An Overview of Statistical Machine Translation

Charles Schafer

David Smith

Johns Hopkins University

Overview of the Overview

- The Translation Problem and Translation Data
  - “What do we have to work with?”
- Modeling
  - “What makes a good translation?”
- Search
  - “What’s the best translation?”
- Training
  - “Which features of data predict good translations?”
- Translation Dictionaries From Minimal Resources
  - “What if I don’t have (much) parallel text?”
- Practical Considerations

The Translation Problem and Translation Data

The Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশ্বে শান্তি, স্বাধীনতা এবং ন্যায়বিচারের ভিত্তি

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world

Why Machine Translation?

* Cheap, universal access to world's online information regardless of original language. (That's the goal)

Why Statistical (or at least Empirical) Machine Translation?

* We want to translate real-world documents. Thus, we should model real-world documents.

* A nice property: design the system once, and extend to new languages automatically by training on existing data.
  F(training data, model) -> parameterized MT system

Ideas that cut across empirical language processing problems and methods

Real-world: don't be (too) prescriptive. Be able to process (translate/summarize/identify/paraphrase) relevant bits of human language as they are, not as they "should be". For instance, genre is important: translating French blogs into English is different from translating French novels into English.

Model: a fully described procedure, generally having variable parameters, that performs some interesting task (for example, translation).

Training data: a set of observed data instances which can be used to find good parameters for a model via a training procedure.

Training procedure: a method that takes observed data and refines the parameters of a model, such that the model is improved according to some objective function.

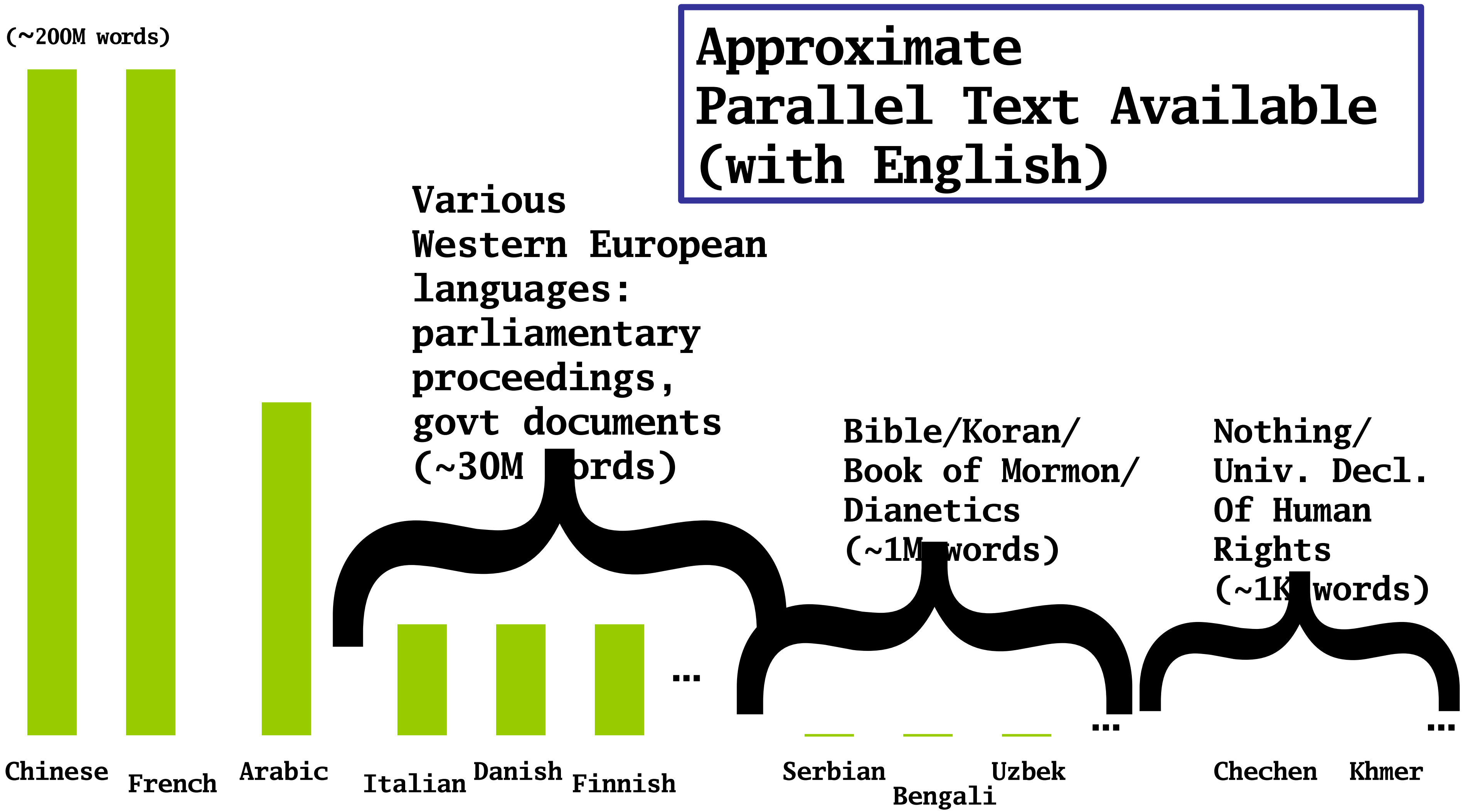
Resource Availability

Most of this tutorial

Most statistical machine translation (SMT) research has focused on a few "high-resource" languages(European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world's languages found on the web.

Most statistical machine translation research has focused on a few high-resource languages (European, Chinese, Japanese, Arabic).



Resource Availability

Most statistical machine translation (SMT) research has focused on a few "high-resource" languages(European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world's languages found on the web.

Romanian Catalan Serbian Slovenian Macedonian Uzbek Turkmen Kyrgyz
Uighur Pashto Tajikh Dari Kurdish Azeri Bengali Punjabi Gujarati
Nepali Urdu Marathi Konkani Oriya Telugu Malayalam Kannada Cebuano

We'll discuss this briefly

The Translation Problem

Document translation? Sentence translation? Word translation?

What to translate? The most common use case is probably document translation.

Most MT work focuses on sentence translation.

What does sentence translation ignore?
- Discourse properties/structure.
- Inter-sentence coreference.

Document Translation: Could Translation Exploit Discourse Structure?

<doc>

<sentence>

Documents usually dont begin with "Therefore"

William Shakespeare was an English poet and playwright widely regarded as the greatest writer of the English language, as well as one of the greatest in Western literature, and the world's pre-eminent dramatist.

<sentence>

He wrote about thirty-eight plays and 154 sonnets, as well as a variety of other poems.

<sentence>

What is the referent of "He"?

. . .

</doc>

Sentence Translation

- SMT has generally ignored extra-sentence structure (good future work direction for the community).

- Instead, we've concentrated on translating individual sentences as well as possible. This is a very hard problem in itself.

- Word translation (knowing the possible English translations of a French word) is not, by itself, sufficient for building readable/useful automatic document translations – though it is an important component in end-to-end SMT systems.

Sentence translation using only a word translation dictionary is called "glossing" or "gisting".

Word Translation (learning from minimal resources)

We'll come back to this later…

and address learning the word translation component (dictionary) of MT systems without using parallel text.

(For languages having little parallel text, this is the best we can do right now)

Sentence Translation

- Training resource: parallel text (bitext).
- Parallel text (with English) on the order of 20M-200M words (roughly, 1M-10M sentences) is available for a number of languages.
- Parallel text is expensive to generate: human translators are expensive ($0.05-$0.25 per word). Millions of words training data needed for high quality SMT results. So we take what is available. This is often of less than optimal genre (laws, parliamentary proceedings, religious texts).

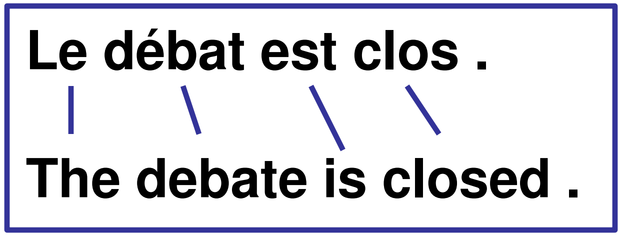
Sentence Translation: examples of more and less literal translations in bitext

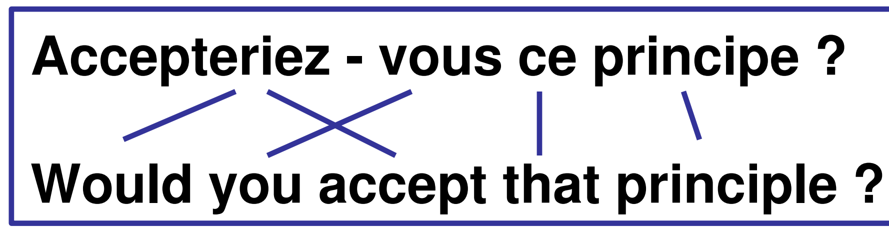
French, English from Bitext	Closely Literal English Translation
Le débat est clos . **The debate is closed .**	*The debate is closed.*
Accepteriez - vous ce principe ? **Would you accept that principle ?**	*Accept-you that principle?*
Merci , chère collègue . **Thank you , Mrs Marinucci .**	*Thank you, dear colleague.*
Avez - vous donc une autre proposition ? **Can you explain ?**	*Have you therefore another proposal?*

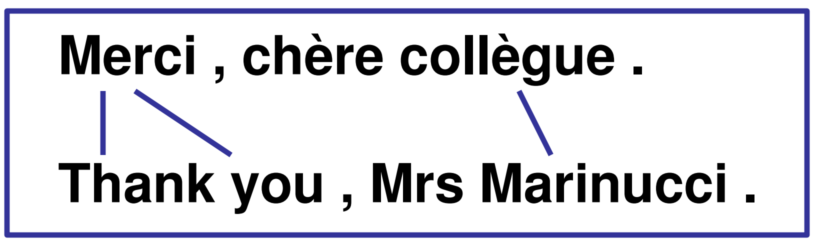
(from French-English European Parliament proceedings)

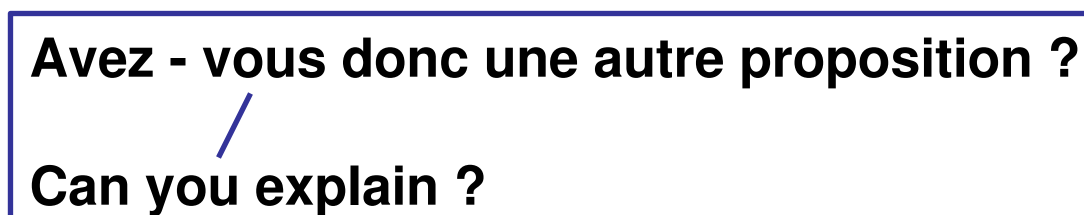
Sentence Translation: examples of more and less literal translations in bitext

Word alignments illustrated. Well-defined for more literal translations.









Translation and Alignment

- As mentioned, translations are expensive to commission and generally SMT research relies on already existing translations

- These typically come in the form of aligned documents.

- A sentence alignment, using pre-existing document boundaries, is performed automatically. Low-scoring or non-one-to-one sentence alignments are discarded. The resulting aligned sentences constitute the training bitext.

- For many modern SMT systems, induction of word alignments between aligned sentences, using algorithms based on the IBM word-based translation models, is one of the first stages of processing. Such induced word alignments are generally treated as part of the observed data and are used to extract aligned phrases or subtrees.

Target Language Models

The translation problem can be described as modeling the probability distribution P(E|F), where F is a string in the source language and E is a string in the target language.

Using Bayes' Rule, this can be rewritten

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

$$= P(F|E)P(E)$$ [since F is observed as the sentence to be translated, P(F)=1]

P(F|E) is called the "translation model" (TM).
P(E) is called the "language model" (LM).
The LM should assign probability to sentences which are "good English".

Target Language Models

- Typically, N-Gram language models are employed

- These are finite state models which predict the next word of a sentence given the previous several words. The most common N-Gram model is the trigram, wherein the next word is predicted based on the previous 2 words.

- The job of the LM is to take the possible next words that are proposed by the TM, and assign a probability reflecting whether or not such words constitute "good English".

---

p(the|went to)   p(the|took the)

p(happy|was feeling)   p(sagacious|was feeling)

p(time|at the)   p(time|on the)

Translating Words in a Sentence

- Models will automatically learn entries in probabilistic translation dictionaries, for instance p(elle|she), from co-occurrences in aligned sentences of a parallel text.

- For some kinds of words/phrases, this is less effective. For example:
  numbers
  dates
  named entities (NE)

The reason: these constitute a large open class of words that will not all occur even in the largest bitext. Plus, there are regularities in translation of numbers/dates/NE.

Handling Named Entities

- For many language pairs, and particularly those which do not share an alphabet, transliteration of person and place names is the desired method of translation.

- General Method:
  1. Identify NE's via classifier
  2. Transliterate name
  3. Translate/reorder honorifics

- Also useful for alignment. Consider the case of Inuktitut-English alignment, where Inuktitut renderings of European names are highly nondeterministic.

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Train a **probabilistic finite-state transducer** to model this ambiguous transformation

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Williams	McLean
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

… Mr. Williams … … mista uialims …

Useful Types of Word Analysis

- Number/Date Handling
- Named Entity Tagging/Transliteration
- Morphological Analysis
  - Analyze a word to its root form (at least for word alignment)

    was -> is   believing -> believe
    ruminerai -> ruminer   ruminiez -> ruminer

  - As a dimensionality reduction technique
  - To allow lookup in existing dictionary

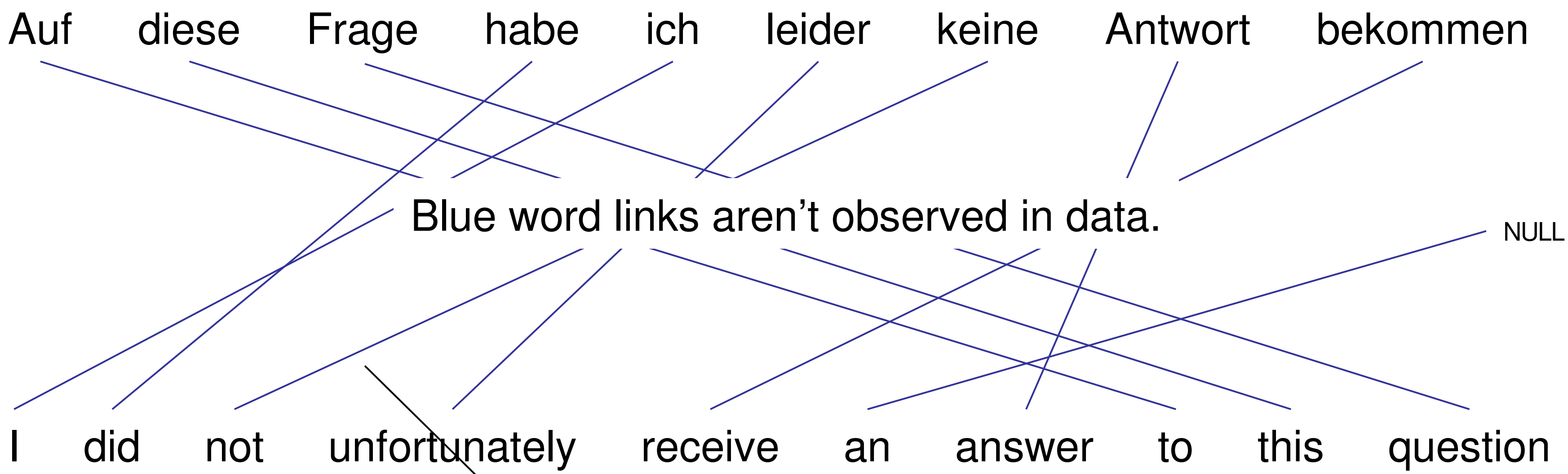
Modeling

What makes a good translation?

Modeling

- Translation models
  - "Adequacy"
  - Assign better scores to accurate (and complete) translations
- Language models
  - "Fluency"
  - Assign better scores to natural target language text

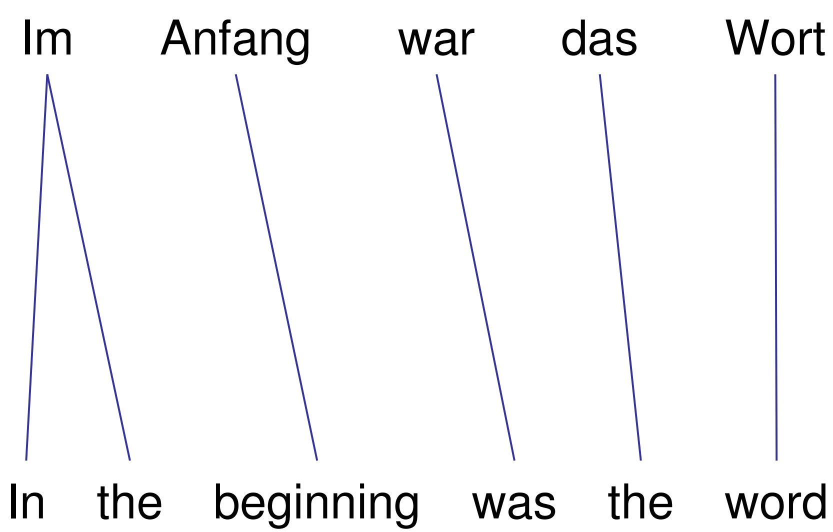
Word Translation Models



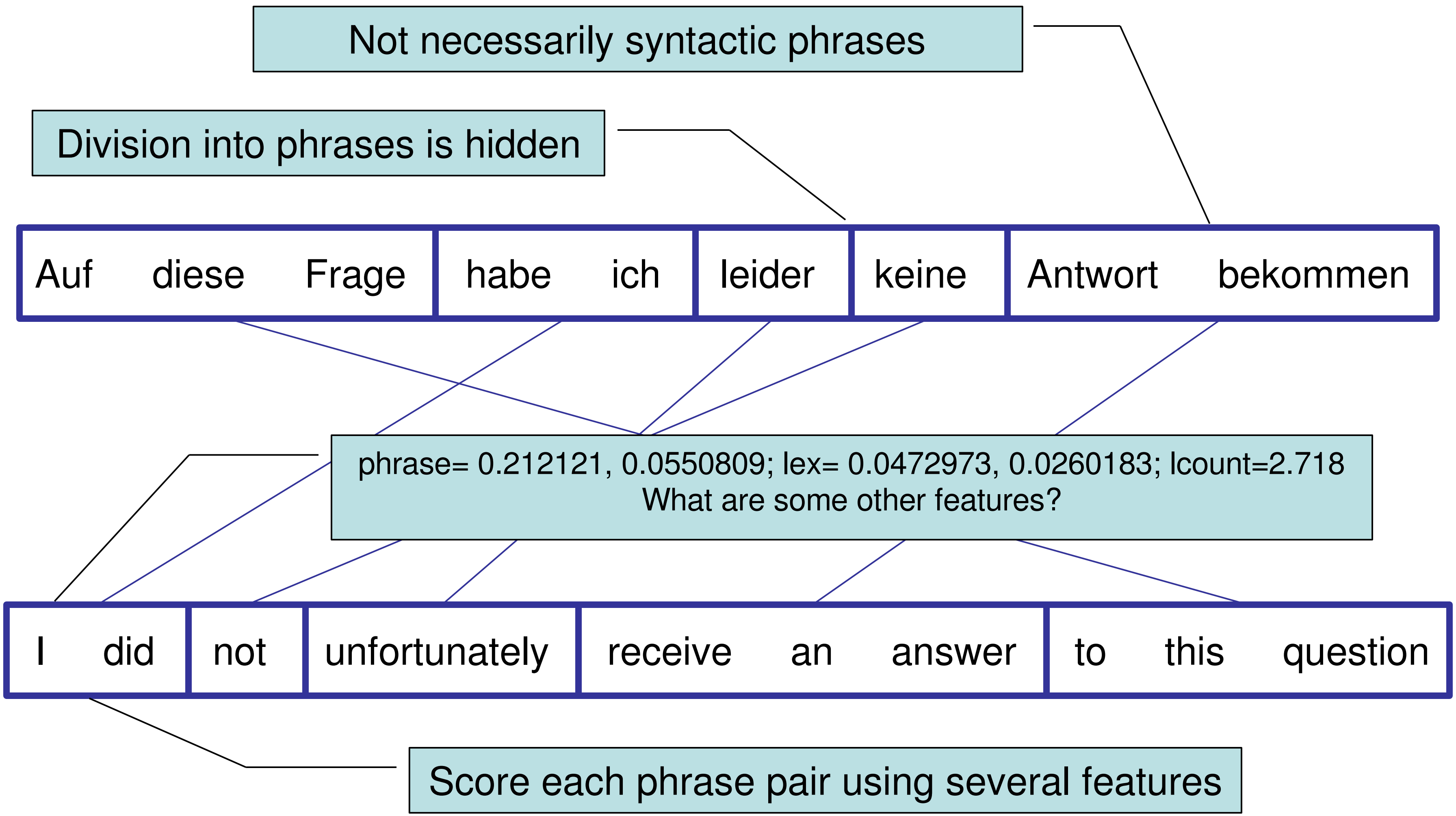
Features for word-word links: lexica, part-of-speech, orthography, etc.

Word Translation Models

- Usually directed: each word in the target generated by one word in the source
- Many-many and null-many links allowed
- Classic IBM models of Brown et al.
- Used now mostly for word alignment, not translation



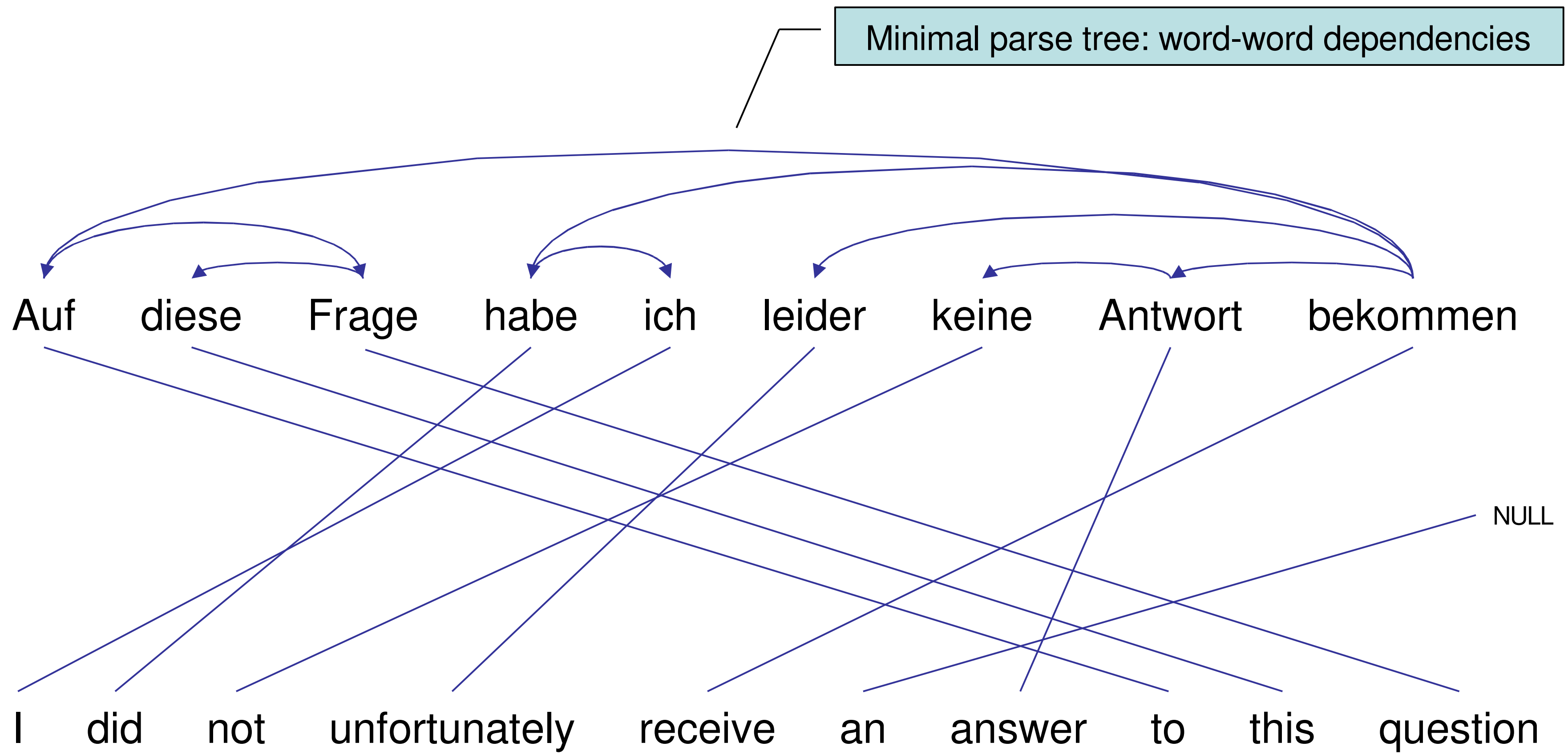
Phrase Translation Models



Phrase Translation Models

- Capture translations **in context**
  - *en Amerique*: **to** America
  - *en anglais*: **in** English
- State-of-the-art for several years
- Each source/target phrase pair is scored by several weighted features.
- The weighted sum of model features is the whole translation's score: $\theta \bullet \mathbf{f}$
- Phrases don't overlap (cf. language models) but have "reordering" features.

Single-Tree Translation Models

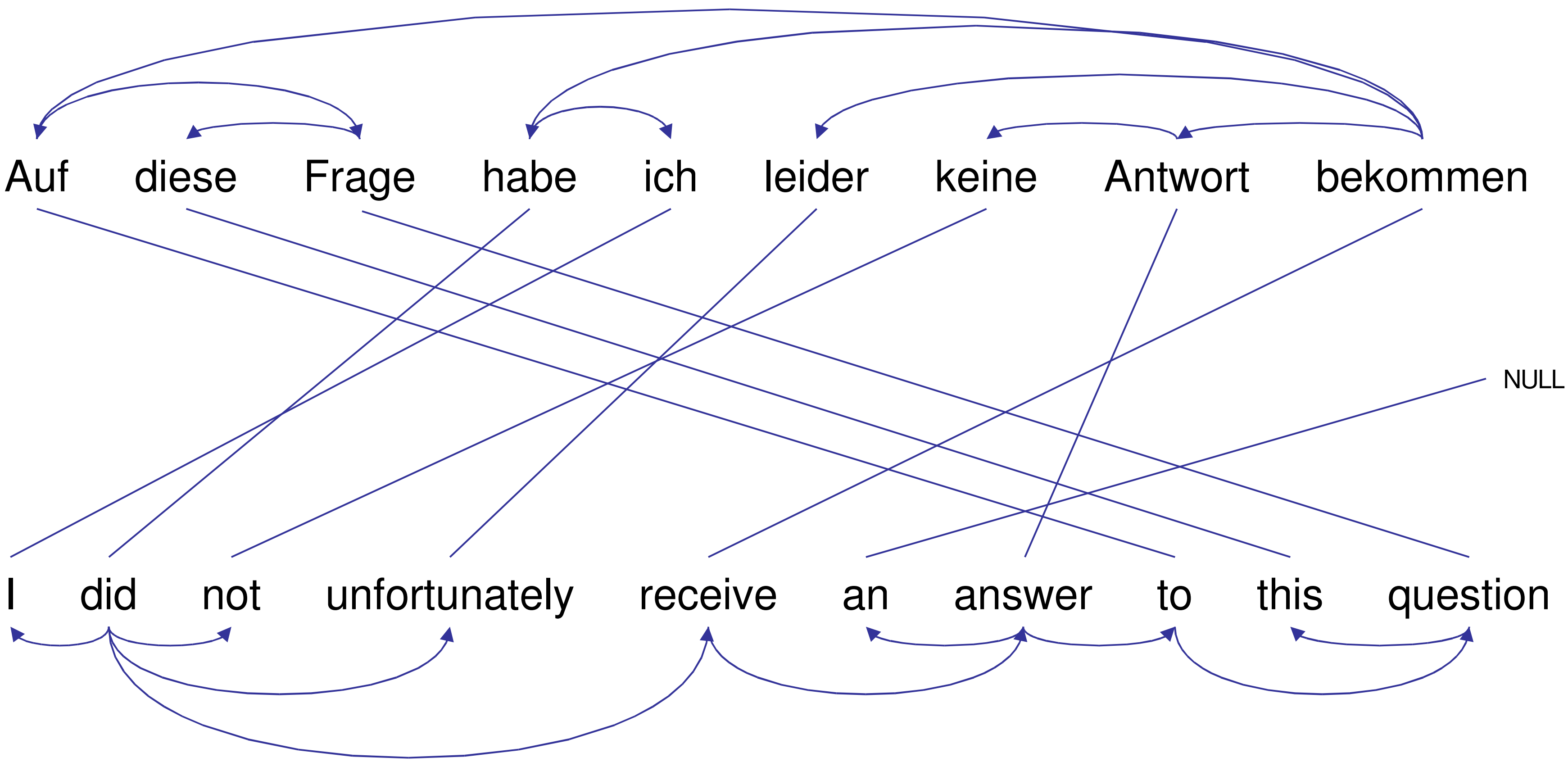


Parse trees with deeper structure have also been used.

Single-Tree Translation Models

- Either source or target has a hidden tree/parse structure
  - Also known as “tree-to-string” or “tree-transducer” models
- The side with the tree generates words/phrases in tree, not string, order.
- Nodes in the tree also generate words/phrases on the other side.
- English side is often parsed, whether it’s source or target, since English parsing is more advanced.

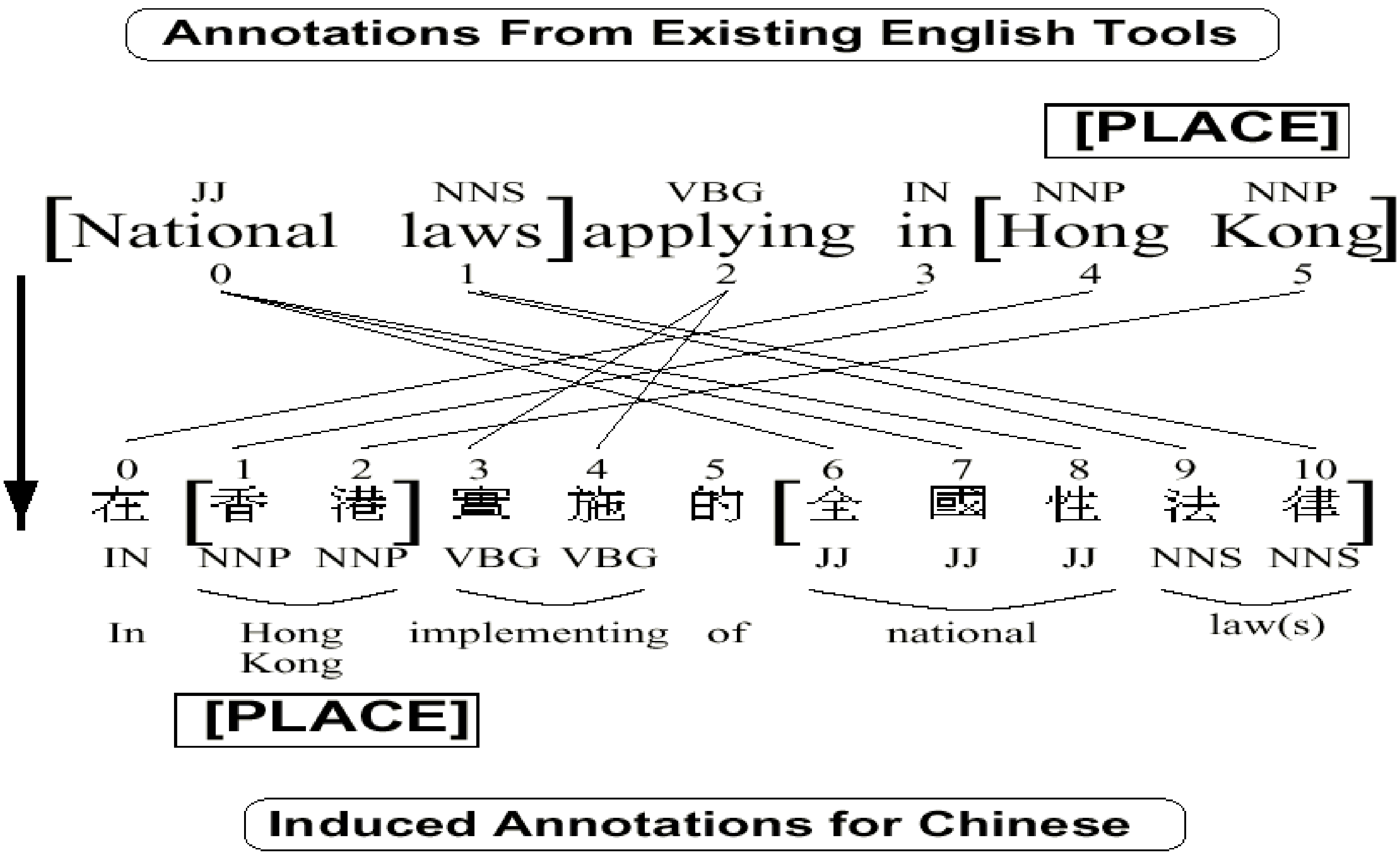
Tree-Tree Translation Models



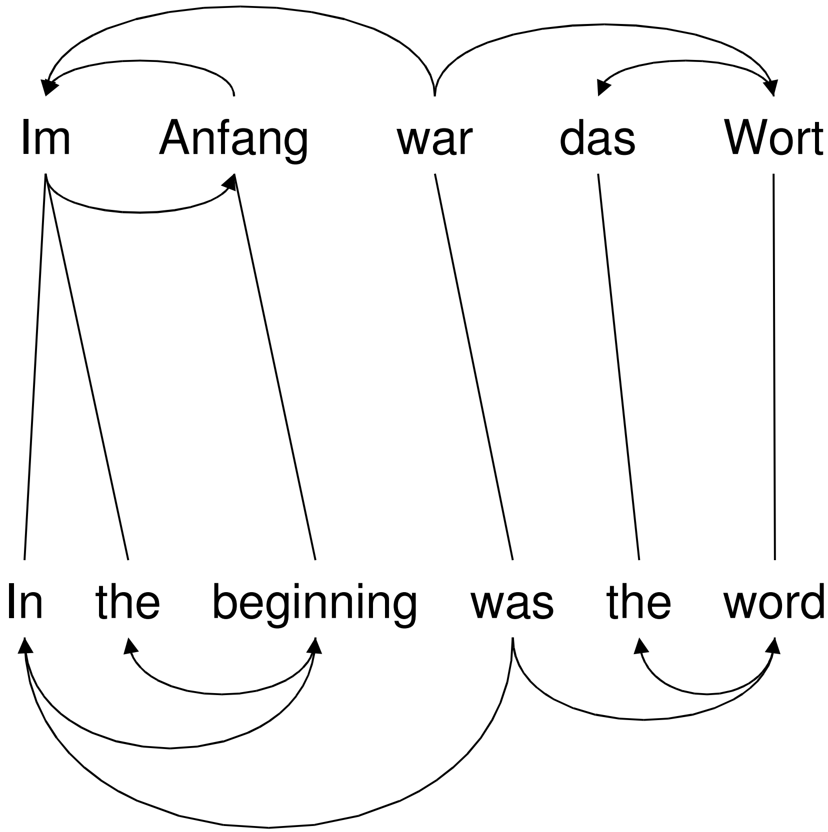
Tree-Tree Translation Models

- Both sides have hidden tree structure
  - Can be represented with a “synchronous” grammar
- Some models assume isomorphic trees, where parent-child relations are preserved; others do not.
- Trees can be fixed in advance by monolingual parsers or induced from data (e.g. Hiero).
- Cheap trees: project from one side to the other

Projecting Hidden Structure

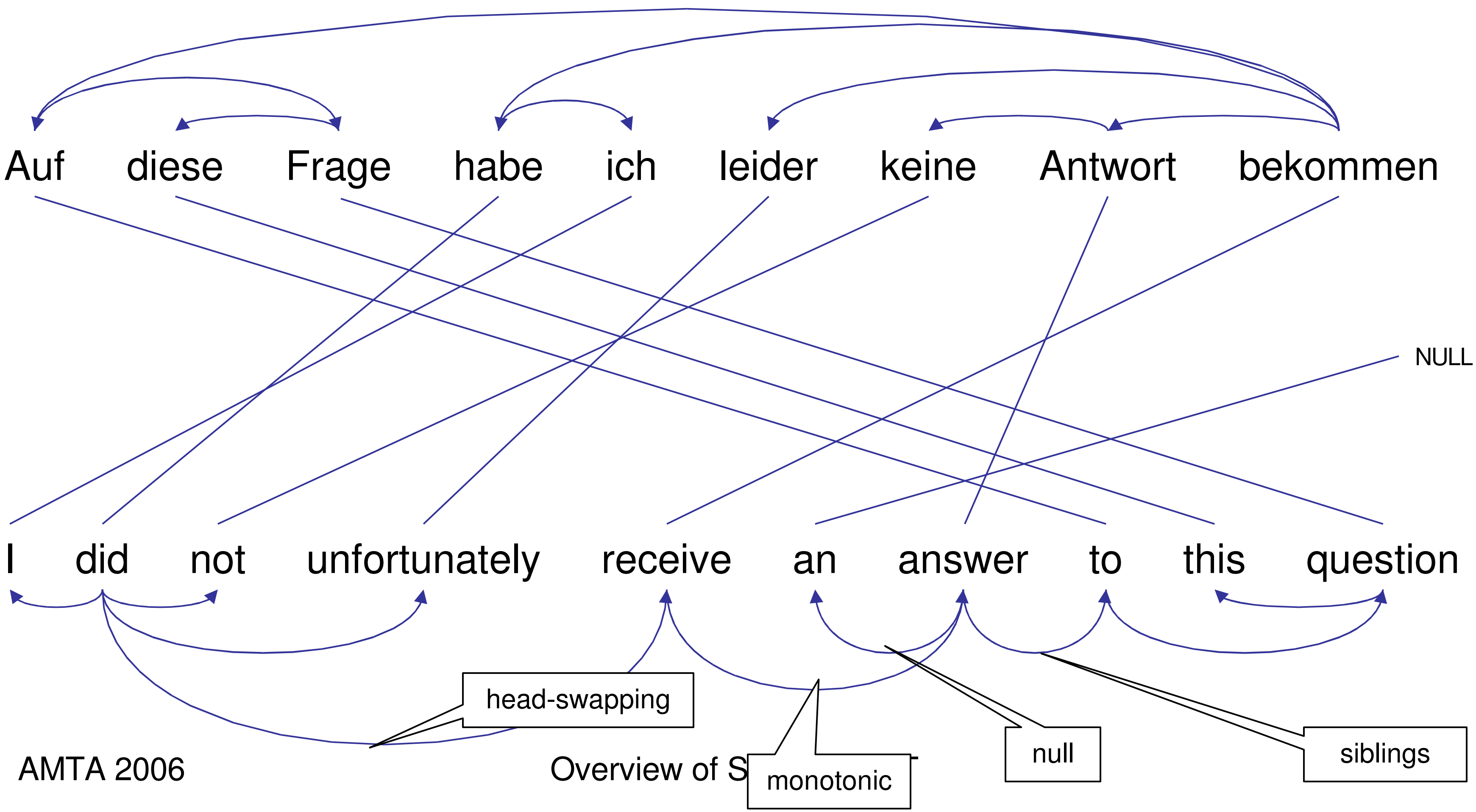


Projection

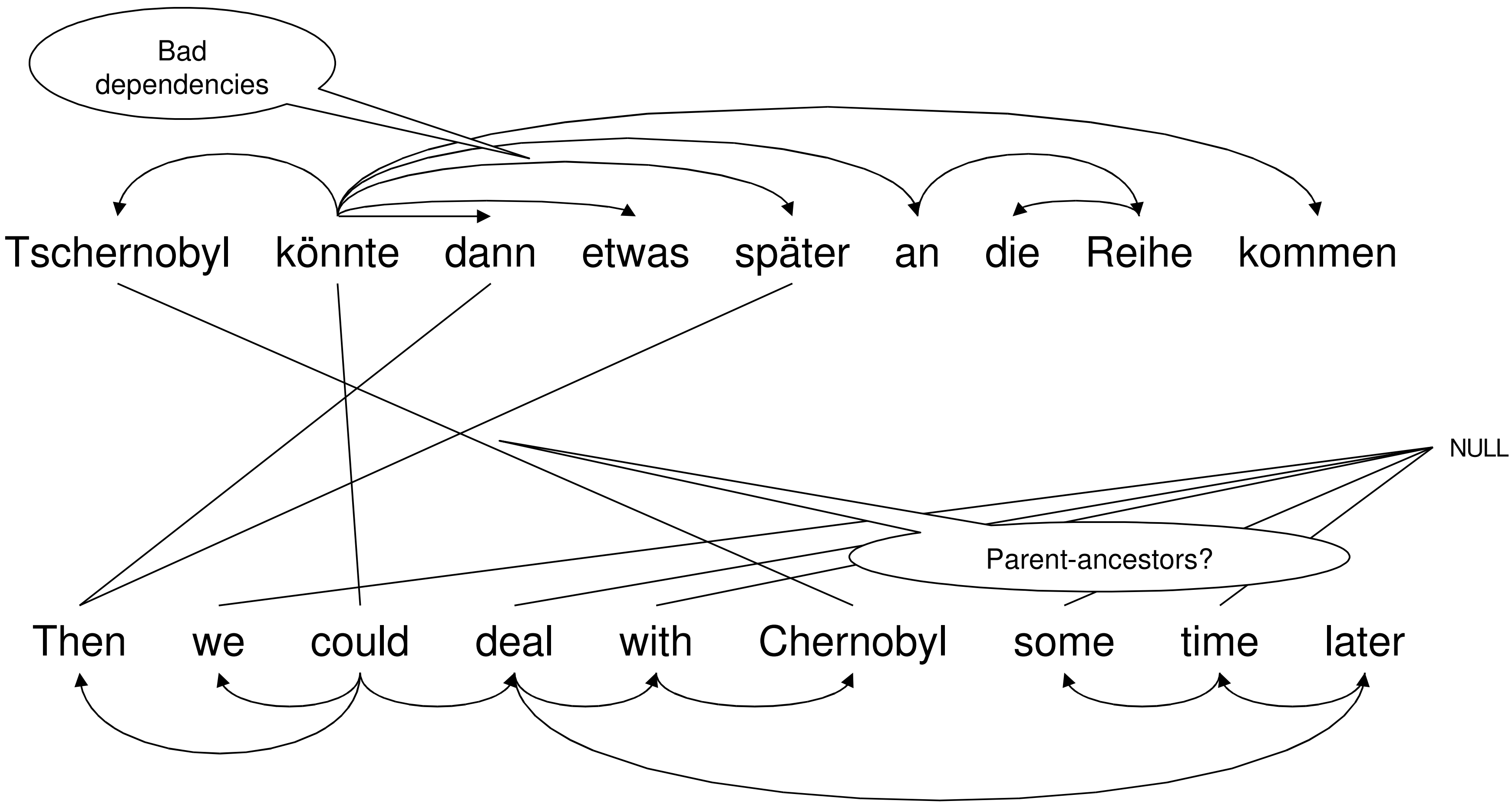


- Train with bitext
- Parse one side
- Align words
- Project dependencies
- Many to one links?
- Non-projective and circular dependencies?

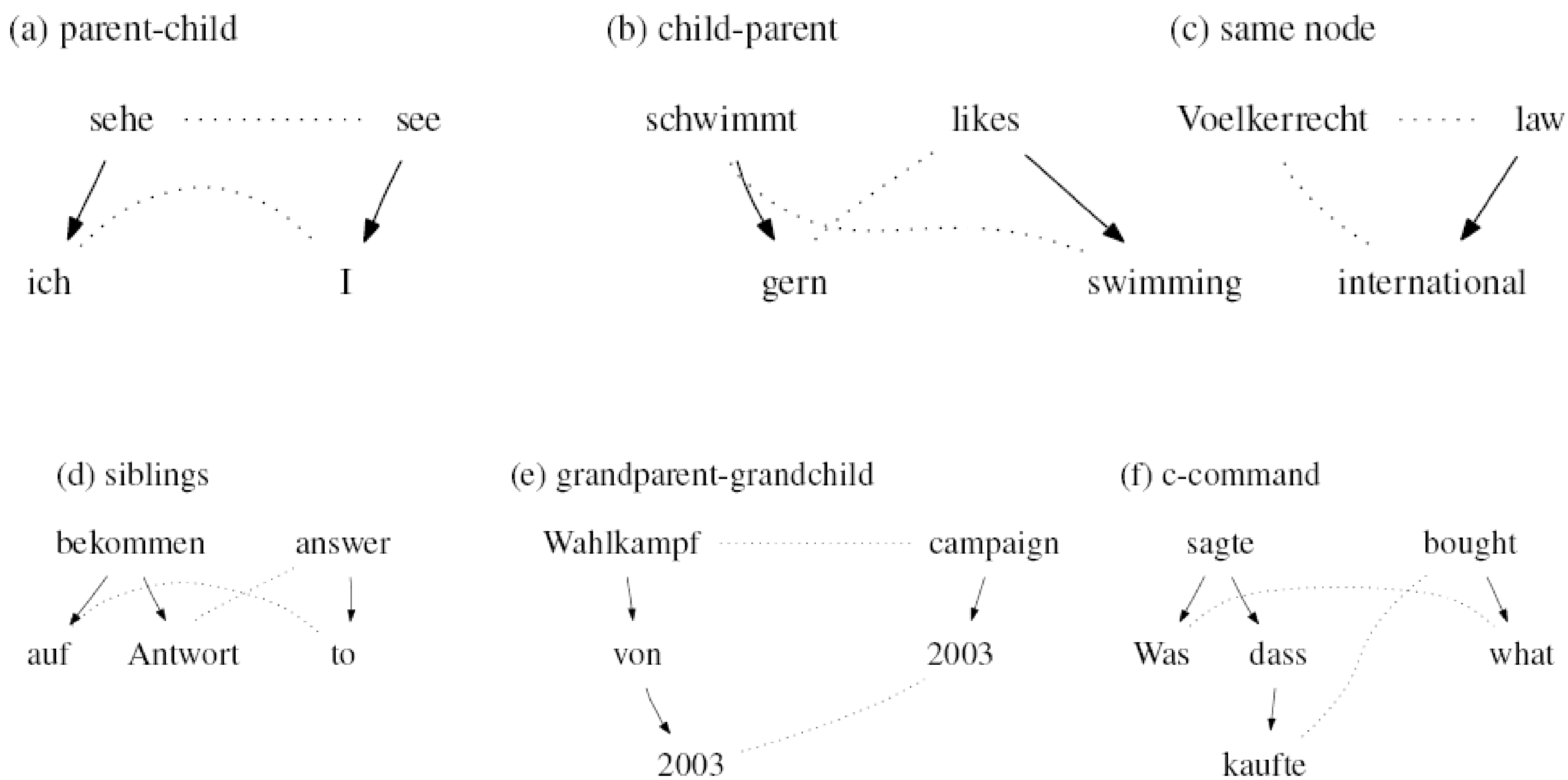
Divergent Projection



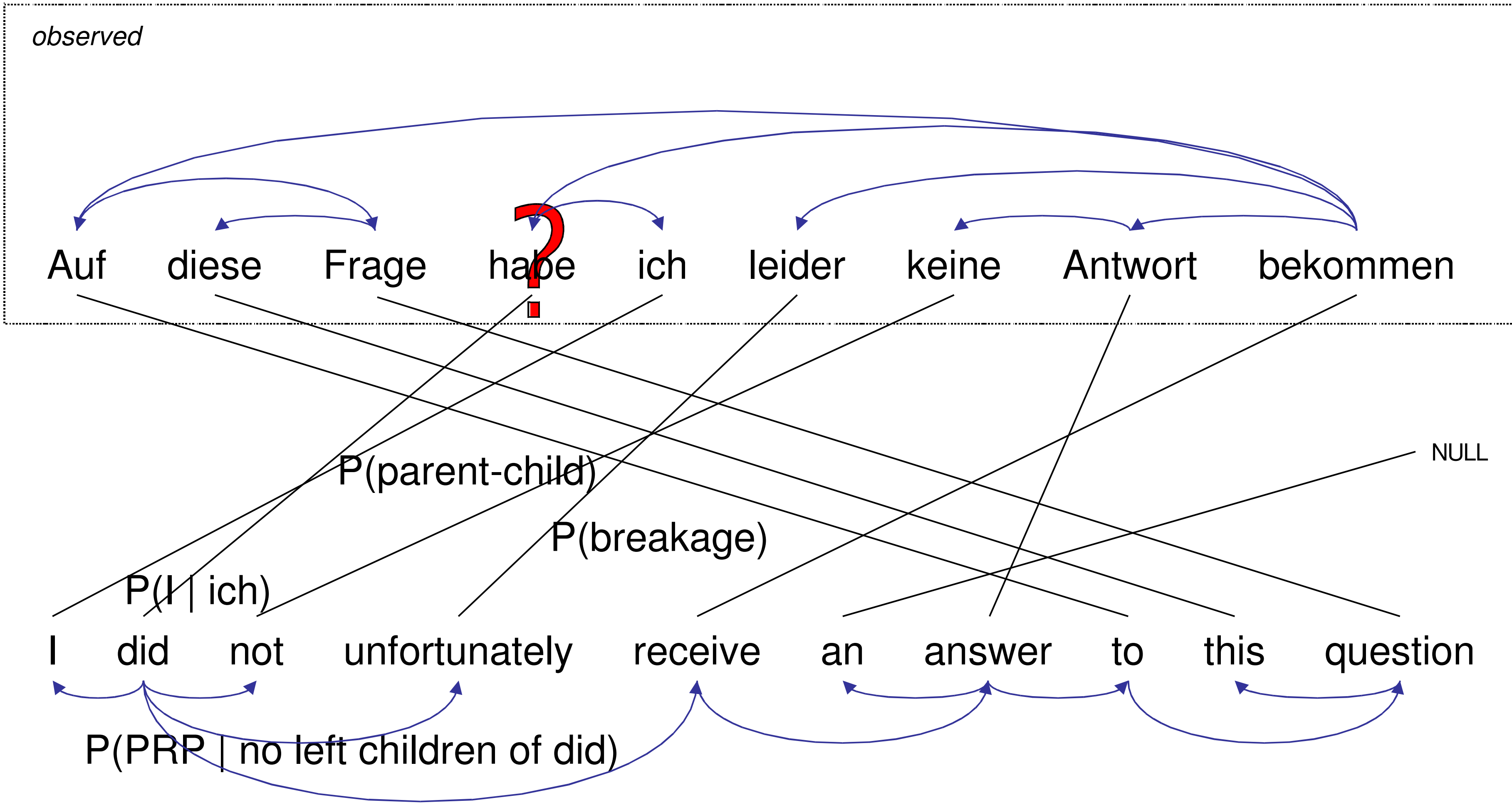
Free Translation



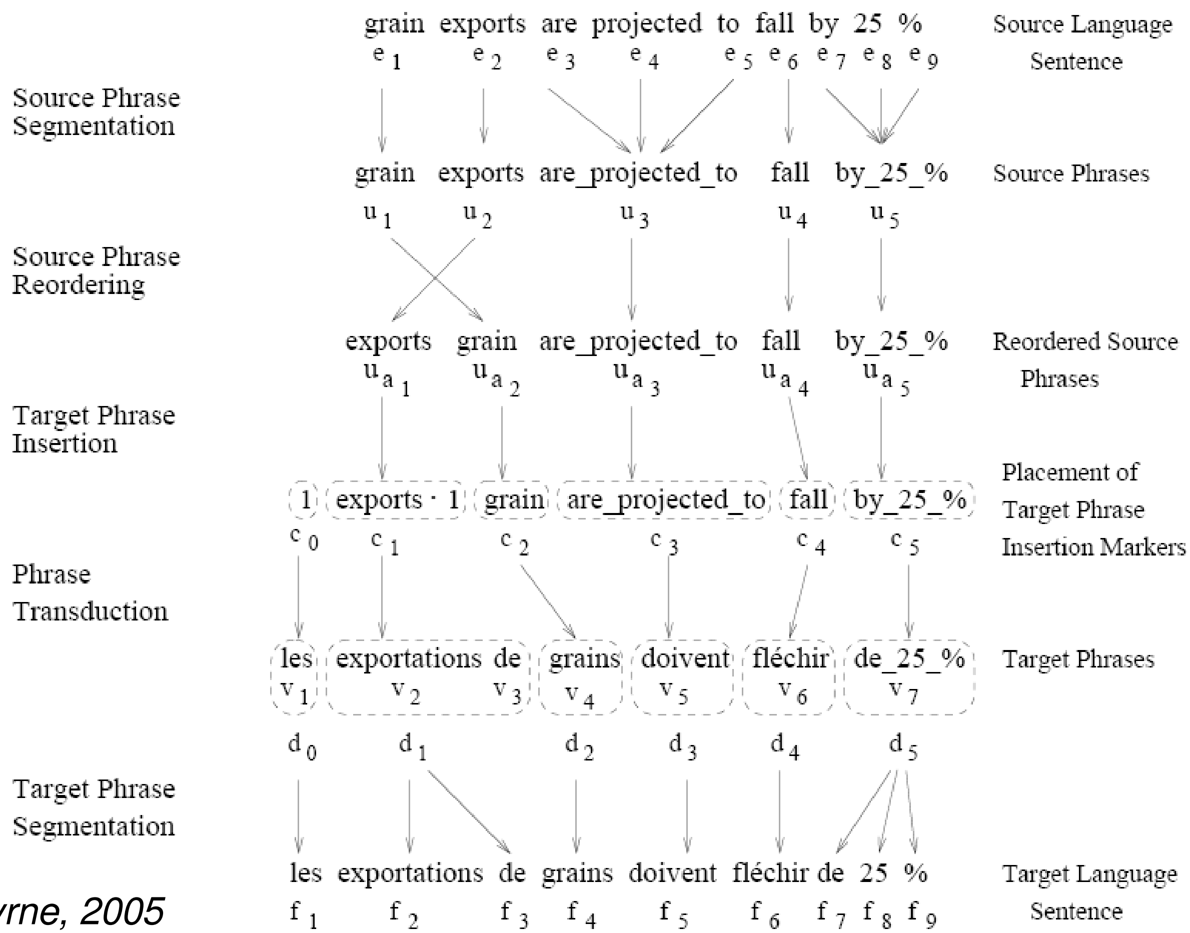
Dependency Menagerie



A Tree-Tree Generative Story



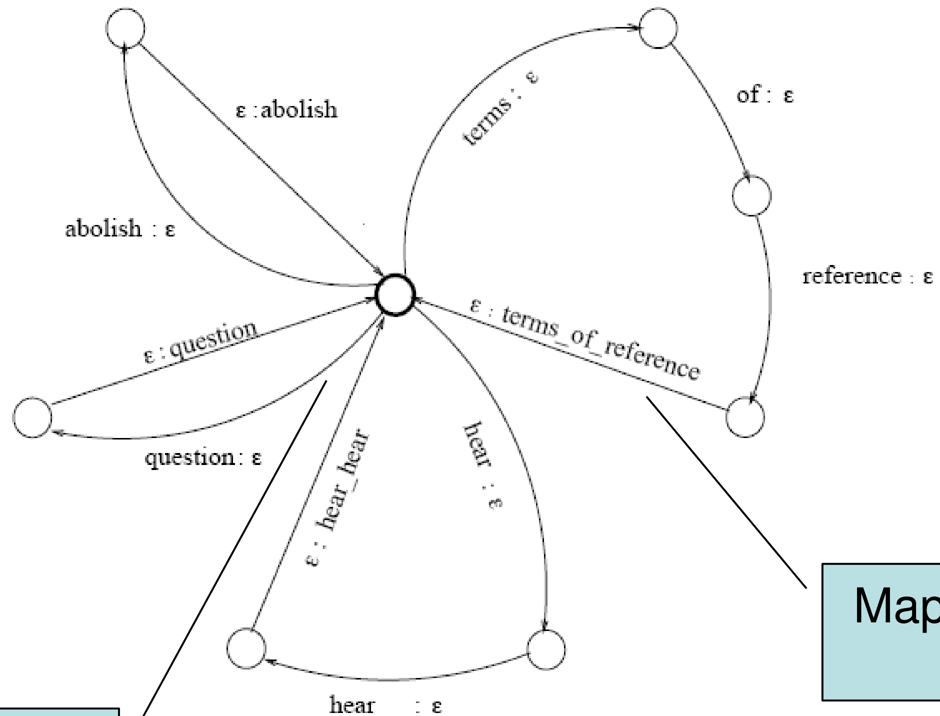
Finite State Models



Kumar, Deng & Byrne, 2005

Finite State Models

First transducer in the pipeline



Here a unigram model of phrases

Map distinct words to phrases

Kumar, Deng & Byrne, 2005

Finite State Models

- Natural composition with other finite state processes, e.g. Chinese word segmentation
- Standard algorithms and widely available tools (e.g. AT&T fsm toolkit)
- Limit reordering to finite offset
- Often impractical to compose all finite state machines offline

Search

What's the best translation
(under our model)?

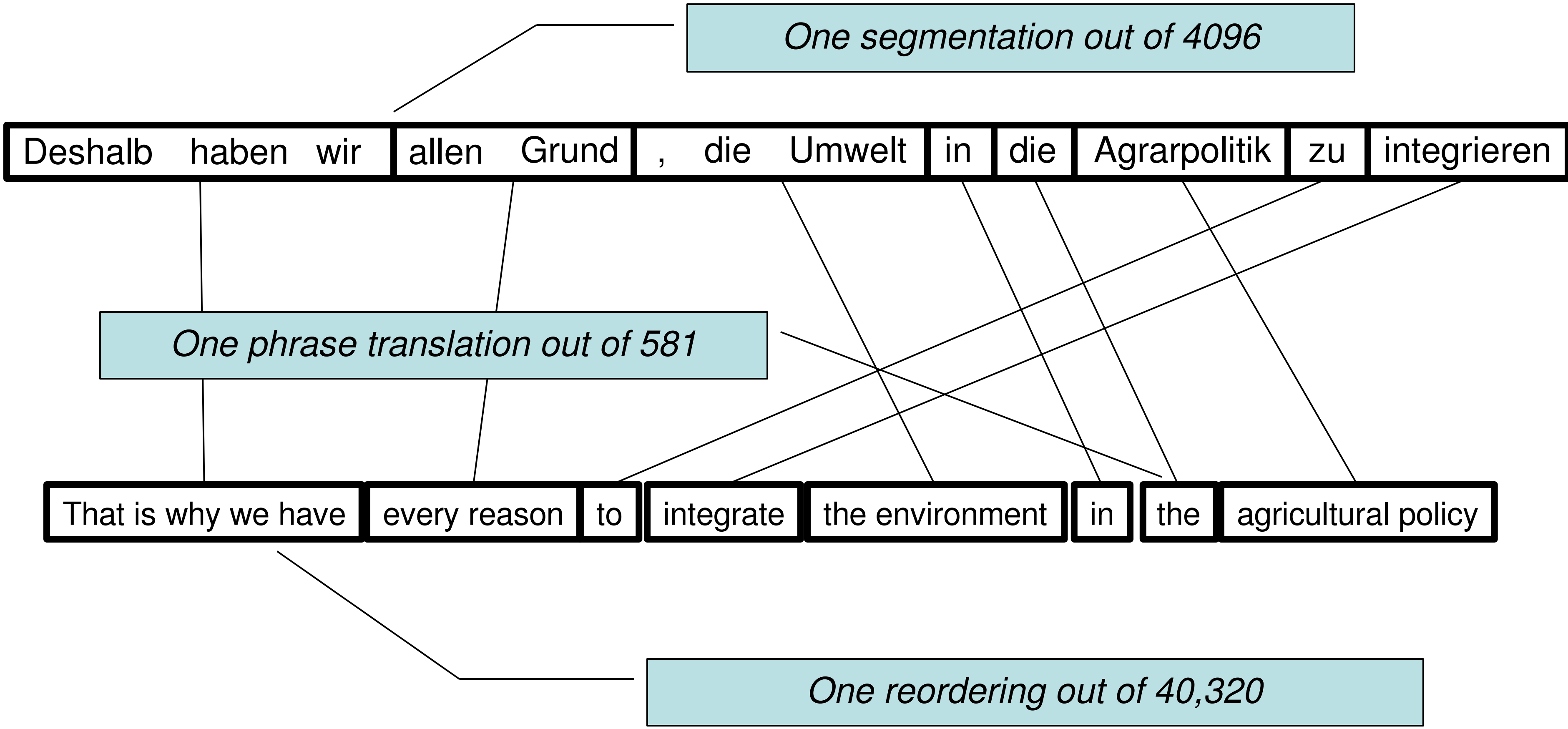
Search

- Even if we know the right words in a translation, there are *n*! permutations.

$10! = 3{,}626{,}800$ $\quad 20! \approx 2.43 \times 10^{18}$ $\quad 30! \approx 2.65 \times 10^{32}$

- We want the translation that gets the highest score under our model
  - Or the best *k* translations
  - Or a random sample from the model's distribution
- But **not** in *n*! time!

Search in Phrase Models



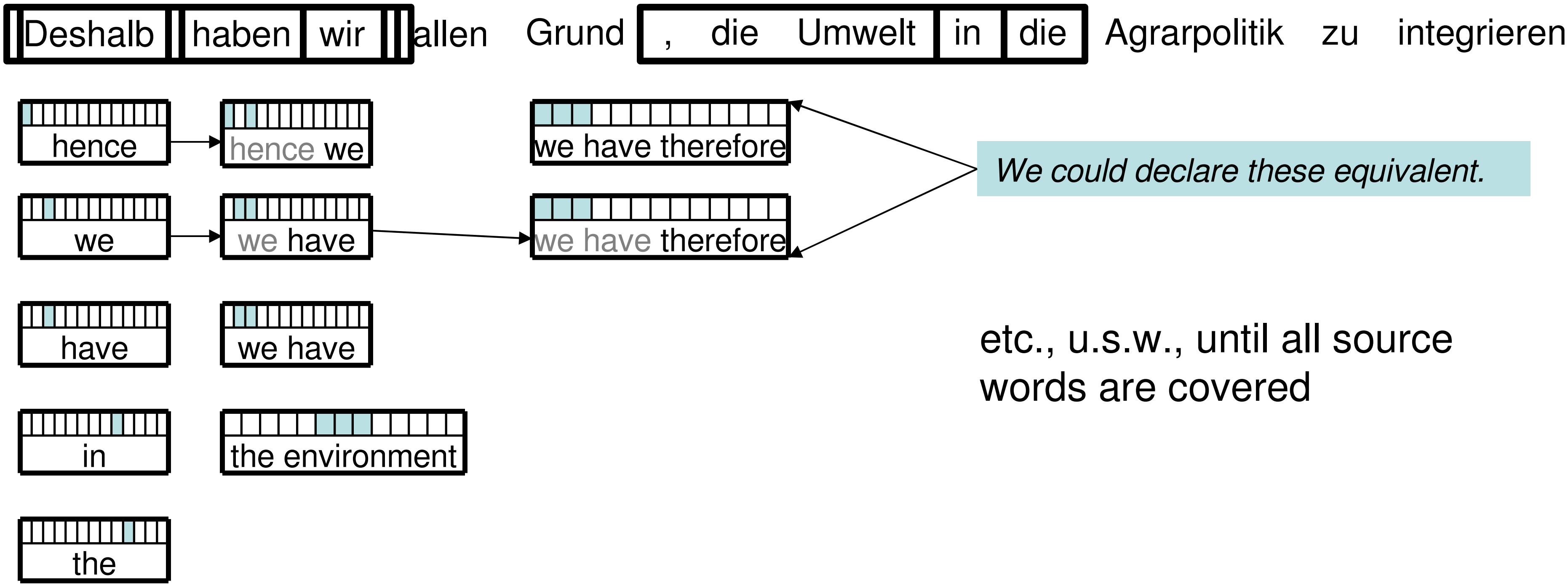
Translate in target language order to ease language modeling.

Search in Phrase Models

Deshalb	haben	wir	allen	Grund	,	die	Umwelt	in	die	Agrarpolitik	zu	integrieren
that is why we have			every reason		the environment			in	the	agricultural policy	to	integrate
therefore	have	we	every reason			the	environment	in the		agricultural policy ,	to integrate	
that is why	we have		all	reason	,	which	environment in		agricultural policy		parliament	
have therefore		us	all the	reason		of the	environment into		the agricultural policy		successfully integrated	
hence		, we	every	reason to make			environmental	on	the cap		be woven together	
we have therefore			everyone	grounds for taking the			the environment	to the		agricultural policy is	on	parliament
so	, we		all of	cause	which		environment ,	to	the cap ,		for	incorporated
hence our			any	why	that		outside	at	agricultural policy		too	woven together
therefore ,		it	of all	reason for	, the		completion	into	that agricultural policy		be	

And many, many more…even before reordering

“Stack Decoding”



etc., u.s.w., until all source words are covered

Search in Phrase Models

- Many ways of segmenting source
- Many ways of translating each segment
- *Restrict* phrases > e.g. 7 words, long-distance reordering
- *Prune* away unpromising partial translations or we'll run out of space and/or run too long
  - How to compare partial translations?
  - Some start with easy stuff: "in", "das", ...
  - Some with hard stuff: "Agrarpolitik", "Entscheidungsproblem", …

What Makes Search Hard?

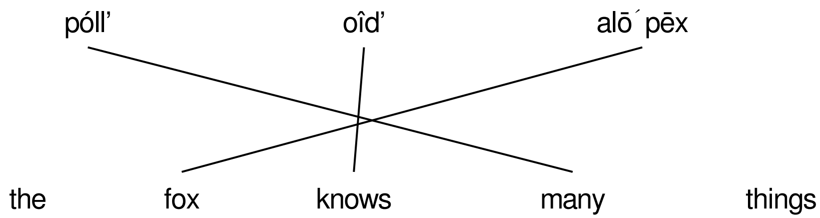
- What we really want: the best (highest-scoring) translation
- What we get: the best translation/phrase segmentation/alignment
  - Even summing over all ways of segmenting *one* translation is hard.
- Most common approaches:
  - Ignore problem
  - Sum over top *j* translation/segmentation/alignment triples to get top *k*<<j translations

Redundancy in *n*-best Lists

Source: Da ich wenig Zeit habe , gehe ich sofort in medias res .

as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in medias res immediately . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am in medias res immediately . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in medias res immediately . | 0-1,0-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am in medias res immediately . | 0-0,0-0 1-1,1-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,10-10 10-10,11-11 11-11,8-8 12-12,12-12
as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i would immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
because i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am immediately in medias res . | 0-0,0-0 1-1,1-1 2-2,4-4 3-3,2-2 4-4,3-3 5-5,5-5 6-6,7-7 7-7,6-6 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in res medias immediately . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,11-11 10-10,10-10 11-11,8-8 12-12,12-12
because i have little time , i am immediately in medias res . | 0-1,0-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,8-8 9-9,9-9 10-10,10-10 11-11,11-11 12-12,12-12
as i have little time , i am in res medias immediately . | 0-0,0-0 1-1,1-1 2-2,4-4 3-4,2-3 5-5,5-5 6-7,6-7 8-8,9-9 9-9,11-11 10-10,10-10 11-11,8-8 12-12,12-12

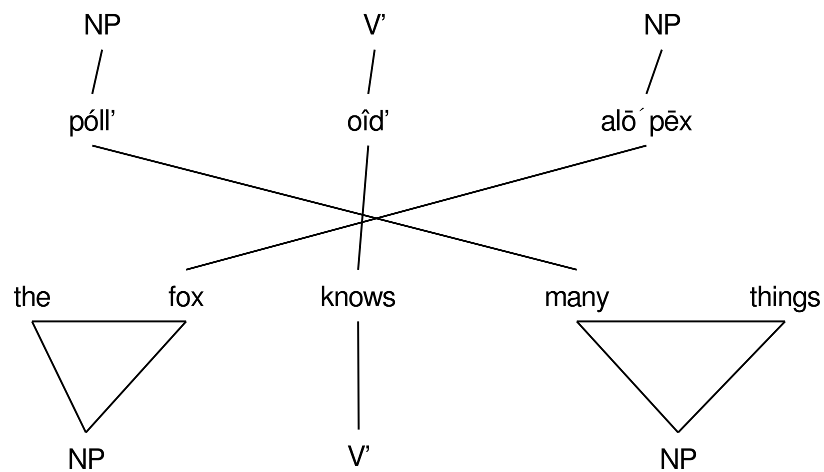
Bilingual Parsing



A variant of CKY chart parsing.

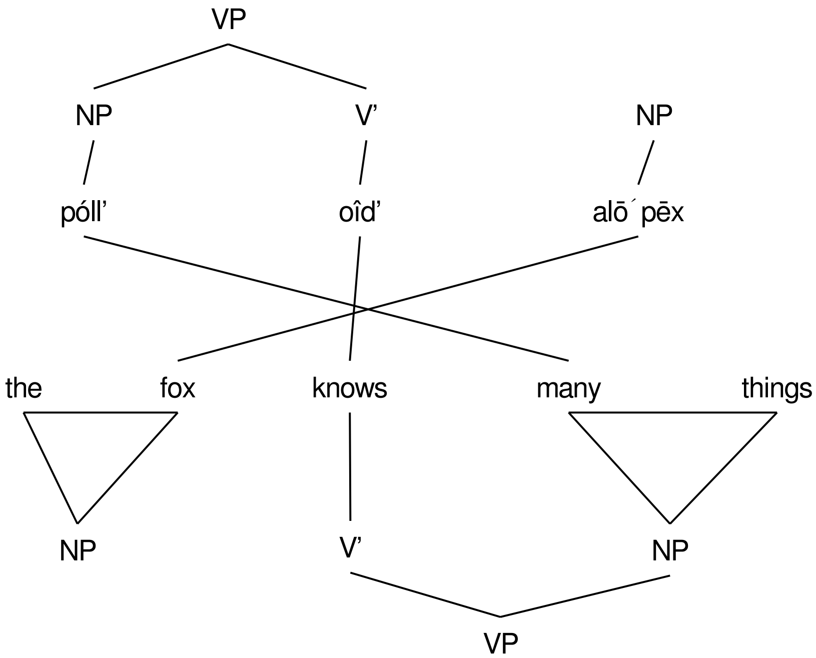
	póll’	oîd’	alō´pē x
the			
fox			NN/NN
knows		VB/VB	
many	JJ/JJ		
things			

Bilingual Parsing



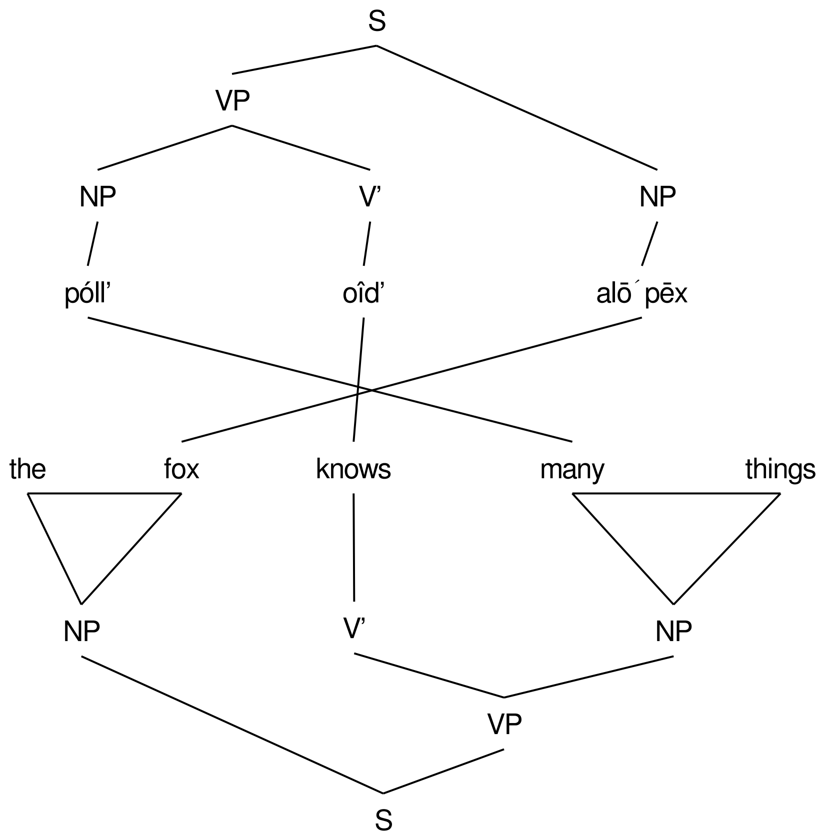
	póll'	oîd'	alō´pē x
the			NP/NP
fox			
knows		VP/VP	
many	NP/NP		
things			

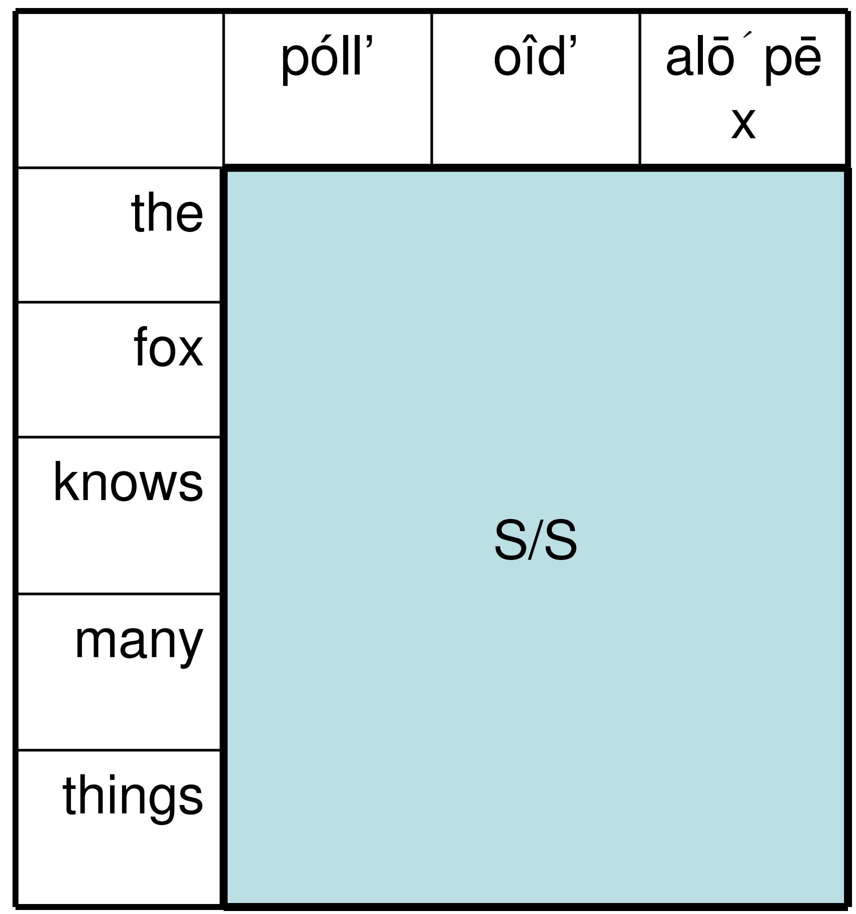
Bilingual Parsing



<table>
<tr><th></th><th>póll'</th><th>oîd'</th><th>alō´pē x</th></tr>
<tr><td>the</td><td></td><td></td><td rowspan="2">NP/NP</td></tr>
<tr><td>fox</td><td></td><td></td></tr>
<tr><td>knows</td><td rowspan="3" colspan="2">VP/VP</td><td></td></tr>
<tr><td>many</td><td></td></tr>
<tr><td>things</td><td></td></tr>
</table>

Bilingual Parsing





MT as Parsing

- If we only have the source, parse it while recording all compatible target language trees.
- Runtime is also multiplied by a *grammar constant*: one string could be a noun and a verb phrase
- Continuing problem of multiple hidden configurations (trees, instead of phrases) for one translation.

Training

Which features of data predict good translations?

Training: Generative/Discriminative

- Generative
  - Maximum likelihood training: max p(data)
  - "Count and normalize"
  - Maximum likelihood with hidden structure
    - Expectation Maximization (EM)
- Discriminative training
  - Maximum conditional likelihood
  - Minimum error/risk training
  - Other criteria: perceptron and max. margin

"Count and Normalize"

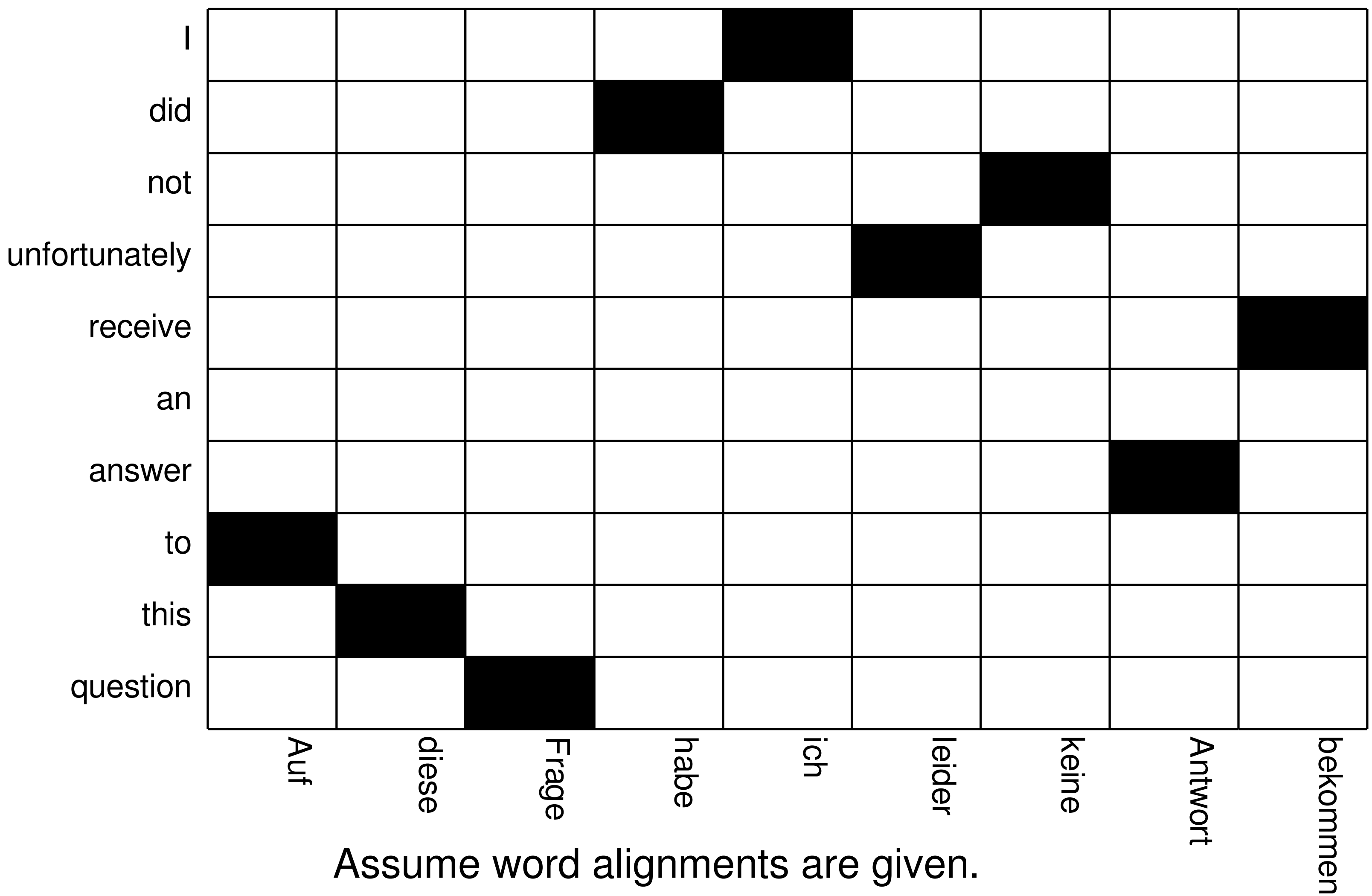
- Language modeling example: assume the probability of a word depends only on the previous 2 words.

$$p(\text{disease} \mid \text{into the}) = \frac{p(\text{into the disease})}{p(\text{into the})}$$

- p(disease|into the) = 3/20 = 0.15
- "Smoothing" reflects a prior belief that p(breech|into the) > 0 despite these 20 examples.

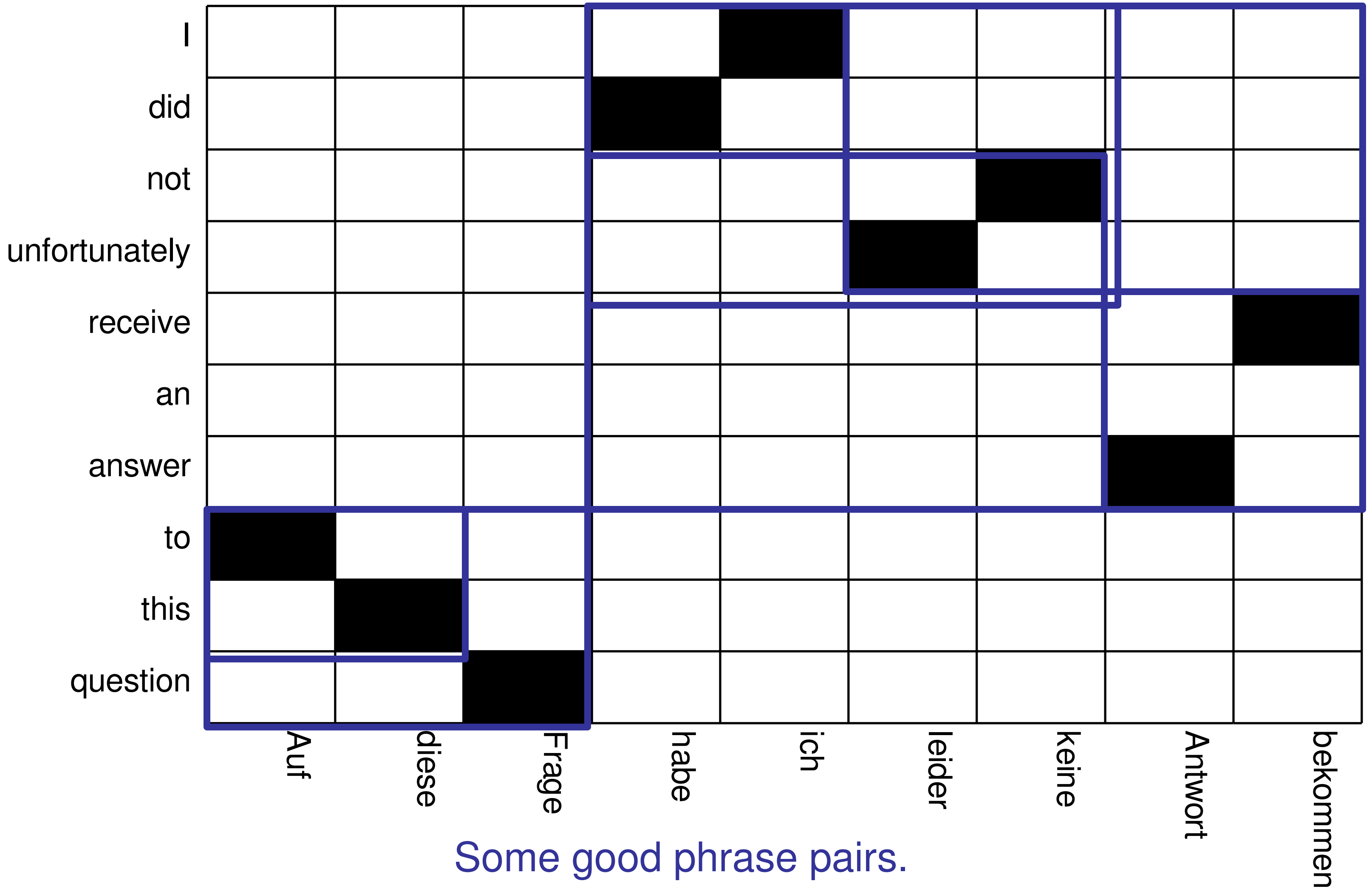
... into the programme ...
... into the **disease** ...
... into the **disease** ...
... into the correct ...
... into the next ...
... into the national ...
... into the integration ...
... into the Union ...
... into the Union ...
... into the Union ...
... into the sort ...
... into the internal ...
... into the general ...
... into the budget ...
... into the **disease** ...
... into the legal …
... into the various ...
... into the nuclear ...
... into the bargain ...
... into the situation ...

Phrase Models



Assume word alignments are given.

Phrase Models



Some good phrase pairs.

Phrase Models

Some bad phrase pairs.

“Count and Normalize”

- Usual approach: treat relative frequencies of source phrase *s* and target phrase *t* as probabilities

$$p(s \mid t) \equiv \frac{count(s,t)}{count(t)} \qquad p(t \mid s) \equiv \frac{count(s,t)}{count(s)}$$

- This leads to overcounting when not all segmentations are legal due to unaligned words.

Hidden Structure

- But really, we don't observe word alignments.
- How are word alignment model parameters estimated?
- Find (all) structures consistent with observed data.
  - Some links are incompatible with others.
  - We need to score complete sets of links.

Hidden Structure and EM

- Expectation Maximization
  - Initialize model parameters (randomly, by some simpler model, or otherwise)
  - Calculate probabilities of hidden structures
  - Adjust parameters to maximize likelihood of observed data given hidden data
  - Iterate
- Summing over *all* hidden structures can be expensive
  - Sum over 1-best, *k*-best, other sampling methods

Discriminative Training

- Given a source sentence, give “good” translations a higher score than “bad” translations.
- We care about good translations, not a high probability of the training data.
- Spend less “energy” modeling bad translations.
- Disadvantages
  - We need to run the translation system at each training step.
  - System is tuned for one task (e.g. translation) and can’t be directly used for others (e.g. alignment)

"Good" Compared to What?

- Compare current translation to
- Idea #1: a human translation. OK, but
  - Good translations can be very dissimilar
  - We'd need to find hidden features (e.g. alignments)
- Idea #2: other top *n* translations (the "n-best list"). Better in practice, but
  - Many entries in n-best list are the same apart from hidden links
- Compare with a **loss function** *L*
  - 0/1: wrong or right; equal to reference or not
  - Task-specific metrics (word error rate, BLEU, …)

MT Evaluation

* Intrinsic

  Human evaluation

  Automatic (machine) evaluation

* Extrinsic

  How useful is MT system output for…

  Deciding whether a foreign language blog is about politics?
  Cross-language information retrieval?
  Flagging news stories about terrorist attacks?
  …

Human Evaluation

Je suis fatigué.

	Adequacy	Fluency
Tired is I.	5	2
Cookies taste good!	1	5
I am exhausted.	5	5

Human Evaluation

PRO

High quality

CON

Expensive!

Person (preferably bilingual) must make a time-consuming judgment per system hypothesis.

Expense prohibits frequent evaluation of incremental system modifications.

Automatic Evaluation

PRO

Cheap. Given available reference translations, free thereafter.

CON

**We can only measure some proxy for translation quality.
(Such as N-Gram overlap or edit distance).**

Automatic Evaluation: Bleu Score

N-Gram precision

$$p_n = \frac{\sum_{n\text{-gram} \in hyp} count_{clip}(n\text{-gram})}{\sum_{n\text{-gram} \in hyp} count(n\text{-gram})}$$

Bounded above by highest count of n-gram in any reference sentence

brevity penalty

$$B = \begin{cases} e^{(1 - |ref| / |hyp|)} & \text{if } |ref| > |hyp| \\ 1 & \text{otherwise} \end{cases}$$

Bleu score: brevity penalty, geometric mean of N-Gram precisions

$$\text{Bleu} = B \cdot \exp\left[\frac{1}{N}\sum_{n=1}^{N} p_n\right]$$

Automatic Evaluation: Bleu Score

hypothesis 1	**I am exhausted**
hypothesis 2	**Tired is I**
reference 1	**I am tired**
reference 2	**I am ready to sleep now**

Automatic Evaluation: Bleu Score

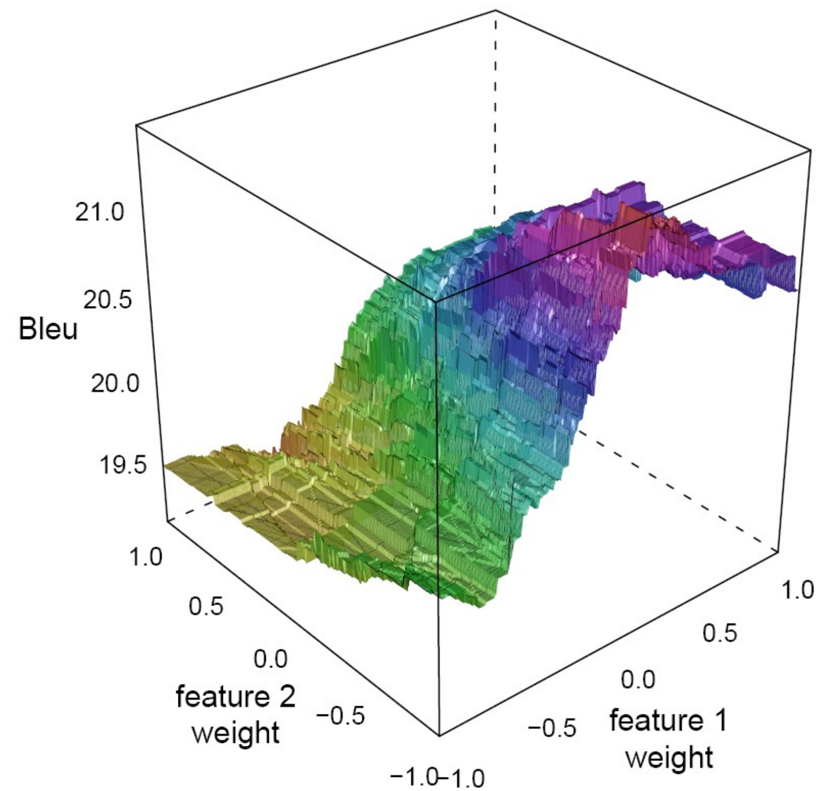
		1-gram	2-gram	3-gram
hypothesis 1	I am exhausted	3/3	1/2	0/1
hypothesis 2	Tired is I	1/3	0/2	0/1
hypothesis 3	I I I	1/3	0/2	0/1

reference 1 — I am tired

reference 2 — I am ready to sleep now and so exhausted

Minimizing Error/Maximizing Bleu

- Adjust parameters to minimize error (*L*) when translating a training set
- Error as a function of parameters is
  - *nonconvex*: not guaranteed to find optimum
  - *piecewise constant*: slight changes in parameters might not change the output.
- Usual method: optimize one parameter at a time with linear programming



Generative/Discriminative Reunion

- Generative models can be cheap to train: “count and normalize” when nothing’s hidden.
- Discriminative models focus on problem: “get better translations”.
- Popular combination
  - Estimate several generative translation and language models using relative frequencies.
  - Find their optimal (log-linear) combination using discriminative techniques.

Generative/Discriminative Reunion

Score each hypothesis with several generative models:

$$\theta_1 p_{phrase}(\bar{s} \mid \bar{t}) + \theta_2 p_{phrase}(\bar{t} \mid \bar{s}) + \theta_3 p_{lexical}(s \mid t) + \cdots + \theta_7 p_{LM}(\bar{t}) + \theta_8 \#\,\text{words} + \cdots$$

If necessary, renormalize into a probability distribution:

$$Z = \sum_k \exp(\theta \bullet \mathbf{f}_k)$$

Unnecessary if thetas sum to 1 and p's are all probabilities.

where *k* ranges over all hypotheses. We then have

$$p(t_i \mid s) = \frac{1}{Z} \exp(\theta \bullet \mathbf{f})$$

Exponentiation makes it positive.

for any given hypothesis *i*.

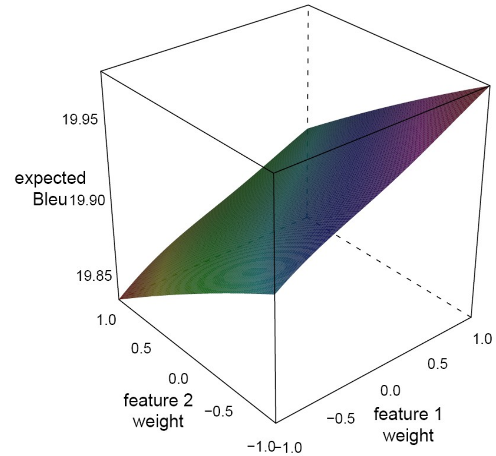
Minimizing Risk

Instead of the error of the 1-best translation, compute **expected error** (risk) using *k*-best translations; this makes the function differentiable.

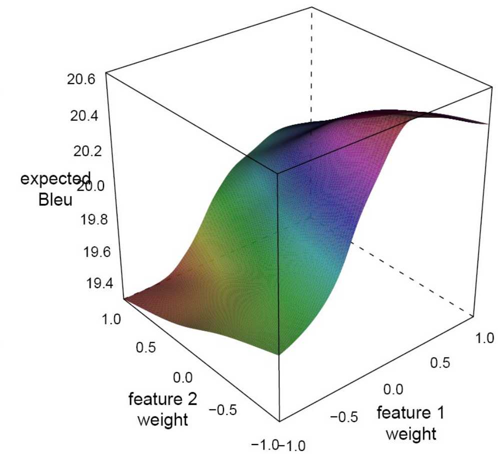
Smooth probability estimates using gamma to even out local bumpiness. Gradually increase gamma to approach the 1-best error.

$$E_{p_{\gamma,\theta}}[L(s,t)]$$

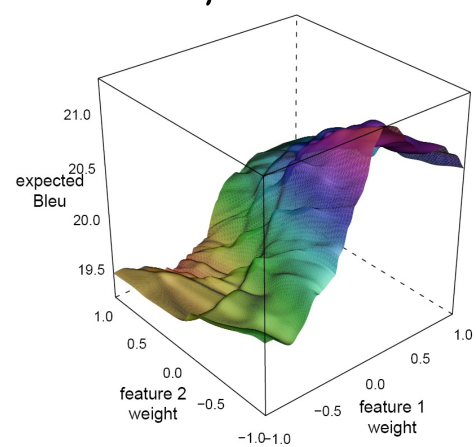
$$p_{\gamma,\theta}(t_i \mid s_i) = \frac{[\exp\ \theta \bullet \mathbf{f}_i]^{\gamma}}{\sum_{k'} [\exp\ \theta \bullet \mathbf{f}_{k'}]^{\gamma}}$$



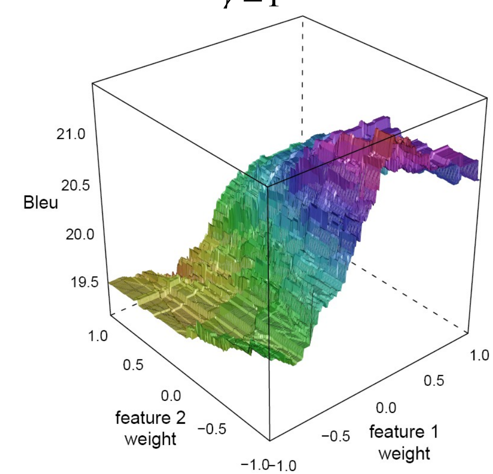
$\gamma = 0.1$



$\gamma = 1$



$\gamma = 10$



$\gamma = \infty$

Learning Word Translation Dictionaries Using Minimal Resources

Learning Translation Lexicons for Low-Resource Languages

{Serbian Uzbek Romanian Bengali} ⟶ English

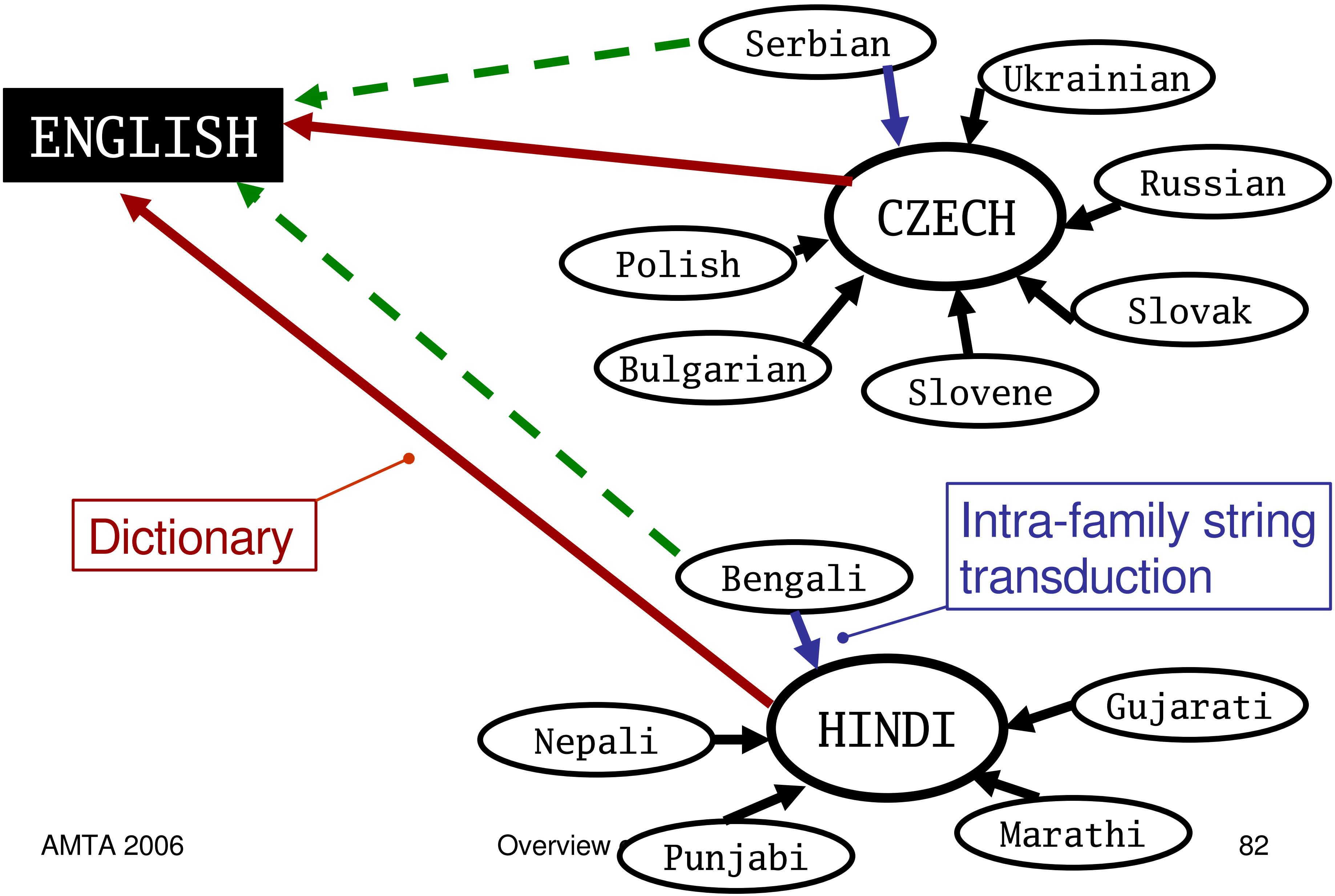
Problem: Scarce resources . . .

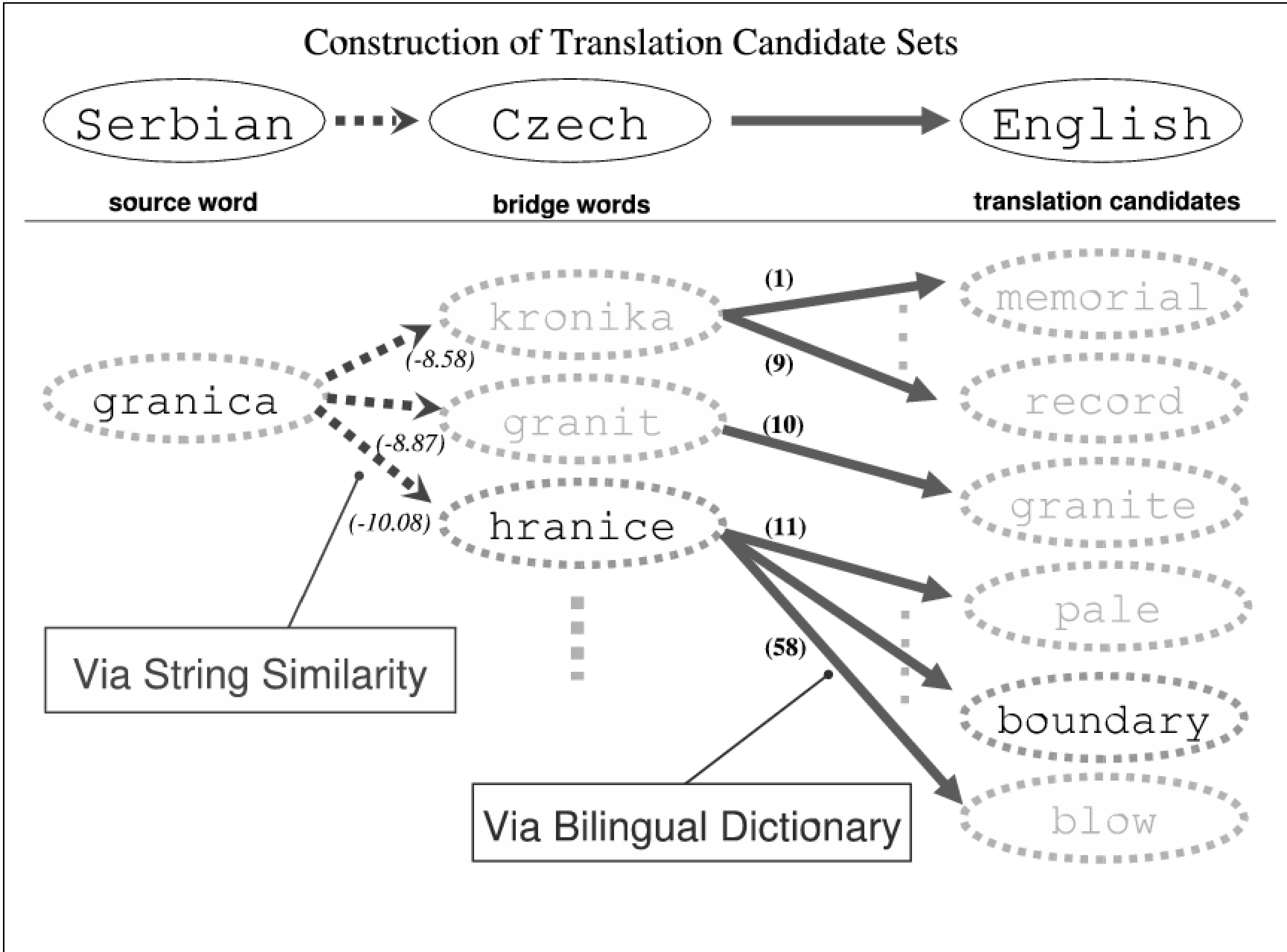
- Large parallel texts are very helpful, but often unavailable
- Often, no "seed" translation lexicon is available
- Neither are resources such as parsers, taggers, thesauri

Solution: Use only monolingual corpora in source, target languages

- But use many information sources to propose and rank translation candidates

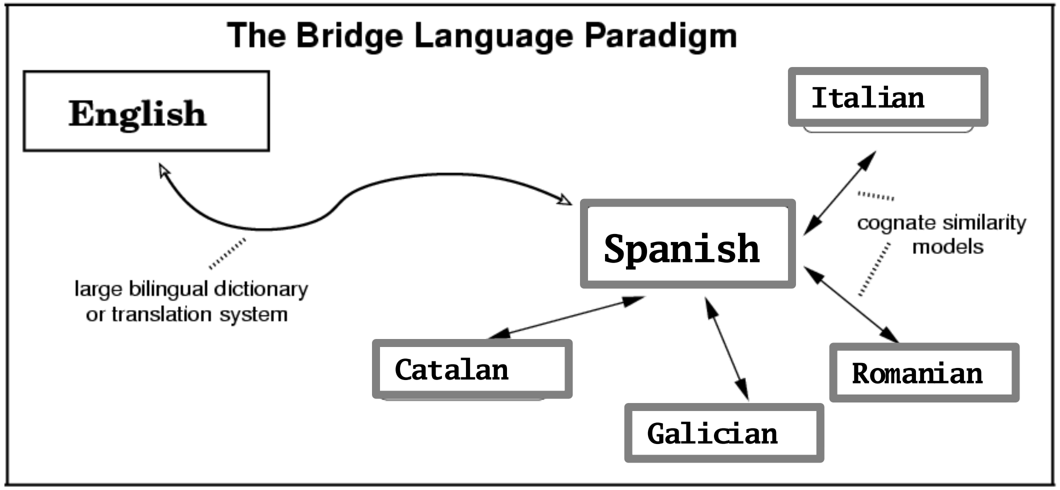
Bridge Languages





*** Constructing translation candidate**

Cognate Selection



some cognates

Spanish-Italian homogenizar omogeneizzare

Polish-Serbian befsztyk biftek

German-Dutch gefestigt gevestigd

Spanish Word	Italian Word	Cognate?
electron	elettrone	
aventurero	avventuriero	
perífrasis	perifrasi	
divul gar	divulgare	
triada	triade	
agresivo	aggressivo	
insertar	inserto	
esprint	sprint	
trópico	tropico	
altimetro	altimetro	
alegato	lista	NO
variado	variato	
cepillar	piallare	
confusin	confusione	
fortificacion	fortificazione	
conjuncion	congiunzione	
encantador	incantatore	
heredero	erede	
vidrio	vetro	
vaciar	variare	NO
talisman	talismano	
sólido	solido	
criptografia	crittografia	
carencia	carenza	
cortesania	cortesia	NO
sadico	sadico	
concentracion	concentrazione	
venida	venuta	
agonizante	agonizzante	
extinguir	estinguere	

The Transliteration Problem

Arabic

Piedade	BEH YEH YEH DAL ALEF DAL YEH
Bolivia	BEH WAW LAM YEH FEH YEH ALEF
Luxembourg	LAM KAF SEEN MEEM BEH WAW REH GHAIN
Zanzibar	ZAIN NOON JEEM YEH BEH ALEF REH

Inuktitut

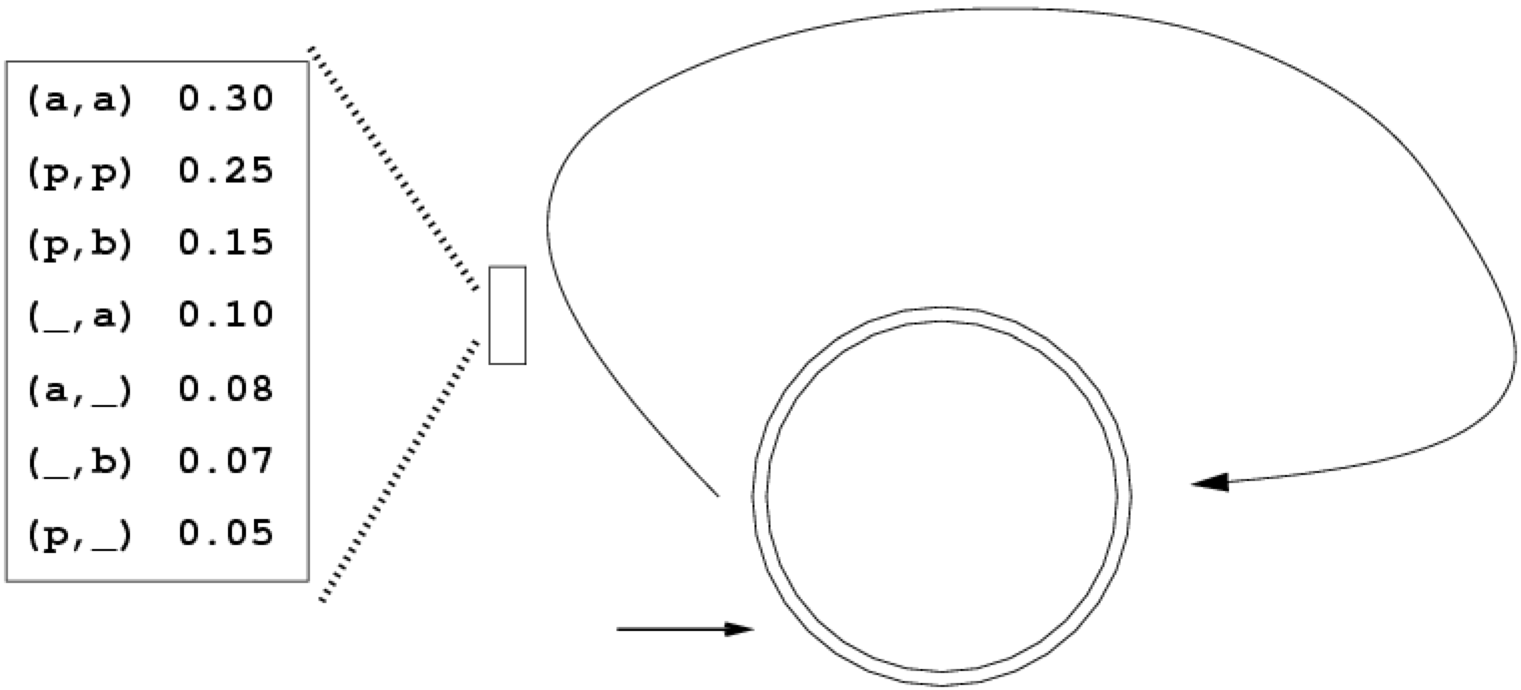
Williams: uialims uilialums uiliammas viliams

Campbell: kaampu kaampul kamvul kaamvul

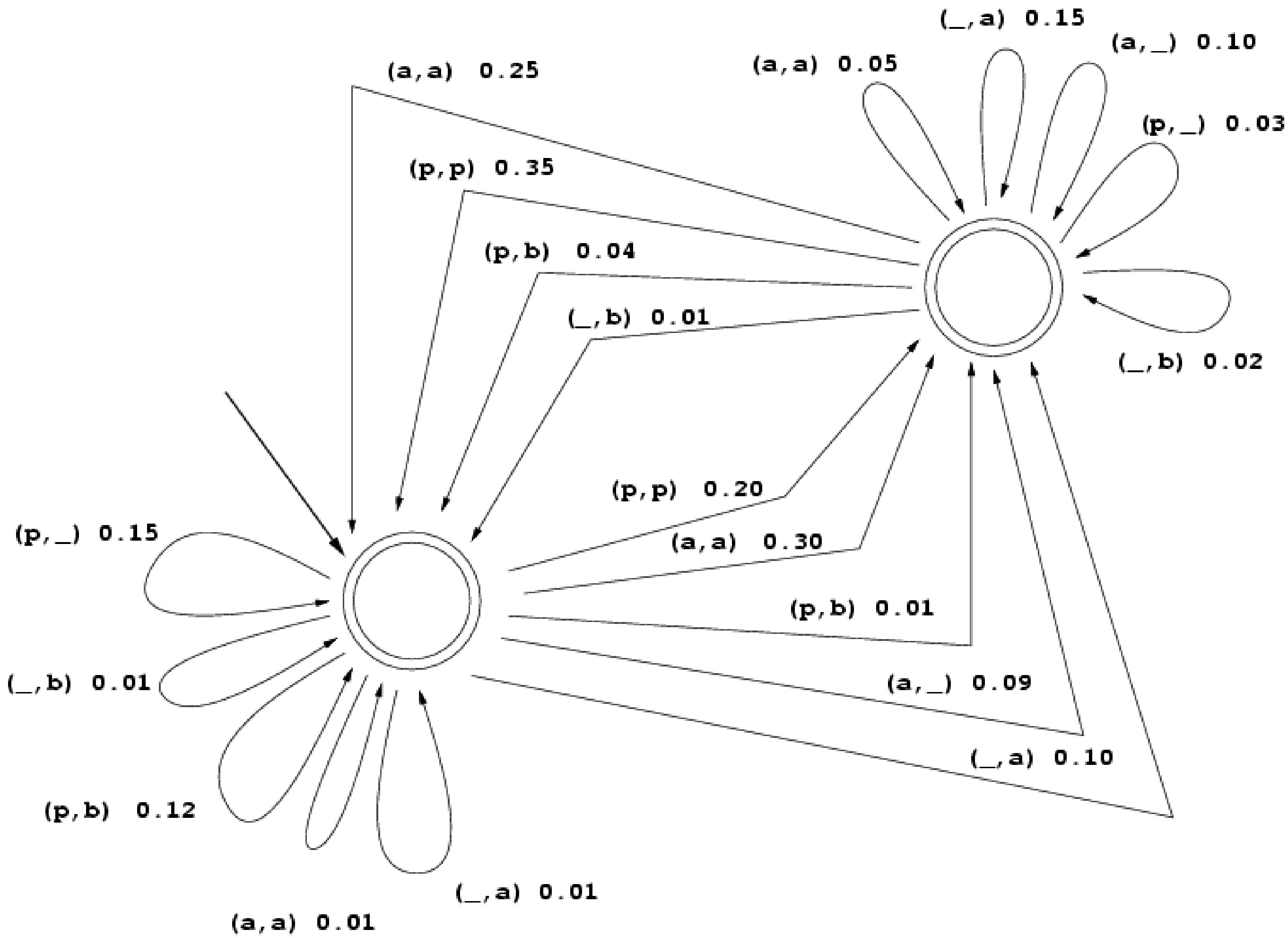
McLean: makalain maklainn makliin makkalain

Memoryless Transducer

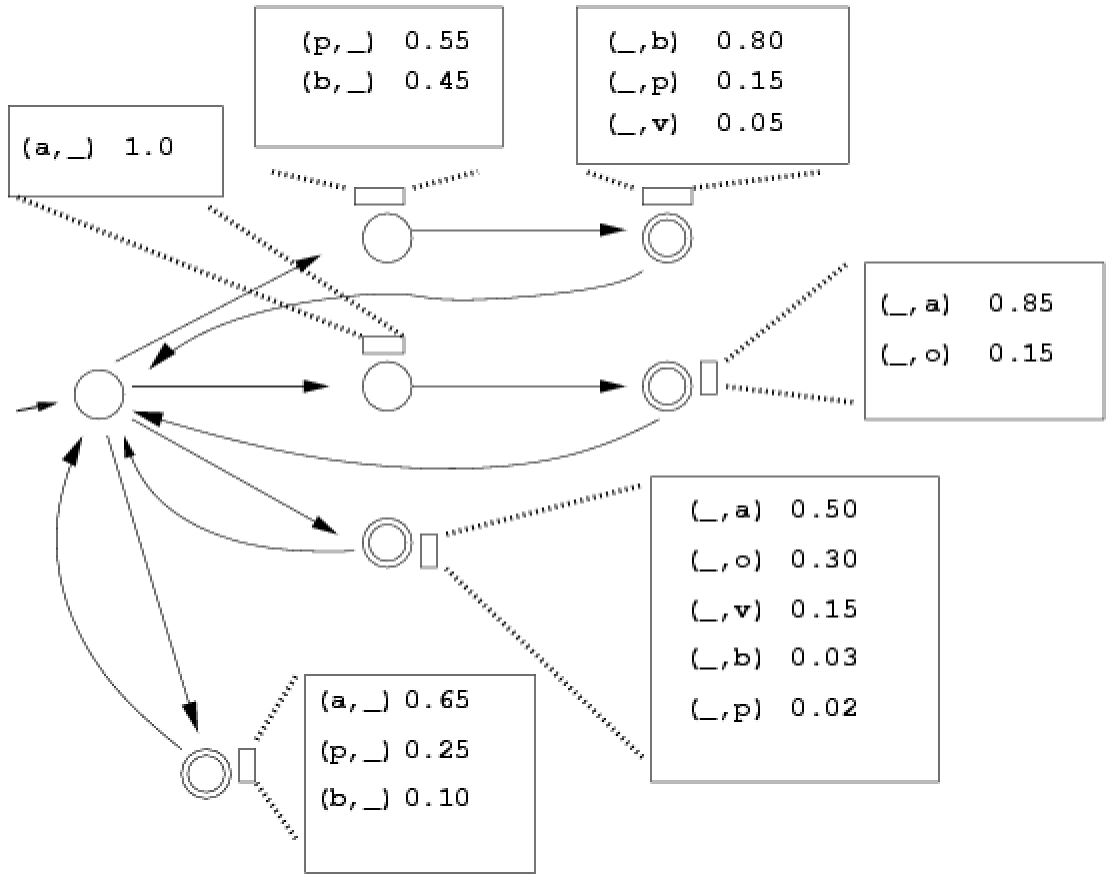
(Ristad & Yianilos 1997)



Two-State Transducer ("Weak Memory")



Unigram Interlingua Transducer



Examples: Possible Cognates Ranked by Various String Models

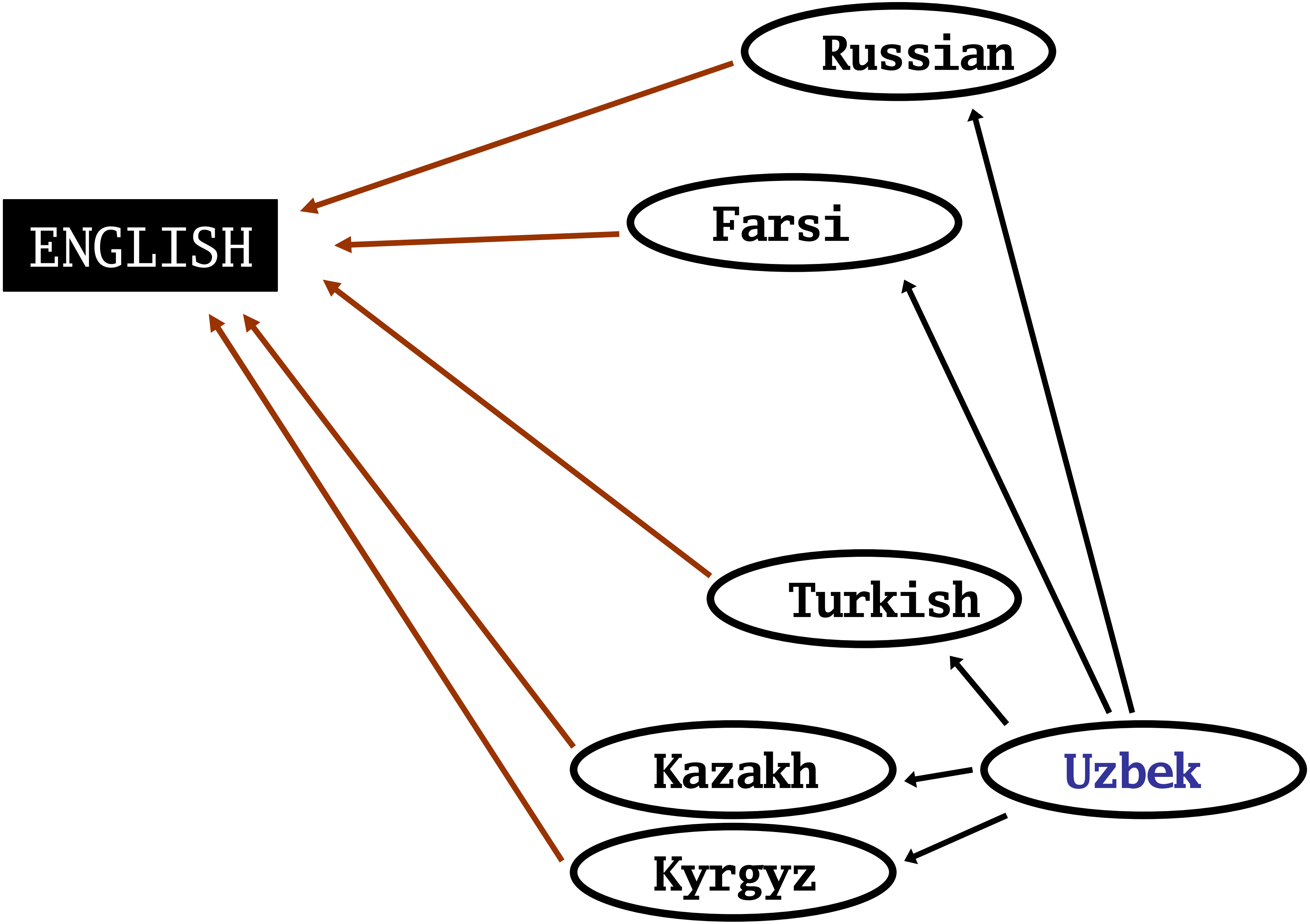
String Transduction Models Ranking Spanish Bridge Words for Romanian Source Word *inghiti*									
C1	**C2**	**C3**	**R&Y**	**2STEF**	**UIT**	**SN**	**AI**	**CDUI**	**JDCO**
S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato
S:ingerir	**S:ingerir**	S:engaste	S:grito	S:negrito	**S:ingerir**	S:ingente	S:negrito	S:infarto	S:engaste
S:engaste	S:engaste	**S:ingerir**	S:gaita	S:grito	S:grito	**S:ingerir**	S:negrita	S:engaste	S:anguila
S:ingreso	S:ingreso	S:inglete	S:grita	**S:ingerir**	S:grita	S:ingle	**S:ingerir**	S:ingreso	S:infarto
S:ingerido	S:ingerido	S:ingreso	S:negrito	S:negrita	S:inglete	S:angra	S:grito	S:introito	S:aguita
S:inglete	S:grito	S:ingerido	S:infarto	S:grita	S:gaita	S:ingerido	S:grita	S:negrito	S:ingreso
S:grito	S:inglete	S:infarto	S:negrita	S:gaita	S:negrito	S:ingenio	S:gaita	S:ingerido	S:intriga
S:infarto	S:infarto	S:grito	**S:ingerir**	S:ingerido	S:infarto	S:engan	S:ingenito	S:negrita	S:intuir
S:grita	S:negrito	S:introito	S:engaste	S:ingreso	S:introito	S:engatado	S:inglete	**S:ingerir**	S:indulto
S:introito	S:grita	S:engreir	S:haiti	S:haiti	S:engreir	S:invita	S:tahiti	S:inglete	S:inglete

String Transduction Models Ranking Turkish Bridge Words for Uzbek Source Word аввалги									
C1	**C2**	**C3**	**R&Y**	**2STEF**	**UIT**	**SN**	**AI**	**CDUI**	**JDCO**
T:evvelki	**T:evvelki**	**T:evvelki**	**T:evvelki**	T:vali	**T:evvelki**	T:edilgi	**T:evvelki**	**T:evvelki**	**T:evvelki**
T:evvelce	**T:evvelce**	**T:evvelce**	T:evveli	T:veli	**T:evvelce**	T:dalga	T:evveli	**T:evvelce**	**T:evvelce**
T:kalga	**T:evvelkí**	T:kalga	T:evvela	T:vals	T:edilgi	T:delgi	T:aval	T:evveli	**T:evvelkí**
T:evvelkí	T:kalga	T:salgi	**T:evvel**	T:delgi	T:algi	T:kalga	T:algi	T:evvela	T:ilkelci
T:vals	T:salgi	T:vals	T:algi	**T:evvelki**	T:salgi	T:evel	**T:evvel**	T:ilkelci	T:sivilce
T:salgi	T:vals	**T:evvelkí**	**T:evvelce**	T:kalga	T:vals	T:dalgl	T:evvela	T:eksilti	T:ilkelce
T:villa	T:villa	T:delgi	T:edilgi	T:dalga	T:delgi	**T:evvelki**	T:salgi	T:zavalli	T:akilci
T:silgi	T:silgi	T:villa	T:aval	T:villa	T:silgi	T:evlat	T:vali	**T:evvelkí**	T:eksilti
T:edilgi	T:ilkelci	T:evveli	T:evel	T:vale	T:kalga	T:dolgu	**T:evvelce**	**T:evvel**	T:asilce
T:volta	T:akilci	T:silgi	T:delgi	T:yilgi	T:dalga	T:veli	**T:evvelkí**	T:ilkelce	T:otelci

Romanian *inghiti* **(ingest)**

Uzbek *avvalgi* **(previous/former)**

* **Effectiveness of cognate**



* **Multi-family bridge**

Similarity Measures

for re-ranking cognate/transliteration hypotheses

1. **Probabilistic string transducers**
2. **Context similarity**
3. **Date distribution similarity**
4. **Similarities based on monolingual word properties**

Similarity Measures

1. Probabilistic string transducers
2. Context similarity
3. Date distribution similarity
4. Similarities based on monolingual word properties

Compare Vectors

nezavisnost vector
Projection of context vector from Serbian to English term space

expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
0	0	2	1.5	1.5	1.5	4	1.5

independence vector
Construction of context term vector

expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
3	1	10	0	479	836	191	0

freedom vector
Construction of context term vector

expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
681	184	104	0	21	4	141	0

Compute **cosine similarity** between *nezavisnost* and “independence”

… and between *nezavisnost* and “freedom”

Similarity Measures

1. Probabilistic string transducers

2. Context similarity

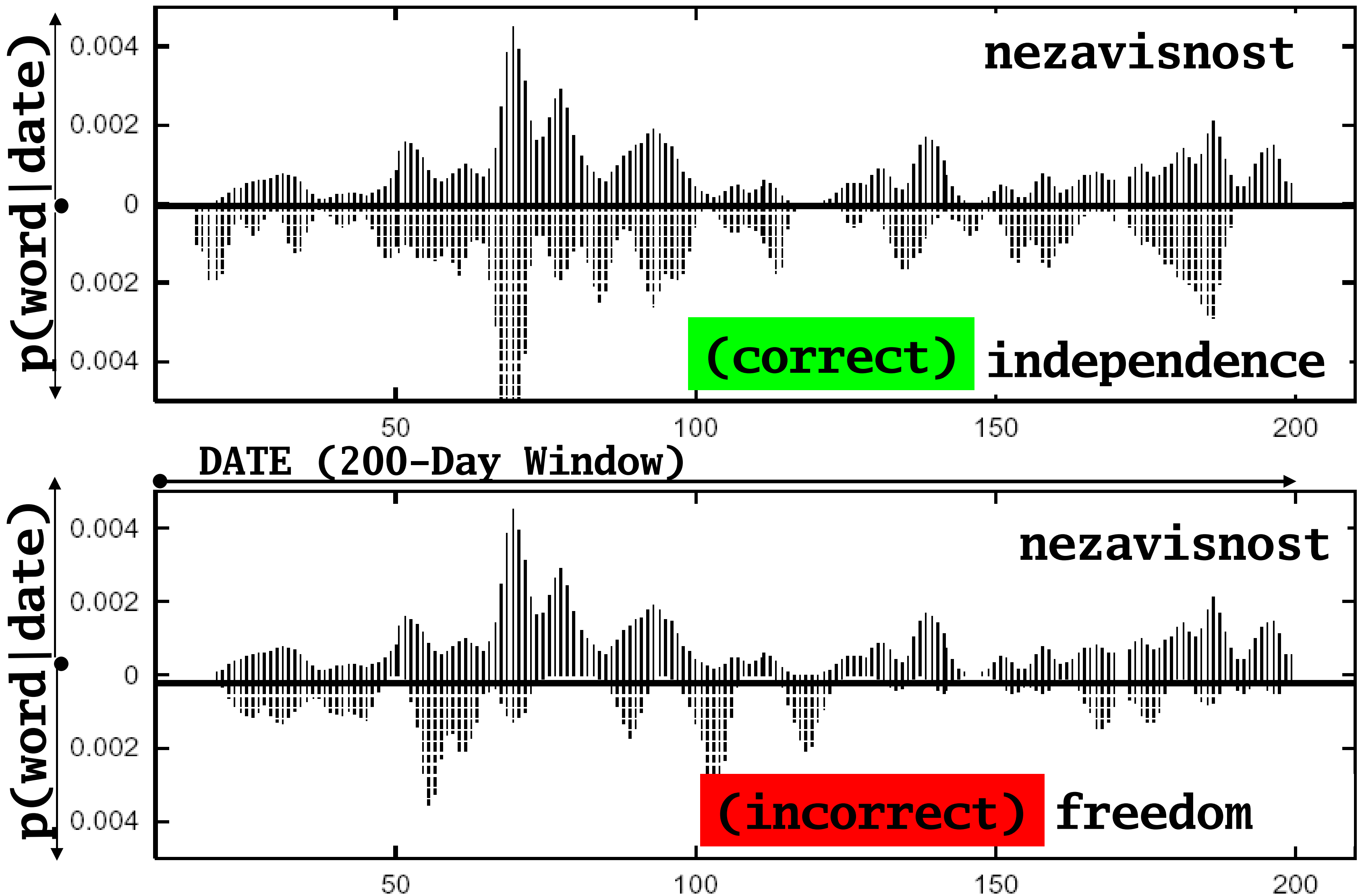
3. Date distribution similarity

4. Similarities based on monolingual word properties

Date Distribution Similarity

- Topical words associated with real-world events appear within news articles in bursts following the date of the event
- Synonymous topical words in different languages, then, display similar distributions across dates in news text: this can be measured
- We use cosine similarity on date term vectors, with term values `p(word|date)`, to quantify this notion of similarity

Date Distribution Similarity - Example



Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties

Relative Frequency

$$rf(w_F) = \frac{f_{C_F}(w_F)}{|C_F|}$$

$$rf(w_E) = \frac{f_{C_E}(w_E)}{|C_E|}$$

Cross-Language Comparison:

$$\min\left(\frac{rf(w_F)}{rf(w_E)}, \frac{rf(w_E)}{rf(w_F)}\right)$$

[min-ratio method]

Precedent in Yarowsky & Wicentowski (2000); used relative frequency similarity for morphological analysis

Combining Similarities: Uzbek

Individual Bridge Language Results For Uzbek Using Combined Similarity Measures				
Rank	Turkish	Russian	Farsi	Kyrgyz
1	0.04	**0.12**	0.03	0.06
5	0.10	**0.23**	0.05	0.08
10	0.13	**0.26**	0.07	0.10
20	0.16	**0.28**	0.08	0.11
50	0.21	**0.30**	0.12	0.13
100	0.24	**0.31**	0.15	0.16
200	0.26	**0.32**	0.19	0.19

Multiple Bridge Language Results For Uzbek Using Combined Similarity Measures					
Rank	Tur+Rus	Tur+Rus +Farsi	Tur+Rus +Eng	Tur+Rus +Farsi +Kaz+Kyr	Tur+Rus +Farsi +Kaz+Kyr+Eng
1	0.12	0.13	0.13	**0.14**	**0.14**
5	0.26	0.27	0.26	0.28	**0.29**
10	0.30	0.31	0.31	**0.34**	**0.34**
20	0.35	0.37	0.35	**0.39**	**0.39**
50	0.39	0.41	0.39	0.42	**0.43**
100	0.41	0.43	0.41	**0.46**	0.45
200	0.43	0.45	0.42	**0.48**	0.46

Combining Similarities: Romanian, Serbian, & Bengali

Multiple Bridge Language Results For Romanian Using Combined Similarity Measures				
Rank	Spanish	Spanish +Russian	Spanish +English	Spanish +Russian +English
1	0.17	0.18	**0.19**	**0.19**
5	0.31	0.35	0.34	**0.37**
10	0.37	0.41	0.41	**0.43**
20	0.43	0.46	0.46	**0.48**
50	0.51	0.53	0.53	**0.55**
100	0.57	0.60	0.58	**0.61**
200	0.60	**0.62**	0.59	**0.62**

Multiple Bridge Language Results For Serbian Using Combined Similarity Measures						
Rank	Cz	Rus	Bulg	Cz +English	Cz+Slovak +Rus+Bulg	Cz+Slovak +Rus+Bulg +English
1	0.13	0.15	**0.19**	0.13	**0.19**	**0.19**
5	0.24	0.24	0.31	0.25	**0.38**	**0.38**
10	0.29	0.28	0.35	0.30	0.42	**0.43**
20	0.32	0.31	0.40	0.34	**0.48**	**0.48**
50	0.38	0.36	0.44	0.39	0.54	**0.55**
100	0.40	0.40	0.48	0.42	**0.59**	**0.59**
200	0.41	0.42	0.50	0.43	**0.60**	**0.60**

Bridge Language Results for Bengali Using Combined Similarity Measures		
Rank	Hindi	Hindi +English
1	0.03	**0.05**
5	0.11	**0.14**
10	0.13	**0.17**
20	0.16	**0.21**
50	0.19	**0.25**
100	0.22	**0.28**
200	0.23	**0.29**

Observations

* With no Uzbek-specific supervision, we can produce an Uzbek-English dictionary which is 14% exact-match correct

* Or, we can put a correct translation in the top-10 list 34% of the time (useful for end-to-end machine translation or cross-language information retrieval)

* Adding more bridge languages helps

Multiple Bridge Language Results For Uzbek Using Combined Similarity Measures

Rank	Tur+Rus	Tur+Rus +Farsi	Tur+Rus +Eng	Tur+Rus +Farsi +Kaz+Kyr	Tur+Rus +Farsi +Kaz+Kyr+Eng
1	0.12	0.13	0.13	**0.14**	**0.14**
5	0.26	0.27	0.26	0.28	**0.29**
10	0.30	0.31	0.31	**0.34**	**0.34**
20	0.35	0.37	0.35	**0.39**	**0.39**
50	0.39	0.41	0.39	0.42	**0.43**
100	0.41	0.43	0.41	**0.46**	0.45
200	0.43	0.45	0.42	**0.48**	0.46

Practical Considerations

Empirical Translation in Practice: System Building

1. Data collection
   - Bitext
   - Monolingual text for language model (LM)

2. Bitext sentence alignment, if necessary

3. Tokenization
   - Separation of punctuation
   - Handling of contractions

4. Named entity, number, date normalization/translation

5. Additional filtering
   - Sentence length
   - Removal of free translations

6. Training…

Some Freely Available Tools

- Sentence alignment
  - http://research.microsoft.com/~bobmoore/
- Word alignment
  - http://www.fjoch.com/GIZA++.html
- Training phrase models
  - http://www.iccs.inf.ed.ac.uk/~pkoehn/training.tgz
- Translating with phrase models
  - http://www.isi.edu/licensed-sw/pharaoh/
- Language modeling
  - http://www.speech.sri.com/projects/srilm/
- Evaluation
  - http://www.nist.gov/speech/tests/mt/resources/scoring.htm
- See also http://www.statmt.org/