

# A Compact Data Structure for Searchable TMs

Chris Callison-Burch,  
Colin Bannard, and Josh Schroeder

**LINEAR B**

# Talk Overview

- Searchable translation memories
- Phrase extraction with SMT
- Indexing using suffix arrays
- Evaluation and conclusion

# Technology for Human Translation

- Market for human translation much larger than for machine translation
- Technology must be useful and simple
- Translation memories most used tool

# The Future of TMs

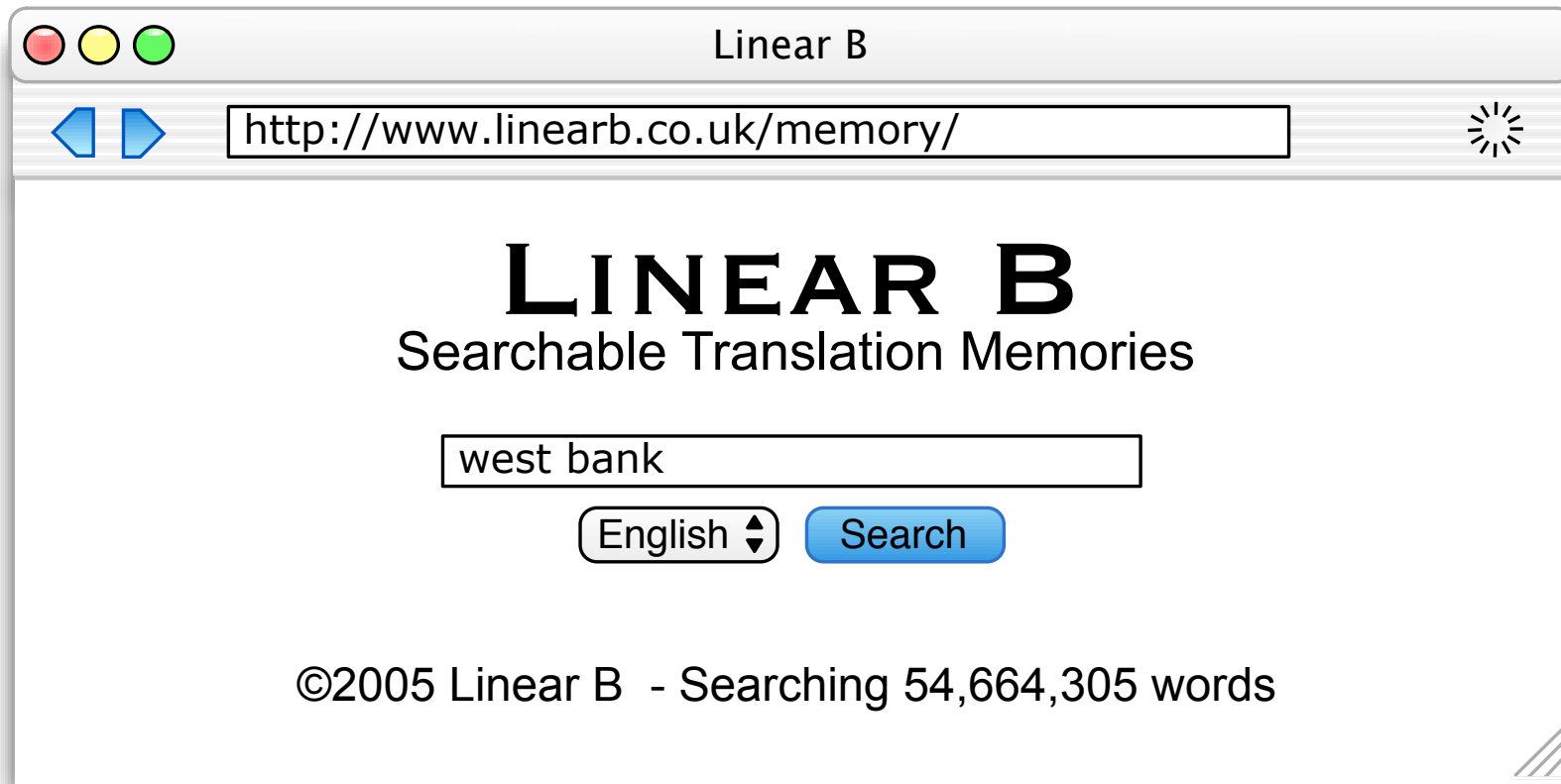
- "Certainly there are improvements that can be made to current TM applications and their underlying technologies ... the next new wave of language technology innovation will come from extending the recycling of translated material to different levels of granularity, and to combine them. Today we mainly recycle on the term level, sentence level, or document level. In the future, we want to extend this to recycling on the phrase level. This will require its own smart linguistic algorithms."

(Trados CTO, article in Clientside News)

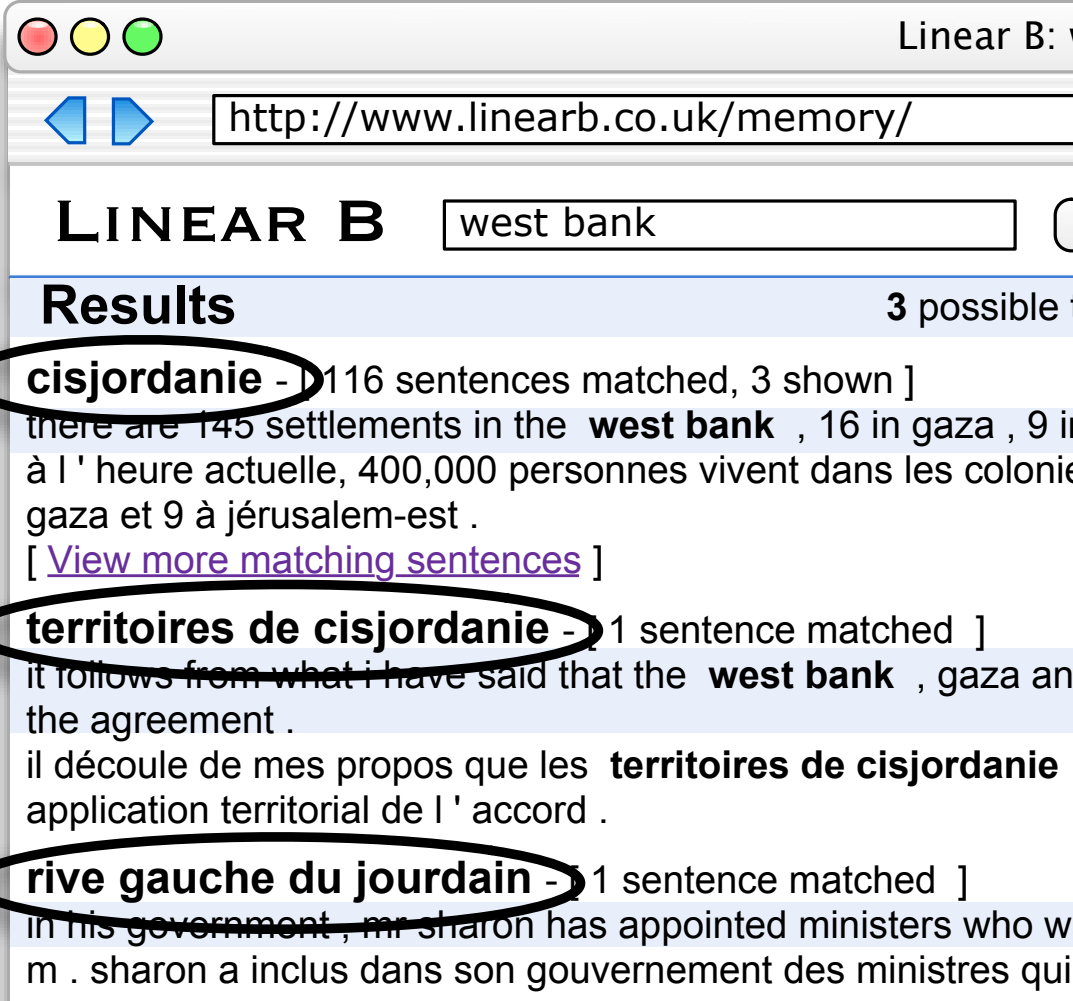
# Searchable Translation Memories

- Simple Google-style searching
- Retrieves translations of phrases
- Ranks them by frequency / probability
- Highlights translations in context

# Simple interface



# Retrieves translations



Linear B: v

http://www.linearb.co.uk/memory/

**LINEAR B**  (

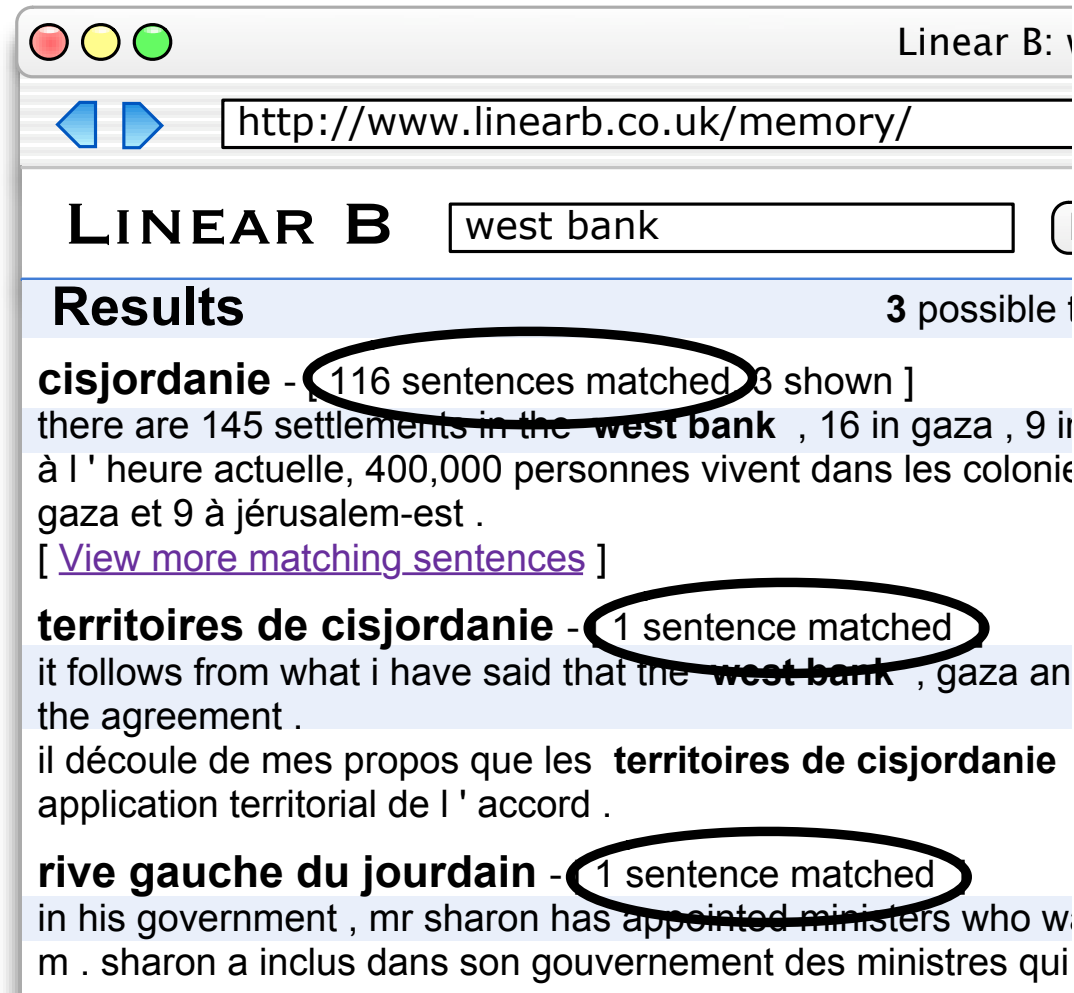
**Results** 3 possible

**cisjordanie** - [ 116 sentences matched, 3 shown ]  
there are 145 settlements in the **west bank** , 16 in gaza , 9 in  
à l ' heure actuelle, 400,000 personnes vivent dans les colonies  
gaza et 9 à jérusalem-est .  
[ [View more matching sentences](#) ]

**territoires de cisjordanie** - [ 1 sentence matched ]  
it follows from what i have said that the **west bank** , gaza and  
the agreement .  
il découle de mes propos que les **territoires de cisjordanie**  
application territoriale de l ' accord .

**rive gauche du jourdain** - [ 1 sentence matched ]  
in his government , mr sharon has appointed ministers who w  
m . sharon a inclus dans son gouvernement des ministres qui

# Ranked by frequency



Linear B: v

http://www.linearb.co.uk/memory/

**LINEAR B**

**Results** 3 possible t

**cisjordanie** - 116 sentences matched [ 3 shown ]  
there are 145 settlements in the **west bank** , 16 in gaza , 9 in  
à l ' heure actuelle, 400,000 personnes vivent dans les colonies  
gaza et 9 à jérusalem-est .  
[ [View more matching sentences](#) ]

**territoires de cisjordanie** - 1 sentence matched  
it follows from what i have said that the **west bank** , gaza and  
the agreement .  
il découle de mes propos que les **territoires de cisjordanie**  
application territoriale de l ' accord .

**rive gauche du jourdain** - 1 sentence matched  
in his government , mr sharon has appointed ministers who w  
m . sharon a inclus dans son gouvernement des ministres qui

# Highlights translations in context

**cisjordanie** - [ 116 sentences matched, 1 shown ]

there are 145 settlements in the **west bank** , 16 in gaza , 9 in east jerusalem ; 400,000 people live in them .

à l ' heure actuelle, 400,000 personnes vivent dans les colonies . on dénombre 145 colonies en **cisjordanie** , 16 à gaza et 9 à jérusalem-est .

[ [View more matching sentences](#) ]

**territoires de cisjordanie** - [ 1 sentence matched ]

it follows from what i have said that the **west bank** , gaza and the golan heights fall outside the territorial scope of the agreement .

il découle de mes propos que les **territoires de cisjordanie** , de gaza et du plateau du golan sortent du champ d ' application territorial de l ' accord .

**rive gauche du jourdain** - [ 1 sentence matched ]

in his government , mr sharon has appointed ministers who want to reclaim the **west bank** .

m . sharon a inclus dans son gouvernement des ministres qui veulent annexer la **rive gauche du jourdain** .

# Technical Challenges

- Identifying phrasal translations
- Compact storage of translation index

Solutions:

- Statistical machine translation
- Suffix arrays

# Statistical MT

- Data-driven
- Learns word and phrase alignments
- Language independent
- Low cost, rapid development

# Parallel corpus

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .





# Phrase Extraction 2

	we	owe	it	to	the	taxpayers	to	keep	the	costs	in	check
wir	■										■	■
sind											■	■
es			■								■	■
den				■	■						■	■
steuerzahlern						■					■	■
schuldig		■									■	■
die									■		■	■
kosten										■	■	■
unter	■	■	■	■	■	■	■	■	■	■	■	■
kontrolle	■	■	■	■	■	■	■	■	■	■	■	■
zu						■						
haben							■					

# Probabilities

- Can calculate probabilities by counting

- $$p(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})}$$

# Statistical MT in Searchable TMs

- Allows us to align phrases, rank them
- Can build an index of extracted phrases, pointers to sentences:

## **unter kontrolle**

under control	1005, 1017, 1115, ...
in check	732, 5914
curbed	3201
controlled	61734

# Practical Problem: Memory

- Index too large for personal computers
- Possible solution: limit phrase length
- Alternative: suffix arrays

# Suffix Arrays for monolingual corpora

- Quick retrieval of any subphrase in corpus
- Alphabetically sorted list of suffixes
- Compact = size of corpus

# Suffix Array

Index of words:

Corpus

0	1	2	3	4	5	6	7	8	9
Spain	declined	to	confirm	that	Spain	declined	to	aid	Morocco

Initialized, unsorted  
Suffix Array

Suffixes denoted by  $s[i]$

s[0]	0	Spain declined to confirm that Spain declined to aid Morocco
s[1]	1	declined to confirm that Spain declined to aid Morocco
s[2]	2	to confirm that Spain declined to aid Morocco
s[3]	3	confirm that Spain declined to aid Morocco
s[4]	4	that Spain declined to aid Morocco
s[5]	5	Spain declined to aid Morocco
s[6]	6	declined to aid Morocco
s[7]	7	to aid Morocco
s[8]	8	aid Morocco
s[9]	9	Morocco

# Suffix Array

Sorted  
Suffix Array

s[0]	8
s[1]	3
s[2]	6
s[3]	1
s[4]	9
s[5]	5
s[6]	0
s[7]	4
s[8]	7
s[9]	2

Suffixes denoted by s[i]

aid Morocco
confirm that Spain declined to aid Morocco
declined to aid Morocco
declined to confirm that Spain declined to aid Morocco
Morocco
Spain declined to aid Morocco
Spain declined to confirm that Spain declined to aid Morocco
that Spain declined to aid Morocco
to aid Morocco
to confirm that Spain declined to aid Morocco

# Suffix Array

Sorted  
Suffix Array

	s[0]	8
log(n)	s[1]	3
	s[2]	6
	s[3]	1
	s[4]	9
	s[5]	5
	s[6]	0
	s[7]	4
log(n)	s[8]	7
	s[9]	2

Suffixes denoted by s[i]

aid Morocco
confirm that Spain declined to aid Morocco
declined to aid Morocco
declined to confirm that Spain declined to aid Morocco
Morocco
Spain declined to aid Morocco
Spain declined to confirm that Spain declined to aid Morocco
that Spain declined to aid Morocco
to aid Morocco
to confirm that Spain declined to aid Morocco

# Suffix Arrays for Parallel Corpora

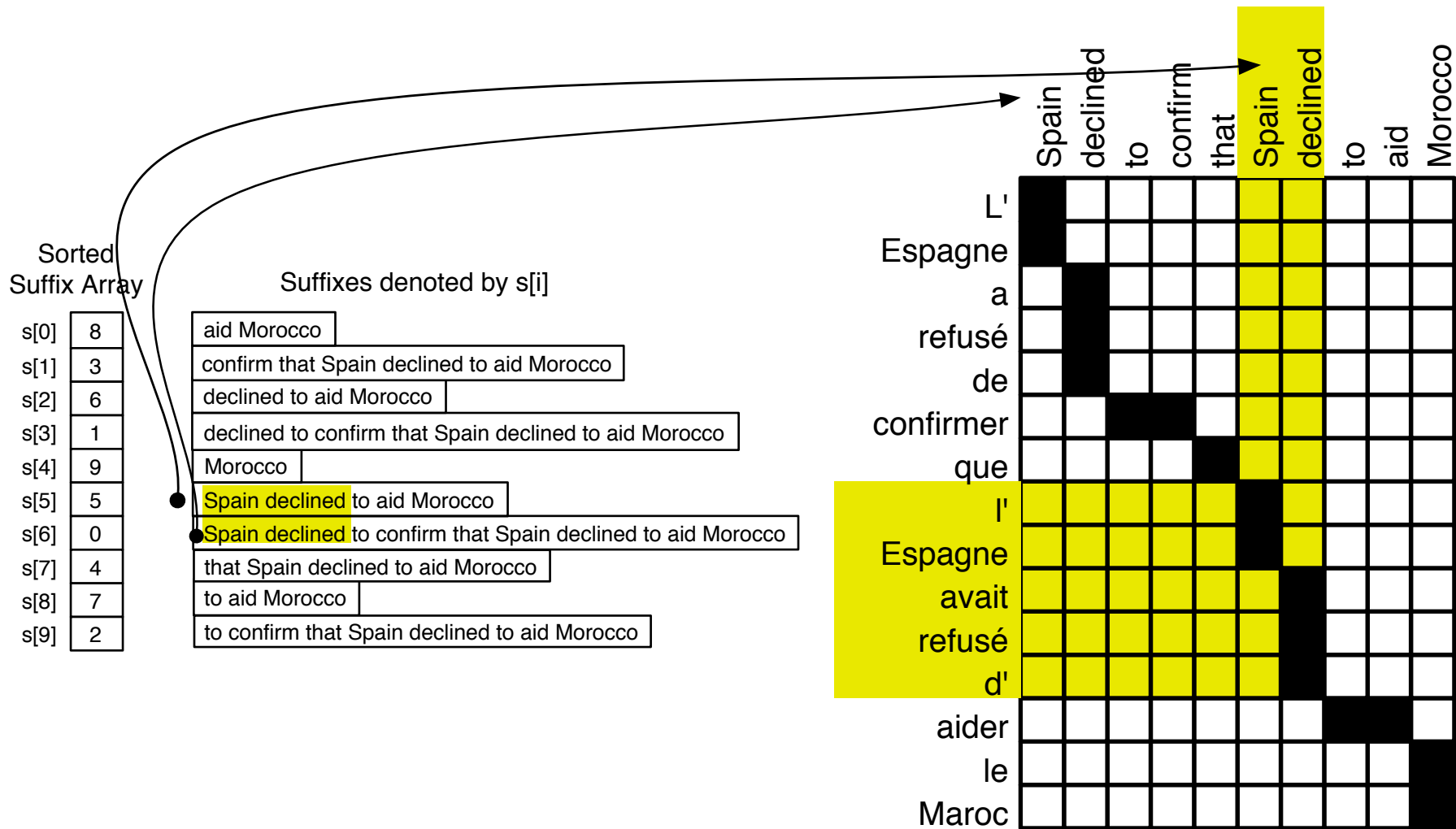
- On the fly lookup
- Very quick for long / infrequent phrases
- Frequent phrases take longer, but can get around this



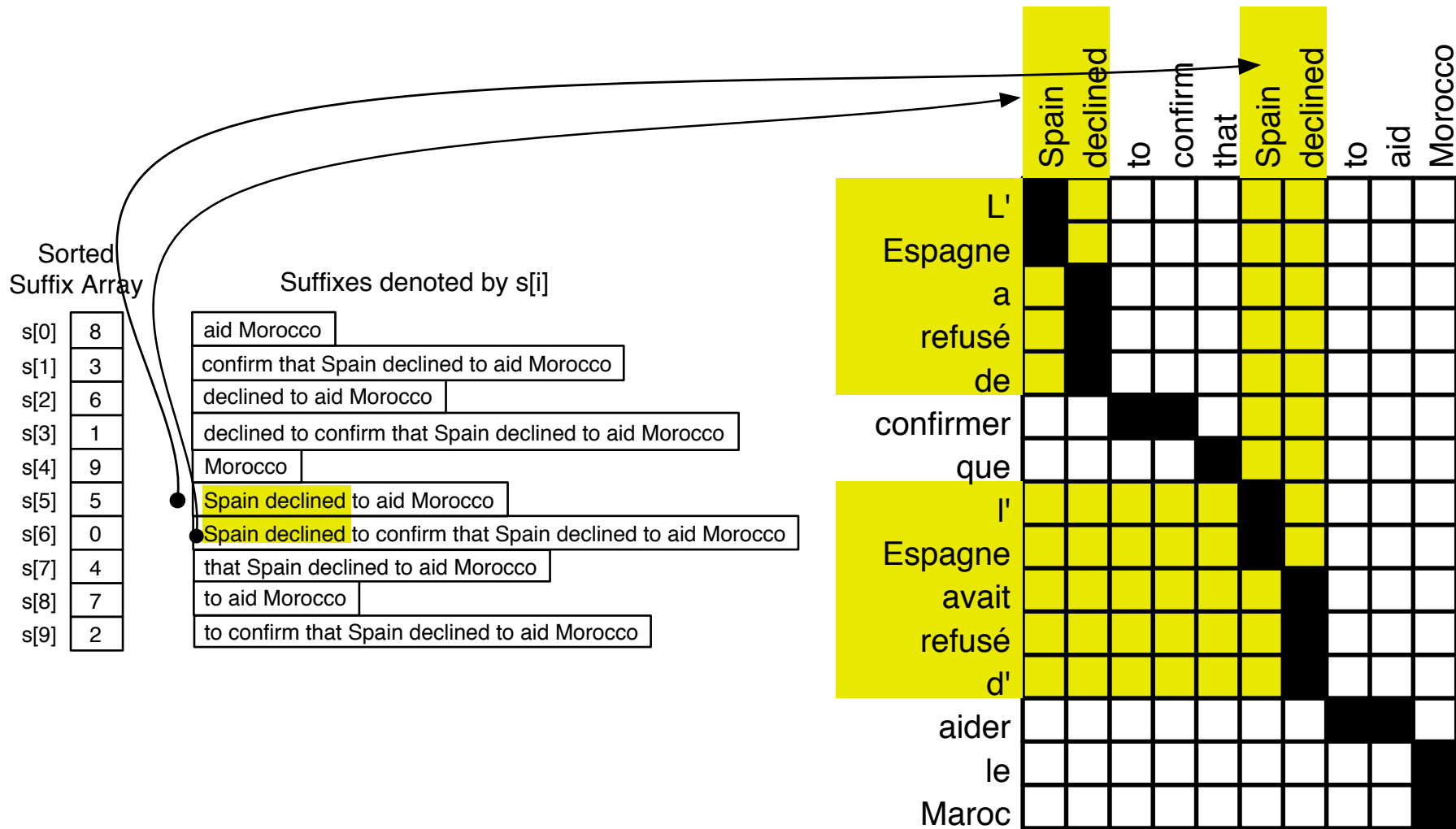




# On the fly lookup



# On the fly lookup



# Suffix Array Summary

- Very compact data structure
- Trade off between size / retrieval time
- Also useable in decoding
- See:
  - Callison-Burch et al (ACL 2005)
  - Zhang and Vogel (EAMT 2005)

# Evaluation

- Created a "gold standard" of correct phrase alignments
- Extracted translations for 120 phrases
- TM with 50,000 German-English sentences
- Evaluated precision/recall

# Results

- 78% of the translations were correct
- 81% of the translations were retrieved
- Top-ranked translation was correct 87% of the time

# Future directions

Linear B: Visualization Tool

<http://www.linearb.co.uk/nist2005/segment1.html>

اردوغان	يؤكد	بأن	تركيا	سترفض	اي	ضغوطات	لحثها	على	الاعتراف	بقبرص
		that turkey that turkey is turkey that turkey had that turkey will						to recognize recognition recognition of to recognize the the recognition		
	confirm that the confirm that affirm that confirms that who assert that our						urge it to to urge them to and urge it to order to urge it to urging			
ardogan ardogan ,	affirms <b>stresses</b> emphasize confirms emphasizes	that that the to , that that ,	turkey of turkey by turkey , turkey turkey ' s	reject will reject refuse not accept would be	any no of any i . e . meaning	pressure to pressure up to pressure	urge it urge and urge it urging urge them to	on to on the for to the	recognition recognition of recognize recognized recognition of the	cyprus cyprus , to cyprus on cyprus cyprus and

# Conclusion

- Presented Linear B's first foray into aides to the human translation process
- Statistical machine translation can augment translation memories
- Suffix arrays make it practical
- Allows us to better exploit the knowledge contained within TMs

# Thank you!

- Try our demo online:  
<http://linearb.co.uk/>