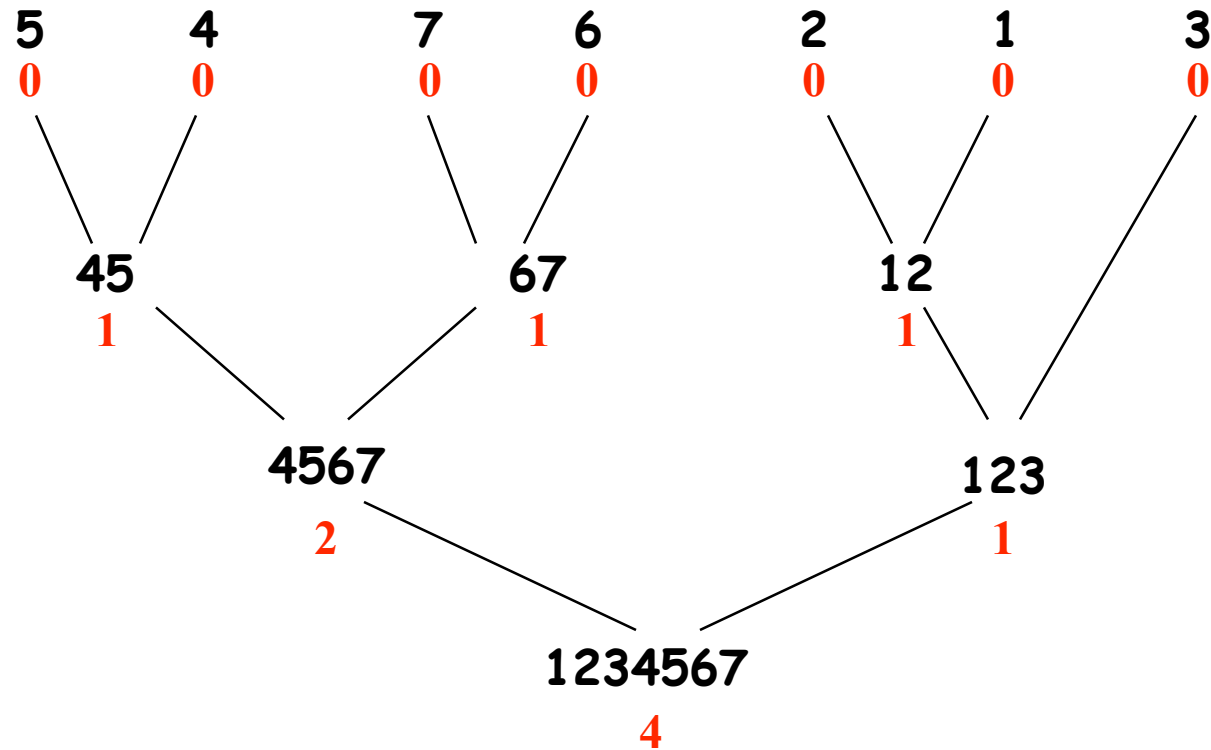
Word-for-word replacement is as good as you can do

# ④ Parse Distance

So etwas interesstantes hätte ich sehen wollen

1 2 3 4 5 6 7

I would\_have wanted to\_see something that interesting

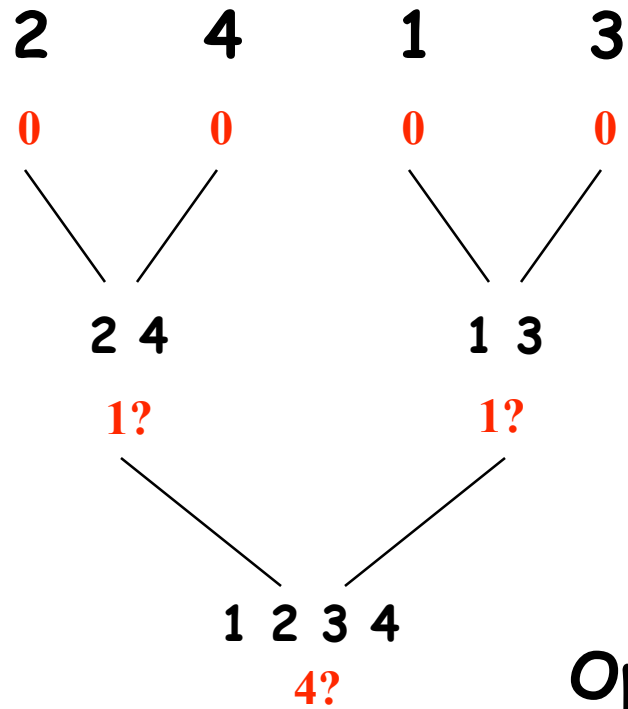


# Deterministic Computation

```
def distortion(string)
  stack = []
  for i in string.reverse
    stack.unshift([i.to_i, i.to_i , 0])
    while stack.length>1 && n = combine(stack[0], stack[1])
      stack = [n]+stack[2..-1]
    end
  end
  puts "Result = " + stack.to_s
end
```

```
def combine(a, b)
  if a[1]+1 == b[0]
    [a[0], b[1], a[2]+b[2]]
  elsif a[0] == b[1]+1
    [b[0], a[1], a[2]+b[2]+1]
  else
    false
  end
end
```

# What about ...



Optimality Scoring

# Combinatorics — "Vowels"

Define vowels as *the minimal set of characters one of which occurs in every word.*

Frequency	
b	1
i	3
l	1
o	3
p	2
r	1
s	1
t	4

rob  
top  
lot  
it  
tip  
is

o v i  $\Rightarrow$

(r v o v b)  $\wedge$   
(t v o v p)  $\wedge$   
(l v o v t)  $\wedge$   
(i v t)  $\wedge$   
(t v i v p)  $\wedge$   
(i v s)

*prime implicants*

## ⑤ Combinatoric MT

Ordered phrase table

$e : f$

Value of a phrase  $v(e:f)$

$f_{ef}$  = number of sentence pairs containing  $e$  and  $f$

$f_e$  = number of english sentences containing  $e$

$f_f$  = number of foreign sentences containing  $f$

$\text{len}(x)$  = length of the string  $x$ .

$$\frac{f_{ef}}{f_e + f_f - f_{ef}} \times \frac{\text{len}(e) + \text{len}(f)}{2}$$

# Combinatoric MT — Training

1. Start with empty rule list
2. Repeat
  1. Add most valuable rule to the end of the rule list
  2. In pairs of sentences to which the rules applies, replace the matching strings by a "blocking" character.

# Combinatoric MT — Translation

Apply rules in order

# The End