Lecture 4:
Machine Translation and IBM Model 1 (Learning)

Intro to NLP, CS585, Fall 2014

http://people.cs.umass.edu/~brenocon/inlp2014/

Brendan O'Connor (http://brenocon.com)

- Homework questions?
- Today
  - Quick LM review
  - Machine translation -- learning
- Next week
  - Machine translation -- decoding, broader issues

Interpolation LM [from last time]

Let $\hat{P}$ be the MLE probability distribution

Interpolation estimate is:

$$P(w|w_{-1}, w_{-2}) = \lambda_2 \hat{P}(w|w_{-1}, w_{-2}) + \lambda_1 \hat{P}(w|w_{-1}) + \lambda_0 \hat{P}(w)$$

Sparse Denser Dense

Mixing weights: $\lambda_0 + \lambda_1 + \lambda_2 = 1$

$$P(w|\text{blacke as})$$

Hamlet from NLTK, V=4812

http://people.cs.umass.edu/~brenocon/inlp2014/lectures/04-hamlet_ngrams.txt

$$= \lambda_2 \hat{P}(w|\text{blacke as}) + \lambda_1 \hat{P}(w|\text{as}) + \lambda_0 \hat{P}(w)$$

$$P(w|\text{blacke as})$$

Hamlet from NLTK, V=4812

http://people.cs.umass.edu/~brenocon/inlp2014/lectures/04-hamlet_ngrams.txt

$$= \lambda_2 \hat{P}(w|\text{blacke as}) + \lambda_1 \hat{P}(w|\text{as}) + \lambda_0 \hat{P}(w)$$

3	**blacke as**
1	blacke as death
1	blacke as hell
1	blacke as his

3 non-zeros

$$P(w|\text{blacke as})$$

Hamlet from NLTK, V=4812

http://people.cs.umass.edu/~brenocon/inlp2014/lectures/04-hamlet_ngrams.txt

$$= \lambda_2 \hat{P}(w|\text{blacke as}) + \lambda_1 \hat{P}(w|\text{as}) + \lambda_0 \hat{P}(w)$$

3	**blacke as**
1	blacke as death
1	blacke as hell
1	blacke as his

3 non-zeros

205	**as**	7	as he	1	as peace
1	as 'twer	1	as healthfull	1	as proper
6	as 'twere	1	as hell	1	as pure
3	as a	1	as her	1	as quick-siluer
1	as actiuely	3	as his	1	as reason
1	as against	1	as how	1	as sharpe
1	as all	1	as hush	1	as sinewes
1	as an	18	as i	1	as sinnes
2	as any	1	as ice	3	as snow
1	as are	6	as if	1	as so
1	as bad	1	as in	1	as some
2	as before	8	as it	1	as swift
3	as by	1	as iust	1	as th'art
1	as chast	2	as kill	13	as the
1	as checking	1	as kinde	1	as therein
1	as common	1	as leuell	2	as they
1	as cunning	1	as liue	3	as this
1	as damn	1	as loue	3	as thou
1	as day	1	as low	1	as thus
2	as death	1	as lying	1	as thy
1	as deepe	1	as mad	5	as to
1	as dicers	1	as made	1	as vnuallued
1	as doth	1	as make	1	as vulcans
1	as easie	1	as many	1	as watchmen
1	as england	2	as may	1	as waxe
1	as ere	1	as meditation	4	as we
1	as false	1	as mine	2	as well
1	as farre	1	as mortall	1	as white
1	as fits	1	as most	2	as will
1	as foule	3	as much	1	as womans
1	as fresh	2	as my	1	as would
2	as from	1	as needfull	12	as you
1	as gaming	2	as of	2	as your
1	as girdle	1	as oft	1	as yours
1	as had	2	as one		
1	as hamlet	2	as our		
1	as hardy	1	as patient		

107 non-zeros

$$P(w|\text{blacke as})$$

Hamlet from NLTK, V=4812

http://people.cs.umass.edu/~brenocon/inlp2014/lectures/04-hamlet_ngrams.txt

$$= \lambda_2 \hat{P}(w|\text{blacke as}) + \lambda_1 \hat{P}(w|\text{as}) + \lambda_0 \hat{P}(w)$$

3	**blacke as**
1	blacke as death
1	blacke as hell
1	blacke as his

3 non-zeros

205	**as**	7	as he	1	as peace
1	as 'twer	1	as	1	as proper
6	as 'twere	healthfull		1	as pure
3	as a	1	as hell	1	as quick-
1	as actiuely	1	as her	siluer	
1	as against	3	as his	1	as reason
1	as all	1	as how	1	as sharpe
1	as an	1	as hush	1	as sinewes
2	as any	18	as i	1	as sinnes
1	as are	1	as ice	3	as snow
1	as bad	6	as if	1	as so
2	as before	1	as in	1	as some
3	as by	8	as it	1	as swift
1	as chast	1	as iust	1	as th'art
1	as	2	as kill	13	as the
checking		1	as kinde	1	as therein
1	as	1	as leuell	2	as they
common		1	as liue	3	as this
1	as cunning	1	as loue	3	as thou
1	as damn	1	as low	1	as thus
1	as day	1	as lying	1	as thy
2	as death	1	as mad	5	as to
1	as deepe	1	as made	1	as
1	as dicers	1	as make	vnuallued	
1	as doth	1	as many	1	as vulcans
1	as easie	2	as may	1	as
1	as england	1	as	watchmen	
1	as ere	meditation		1	as waxe
1	as false	1	as mine	4	as we
1	as farre	1	as mortall	2	as well
1	as fits	1	as most	1	as white
1	as foule	3	as much	2	as will
1	as fresh	2	as my	1	as
2	as from	1	as needfull	womans	
1	as gaming	2	as of	1	as would
1	as girdle	1	as oft	12	as you
1	as had	2	as one	2	as your
1	as hamlet	2	as our	1	as yours
1	as hardy	1	as patient		

107 non-zeros

17	!	1	adoption	1	angle
25	&	1	adores	1	angry
61	'	1	aduanc	1	animals
200	'd	2			
2	'em	aduancement		annexment	
5	'gainst	1		1	annoint
3	'm	aduantage		1	annuall
14	're	5	adue	6	anon
119	's	1		8	another
61	't	aduenturous		1	anothers
1	'tane	3	aduice	9	answer
7	'tis	1	aduise	4	answere
1	'twas	adulterate		1	answered
1	'tweene	1	aeneas	1	answerest
10	'twere	2	afarre	1	answers
4	'twill	affaire		anticipation	
1	'twixt	3	affaires	2	anticke
45	(	1	affayre	1	antike
44	)	affear		1	antiquity
2892	,	1		17	any
1879	.	affectation		1	apale
3	.play	4	affection	2	apart
1	.player	1	affections	2	ape
1	1599	1	afflict	1	apparell
566	:	affliction		2	apparition
298	;	3	affraide	2	appear.
459	?	1	affrighted	1	appear'd
6	[	affront		3	appeare
6	]	13	after	3	appeares
497	a	1		1	appetite
2	a'th	afternoone		1	applaud
1	a-crosse	1		1	appliance
1	a-downe	afterwards		1	appointment
1	a-downe-	34	againe	2	apprehension
a		20	against	1	approue
1	a-dreames	8	age	1	appurtenance
1	a-foot	1	ago	2	apt
1	a-sleepe	1	agreeing	1	ar't
3	a-while	2	ah	1	ardure
1	a-worke	2	aint	121	are
1		1	ake	3	argall
abhominably		4	akers	1	argues
1	abhorred	1	al	3	argument
1	abilitie	3	alacke	4	arm
3	aboord	1	alarme	5	arme
4	aboue	1	alarum	9	armes
19	about	9	alas	1	armie
1		5	alexander	2	armour
abridgements		1	alijs	1	armours
1	abroad	109	all	1	arraigne
1	absent	1		2	arrant
1	absolute	allegeance		3	arras
1		1	allies	1	arrest
abstinence		1	allow	1	arrests
1	abstracts	2	allowance	1	arriued
2	absurd	2	allowed	1	arrow
1	abus	7	almost	2	arrowes
1	abuse	10	alone	16	art
1	abuses	3	along	1	article
2	accent	2	aloofe	1	articles
1	accepts	4	already	1	artire
1	accesse	1	altitude	1	artlesse
3	accident	altogether		205	as
1		2	alwayes	1	asham
accidentall		52	am	1	aside
1	accord	1	amaz	1	ask't
2	according	1	amaze	2	aske
1	account	2		2	asking
1		amazement		1	aslant
accounted		1	amb	1	asleepe
1	accurst	1	ambass	1	aspect
1	accuse	1		1	assaid
1	acquaint	ambassador		1	assaies
1	acquire	3		1	assaile
1		ambassadors		1	assault
acquittance		1		3	assay
9	act	ambassadours		3	asse
4	acte	1	amber	1	asse-
1	acted	1		2	assignes
1	acting	ambiguous		1	assis
10	action	5	ambition	2	assistant
1	actions	3	ambitious	1	
1	actiuely	1	amble	associates	
3	actor	1	amen	3	assume
2	actors	1	amis	1	assumes
2	acts	1	amisse	1	assur
2	actus	1	amities	1	assur'd
1	adaies	1	among	2	assurance
1	adam	53	an	1	assure
1	adams	1	ancient	1	astonish
1	addicted	1	anckle	1	asunder
1	addition	862	and	81	at
1	addresse	1	and't	4	attendant
1	adheres	1	angel	2	
1	adieu	2	angell	attendants	
1	adiew	3	angels	1	attended
1	adioyn	1	anger	1	attends
1	admirable			1	attent
3					
admiration					
2	admit				
1					
admittance					

1	attractiue		
5	audience		
1	audit		
1	augury		
1	aunt		
1	auoid		
1	auouch		
2	auoyd		
1			
auspicious			
2	author		
1			
authorities			
2	awake		
27	away		
2	awe		
4	awhile		
2	axe		
1	ayde		
2	aye		
1	aygre		
1	ayme		
12	ayre		
1	ayres		
1	ayrie		
1	ayry		
1	babe		
2	baby		
1	bace		
2	back		
6	backe		
1	backward		
7	bad		
1	bait		
2	bak		
1	bakers		
1	bakt-		
meats			
1	ban		
1	bands		
1	banke		
3	bap		
1	bapt		
1	baptista		
7	bar		
1	barbars		
2	barbary		
2	bare		
1	bare-foot		
1	barke		
10	barn		
8	barnardo		
1	barr'd		
1	barre		
1	barren		
4	base		
1			
basenesse			
2	baser		
2	basket		
1	bastard		
1	bat		
1	bated		
1	battalians		
1	batten		
1	battery		
1			
battlements			
1	baudry		
1	bawd		
1	bawdy		
191	be		
1	be-ratled		
1	beame		
1	bear't		
6	beard		
1	beards		
13	beare		
1	bearers		
2	beares		
1	bearing		
5	beast		
2	beasts		
1	beaten		
3	beating		
1	beats		
1	beauer		
1			
beauteous			
4	beautie		
1	beautied		
1	beauties		
1	beautifed		
1	beautified		
3	beauty		
1	became		
1	because		
1	becke		
1	beckens		
1	beckons		
2	becomes		
11	bed		
1	bedded		
1	bedrid		
16	bee		
3	been		
13	beene		
1	beer		

4812 non-zeros

$P(w|\text{blacke as})$

Hamlet from NLTK, V=4812

http://people.cs.umass.edu/~brenocon/inlp2014/lectures/04-hamlet_ngrams.txt

$$= \lambda_2 \hat{P}(w|\text{blacke as}) + \lambda_1 \hat{P}(w|\text{as}) + \lambda_0 \hat{P}(w)$$

3	**blacke as**
1	blacke as death
1	blacke as hell
1	blacke as his

3 non-zeros

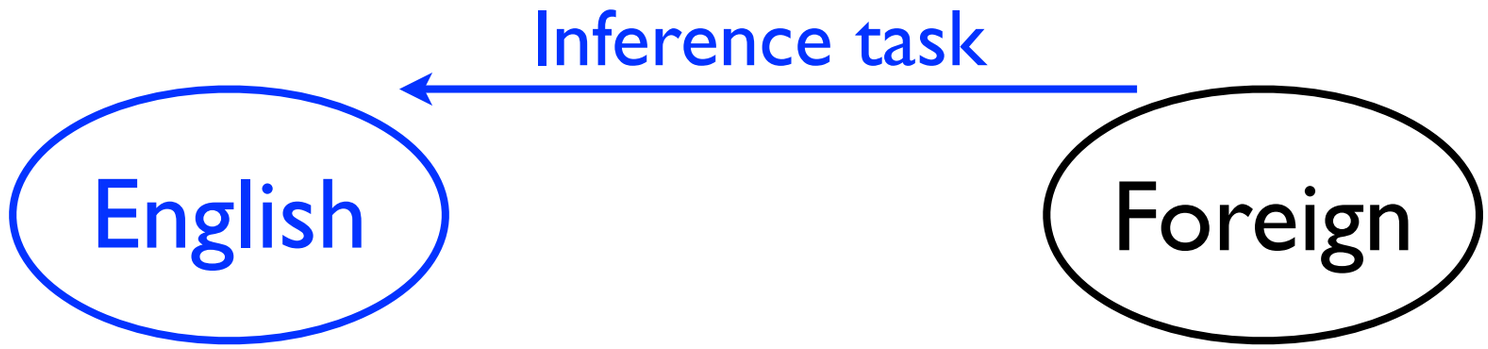
107 non-zeros

4812 non-zeros

- Generative view:
  - Choose history size by prob ($\lambda_2, \lambda_1, \lambda_0$,).
  - Generate *w* by MLE of that history size.
  - Size 1 or 0 are "backoff" choices.
  - Prediction time: marginalize (average) over history size selection.
- Many other methods have different ways of combining different sized histories.
- Sharing statistical strength: "X as"
- Extension: common prefixes require less smoothing. "of the _" has 59 non-zeros, so "the _" less important.
- To think about: shouldn't contexts share statistical strength beyond suffix matching?

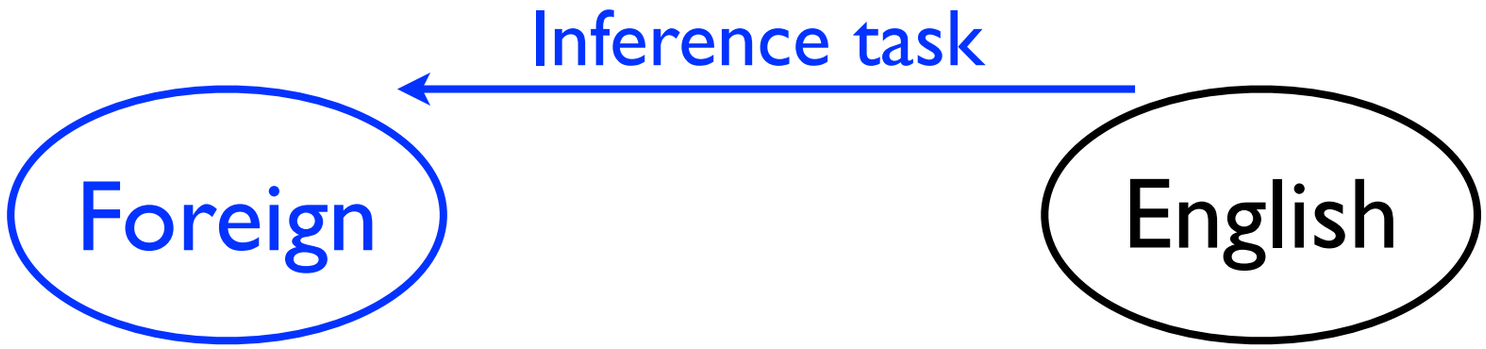
MT as Noisy Channel

Convention in Collins/Knight: translating from ***f*** into ***e***



MT as Noisy Channel

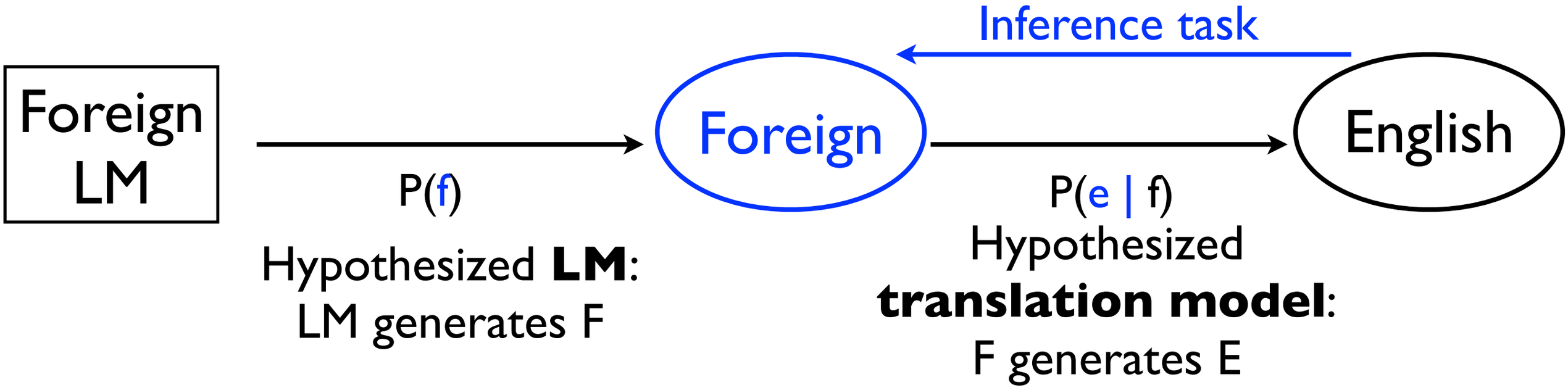
Convention today (sorry!): translating from ***e*** into ***f***



Early 90's, IBM Research

MT as Noisy Channel

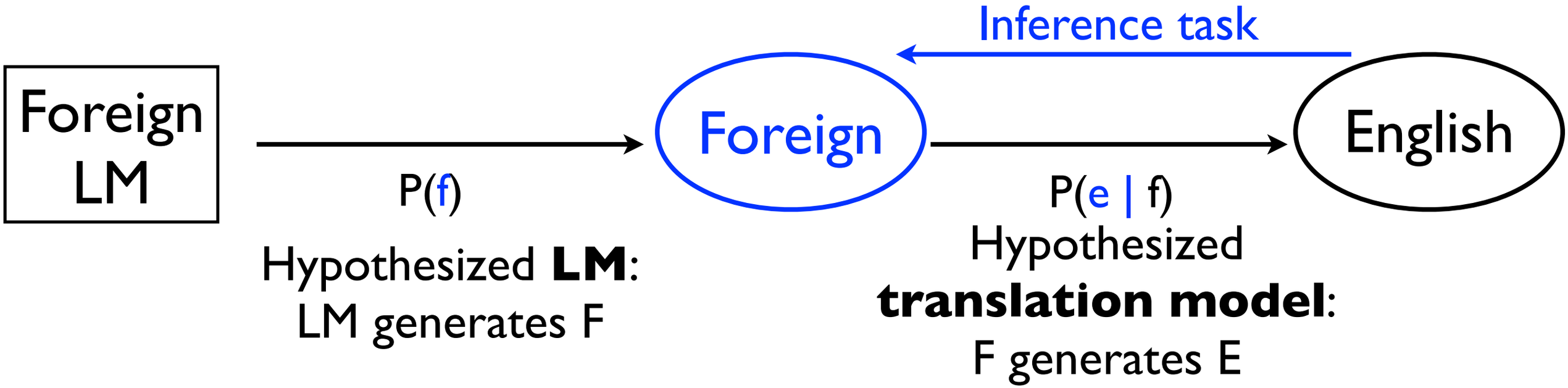
Convention today (sorry!): translating from ***e*** into ***f***



Early 90's, IBM Research

MT as Noisy Channel

Convention today (sorry!): translating from ***e*** into ***f***

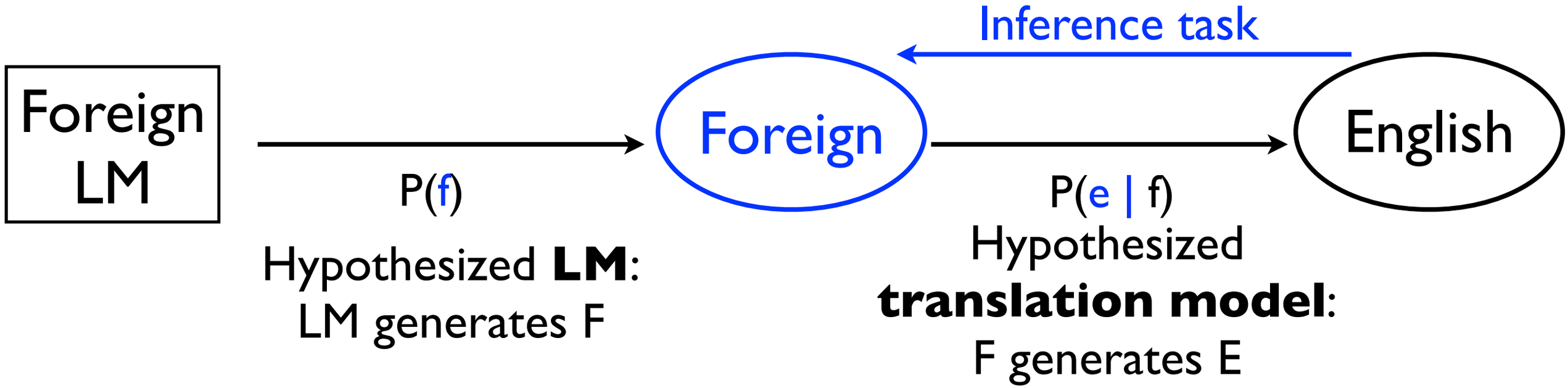


$$P(f \mid e) = P(e \mid f)\, P(f) / P(e)$$

Early 90's, IBM Research

MT as Noisy Channel

Convention today (sorry!): translating from ***e*** into ***f***



$$P(f \mid e) = P(e \mid f)\, P(f) / P(e)$$

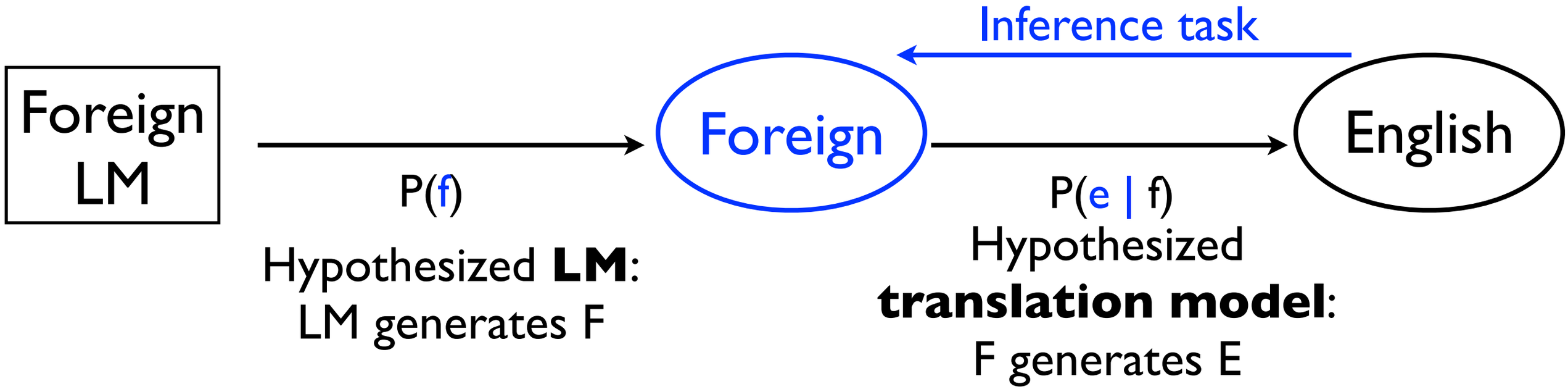
Inference task ("Decoding")

$$\underset{f}{\text{argmax}}\ P(f \mid e) = \underset{f}{\text{argmax}}\ P(e \mid f)\, P(f)$$

Early 90's, IBM Research

MT as Noisy Channel

Convention today (sorry!): translating from ***e*** into ***f***



$$P(f \mid e) = P(e \mid f)\, P(f) / P(e)$$

$$\text{Inference task} \quad \text{argmax}_f\ P(f \mid e) = \text{argmax}_f\ P(e \mid f)\, P(f)$$

("Decoding")

Learning task

P(f): from large monolingual corpus

P(e | f) today: **lexical translation models**

Early 90's, IBM Research

Data to learn statistical MT

- Statistical machine translation is almost entirely based on *parallel corpora*, a.k.a. *bitexts*.
- Training data: human-translated sentence pairs (**e**, **f**)

Lexical Translation

- How do we translate a word? Look it up in the dictionary

  Haus : house, home, shell, household

- Multiple translations
  - Different word senses, different registers, different inflections (?)
  - *house, home* are common
  - *shell* is specialized (the Haus of a snail is a shell)

How common is each?

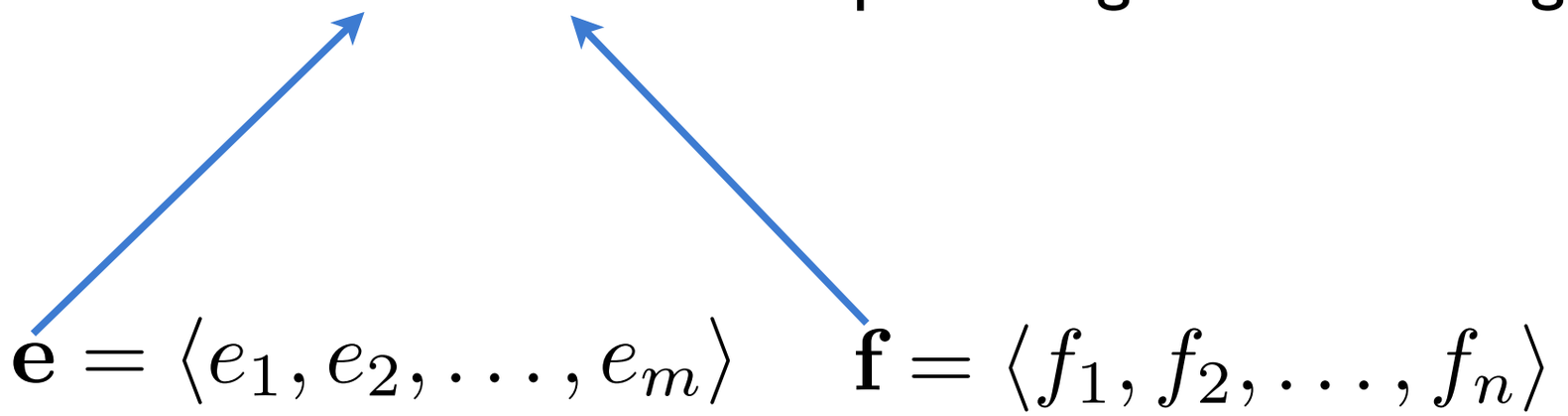
Translation	Count
house	5000
home	2000
shell	100
household	80

MLE

$$\hat{p}_{\text{MLE}}(e \mid \texttt{Haus}) = \begin{cases} 0.696 & \text{if } e = \texttt{house} \\ 0.279 & \text{if } e = \texttt{home} \\ 0.014 & \text{if } e = \texttt{shell} \\ 0.011 & \text{if } e = \texttt{household} \\ 0 & \text{otherwise} \end{cases}$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences



Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences
- Lexical translation makes the following ***assumptions***:
  - Each word in e_i in $\mathbf{e}$ is generated from exactly one word in $\mathbf{f}$
  - Thus, we have an *alignment* a_i that indicates which word e_i "came from", specifically it came from f_{a_i}.
  - Given the alignments $\mathbf{a}$, translation decisions are conditionally independent of each other and depend *only* on the aligned source word f_{a_i}.

Lexical Translation

$$\mathbf{e} = \langle e_1, e_2, ...e_m \rangle \qquad \mathbf{f} = \langle f_1, f_2, ...f_n \rangle$$
$$\mathbf{a} = \langle a_1, a_2, ...a_m \rangle \quad \text{each } a_i \in \{0, 1, ..., n\}$$

Chain rule

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in \{0,1,..,n\}^m} p(\mathbf{a} \mid \mathbf{f}, m) \times p(\mathbf{e} \mid \mathbf{a}, \mathbf{f}, m)$$

Modeling assumptions

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in \{0,1,..,n\}^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

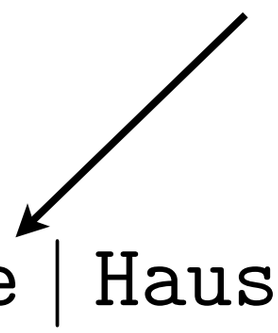
[Alignment] x [Translation | Alignment]

Lexical Translation

$$p(e_i \mid f_{a_i})$$

Lexical Translation

$$p(e_i \mid f_{a_i})$$



$$p(\texttt{house} \mid \texttt{Haus})$$

Lexical Translation

$$p(e_i \mid f_{a_i})$$

$$p(\texttt{house} \mid \texttt{Haus}) \quad p(\texttt{shell} \mid \texttt{Haus})$$

Lexical Translation

$$\mathbf{e} = \langle e_1, e_2, ...e_m \rangle \qquad \mathbf{f} = \langle f_1, f_2, ...f_n \rangle$$

$$\mathbf{a} = \langle a_1, a_2, ...a_m \rangle \quad \text{each } a_i \in \{0, 1, ..., n\}$$

Modeling assumptions

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in \{0,1,..,n\}^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

[Alignment] x [Translation | Alignment]

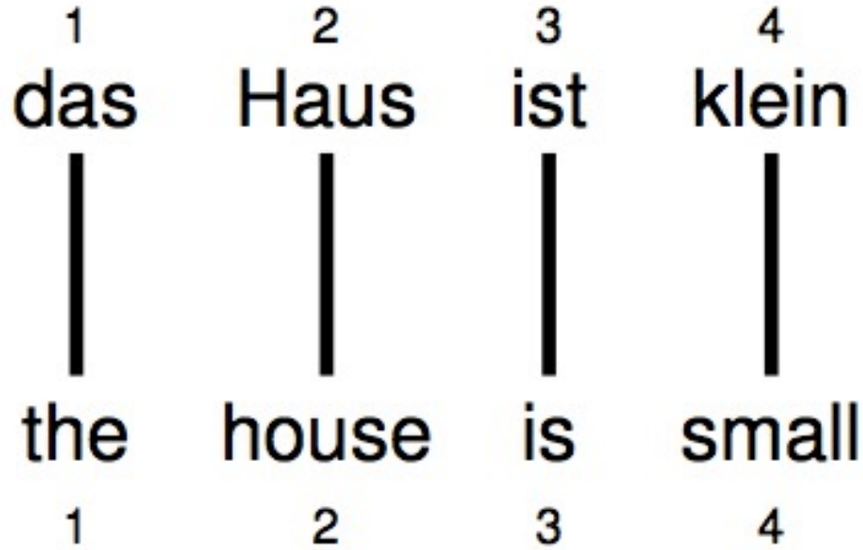
Alignment

$$p(\mathbf{a} \mid \mathbf{f}, m)$$

Most of the action for the first 10 years of MT was here. Words weren't the problem, word *order* was hard.

Alignment

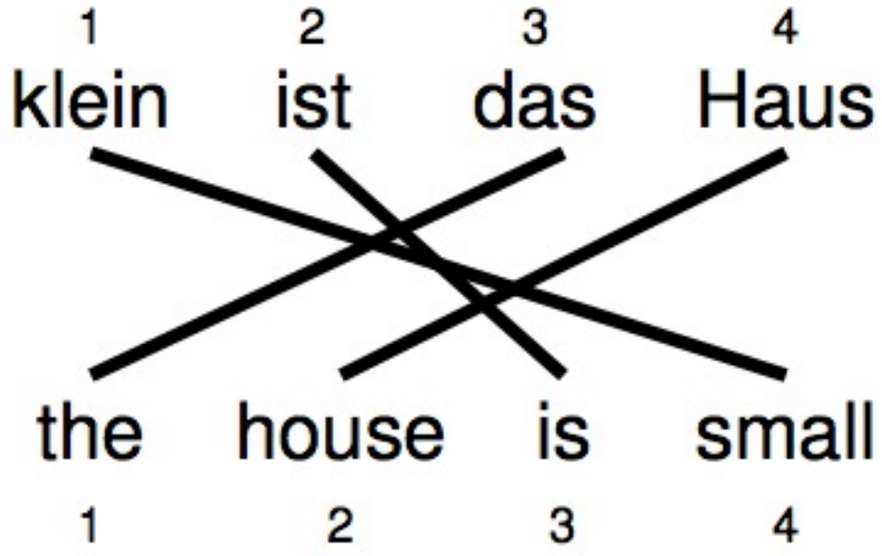
- Alignments can be visualized in by drawing links between two sentences, and they are represented as vectors of positions:



$$\mathbf{a} = (1, 2, 3, 4)^\top$$

Reordering

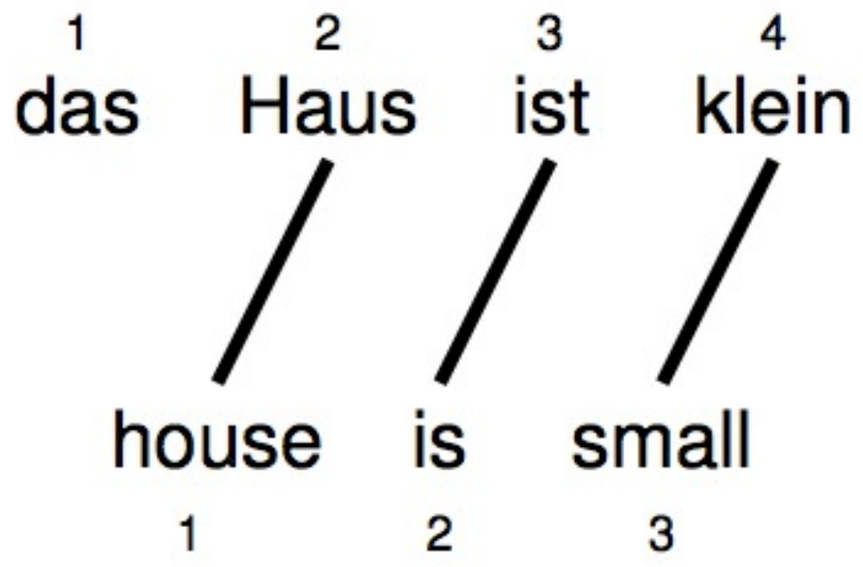
- Words may be reordered during translation.



$$\mathbf{a} = (3, 4, 2, 1)^\top$$

Word Dropping

- A source word may not be translated at all



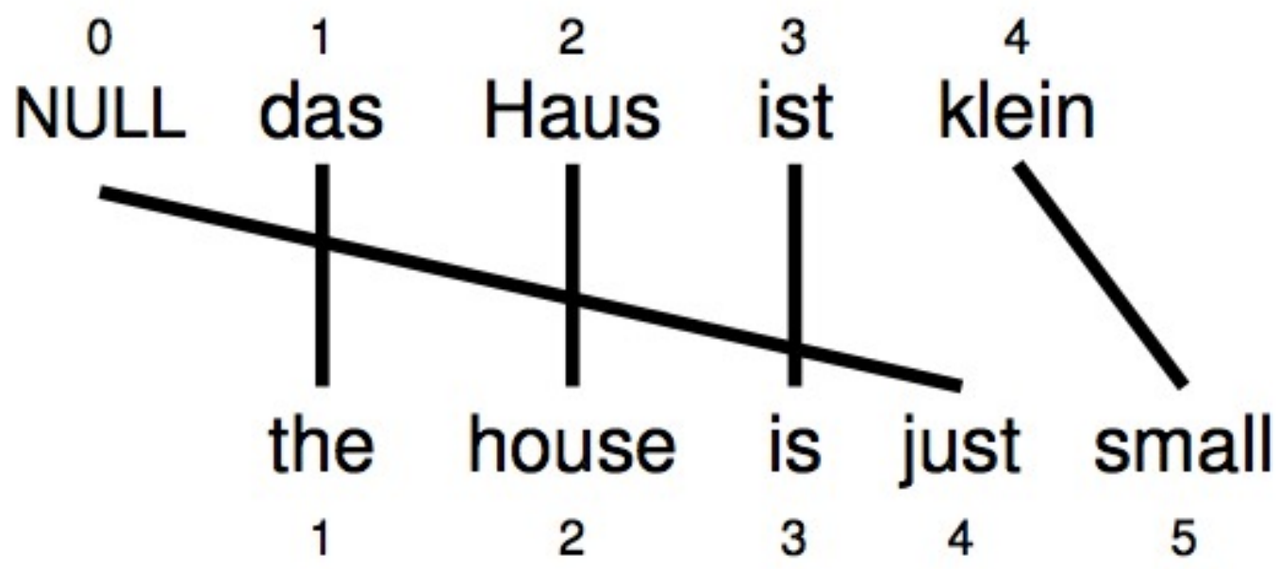
$$\mathbf{a} = (2, 3, 4)^\top$$

Word Insertion

- Words may be inserted during translation

  English just does not have an equivalent

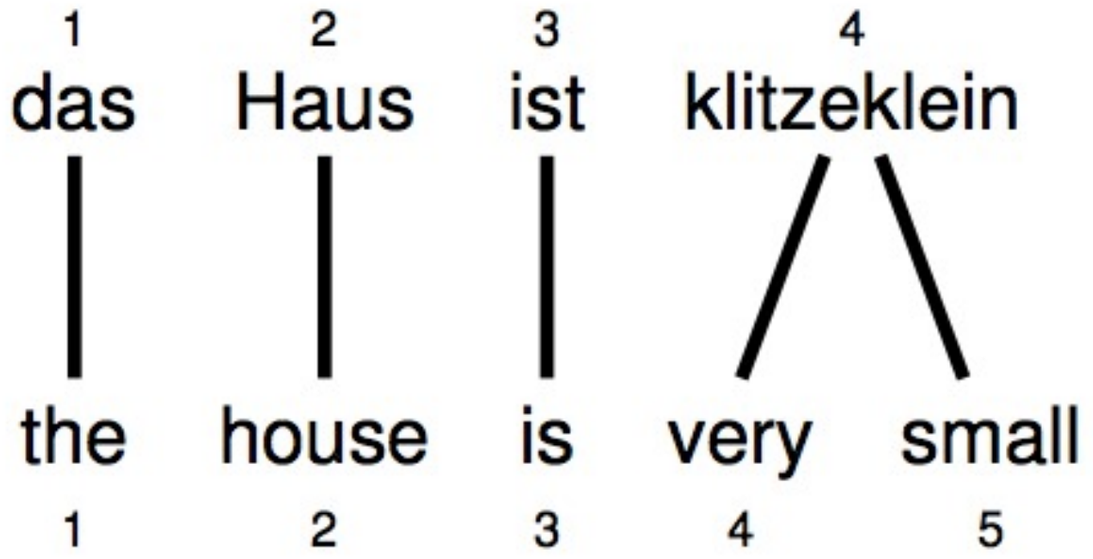
  But it must be explained - we typically assume every source sentence contains a NULL token



$$\mathbf{a} = (1, 2, 3, 0, 4)^\top$$

One-to-many Translation

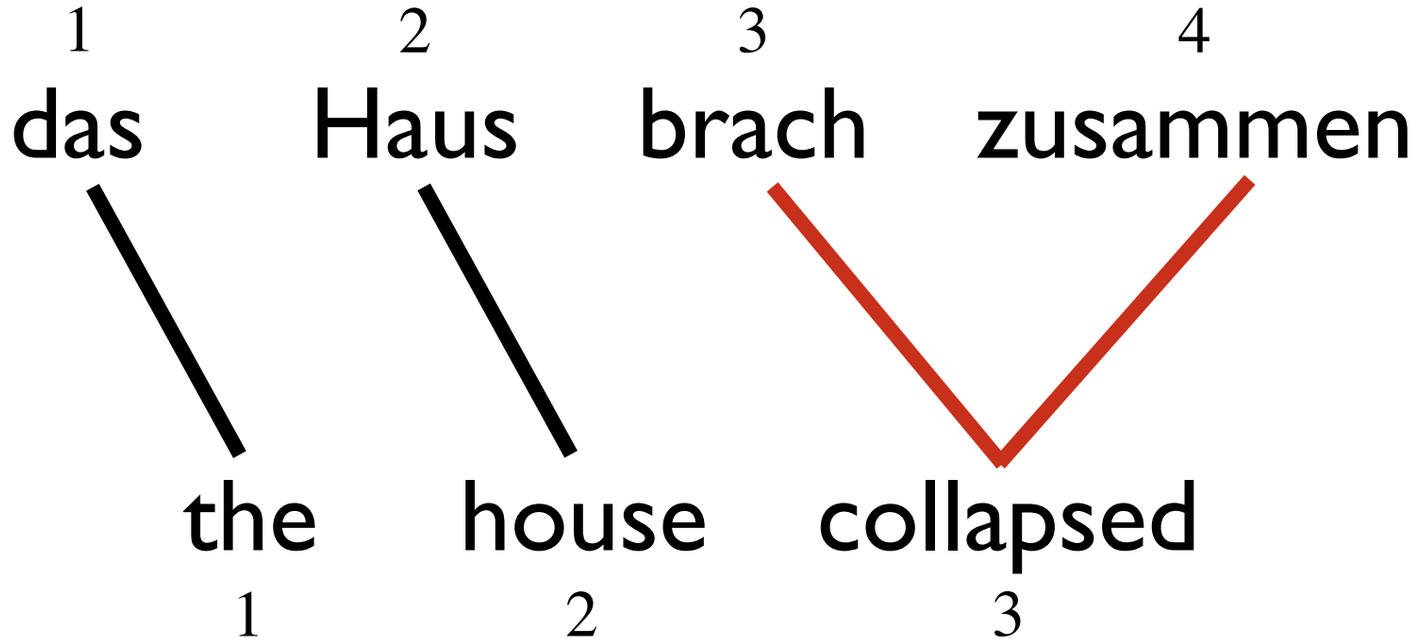
- A source word may translate into **more than one** target word



$$\mathbf{a} = (1, 2, 3, 4, 4)^\top$$

Many-to-one Translation

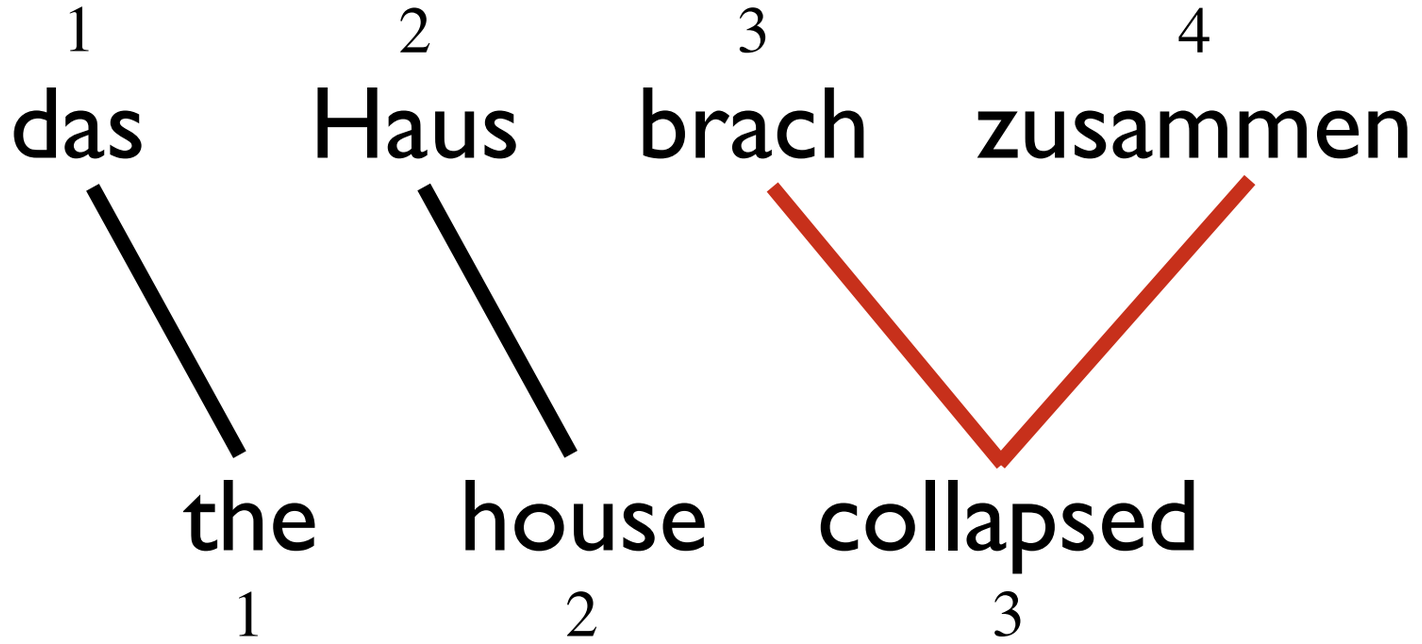
- **More than one source word** may **not** translate as a unit in lexical translation



$\mathbf{a} = ???$

Many-to-one Translation

- **More than one source word** may **not** translate as a unit in lexical translation



$\mathbf{a} = ???$ $\mathbf{a} = (1, 2, (3, 4)^\top)^\top$ **?**

IBM Model 1

- Simplest possible lexical translation model
- Additional assumptions
  - The m alignment decisions are independent
  - The alignment distribution for each a_i is uniform over all source words and NULL

$$\begin{aligned}
&\text{for each } i \in [1, 2, \ldots, m] \\
&\qquad a_i \sim \text{Uniform}(0, 1, 2, \ldots, n) \\
&\qquad e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})
\end{aligned}$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m}$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n}$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i, a_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

$$p(a, b, c, d) = p(a)p(b)p(c)p(d)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

$$= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

$$= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$= \frac{1}{(1+n)^m} \prod_{i=1}^{m} \sum_{a_i=0}^{n} p(e_i \mid f_{a_i})$$

Example

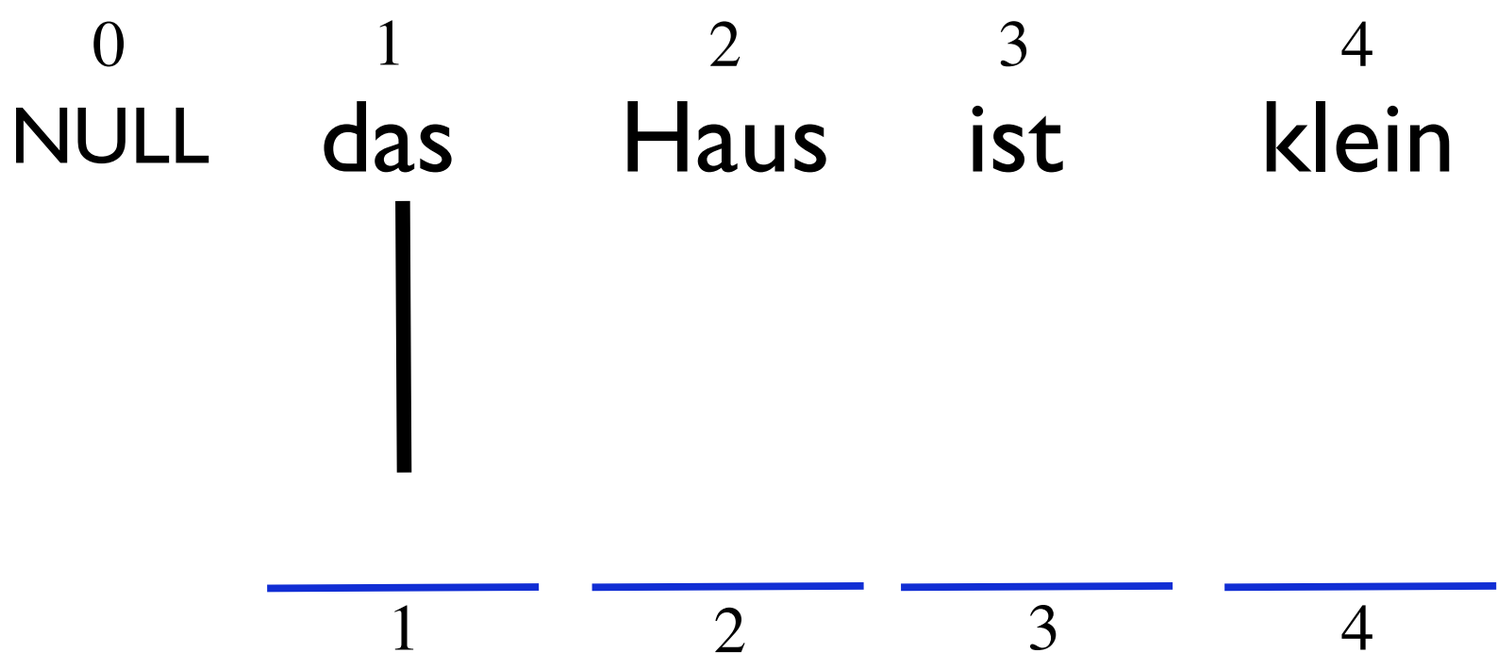
0	1	2	3	4
NULL	das	Haus	ist	klein

Start with a foreign sentence and a target length.

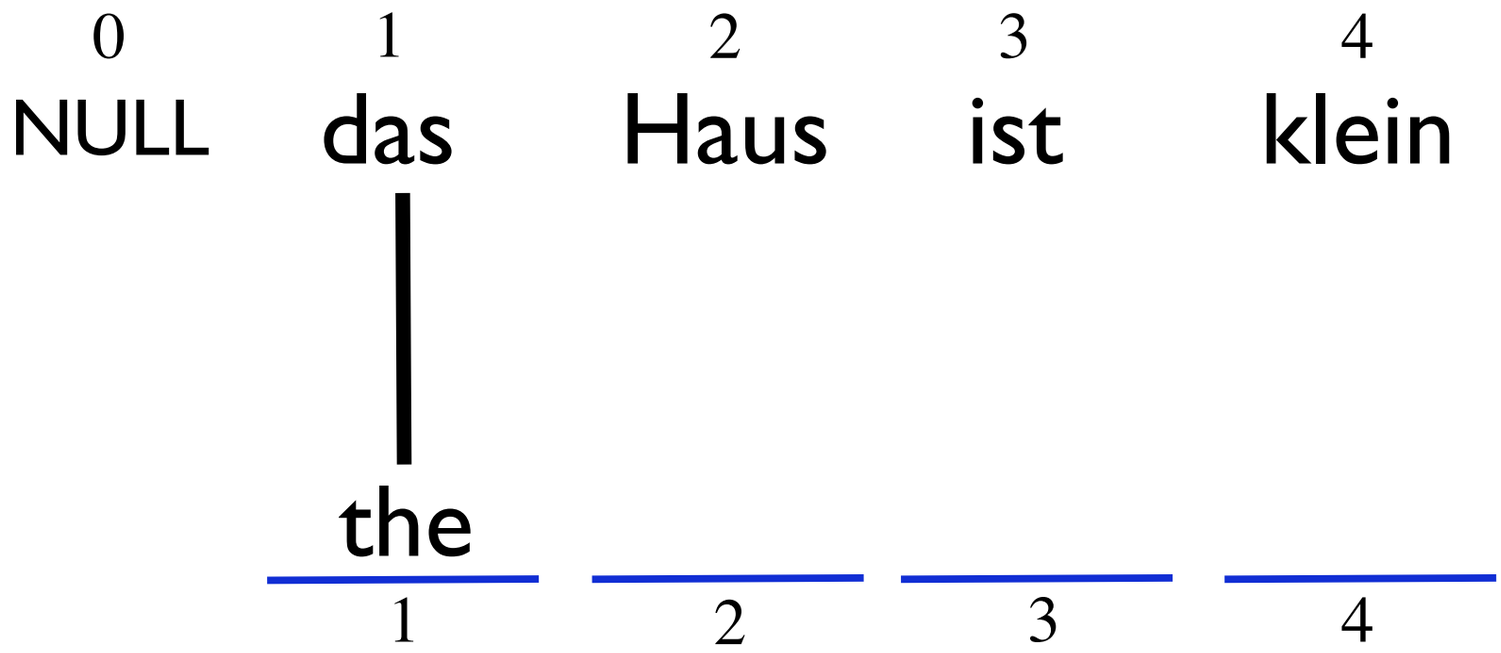
Example

0	1	2	3	4
NULL	das	Haus	ist	klein

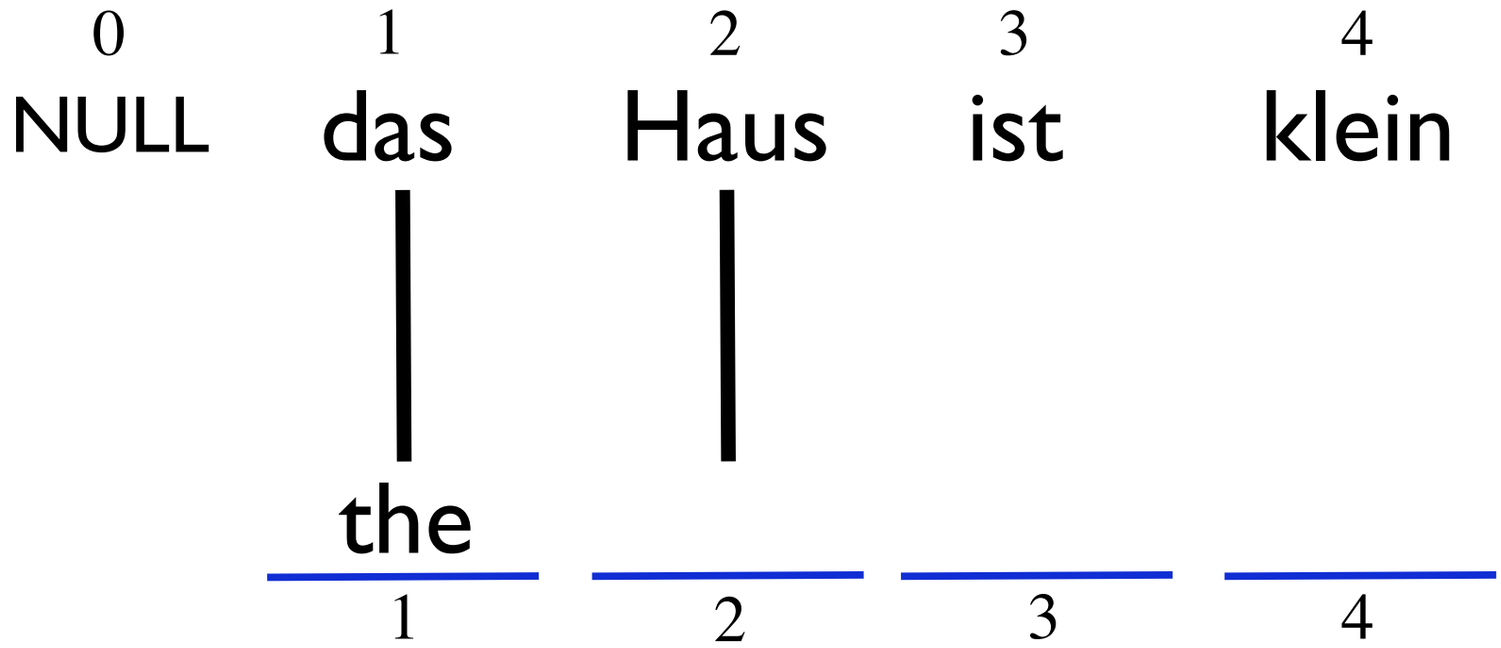
Example



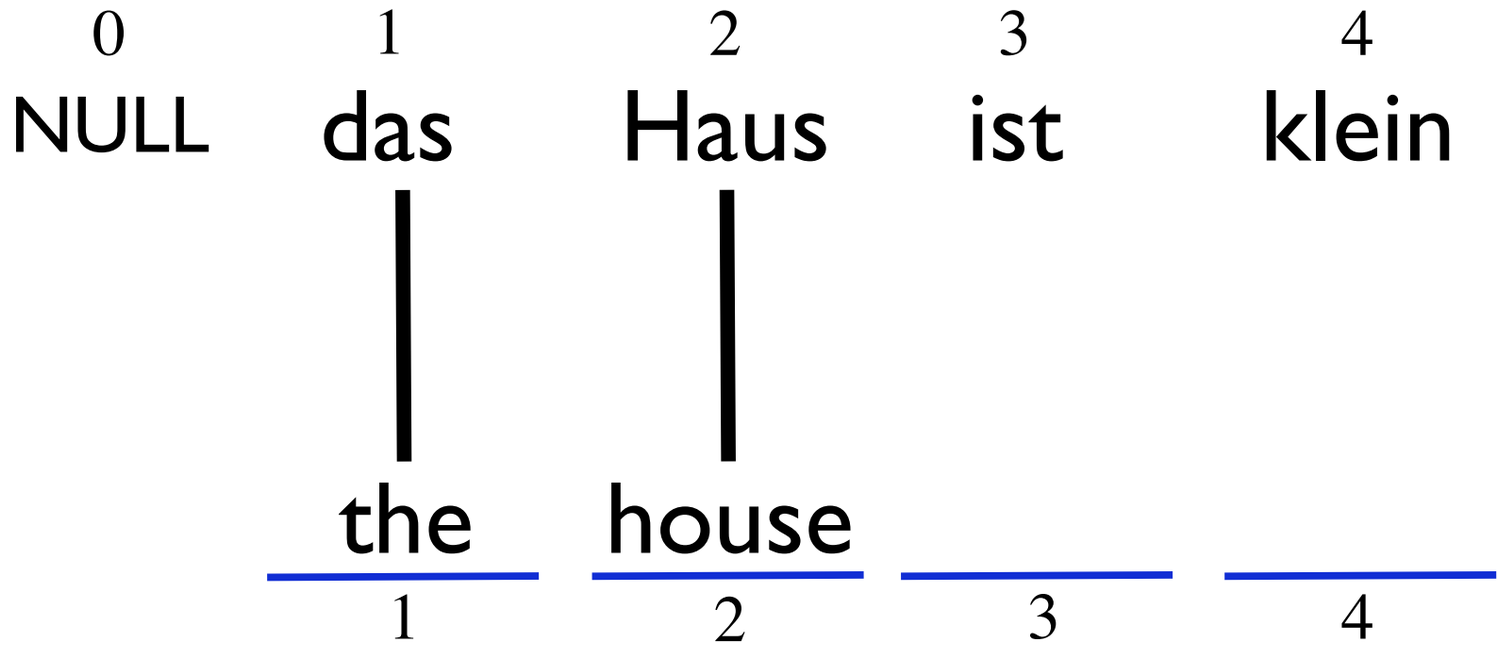
Example



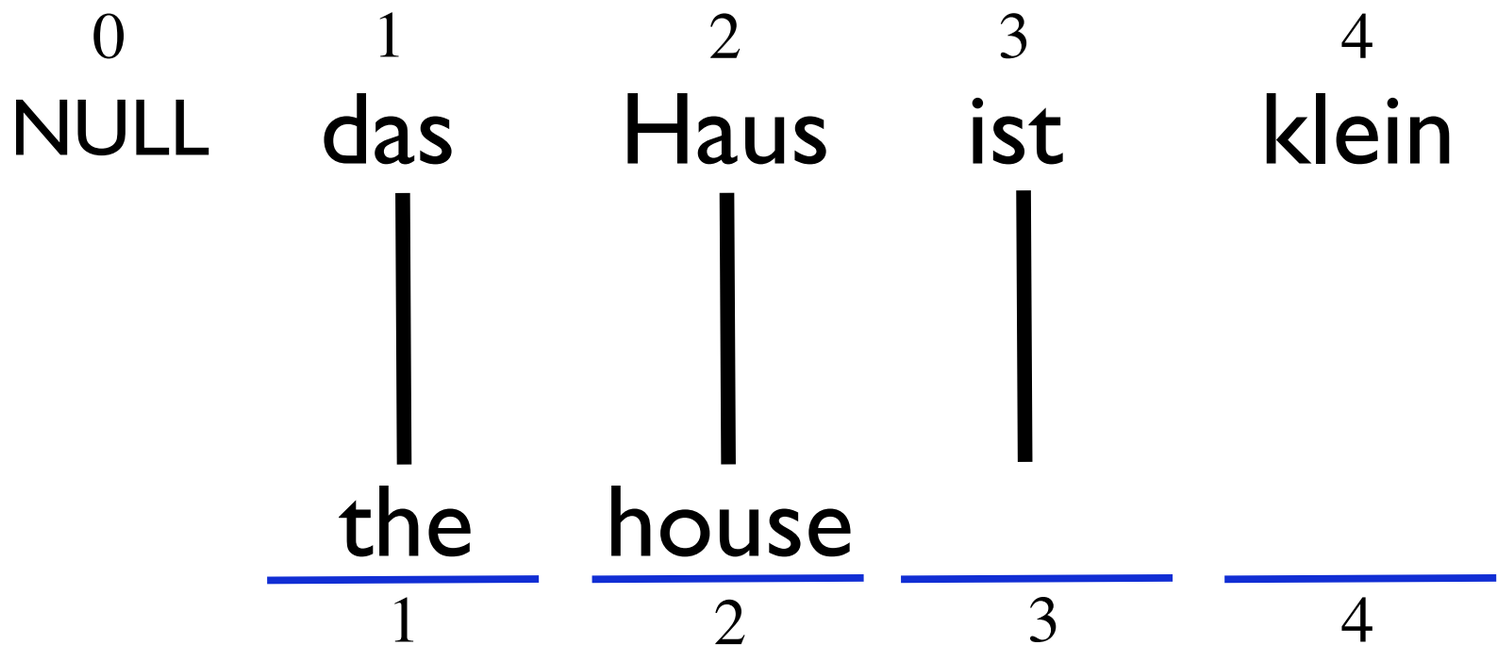
Example



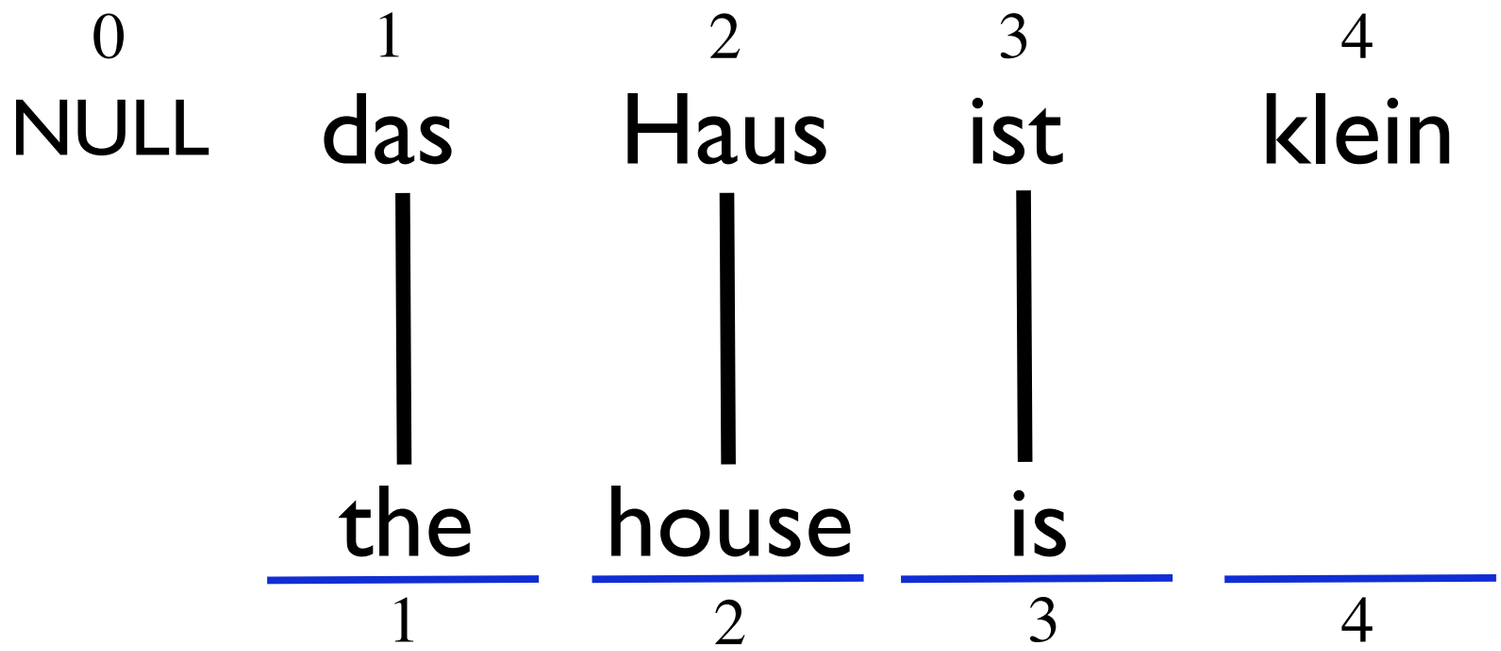
Example



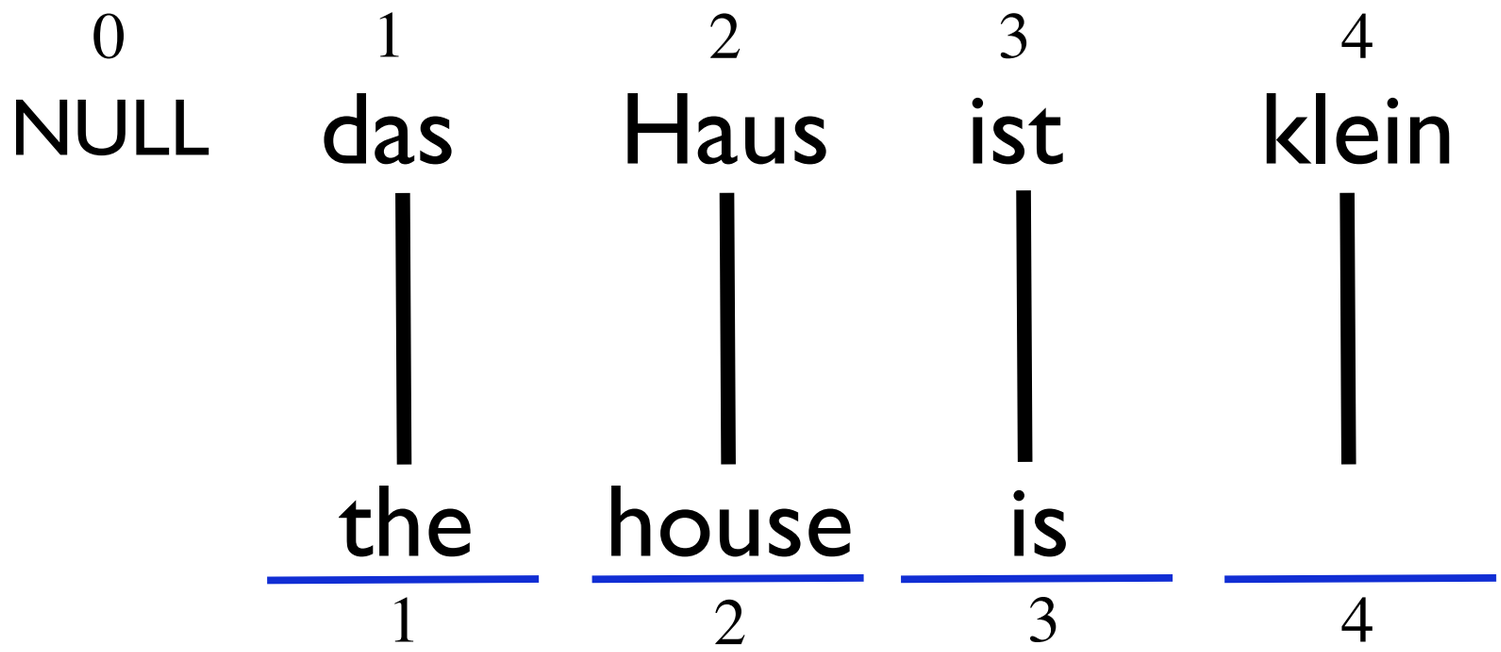
Example



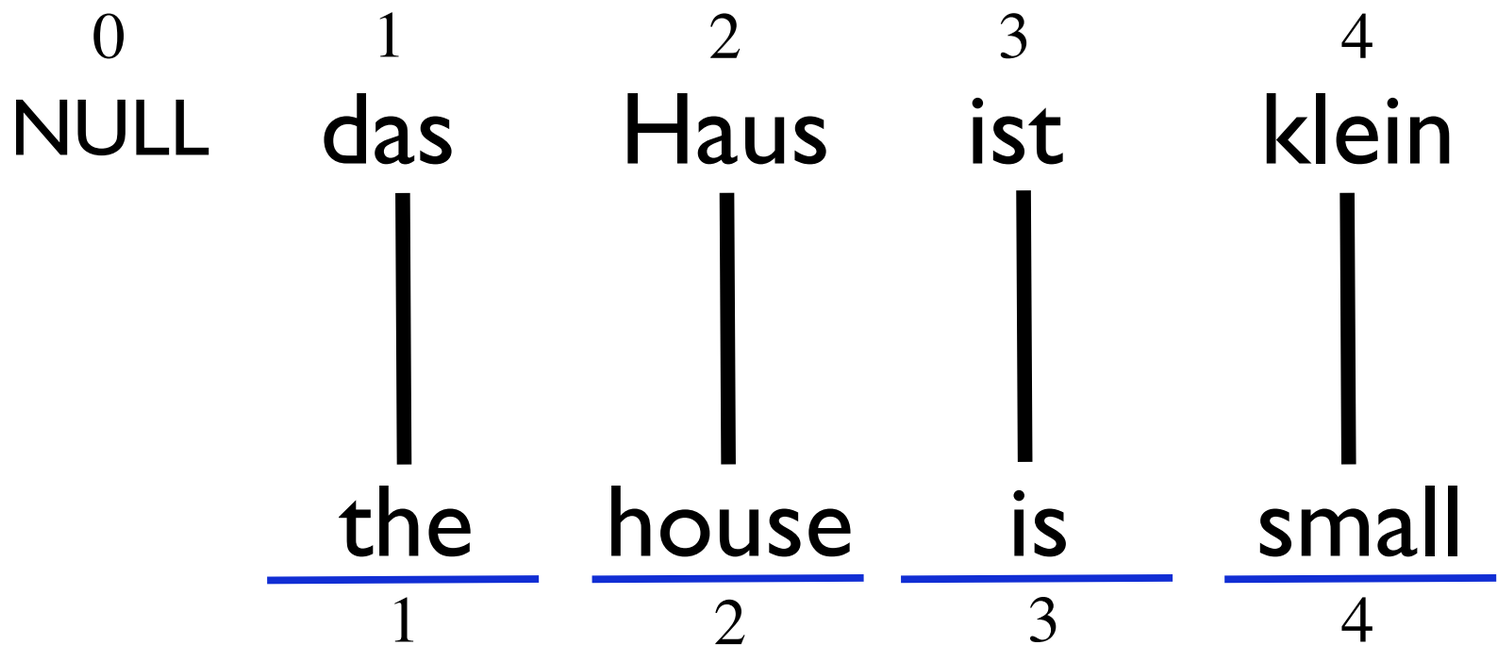
Example



Example



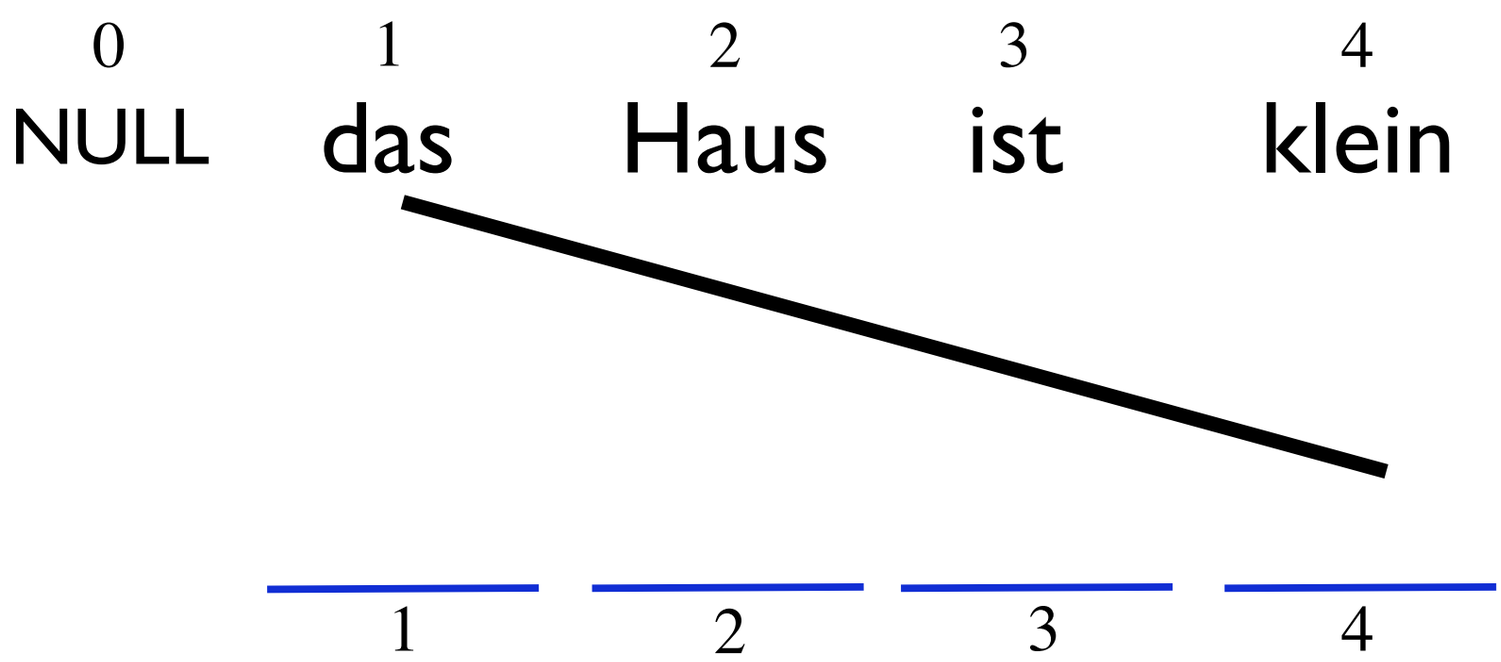
Example



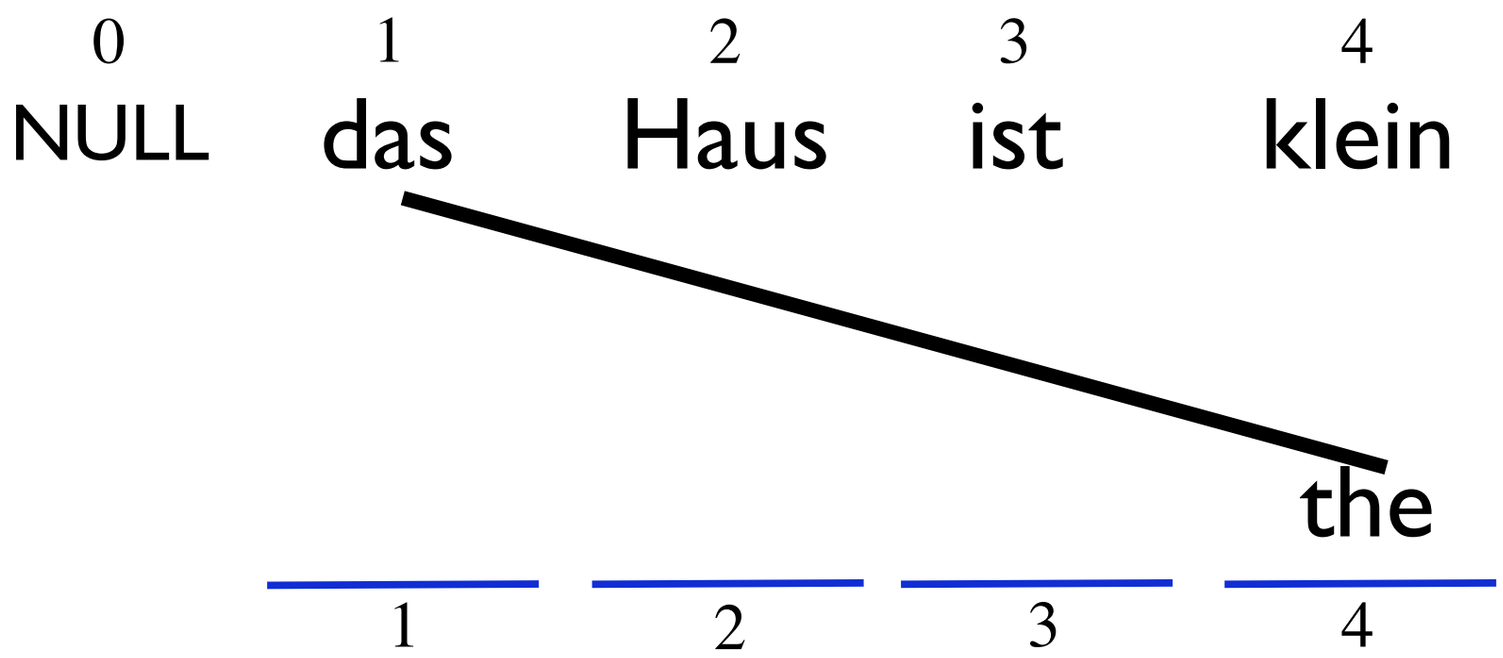
Example

0	1	2	3	4
NULL	das	Haus	ist	klein

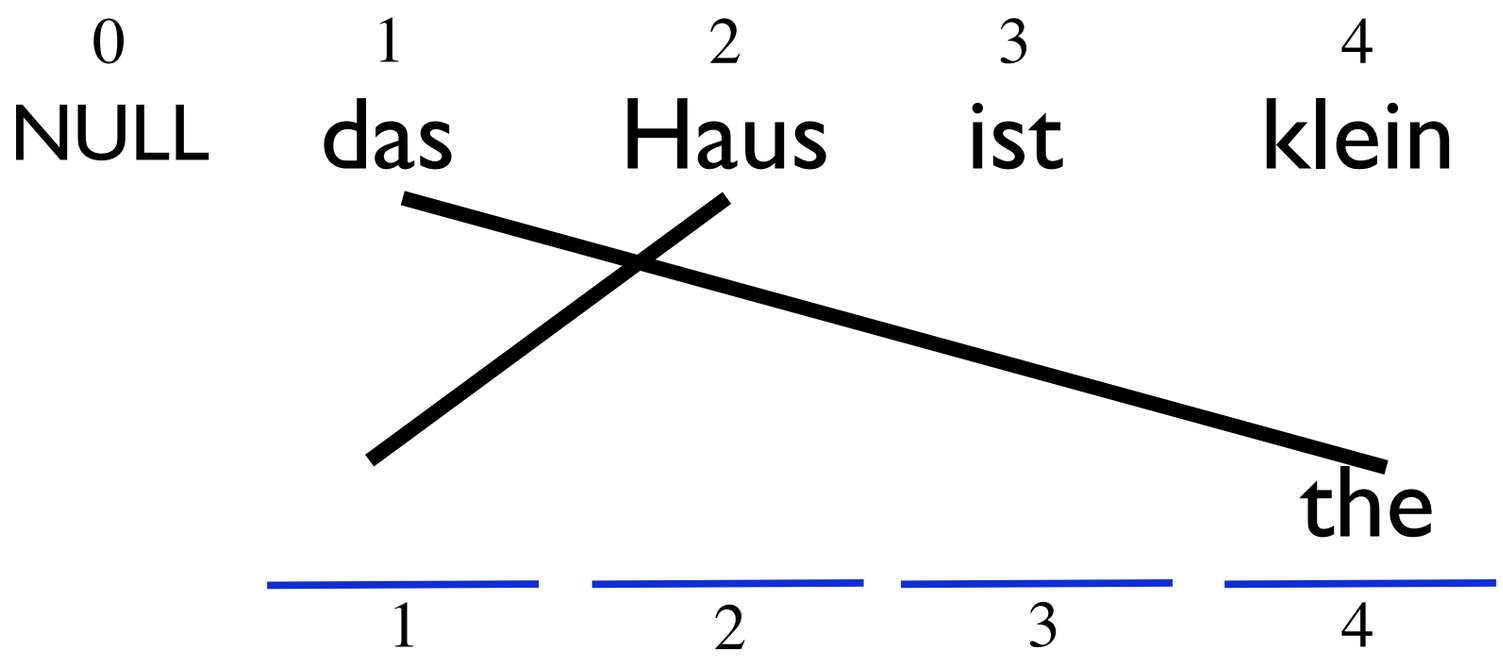
Example



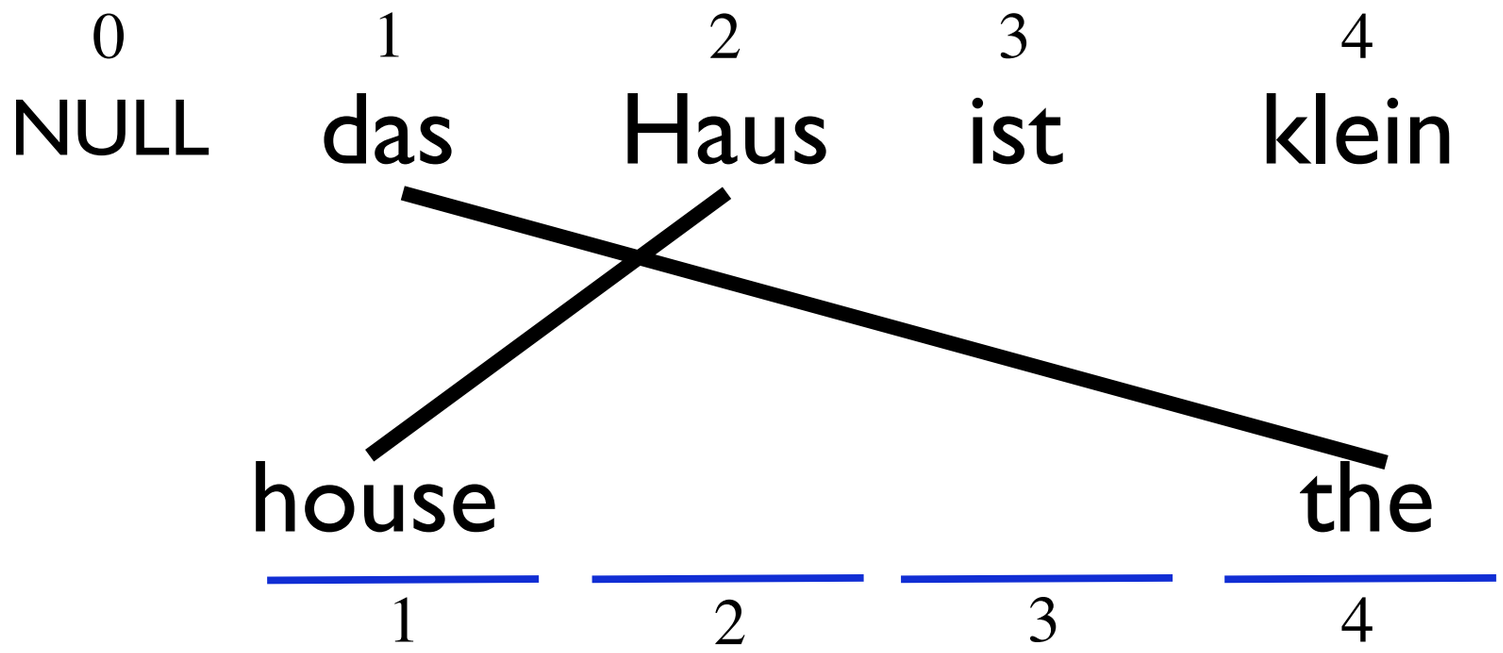
Example



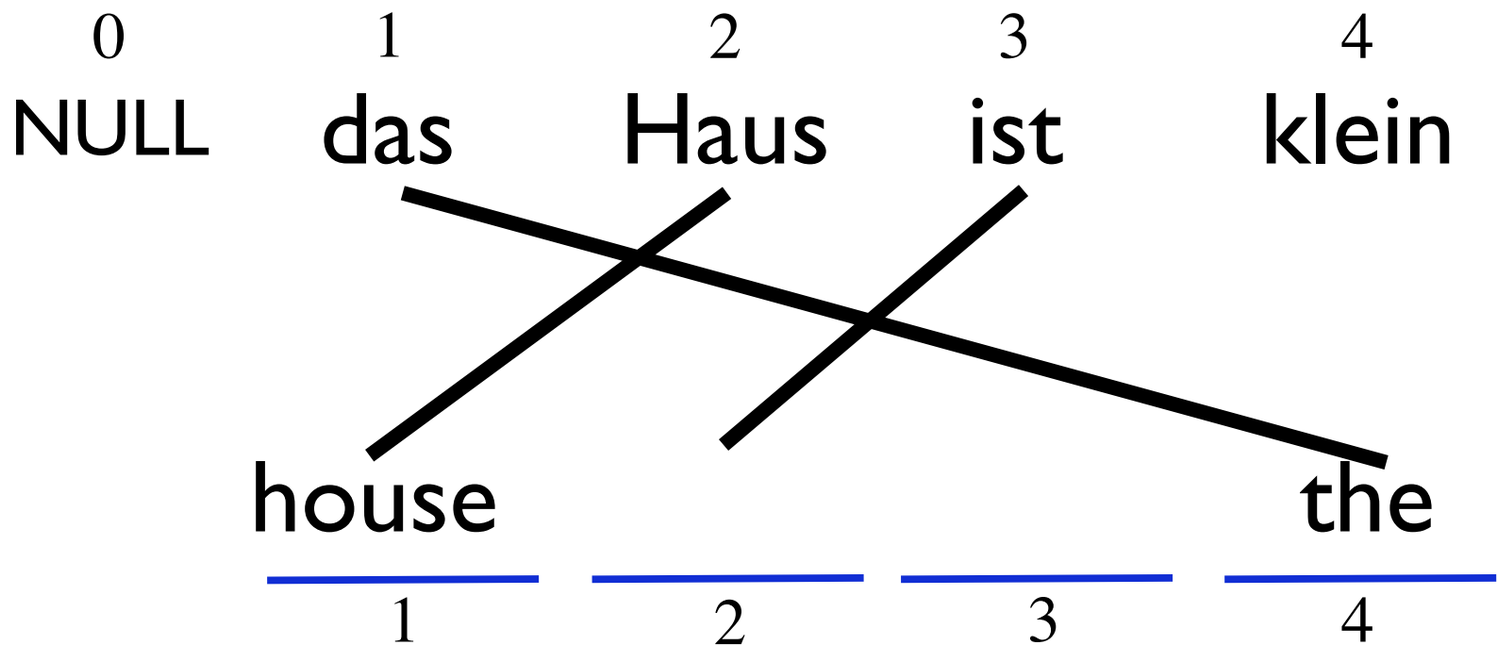
Example



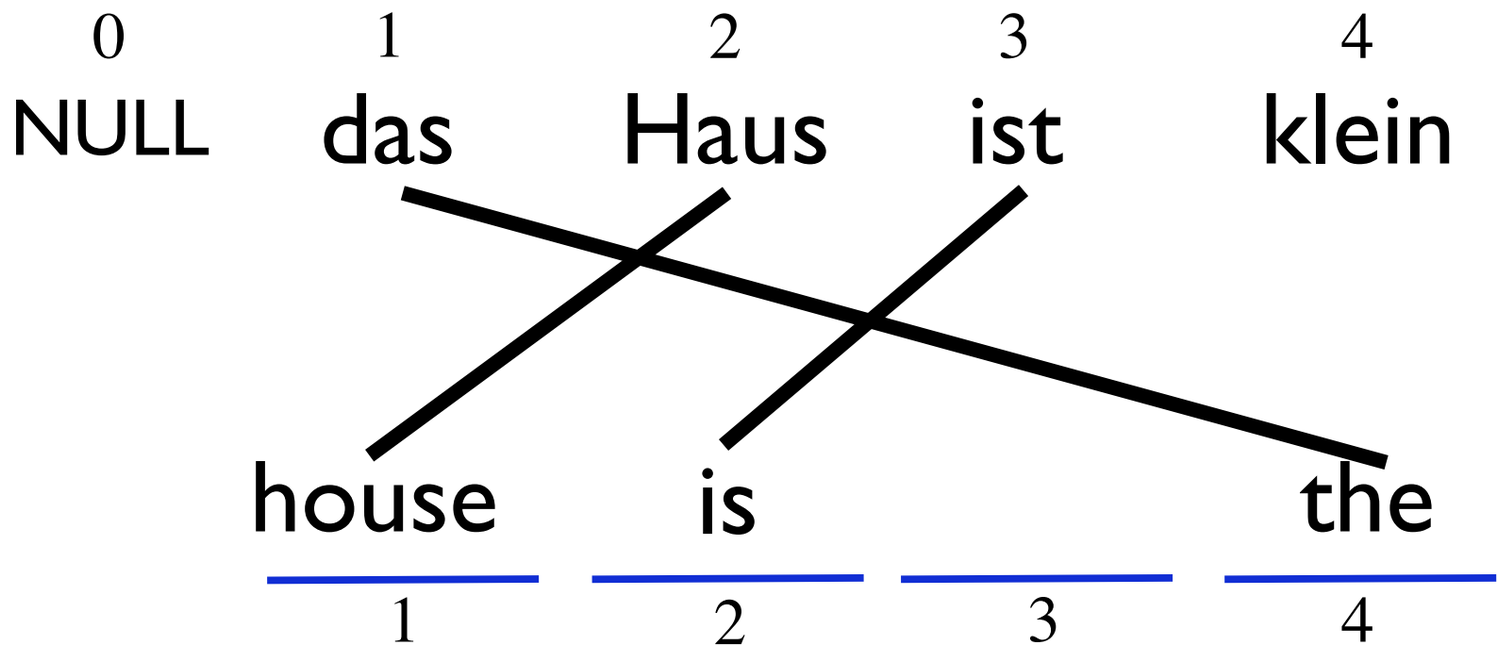
Example



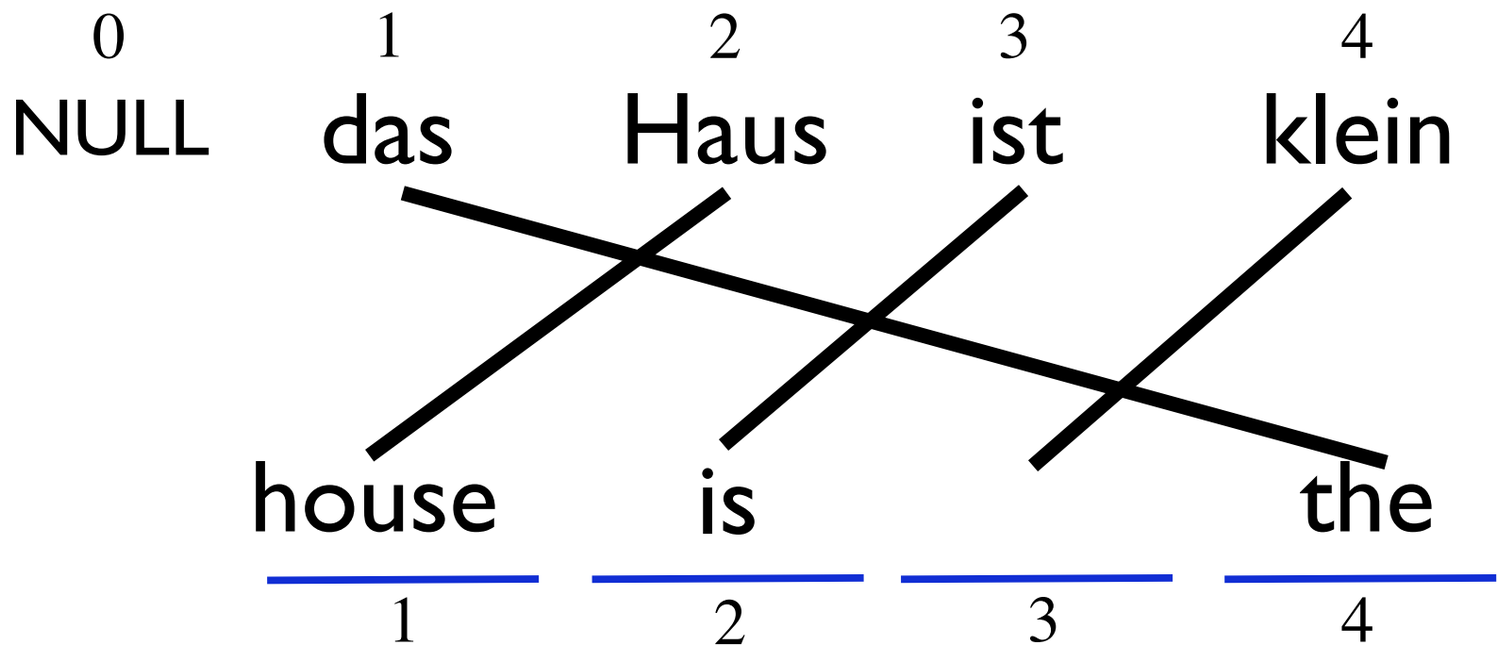
Example



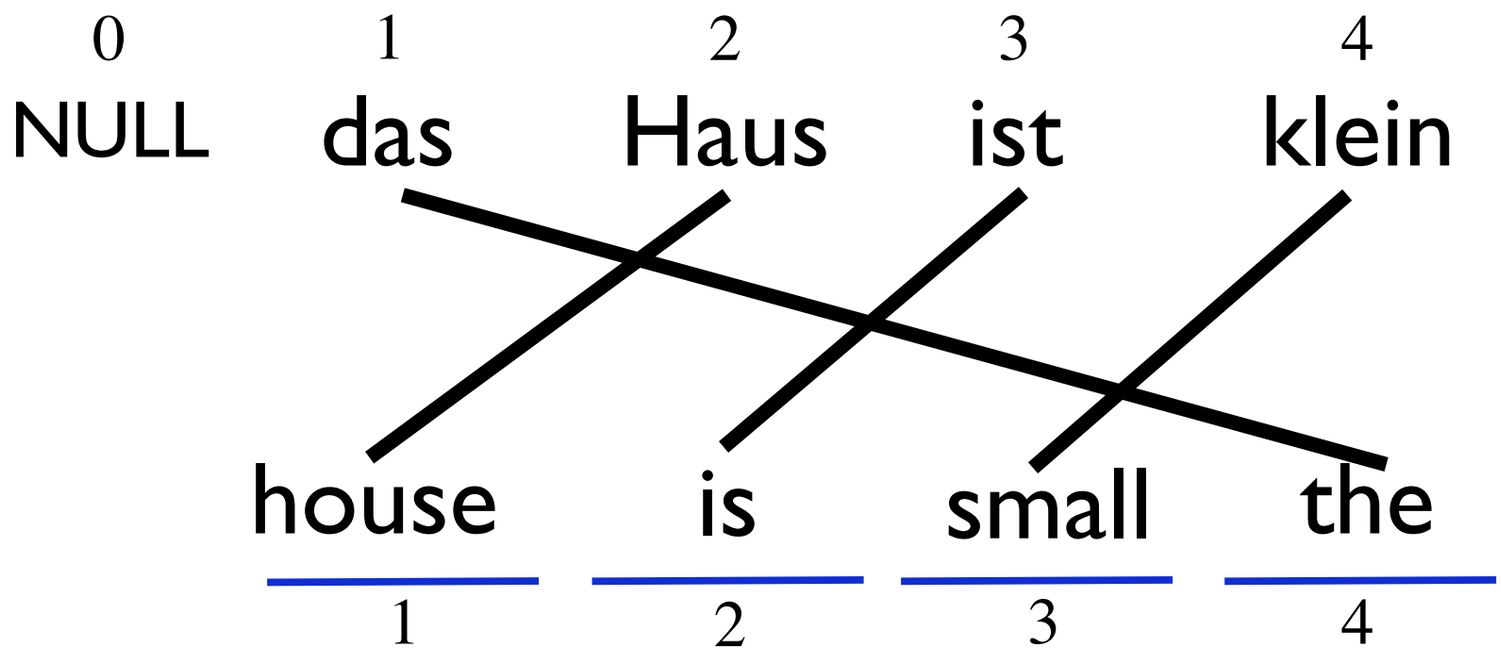
Example



Example



Example



What is this good for?

- 1. Evaluate translation quality
- 2. Find the best alignment
  - The IBM models are still used for this, though not as translation models

Finding the Viterbi Alignment

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in [0,1,\ldots,n]^m} p(\mathbf{a} \mid \mathbf{e}, \mathbf{f})$$

$$= \arg \max_{\mathbf{a} \in [0,1,\ldots,n]^m} \frac{p(\mathbf{e}, \mathbf{a} \mid \mathbf{f})}{\sum_{\mathbf{a}'} p(\mathbf{e}, \mathbf{a}' \mid \mathbf{f})}$$

$$= \arg \max_{\mathbf{a} \in [0,1,\ldots,n]^m} p(\mathbf{e}, \mathbf{a} \mid \mathbf{f})$$

$$= p(\mathbf{a}|\mathbf{f})\ p(\mathbf{e}|\mathbf{a}, \mathbf{f})$$

$$a_i^* = \arg \max_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

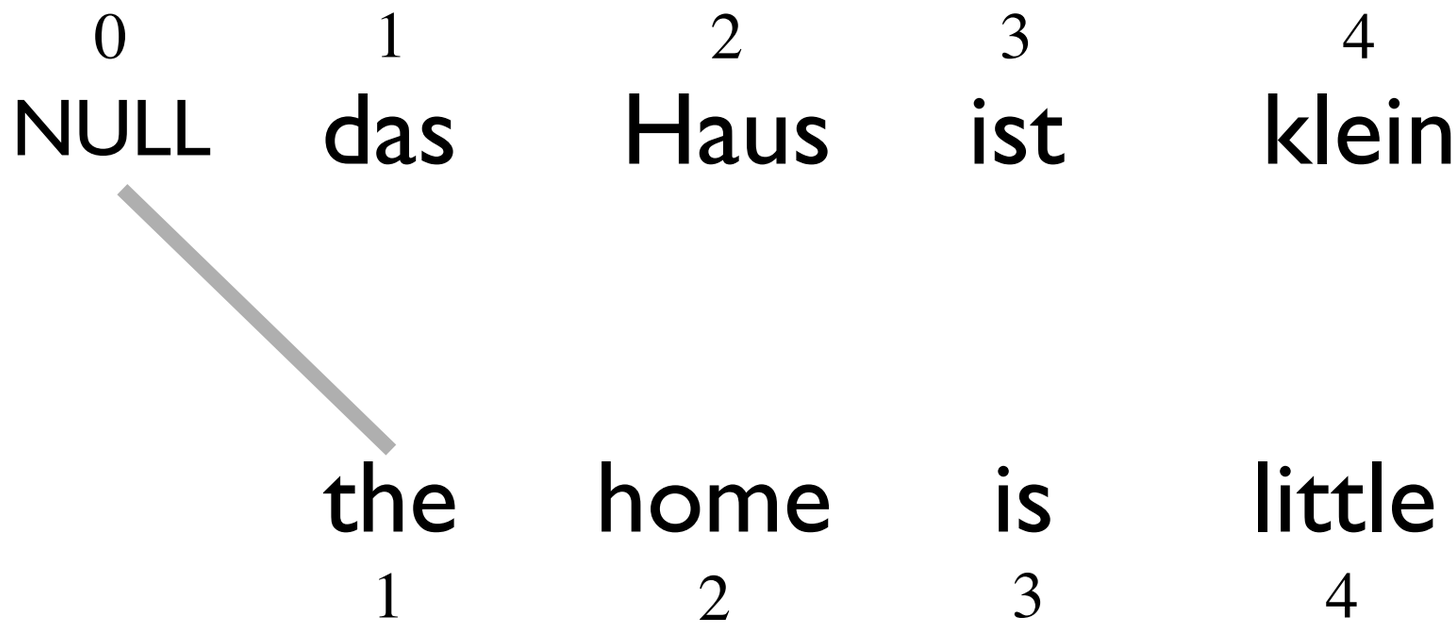
$$= \arg \max_{a_i=0}^{n} p(e_i \mid f_{a_i})$$

Finding the Viterbi Alignment

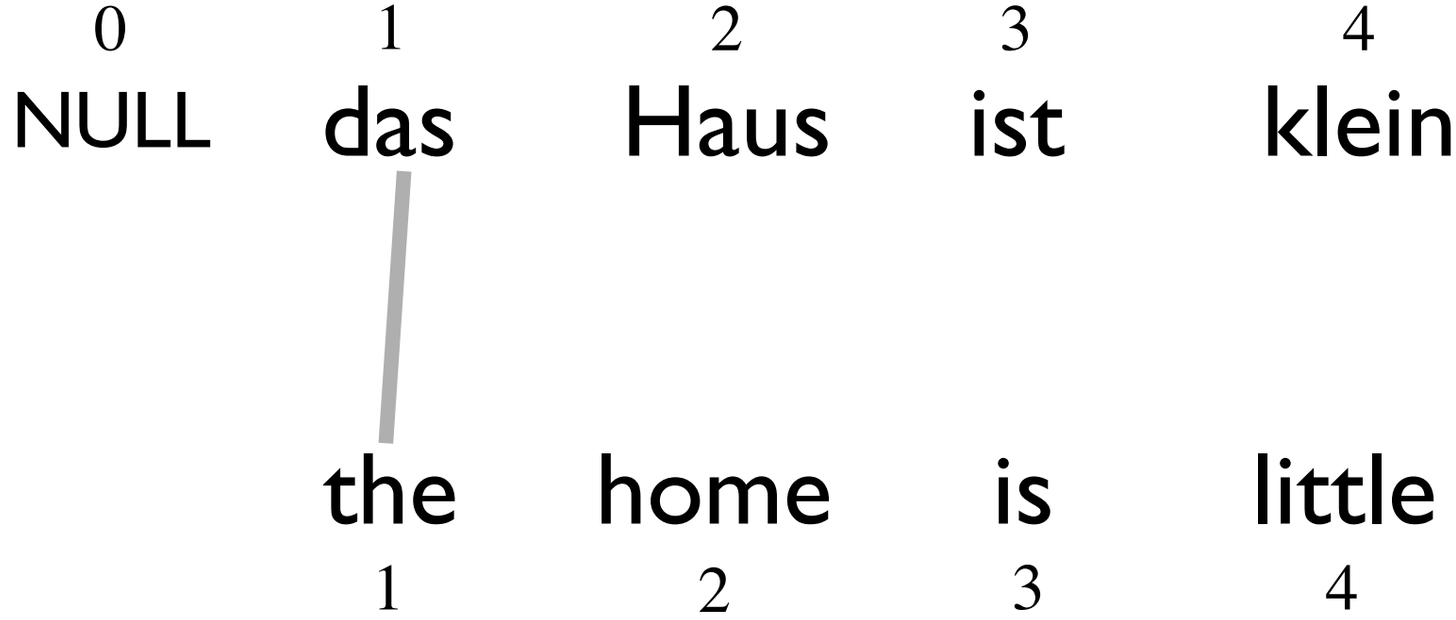
0	1	2	3	4
NULL	das	Haus	ist	klein

the	home	is	little
1	2	3	4

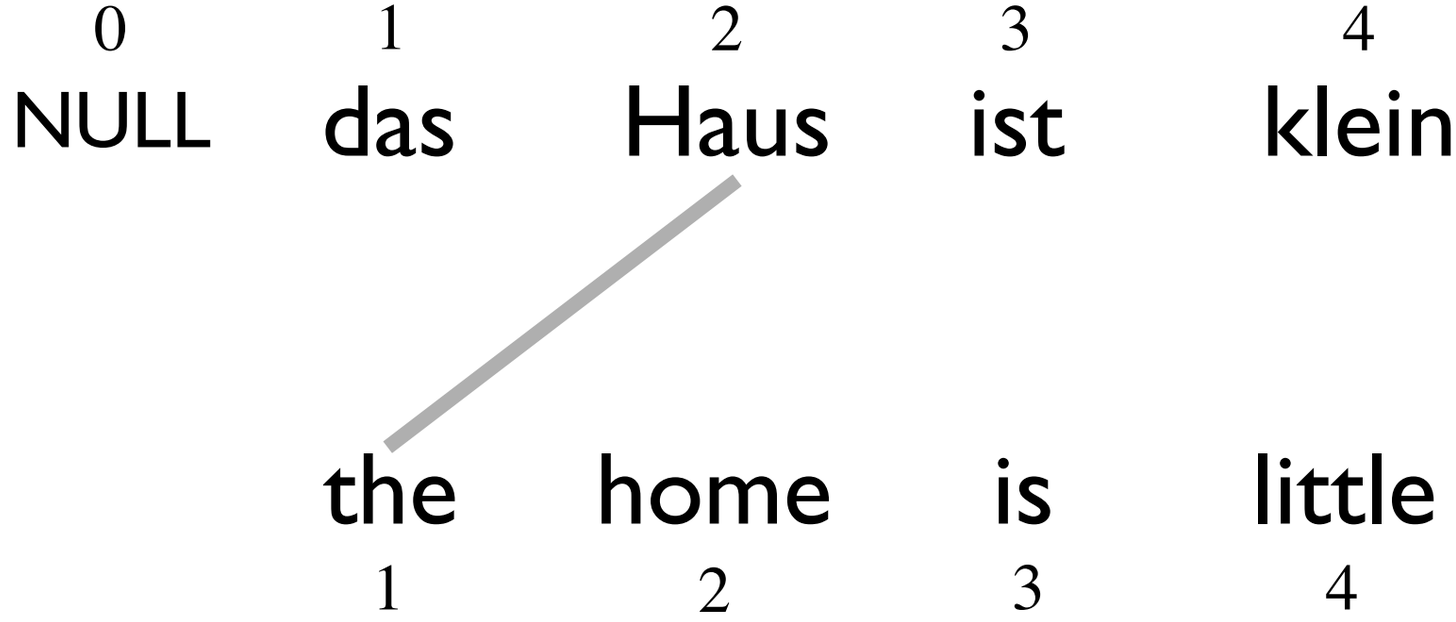
Finding the Viterbi Alignment



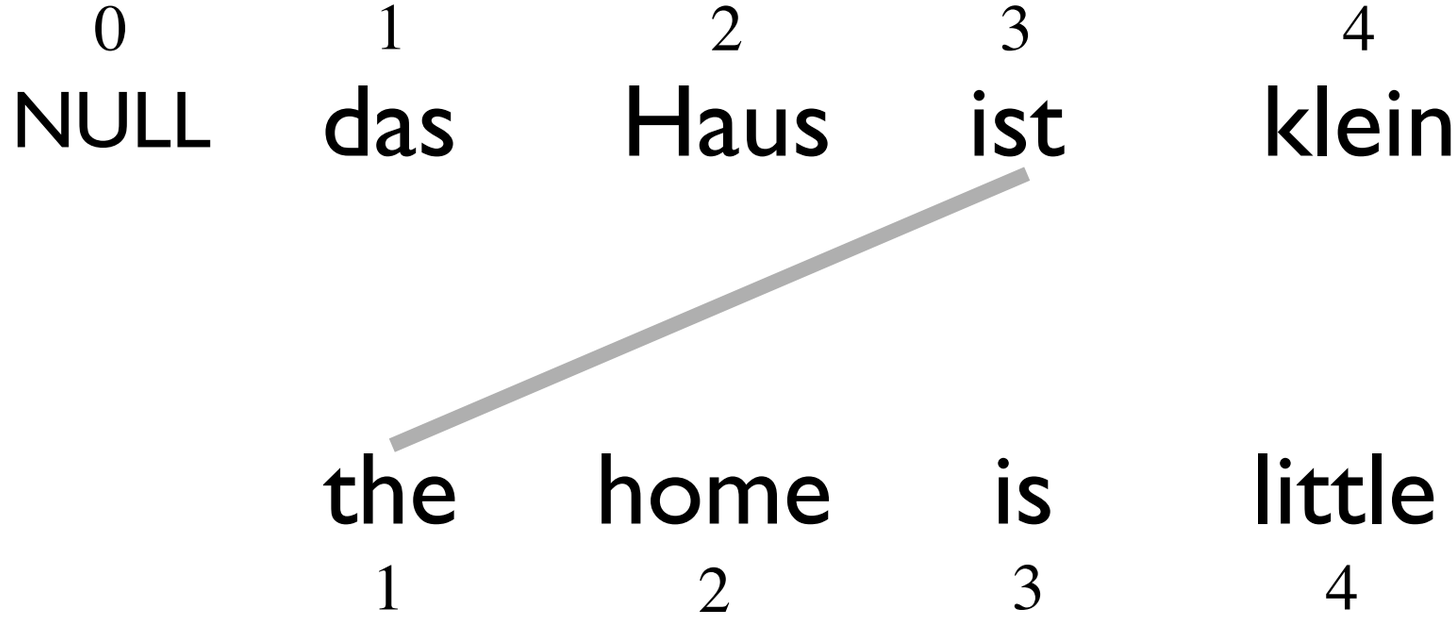
Finding the Viterbi Alignment



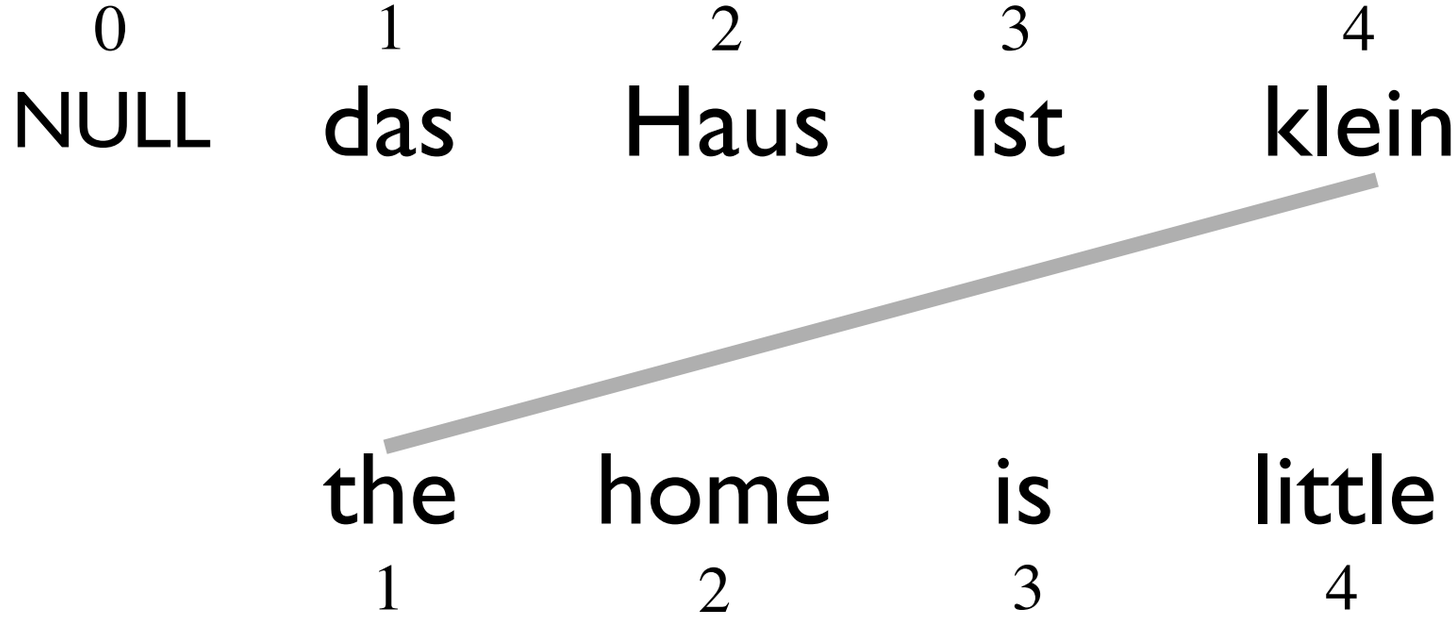
Finding the Viterbi Alignment



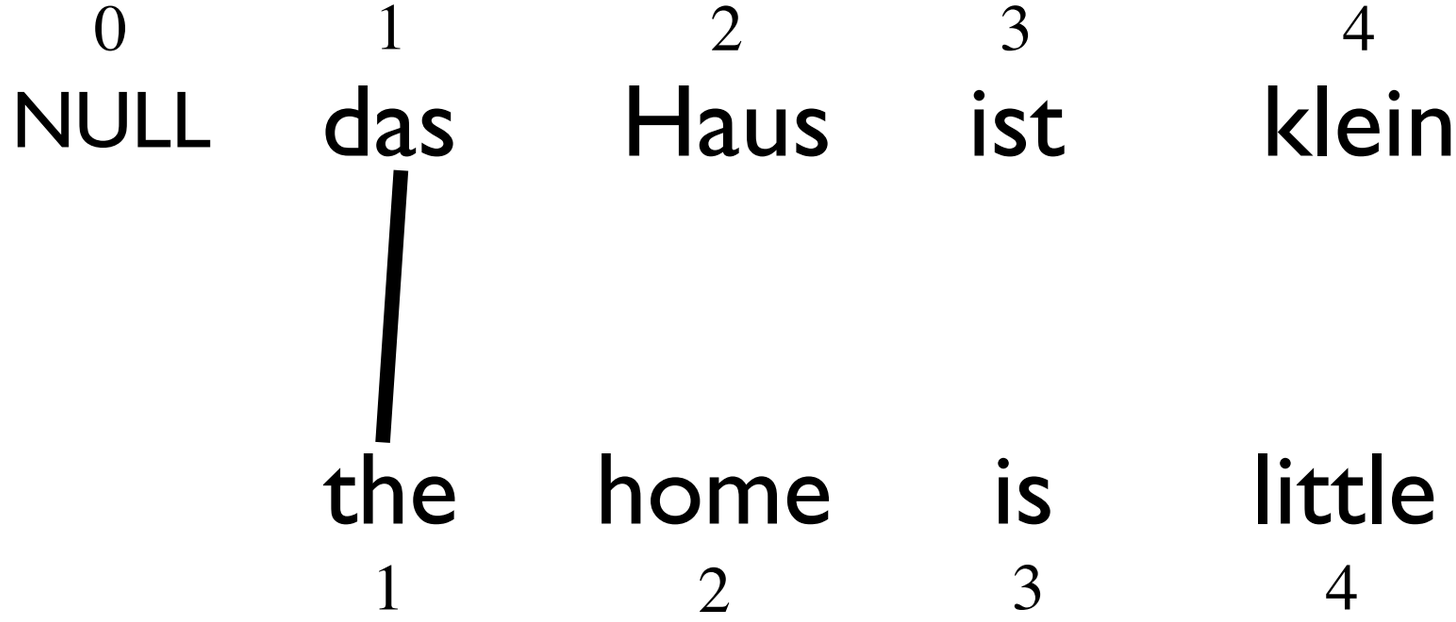
Finding the Viterbi Alignment



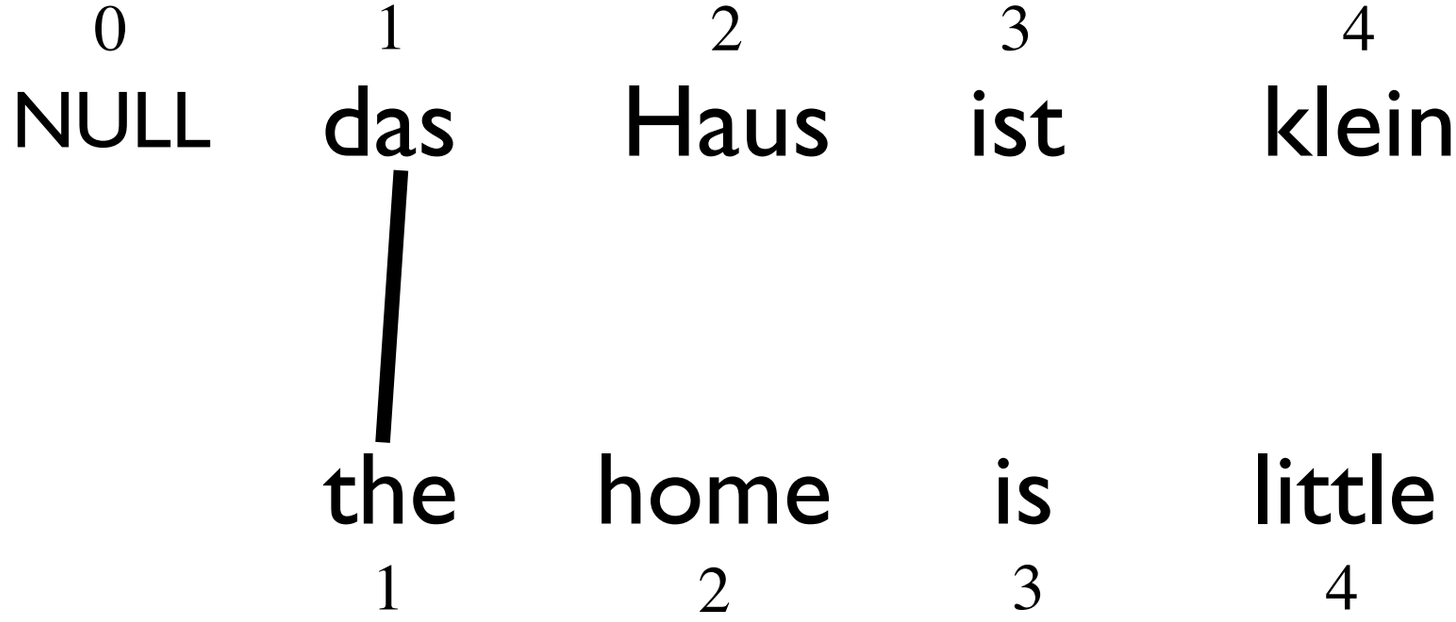
Finding the Viterbi Alignment



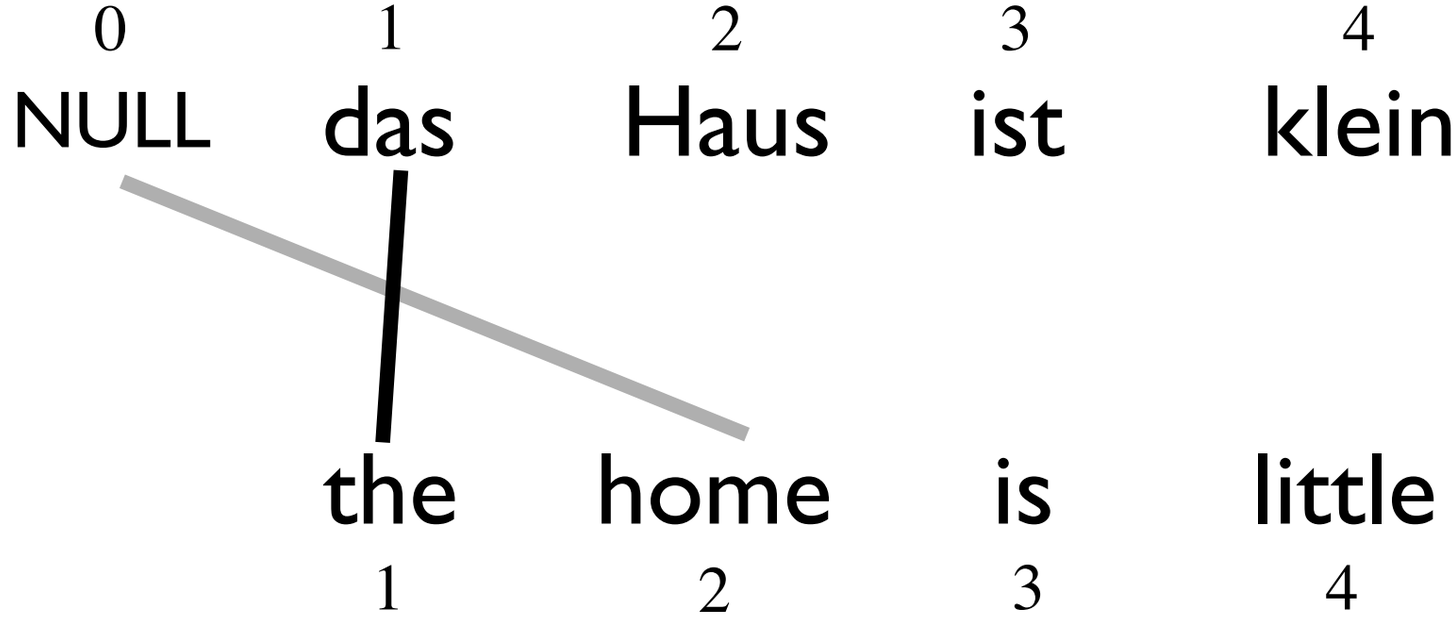
Finding the Viterbi Alignment



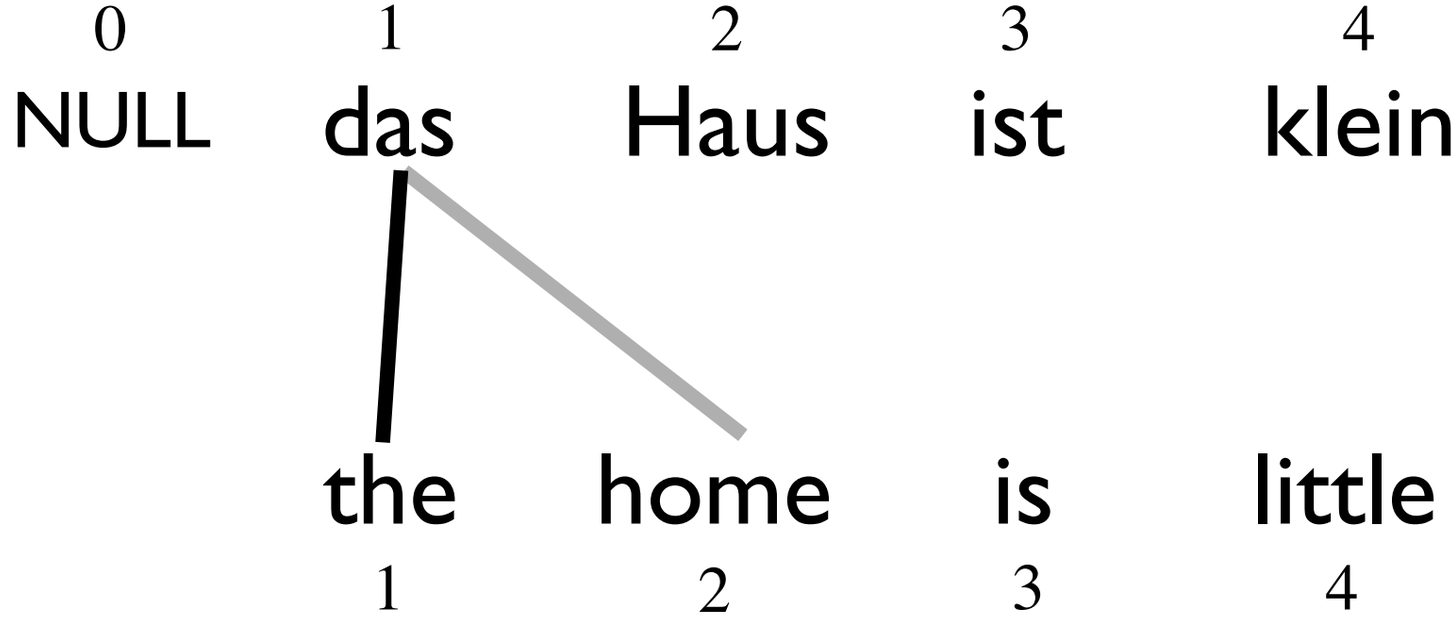
Finding the Viterbi Alignment



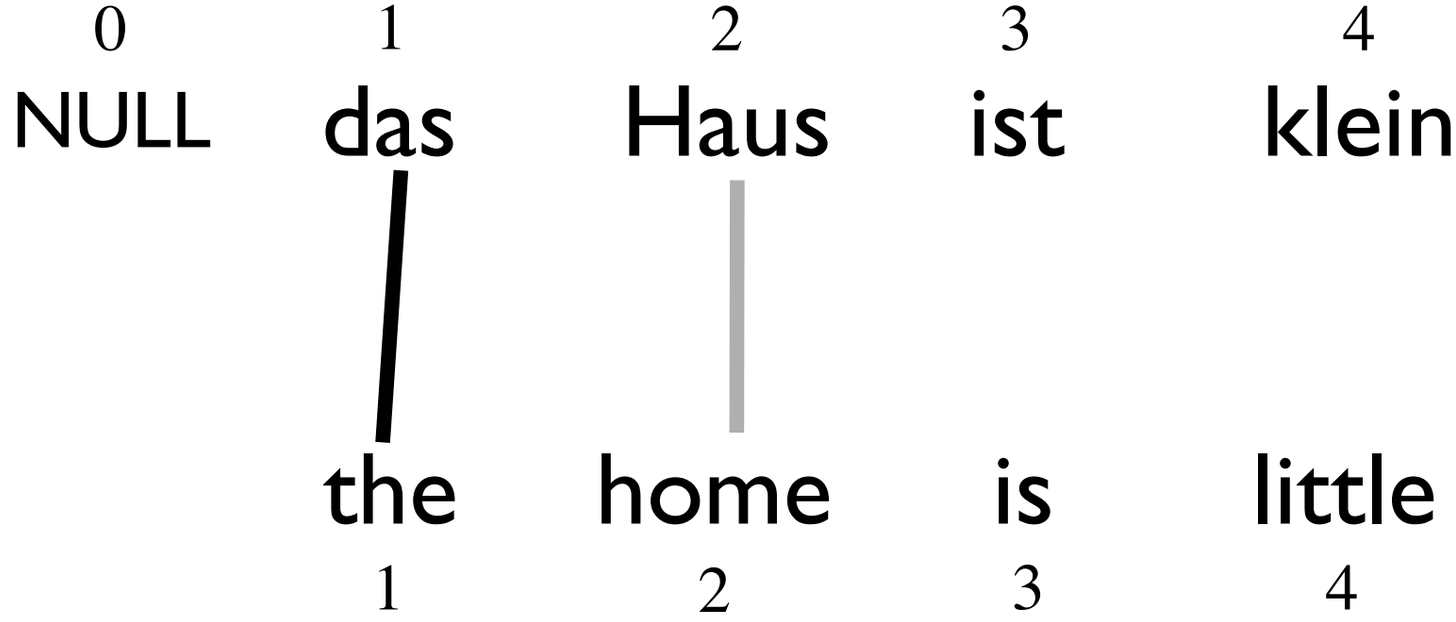
Finding the Viterbi Alignment



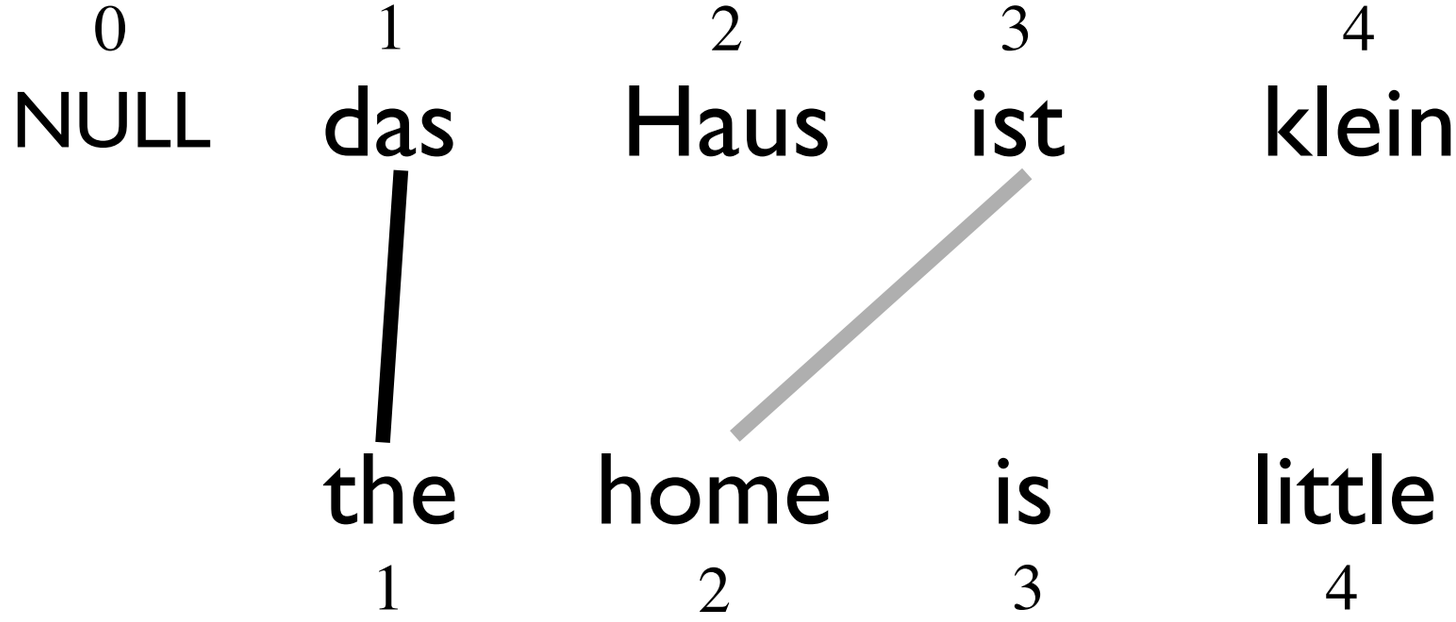
Finding the Viterbi Alignment



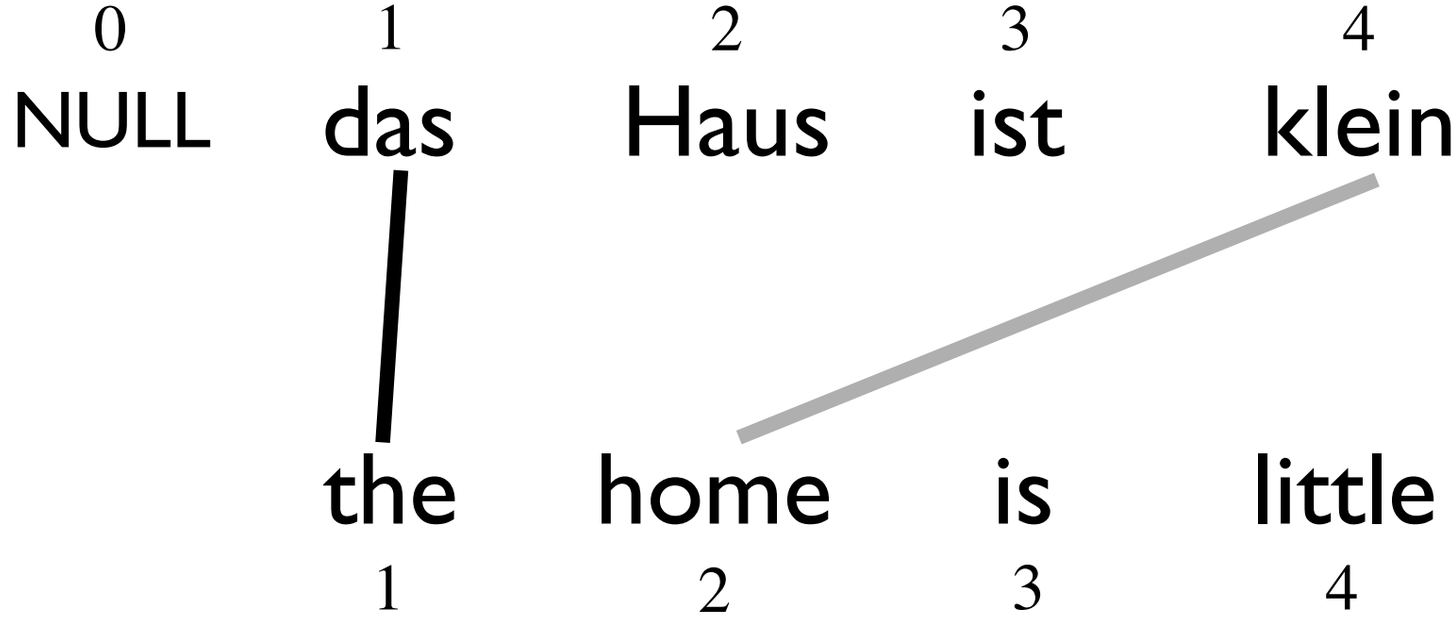
Finding the Viterbi Alignment



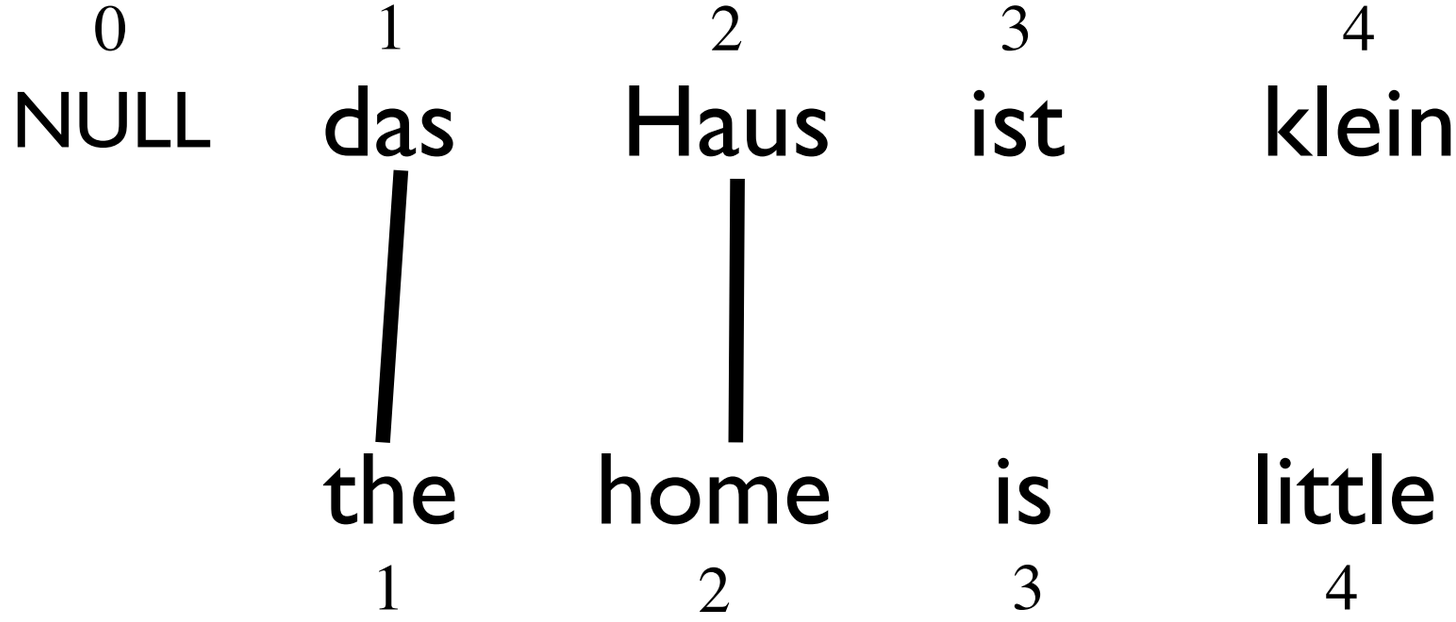
Finding the Viterbi Alignment



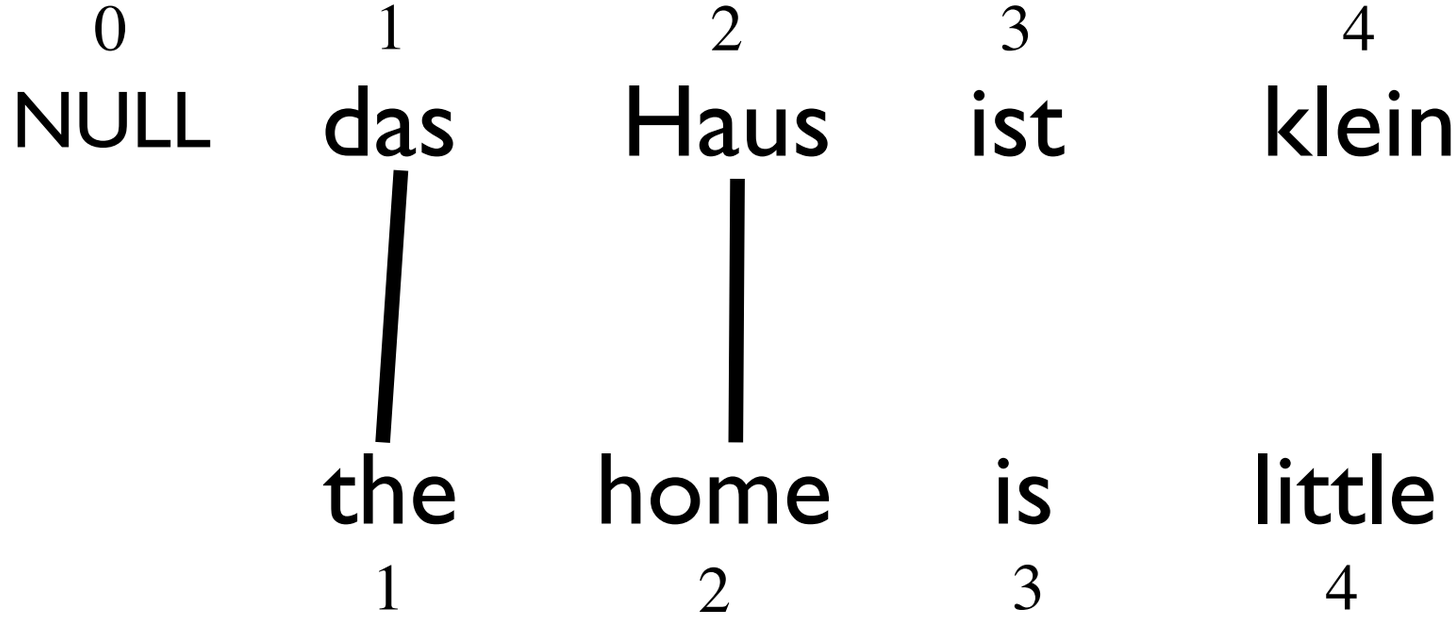
Finding the Viterbi Alignment



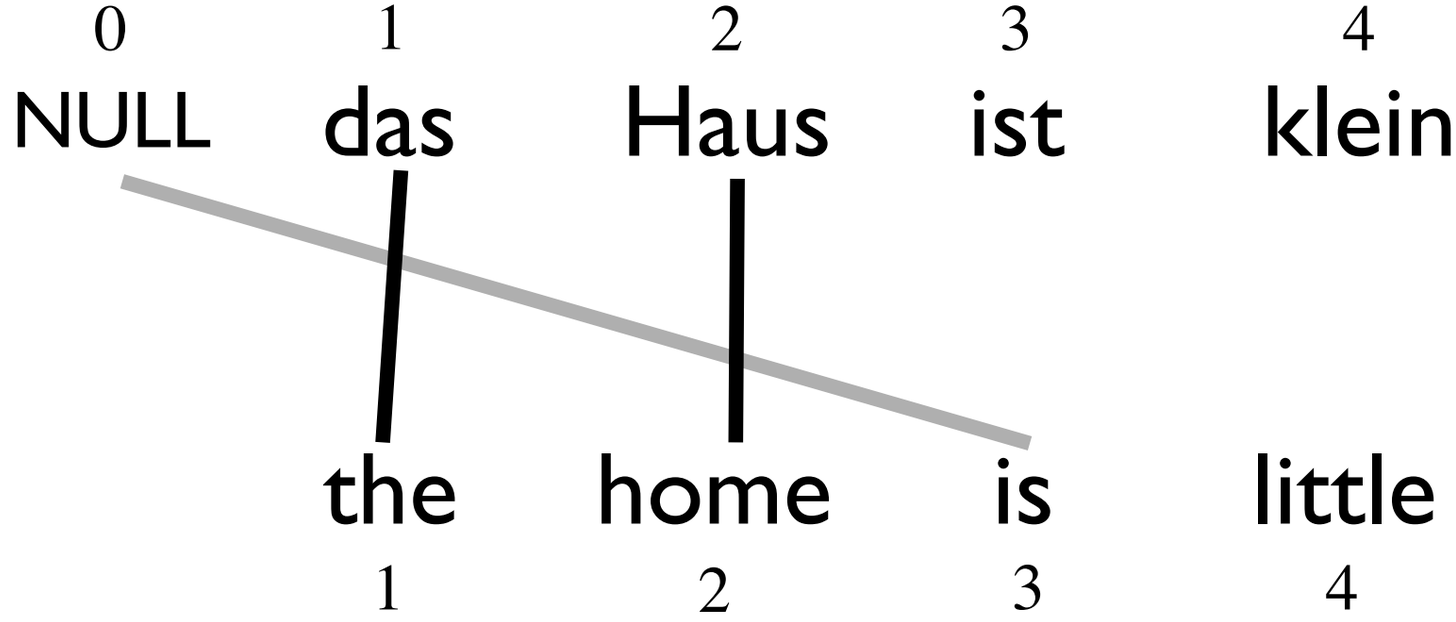
Finding the Viterbi Alignment



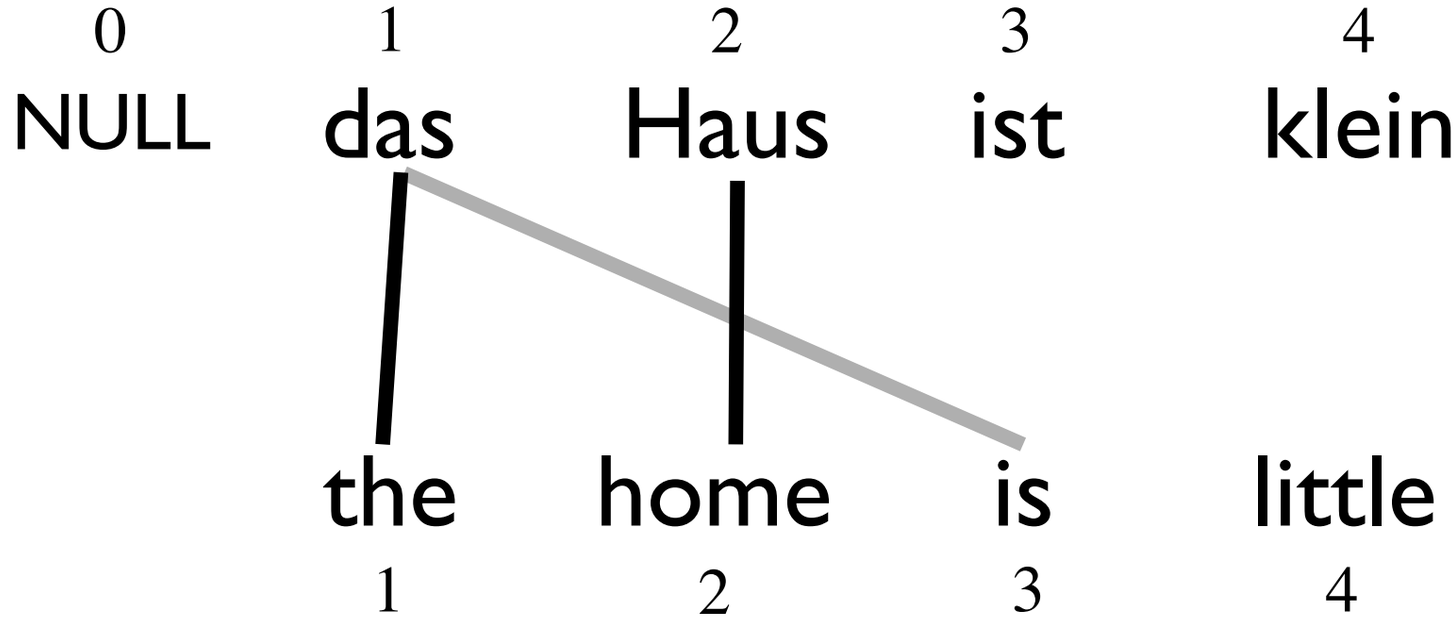
Finding the Viterbi Alignment



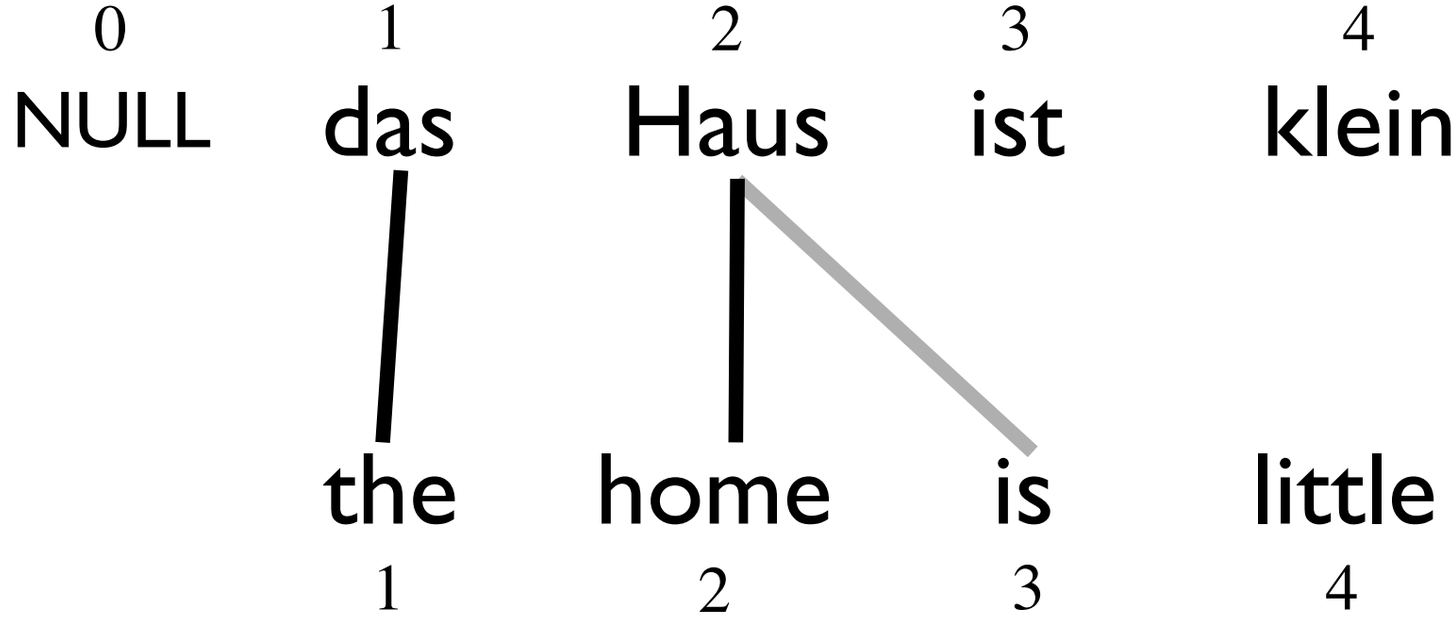
Finding the Viterbi Alignment



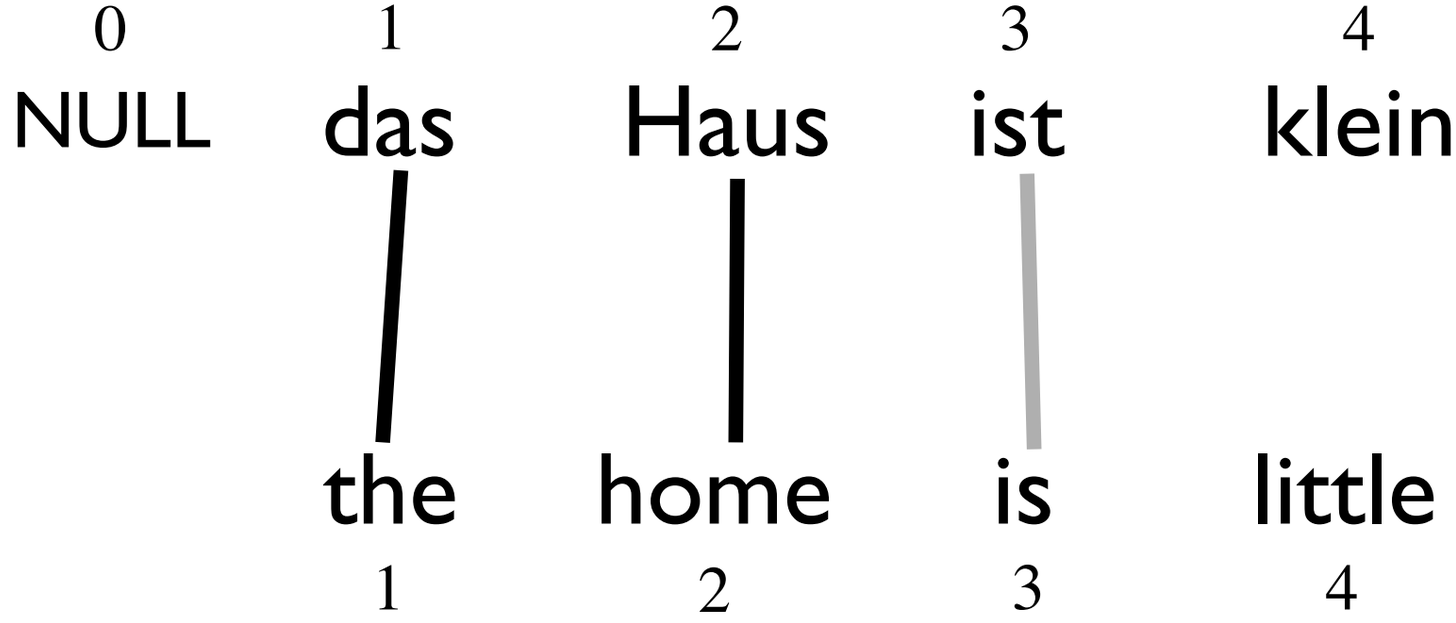
Finding the Viterbi Alignment



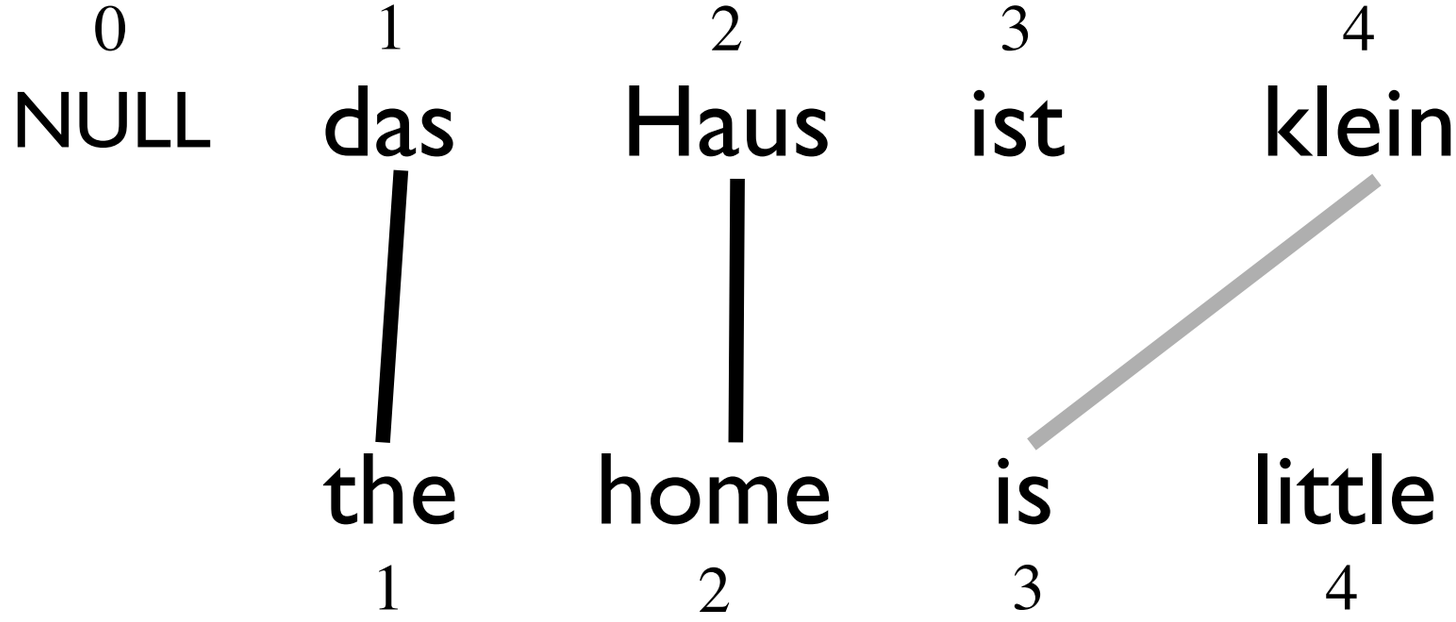
Finding the Viterbi Alignment



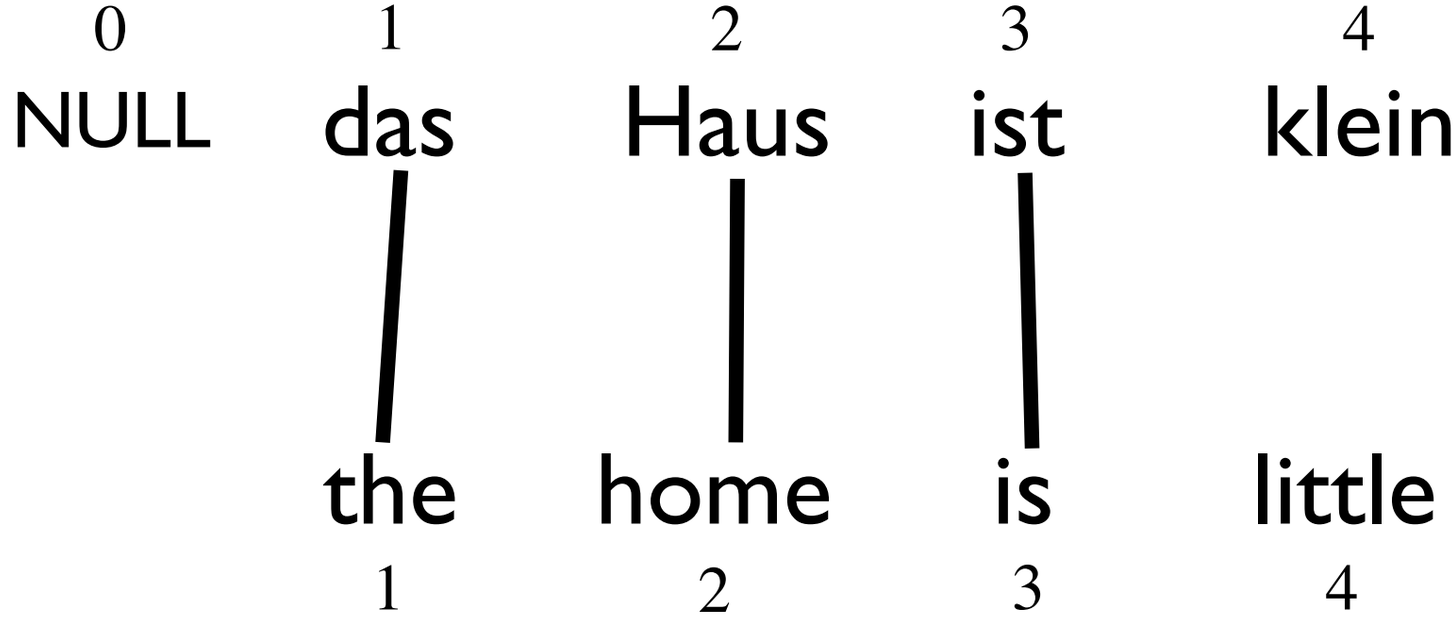
Finding the Viterbi Alignment



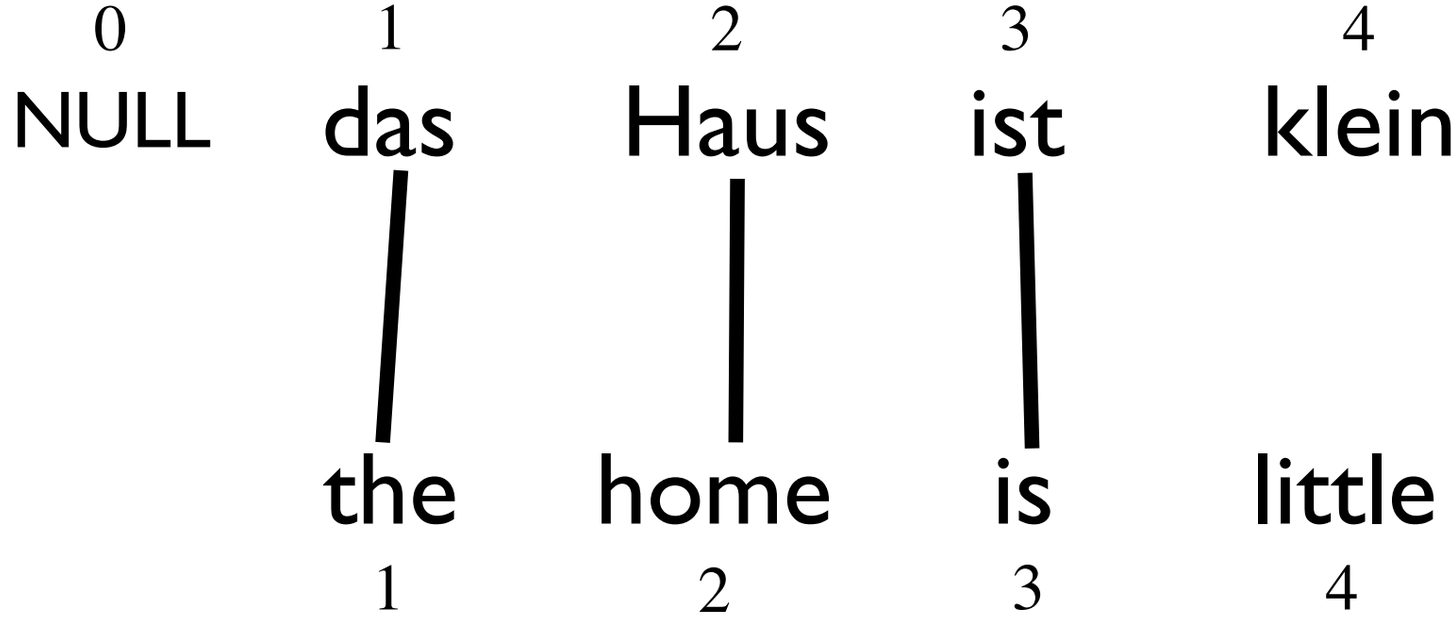
Finding the Viterbi Alignment



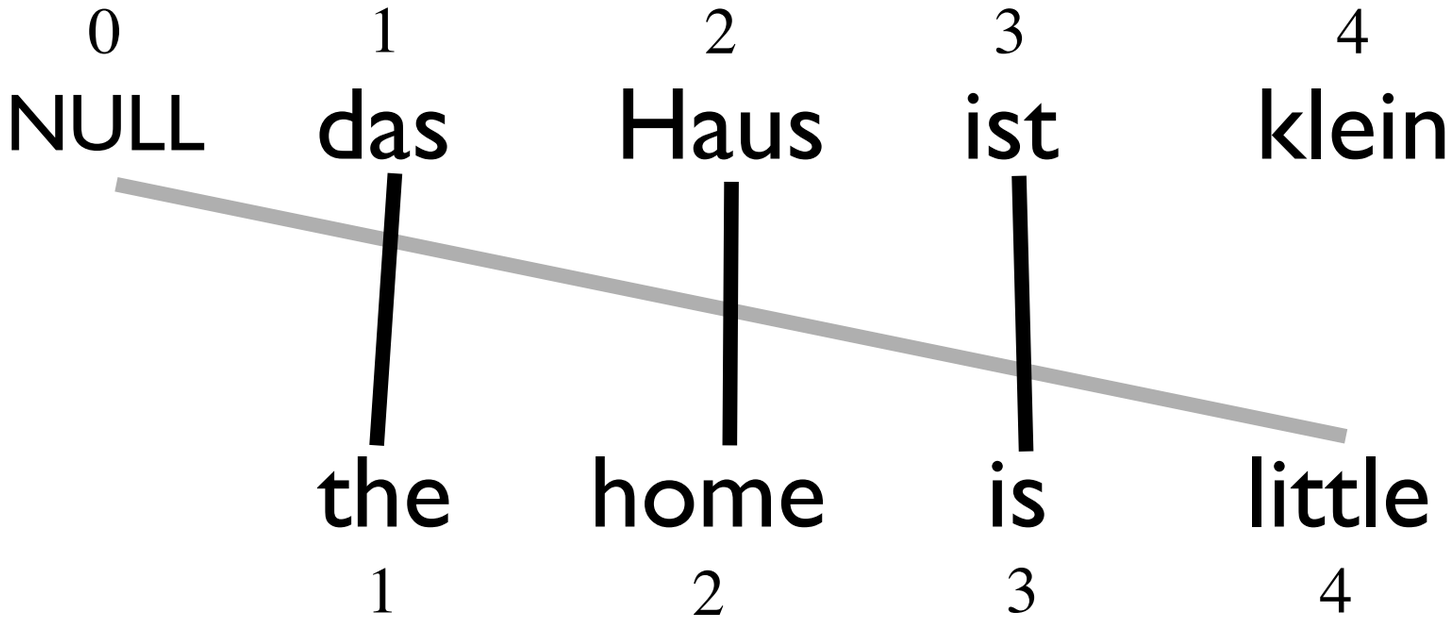
Finding the Viterbi Alignment



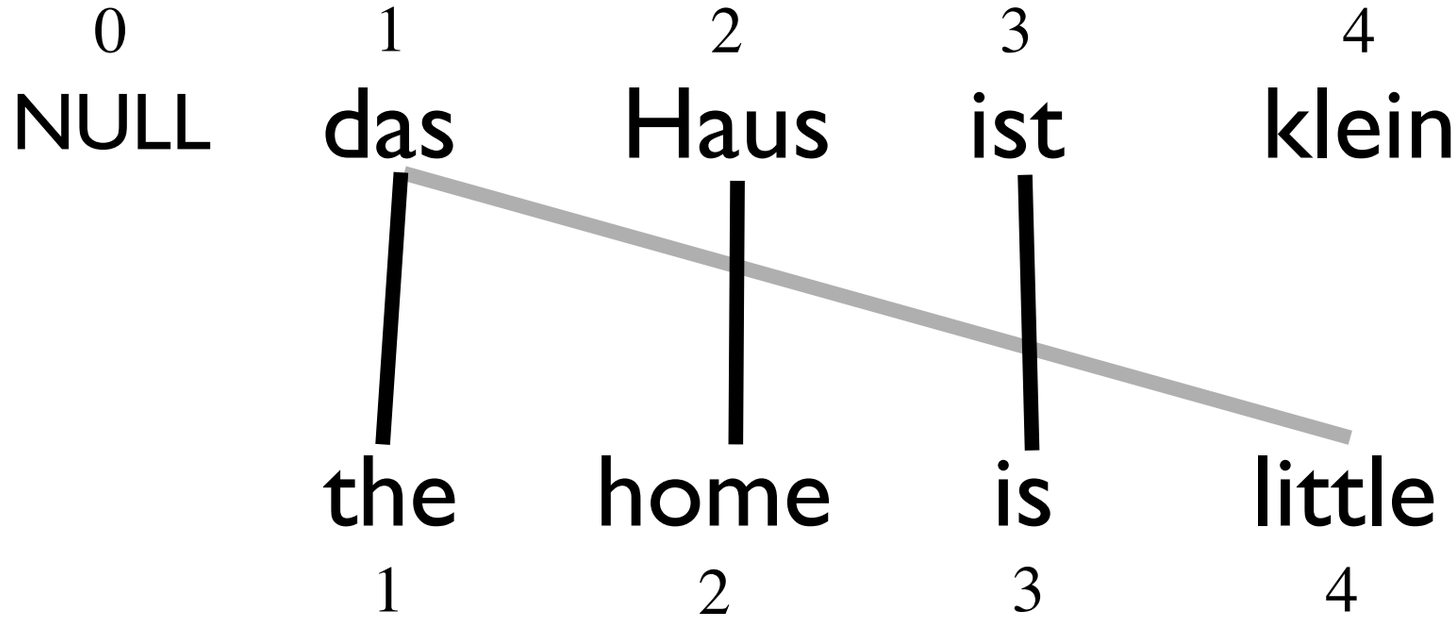
Finding the Viterbi Alignment



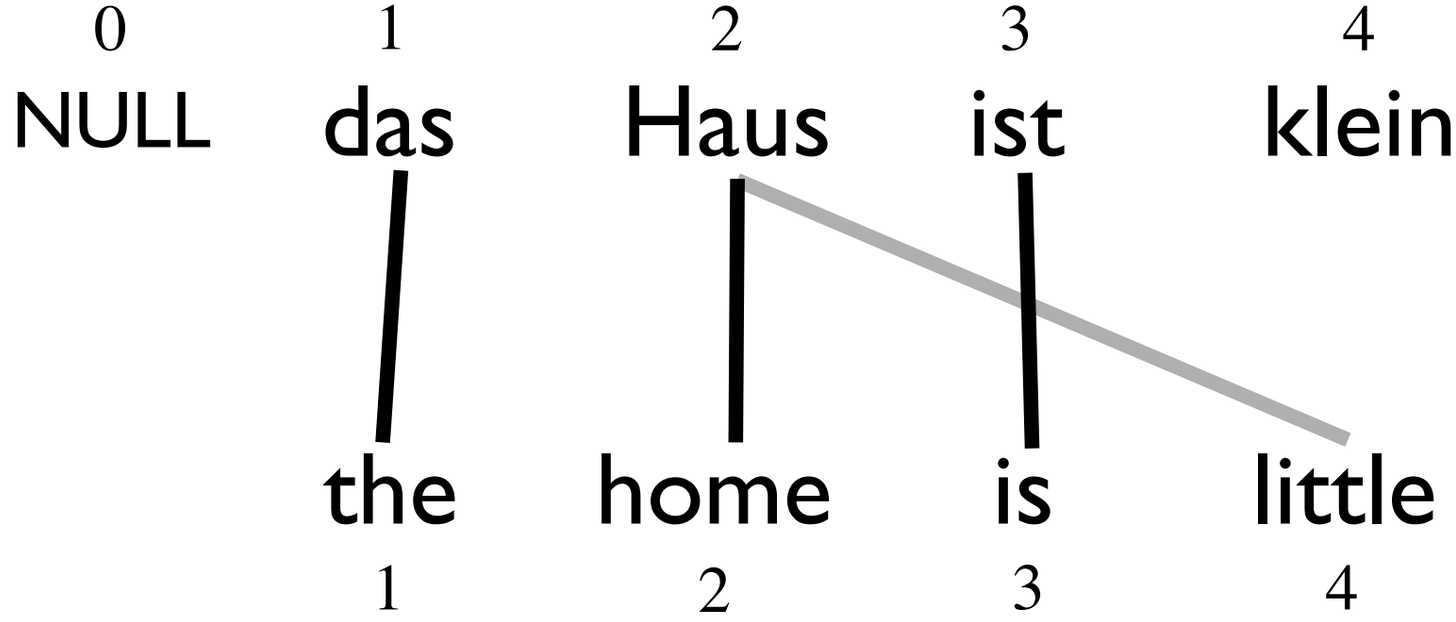
Finding the Viterbi Alignment



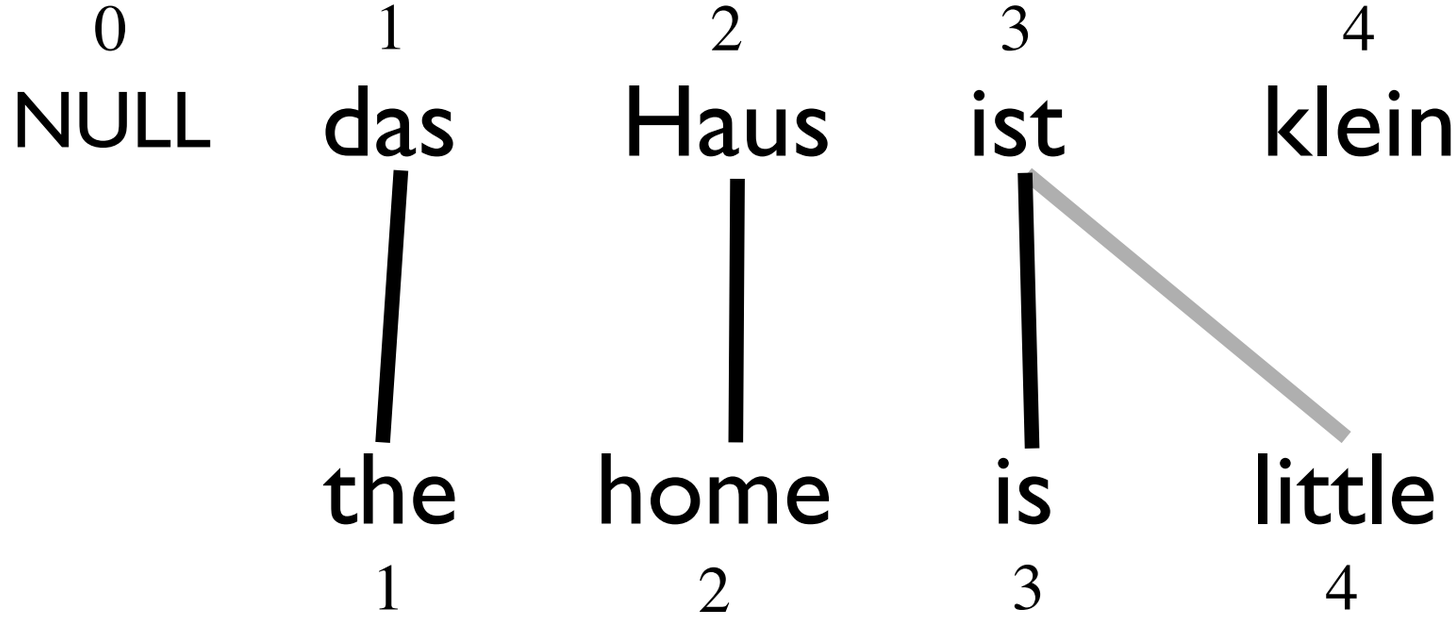
Finding the Viterbi Alignment



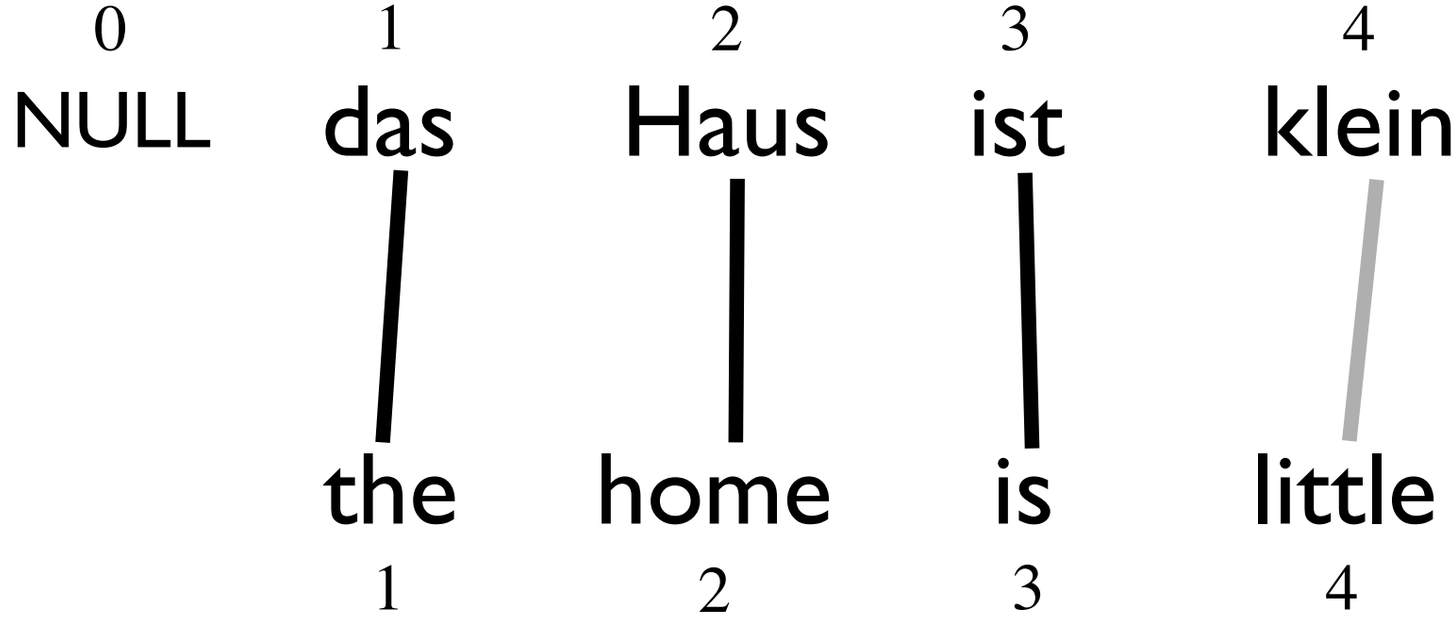
Finding the Viterbi Alignment



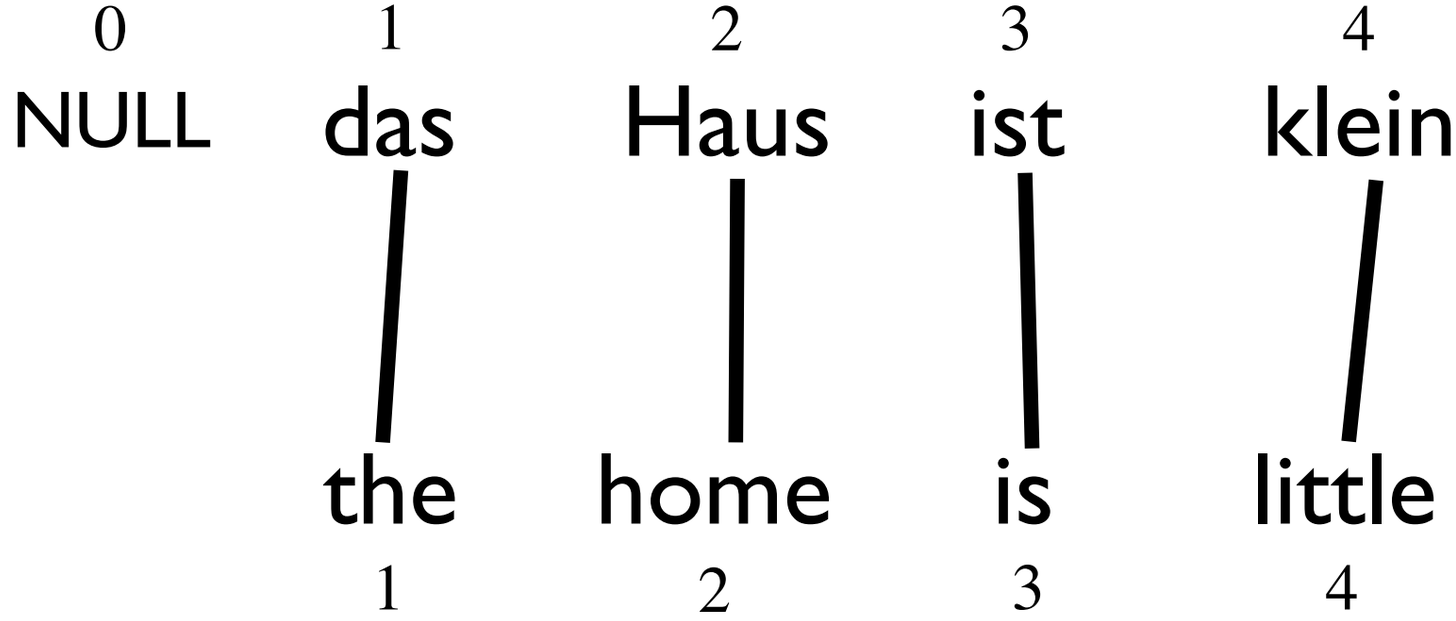
Finding the Viterbi Alignment



Finding the Viterbi Alignment



Finding the Viterbi Alignment



Learning Lexical Translation Models

- How do we learn the parameters $p(e \mid f)$
- "Chicken and egg" problem
  - If we had the alignments, we could estimate the parameters (MLE)
  - If we had parameters, we could find the most likely alignments

How to learn?

- Training data: tons of (**e**,**f**) pairs. Want to learn *theta*: lexical translation probabilities.
- If knew the alignments **a**, MLE would be easy!

$$\max_{\theta} \sum_{(\mathbf{e},\mathbf{f})} \log p(\mathbf{e} \mid \mathbf{a}, \mathbf{f}; \theta)$$

- But **a** is a latent variable. The marginal log-likelihood needs to sum-out the alignments. No closed form.

$$\max_{\theta} \sum_{(\mathbf{e},\mathbf{f})} \log p(\mathbf{e} \mid \mathbf{f}; \theta)$$

$$= \sum_{(\mathbf{e},\mathbf{f})} \log \left(\sum_{\mathbf{a}} p(\mathbf{a} \mid \mathbf{f}; \theta) \times p(\mathbf{e} \mid \mathbf{a}, \mathbf{f}; \theta) \right)$$

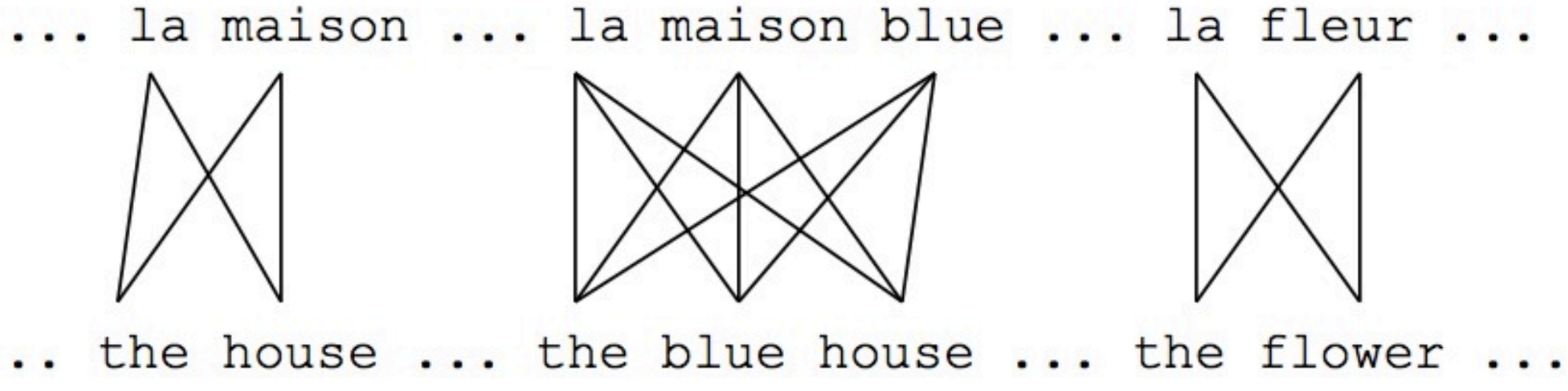
EM Algorithm

- pick some random (or uniform) parameters
- Repeat until you get bored (~ 5 iterations for lexical translation models)
  - using your current parameters, compute "expected" alignments for every target word token in the training data

$$p(a_i \mid \mathbf{e}, \mathbf{f})$$ (on board)

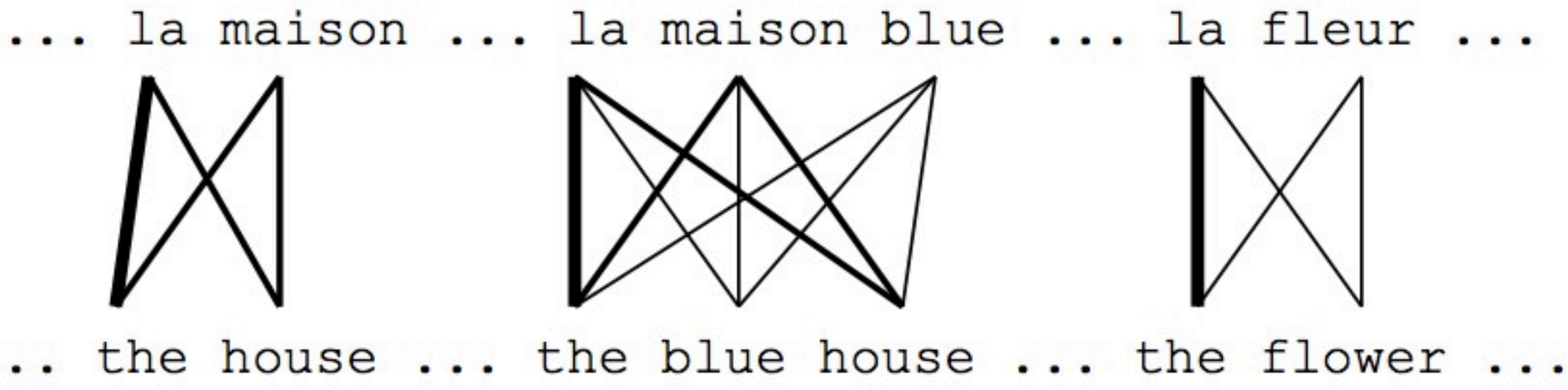
  - keep track of the expected number of times f translates into e throughout the whole corpus
  - keep track of the expected number of times that f is used as the source of any translation
  - use these expected counts as if they were "real" counts in the standard MLE equation

EM for Model 1



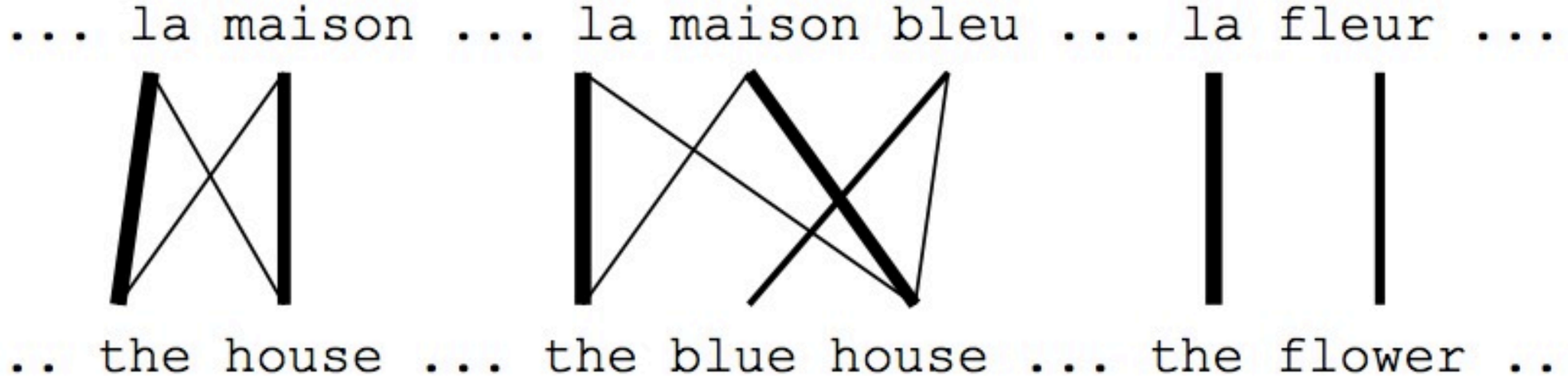
- Initial step: all alignments equally likely
- Model learns that, e.g., la is often aligned with the

EM for Model 1



- After one iteration
- Alignments, e.g., between la and the are more likely

EM for Model 1

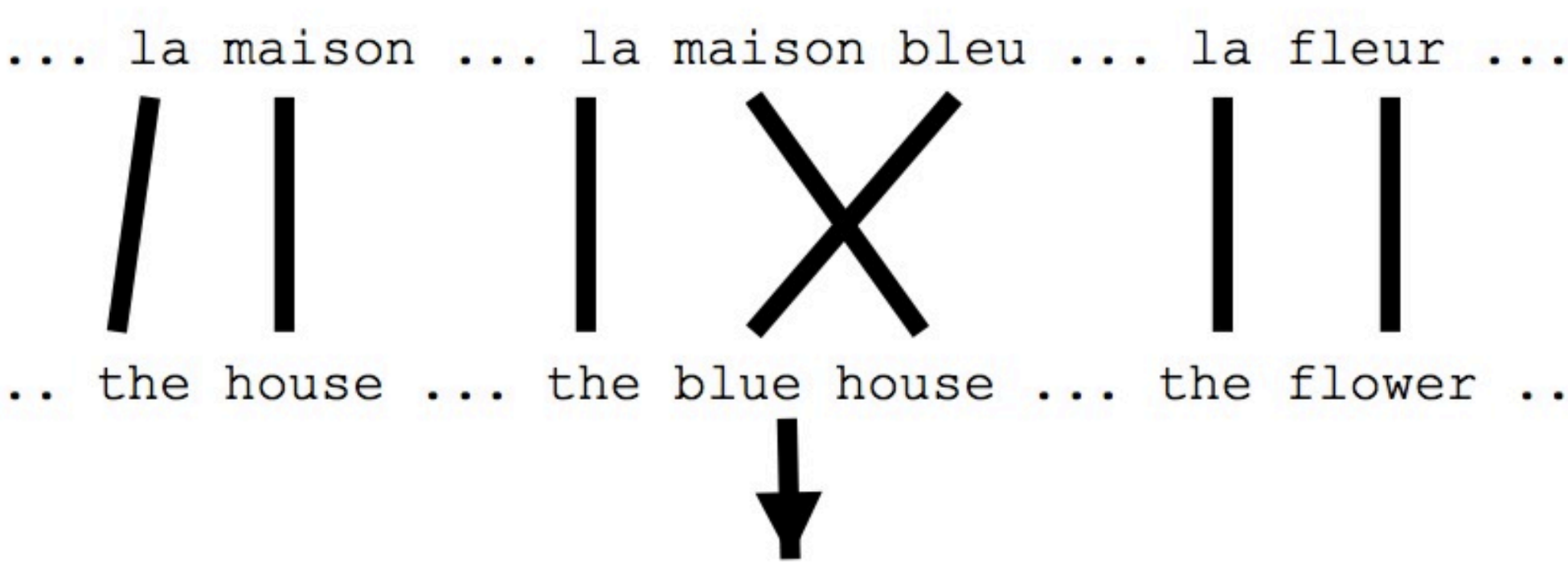


- After another iteration
- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)

EM for Model 1

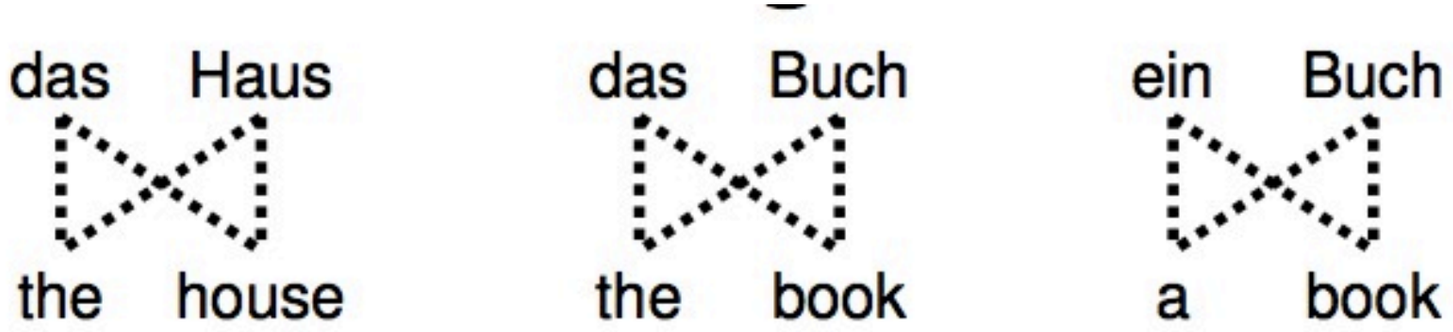
- Convergence
- Inherent hidden structure revealed by EM

EM for Model 1



- Parameter estimation from the aligned corpus

Convergence



e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

Now what?

- Decoding: find the best translation for a sentence
- Alignments: find the best alignment
- Evaluation?
- Other approaches?