Beyond Bag of Words

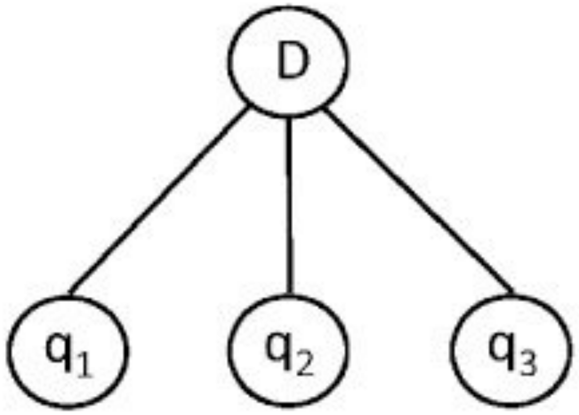
CS6200
Information Retrieval

Bags of Words

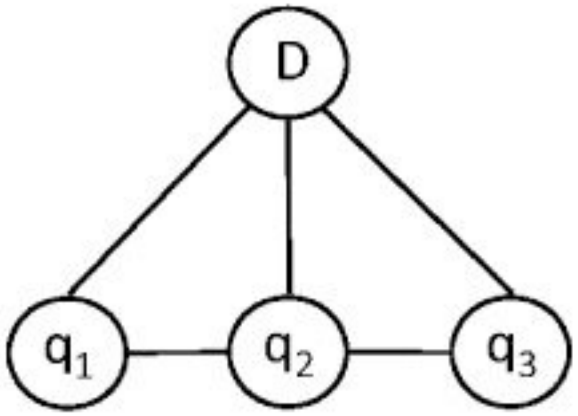
- Most efficient (and still very effective) retrieval models treat words/terms as independent
- Generalized models based on (log-)linear combination of **features** of both query and document

$$S_W(D;Q)=\sum_j w_j \cdot f_j(D,Q)$$

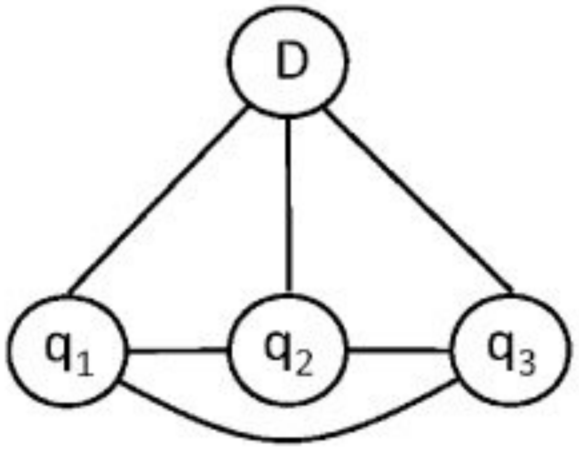
Term Dependence Models



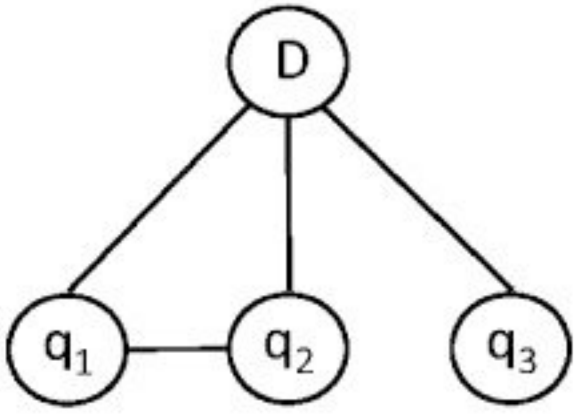
Full independence



Sequential dependence

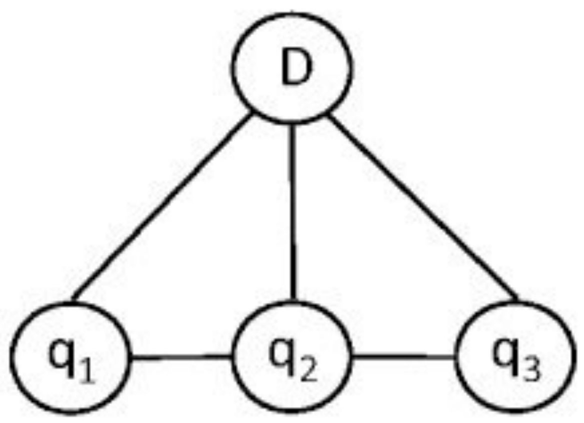


Full dependence



General dependence

Term Dependence Models

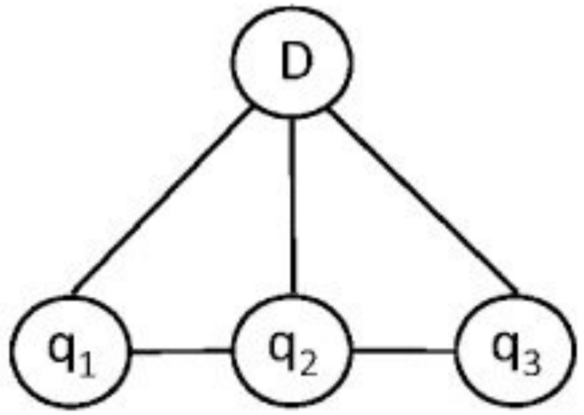


Sequential dependence

$$S_W(D;Q)=\sum_j w_j \cdot f_j(D,Q)$$

```
#weight(0.8 #combine(president abraham lincoln)
        0.1 #combine(#od:1(president abraham)
                     #od:1(abraham lincoln))
        0.1 #combine(#uw:8(president abraham)
                     #uw:8(abraham lincoln)))
```

Term Dependence Models



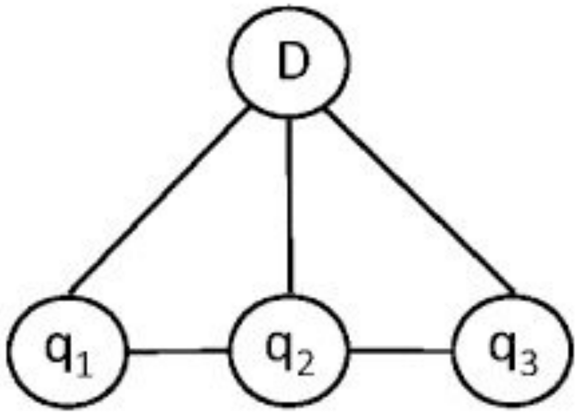
$$S_W(D;Q)=\sum_j w_j \cdot f_j(D,Q)$$

Sequential dependence

Weights of different bigrams are **tied**

```
#weight(0.8 #combine(president abraham lincoln)
        0.1 #combine(#od:1(president abraham)
                     #od:1(abraham lincoln))
        0.1 #combine(#uw:8(president abraham)
                     #uw:8(abraham lincoln)))
```

Term Dependence Models



$$S_W(D;Q)=\sum_j w_j \cdot f_j(D,Q)$$

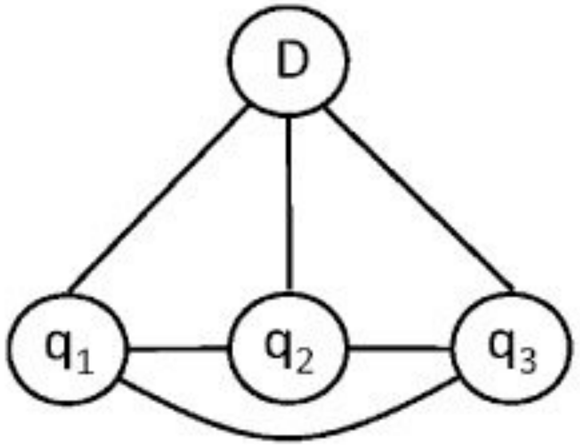
Sequential dependence

Weights of different bigrams are **tied**

```
#weight(0.8 #combine(president abraham lincoln)
        0.1 #combine(#od:1(president abraham)
                     #od:1(abraham lincoln))
        0.1 #combine(#uw:8(president abraham)
                     #uw:8(abraham lincoln)))
```

Therefore, estimate weights for **classes** of features, not for each individual bigram

Term Dependence Models

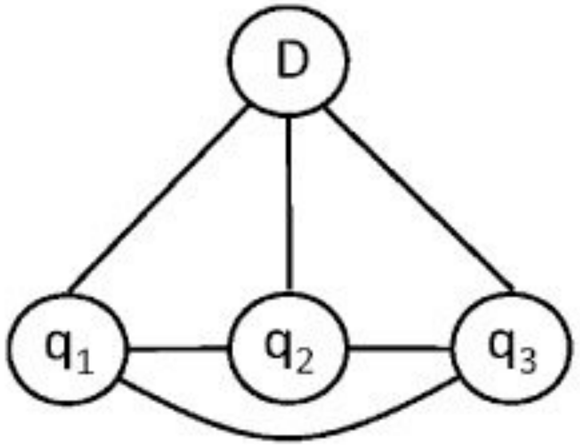


Full dependence

$$S_W(D;Q)=\sum_j w_j \cdot f_j(D,Q)$$

```
#weight(0.8 #combine(president abraham lincoln)
        0.1 #combine(#od:1(president abraham)
                     #od:1(abraham lincoln)
                     #od:1(president abraham lincoln))
        0.1 #combine(#uw:8(president abraham)
                     #uw:8(abraham lincoln)
                     #uw:8(president lincoln)
                     #uw:12(president abraham lincoln)))
```

Term Dependence Models



Full dependence

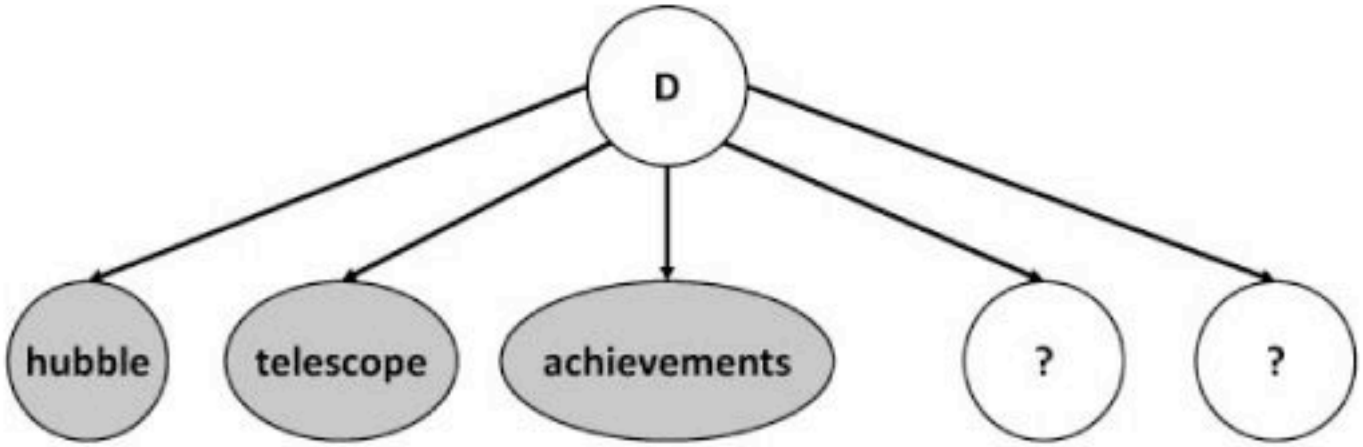
$$S_W(D;Q)=\sum_j w_j \cdot f_j(D,Q)$$

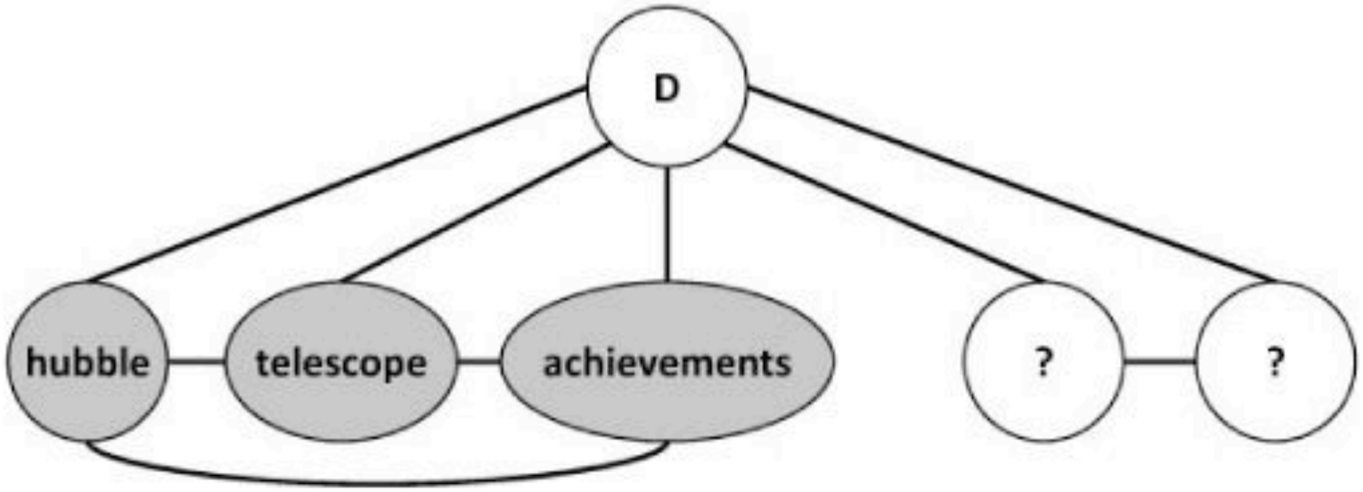
```
#weight(0.8 #combine(president abraham lincoln)
        0.1 #combine(#od:1(president abraham)
                     #od:1(abraham lincoln)
                     #od:1(president abraham lincoln))
        0.1 #combine(#uw:8(president abraham)
                     #uw:8(abraham lincoln)
                     #uw:8(president lincoln)
                     #uw:12(president abraham lincoln)))
```

Features (with tied weights) for bigrams, trigrams, etc.

Term Dependence Models

Unigram relevance model





Latent concept expansion

1-word concepts	*2-word concepts*
telescope	hubble telescope
hubble	space telescope
space	hubble space
mirror	telescope mirror
NASA	telescope hubble
launch	mirror telescope
astronomy	telescope NASA
shuttle	telescope space
test	hubble mirror
new	NASA hubble
discovery	telescope astronomy
time	telescope optical
universe	hubble optical
optical	telescope discovery
light	telescope shuttle

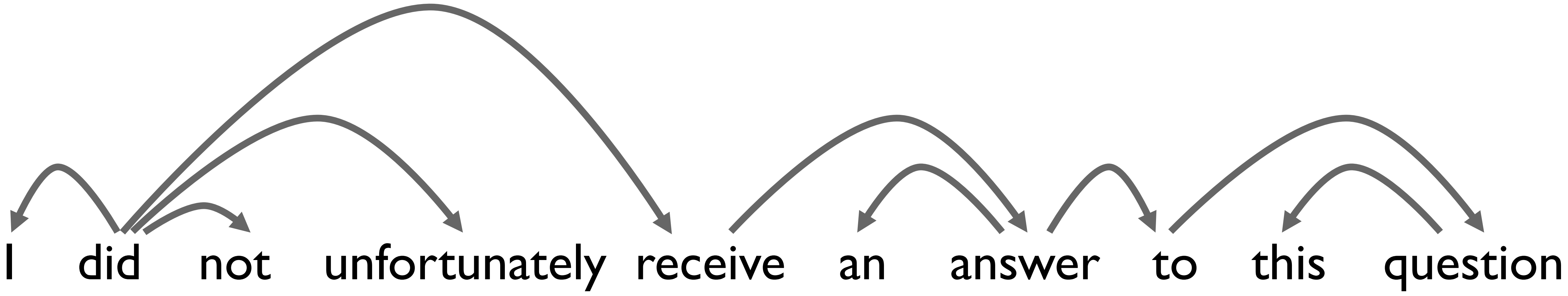
Syntactic Dependencies

Maximum directed spanning tree

I did not unfortunately receive an answer to this question

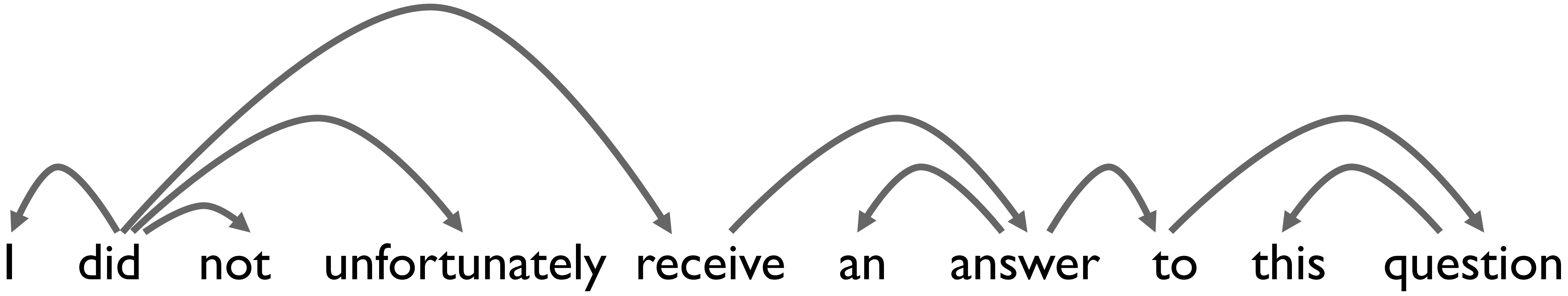
Syntactic Dependencies

Maximum directed spanning tree



Syntactic Dependencies

Maximum directed spanning tree



I did not unfortunately receive an answer to this question

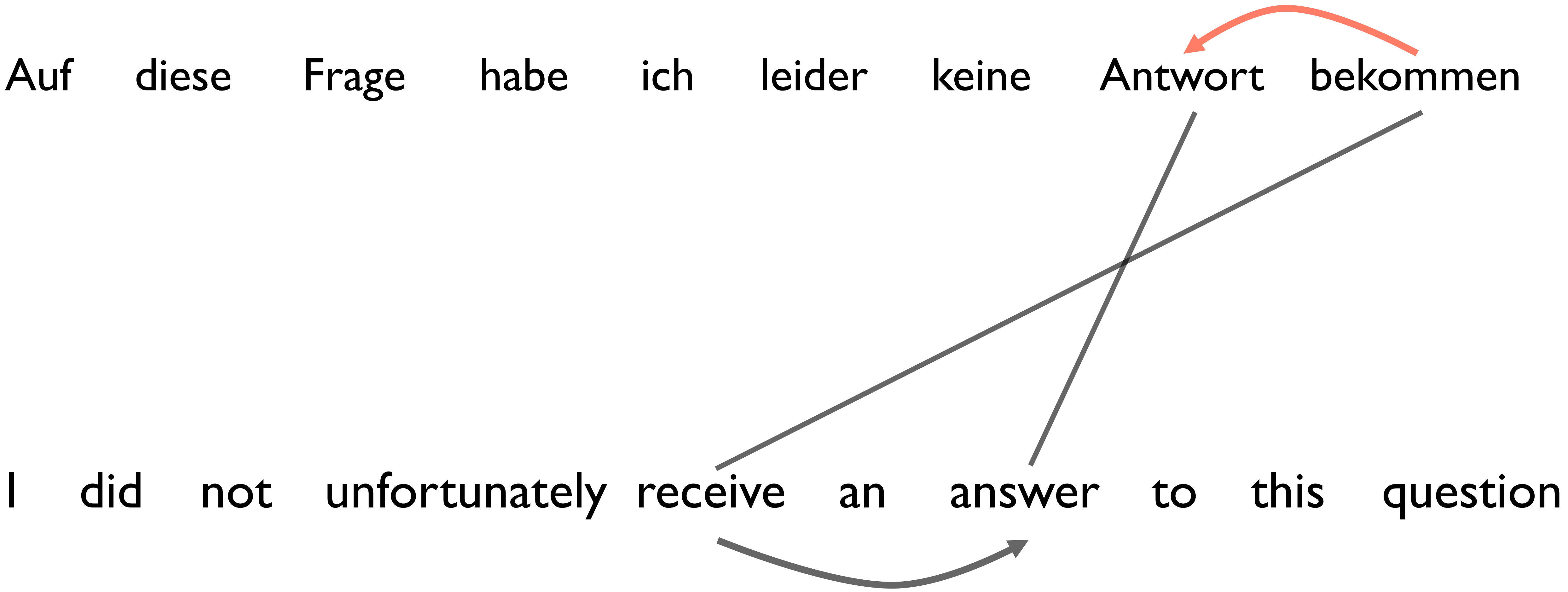
$n^{n-1} = 10^9 = 1$ billion

possible trees!

Cross-Language Syntax

Parser projection

Target language



Auf diese Frage habe ich leider keine Antwort bekommen

I did not unfortunately receive an answer to this question

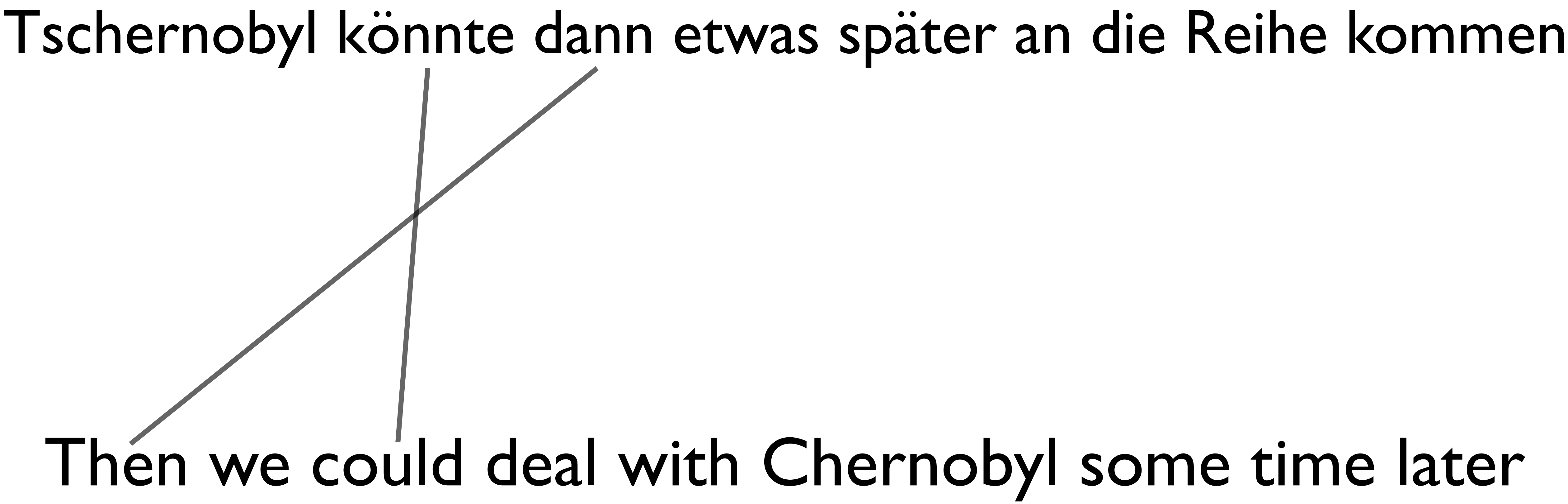
Source language

Cross-Language Syntax

Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

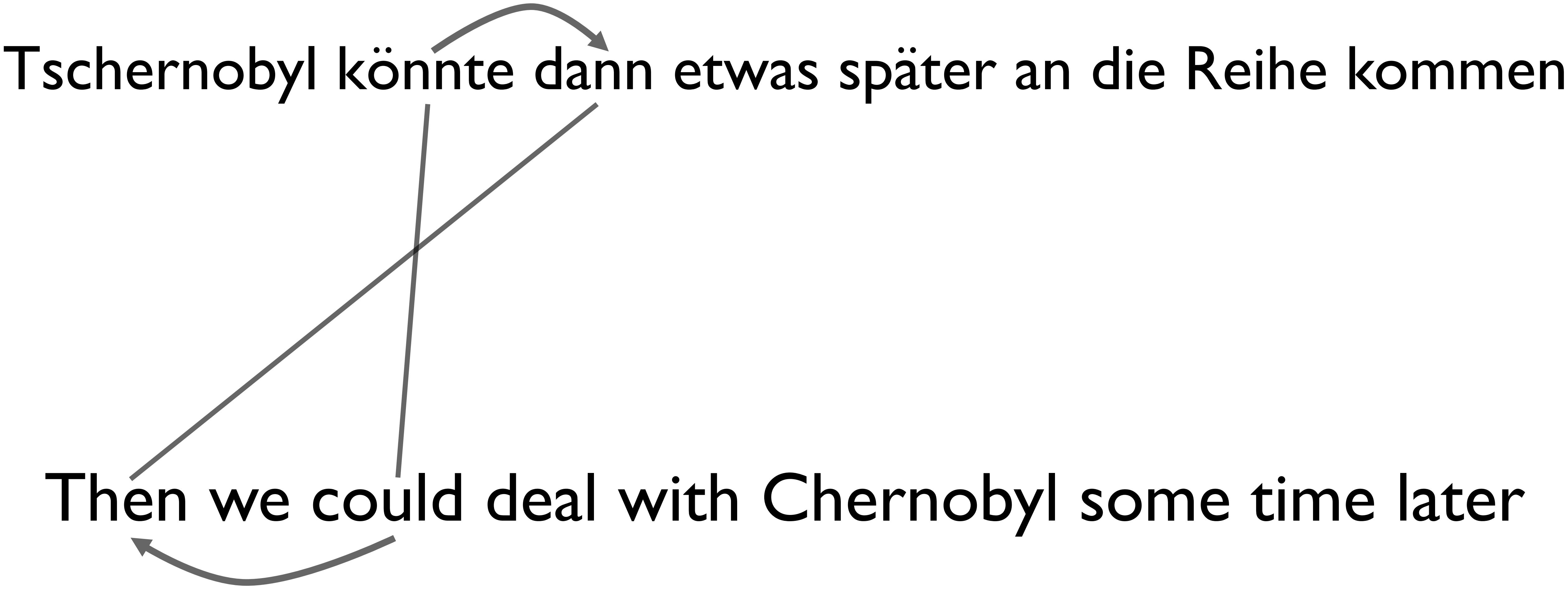
Cross-Language Syntax



Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

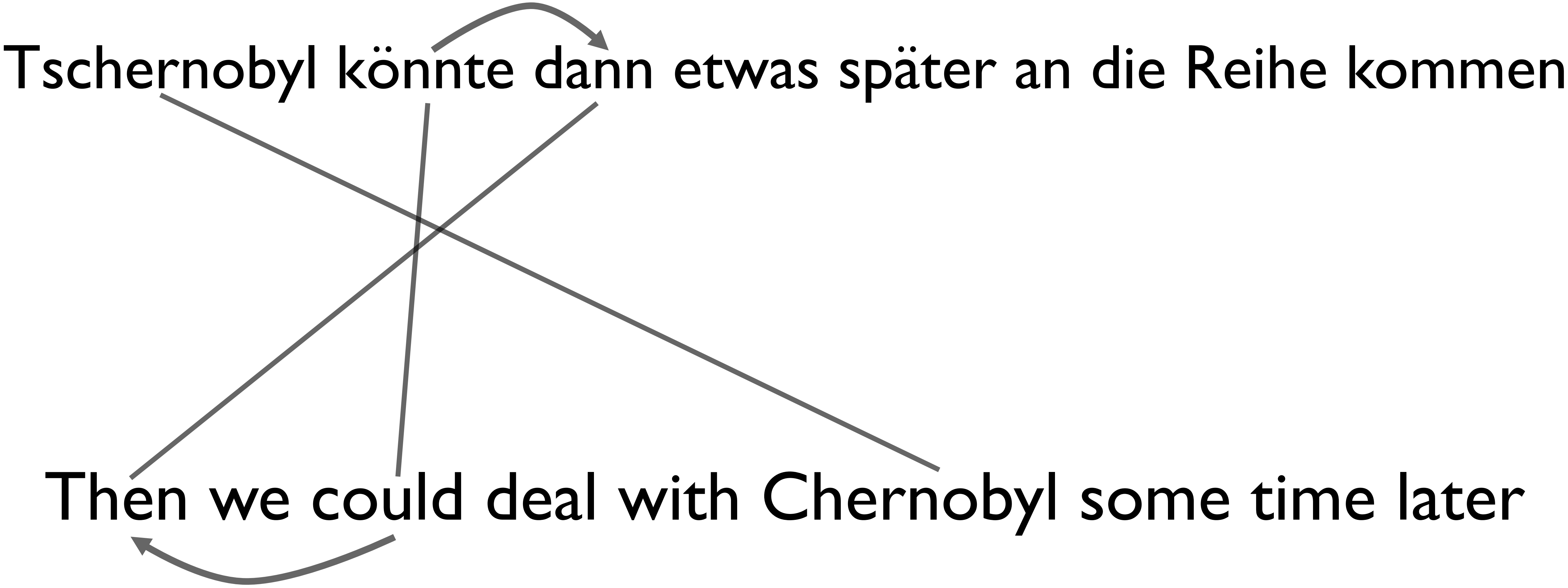
Cross-Language Syntax



Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

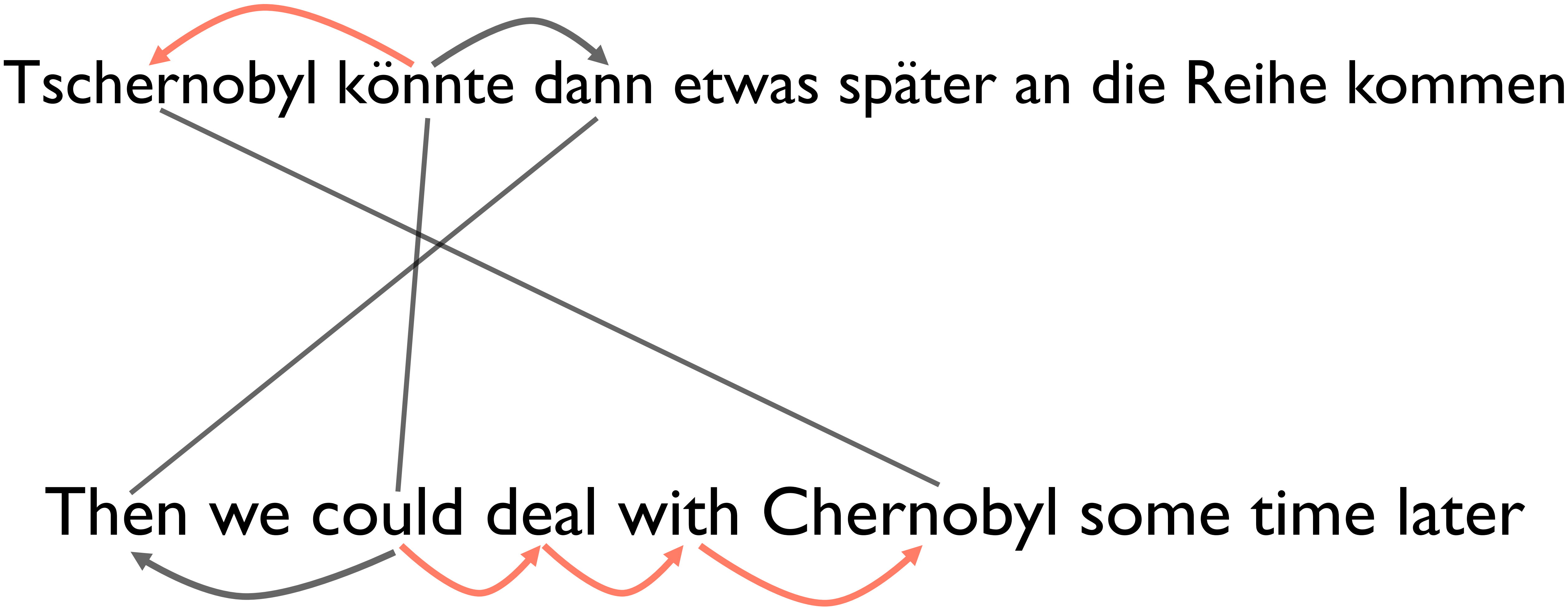
Cross-Language Syntax



Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

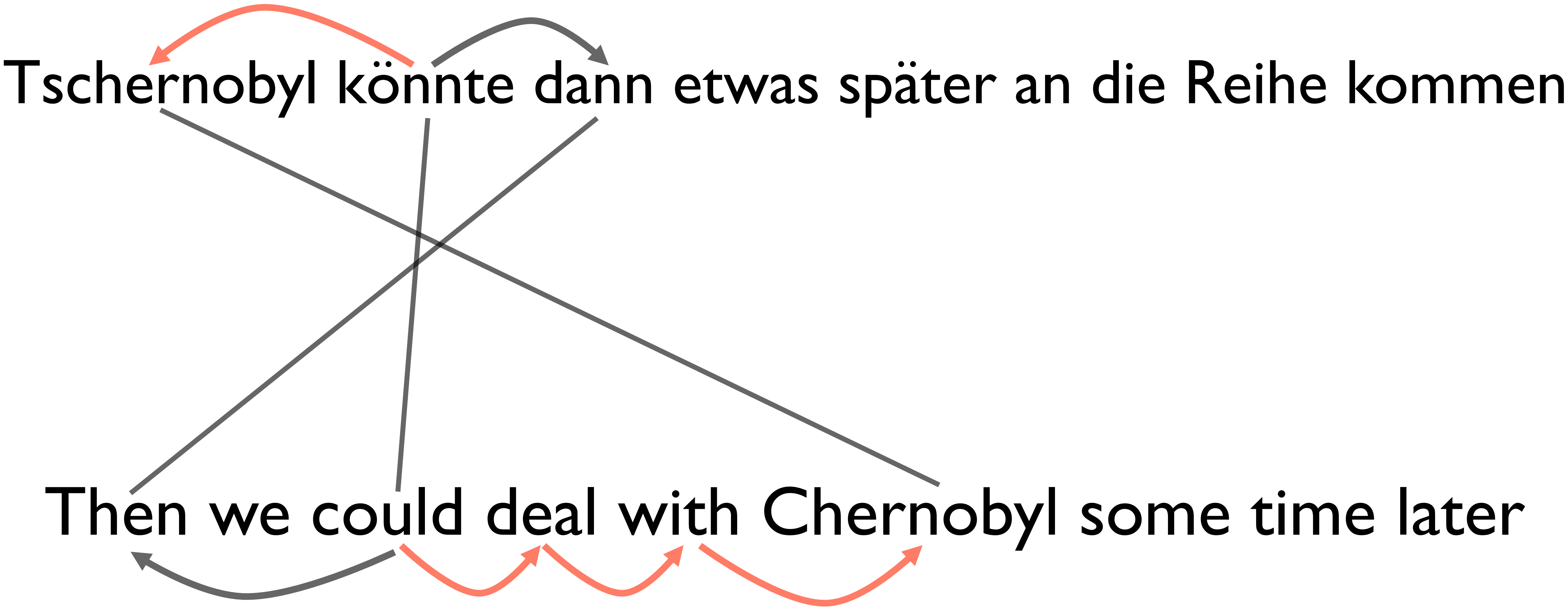
Cross-Language Syntax



Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

Cross-Language Syntax



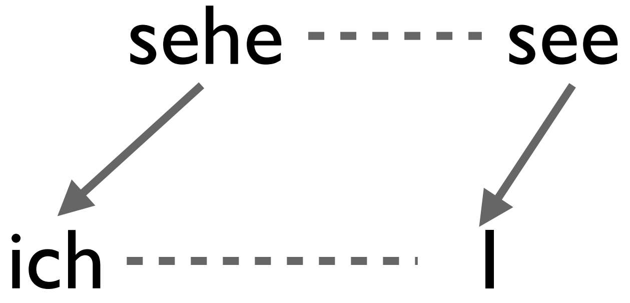
Tschernobyl könnte dann etwas später an die Reihe kommen

Then we could deal with Chernobyl some time later

Structures not isomorphic:
use quasi-synchronous alignment features

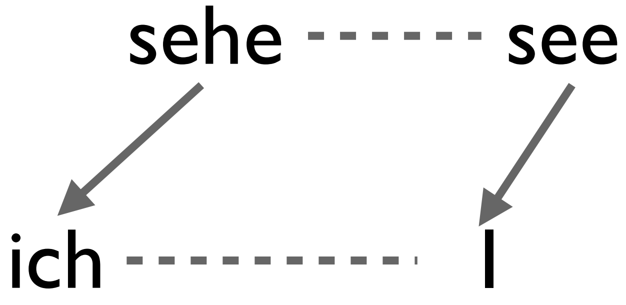
Alignment Configurations

Alignment Configurations

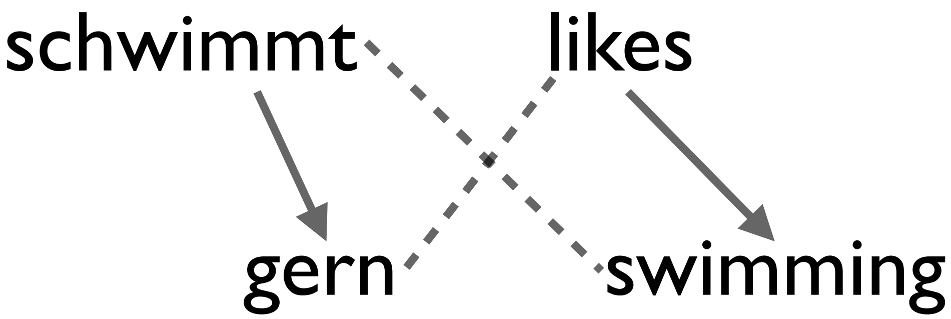


monotonic (parent-child)

Alignment Configurations

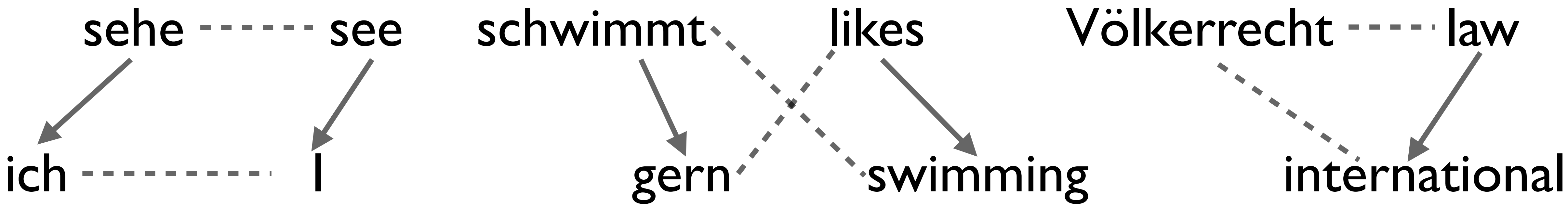


monotonic
(parent-child)



head swapping

Alignment Configurations



sehe see

ich I

monotonic (parent-child)

schwimmt likes

gern swimming

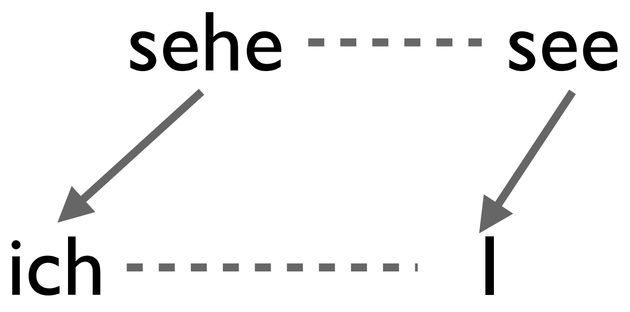
head swapping

Völkerrecht law

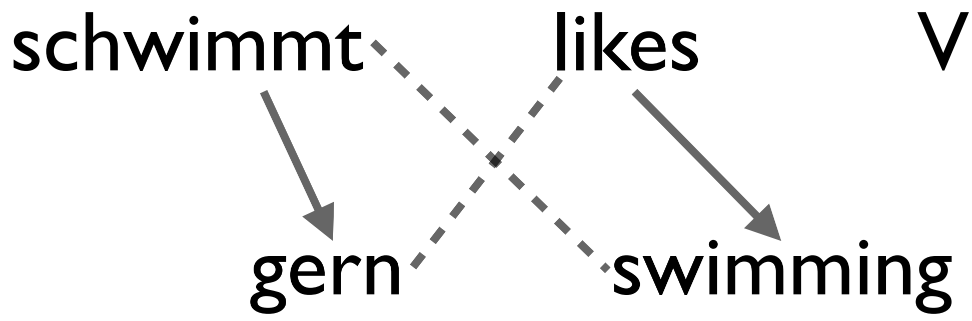
international

two-to-one

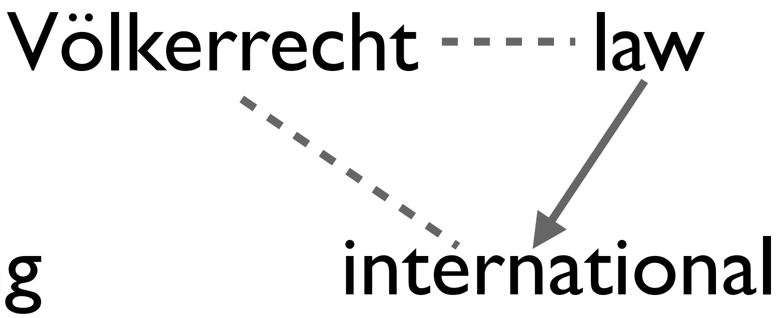
Alignment Configurations


sehe see
ich I

monotonic (parent-child)

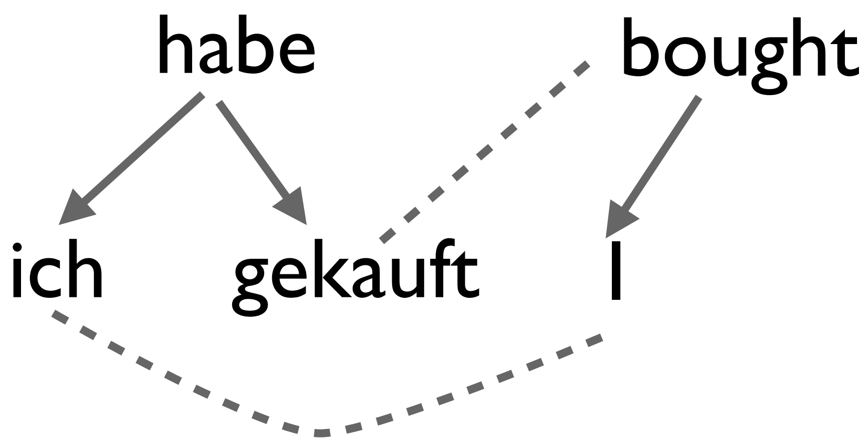

schwimmt likes
gern swimming

head swapping

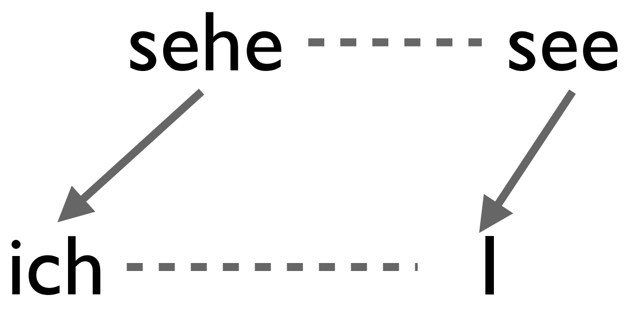

Völkerrecht law
international

two-to-one

siblings


habe bought
ich gekauft I

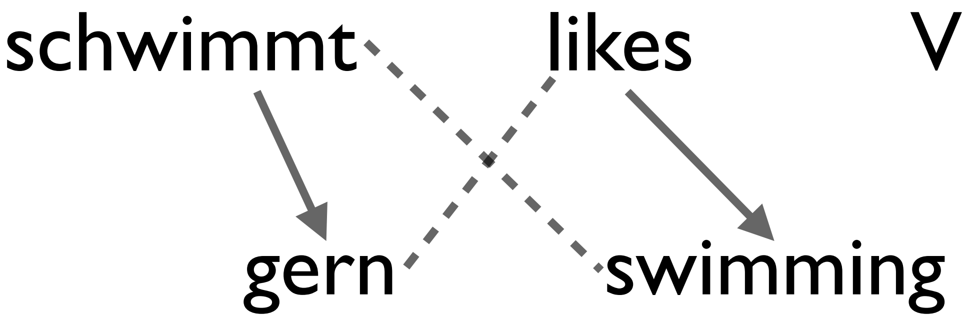
Alignment Configurations



sehe see

ich I

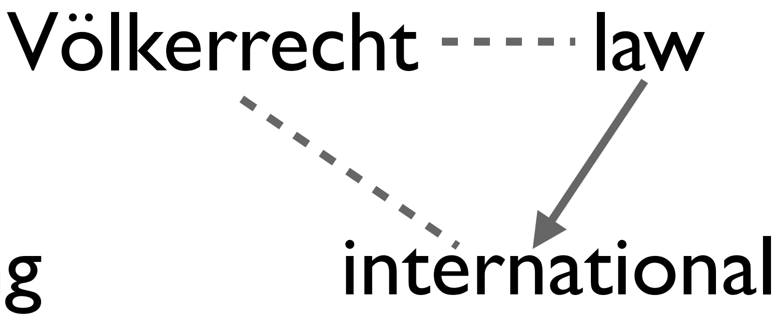
monotonic (parent-child)



schwimmt likes

gern swimming

head swapping

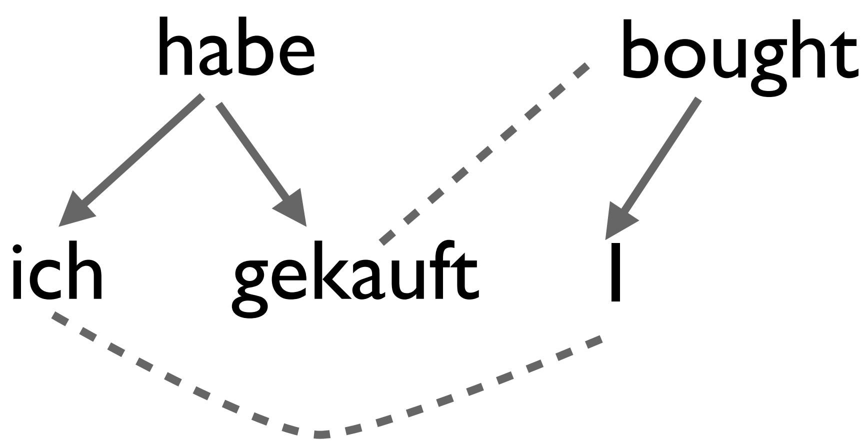


Völkerrecht law

international

two-to-one

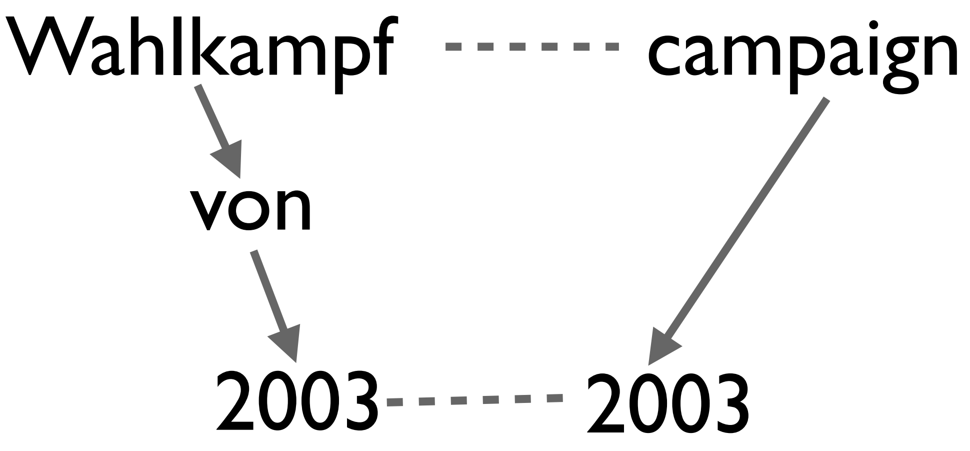
siblings



habe bought

ich gekauft I

grandparent-grandchild

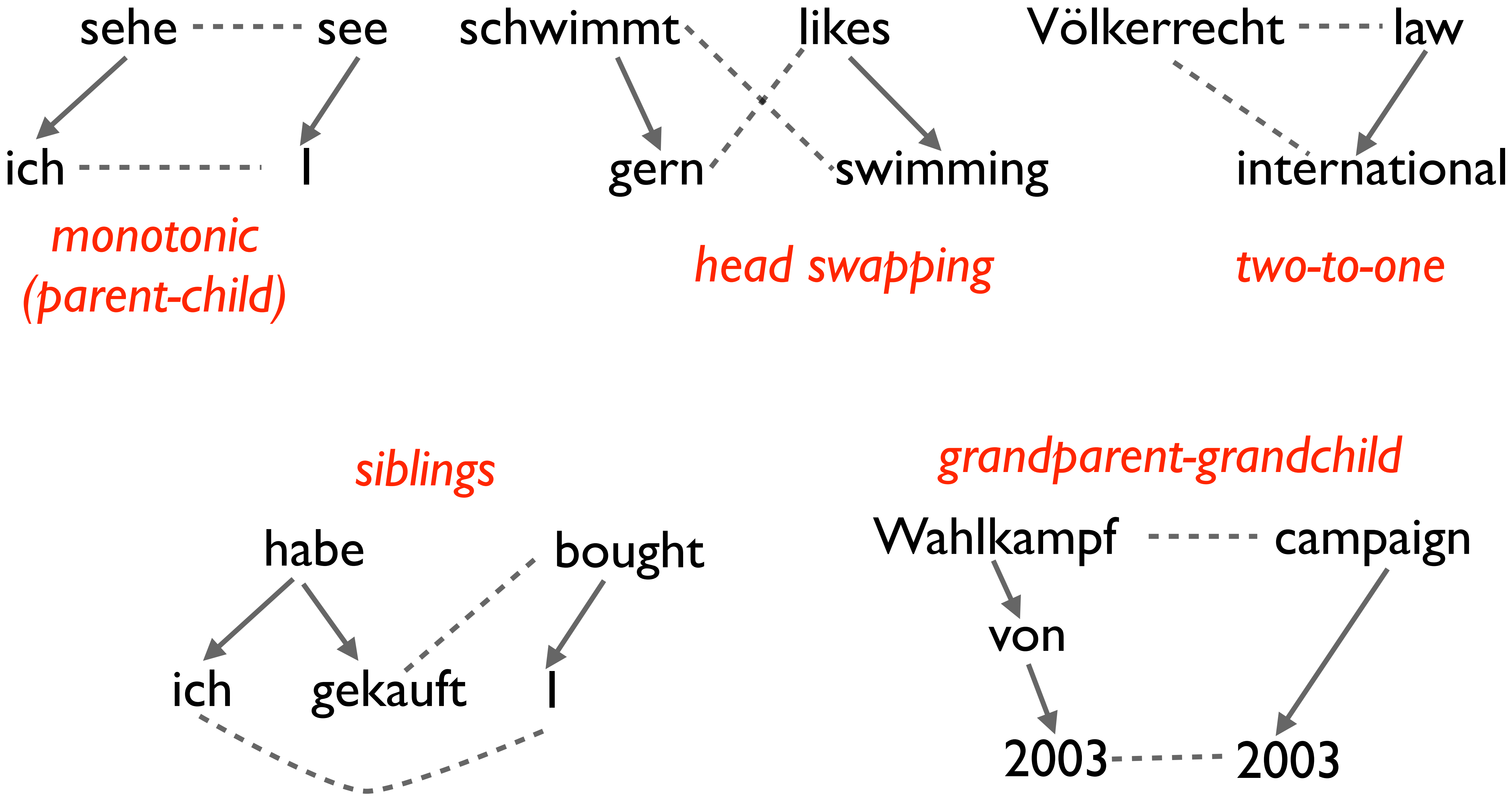


Wahlkampf campaign

von

2003 2003

Alignment Configurations



sehe see ich I

monotonic (parent-child)

schwimmt likes gern swimming

head swapping

Völkerrecht law international

two-to-one

siblings

habe bought ich gekauft I

grandparent-grandchild

Wahlkampf campaign von 2003 2003

Also C-command, descendant, "none of the above"

Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT* 2006; *EMNLP* 2009)

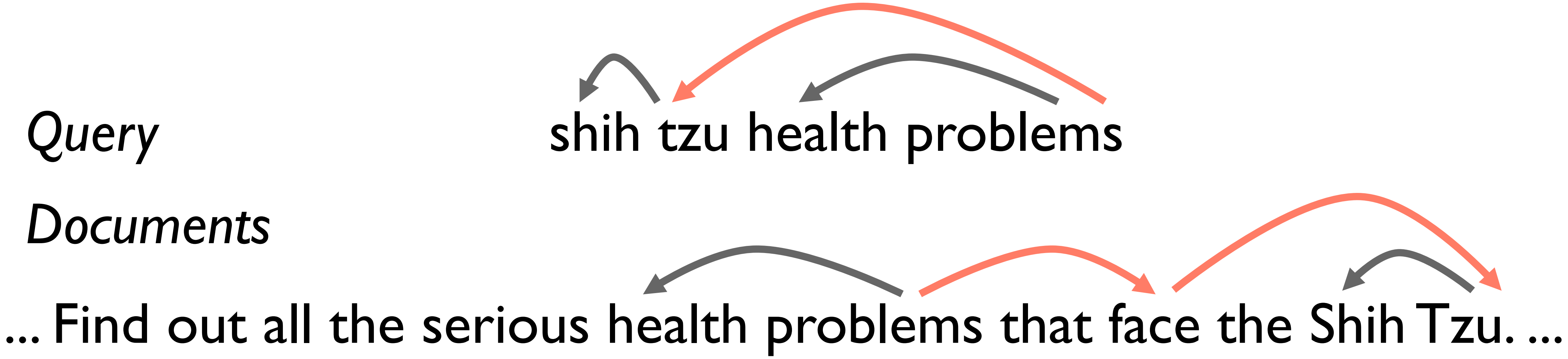
Query shih tzu health problems

Documents

... Find out all the serious health problems that face the Shih Tzu. ...

Quasi-Synchronous Dependence

- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT* 2006; *EMNLP* 2009)



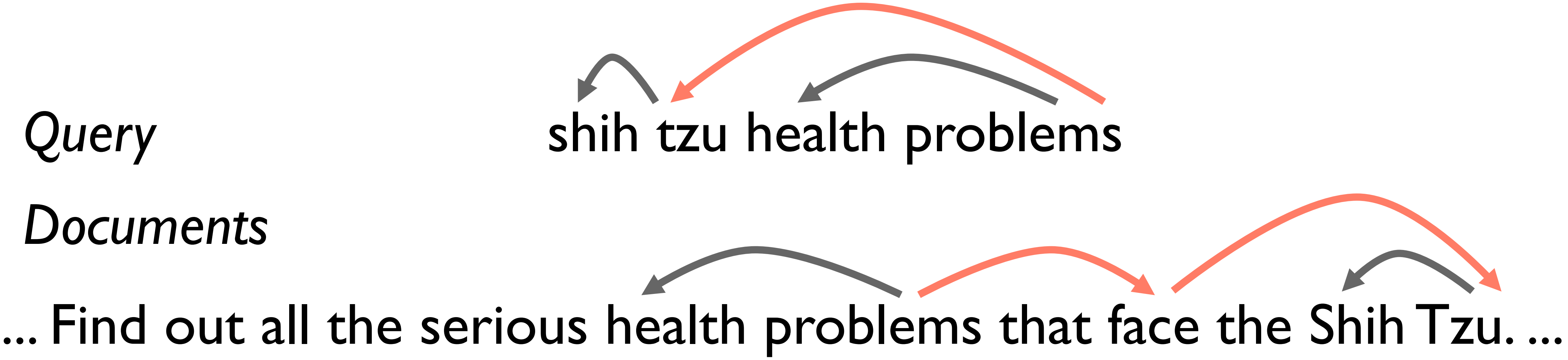
Query shih tzu health problems

Documents

... Find out all the serious health problems that face the Shih Tzu. ...

Quasi-Synchronous Dependence

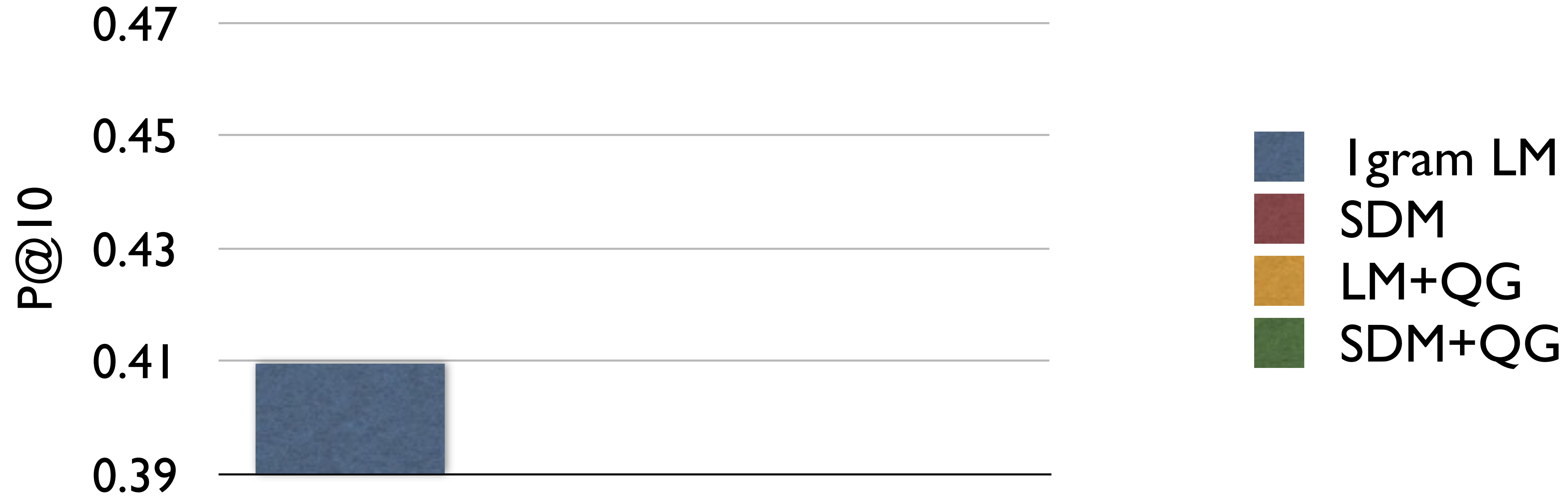
- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT* 2006; *EMNLP* 2009)



Query shih tzu health problems

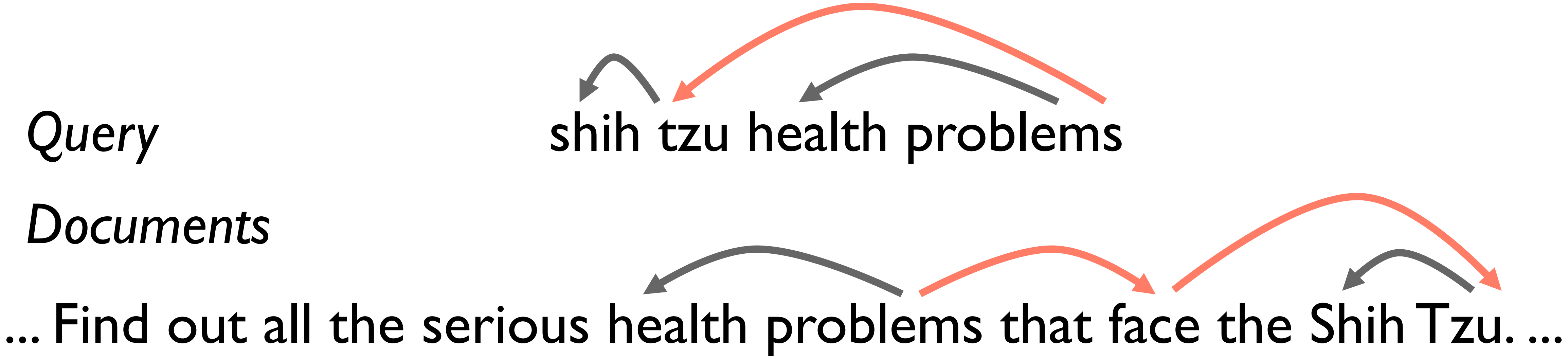
Documents

... Find out all the serious health problems that face the Shih Tzu. ...



Quasi-Synchronous Dependence

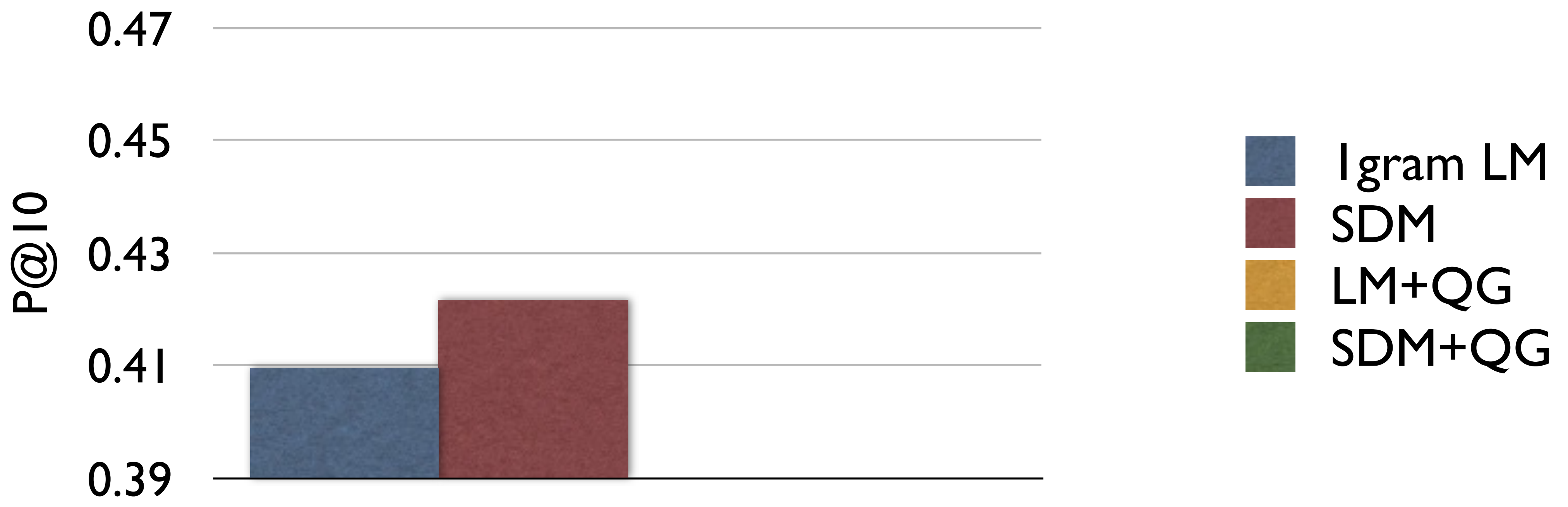
- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT* 2006; *EMNLP* 2009)



Query shih tzu health problems

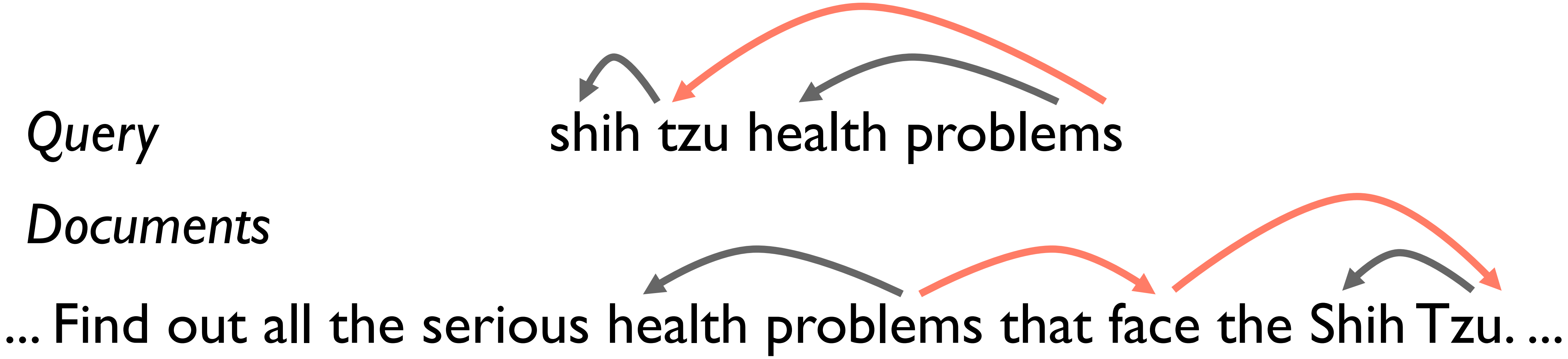
Documents

... Find out all the serious health problems that face the Shih Tzu. ...



Quasi-Synchronous Dependence

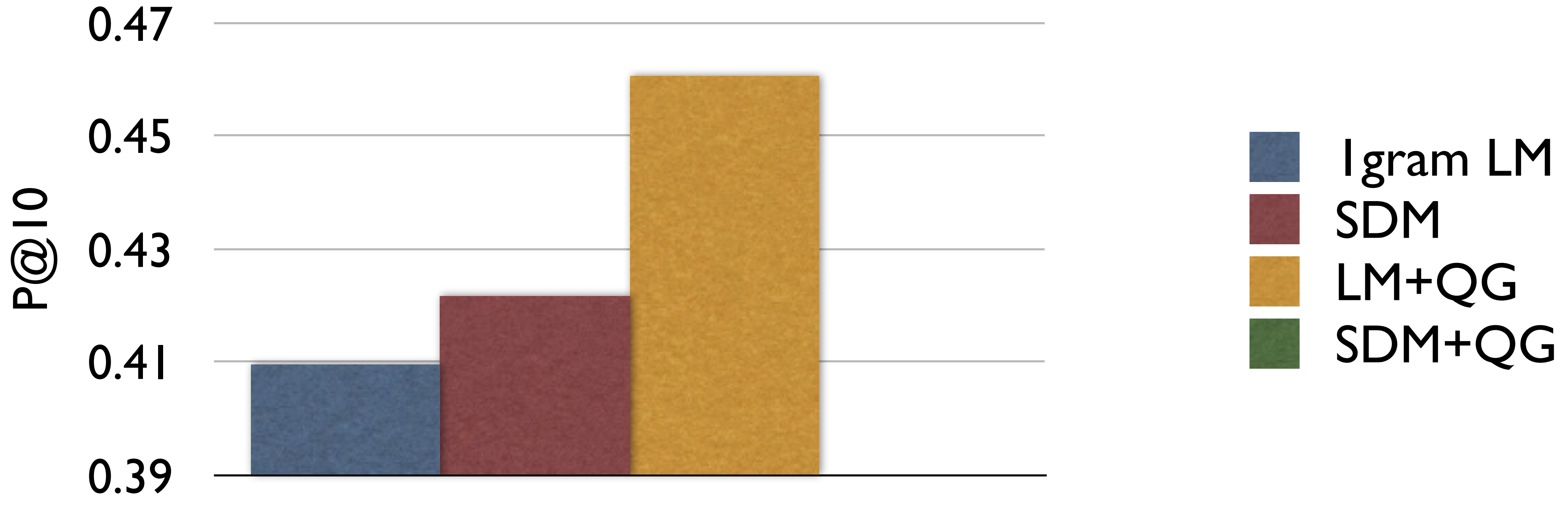
- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT* 2006; *EMNLP* 2009)



Query shih tzu health problems

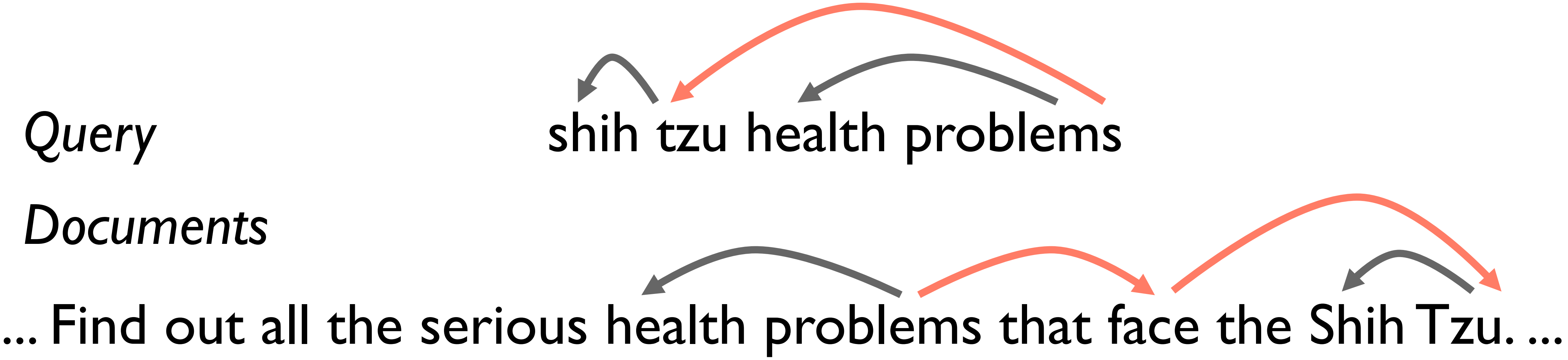
Documents

... Find out all the serious health problems that face the Shih Tzu. ...



Quasi-Synchronous Dependence

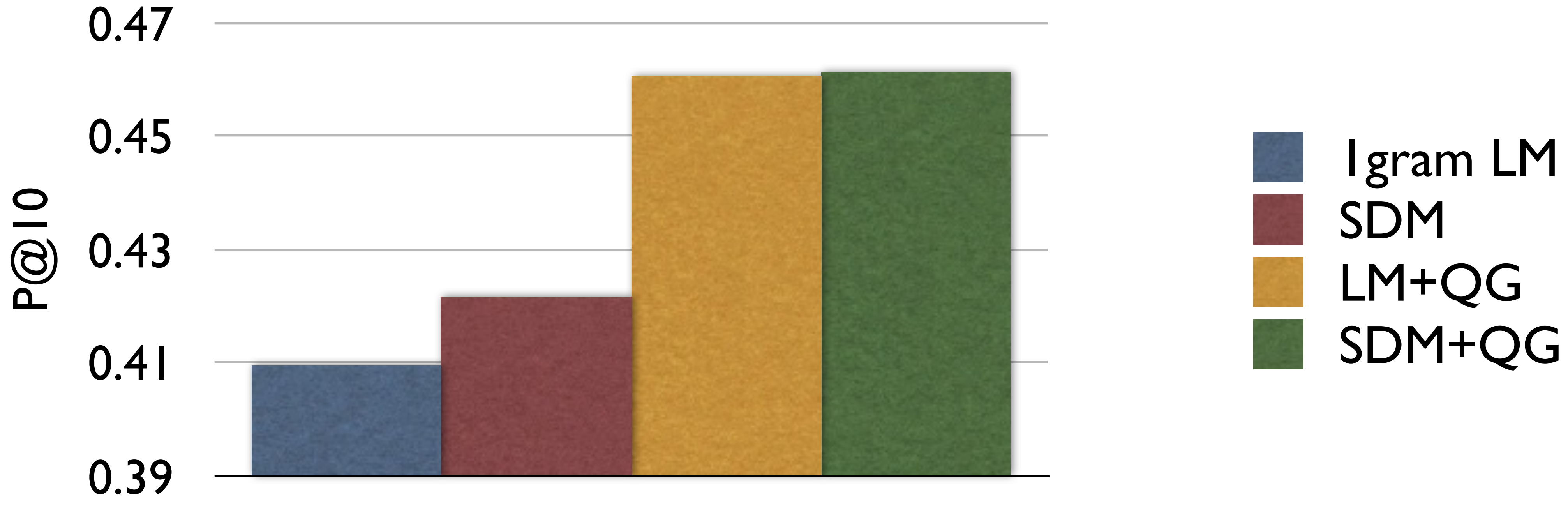
- Syntactic models of document mismatch
  - Developed (by me) for MT (*SMT* 2006; *EMNLP* 2009)



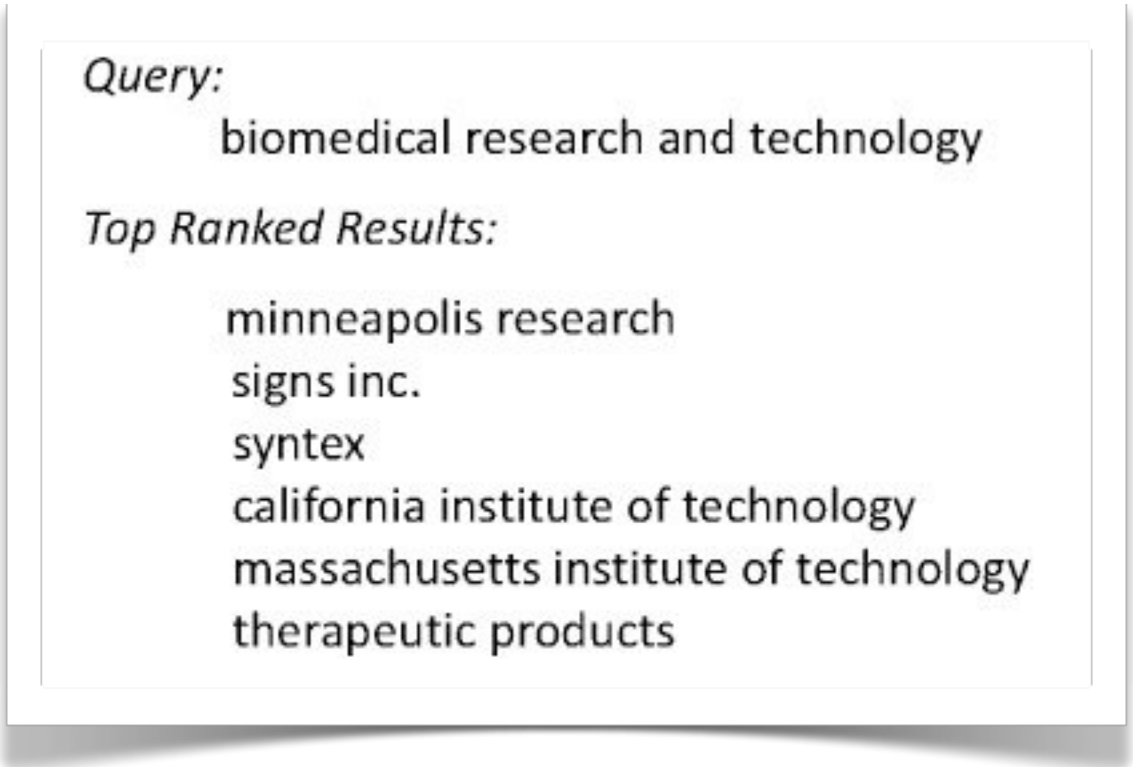
Query shih tzu health problems

Documents

... Find out all the serious health problems that face the Shih Tzu. ...



Entity Search

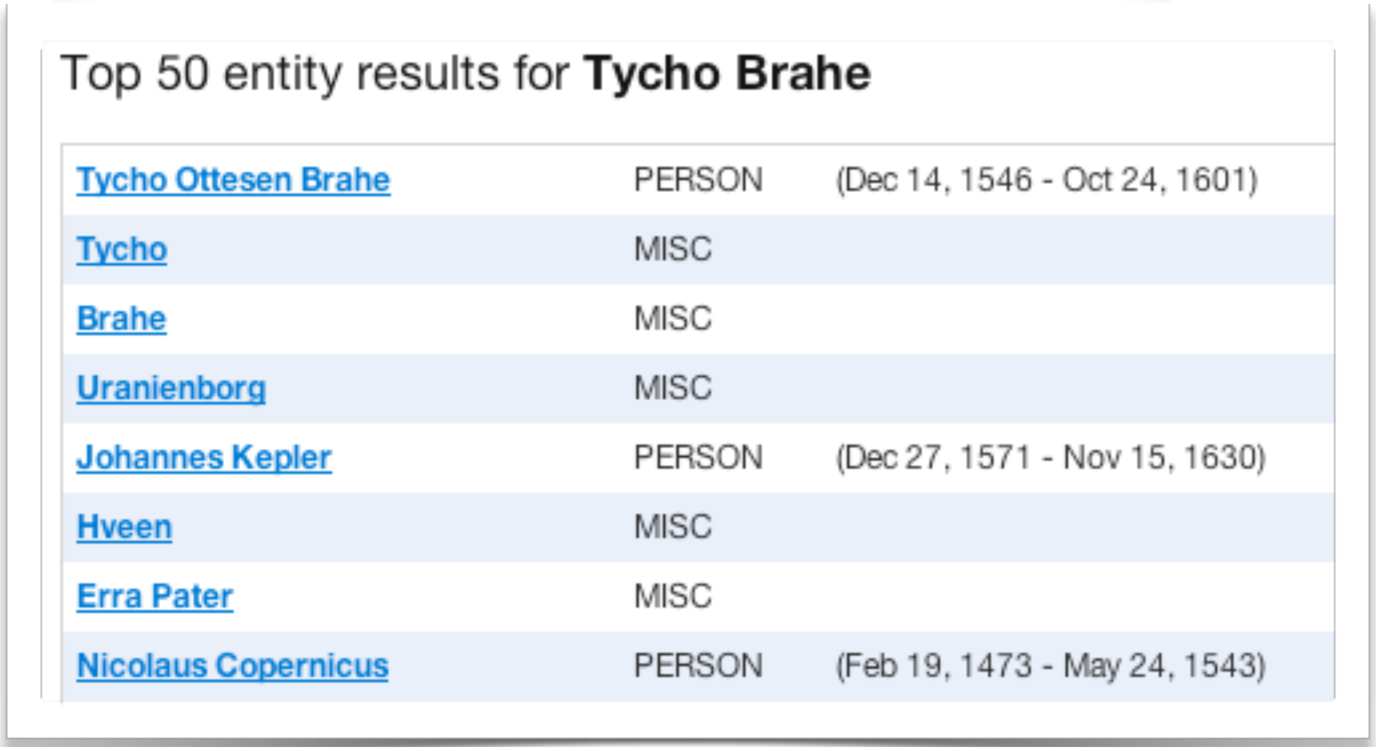


Query:

biomedical research and technology

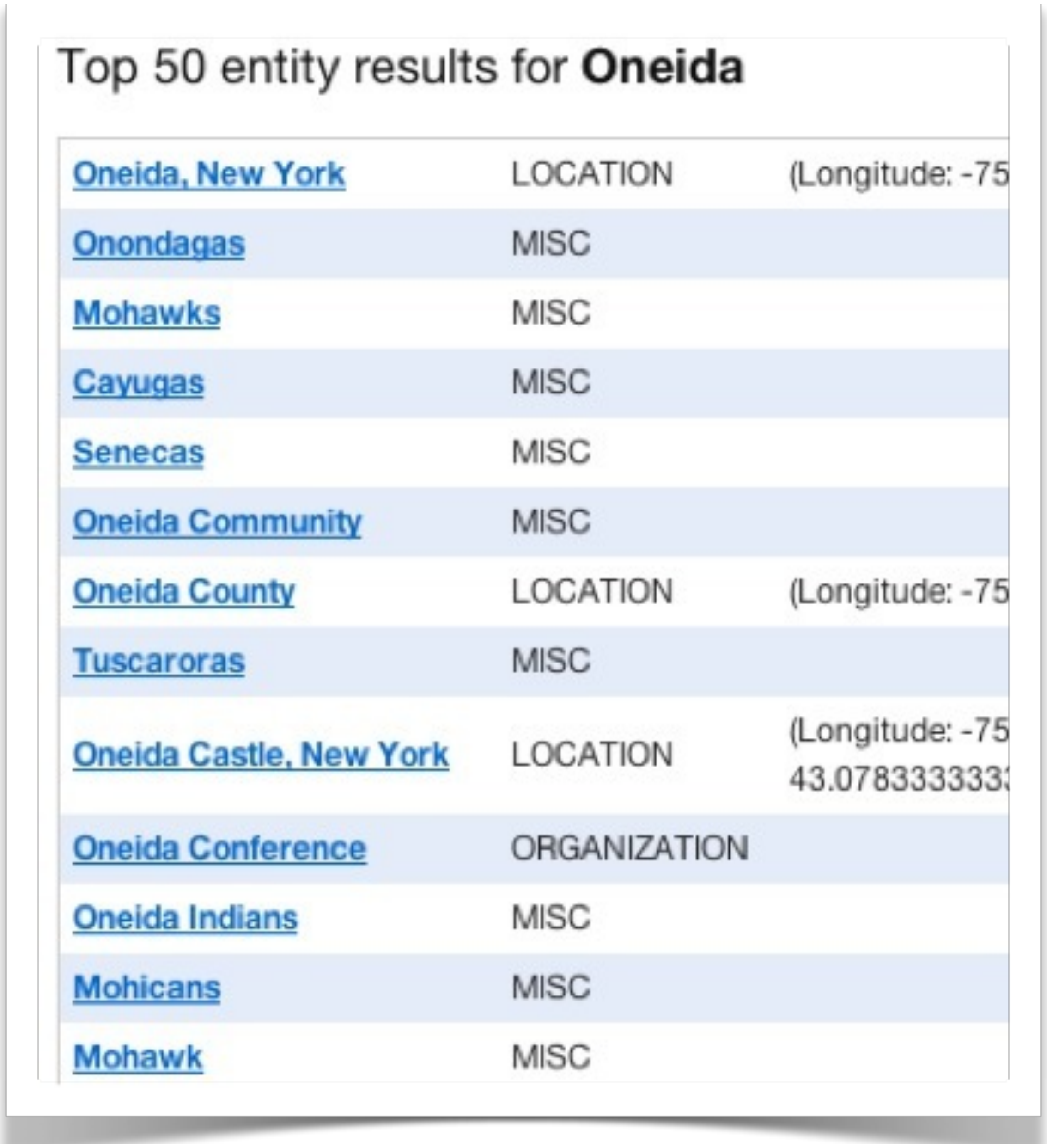
Top Ranked Results:

minneapolis research
signs inc.
syntex
california institute of technology
massachusetts institute of technology
therapeutic products



Top 50 entity results for **Tycho Brahe**

Tycho Ottesen Brahe	PERSON	(Dec 14, 1546 - Oct 24, 1601)
Tycho	MISC	
Brahe	MISC	
Uranienborg	MISC	
Johannes Kepler	PERSON	(Dec 27, 1571 - Nov 15, 1630)
Hveen	MISC	
Erra Pater	MISC	
Nicolaus Copernicus	PERSON	(Feb 19, 1473 - May 24, 1543)



Top 50 entity results for **Oneida**

Oneida, New York	LOCATION	(Longitude: -75
Onondagas	MISC	
Mohawks	MISC	
Cayugas	MISC	
Senecas	MISC	
Oneida Community	MISC	
Oneida County	LOCATION	(Longitude: -75
Tuscaroras	MISC	
Oneida Castle, New York	LOCATION	(Longitude: -75 43.078333333
Oneida Conference	ORGANIZATION	
Oneida Indians	MISC	
Mohicans	MISC	
Mohawk	MISC	

Entity Search

$$P(e, q) = \sum_{d \in D} P(e, q|d)P(d)$$

Joint distribution over queries and entities

$$P(e, q|d) = P(q|e, d)P(e|d)$$

Factorized dist’n

$$P(q|e, d) = \frac{1}{Z}\sum_{i=1}^{N} \delta_d(i, q)k(q, e)$$

Probability by proximity

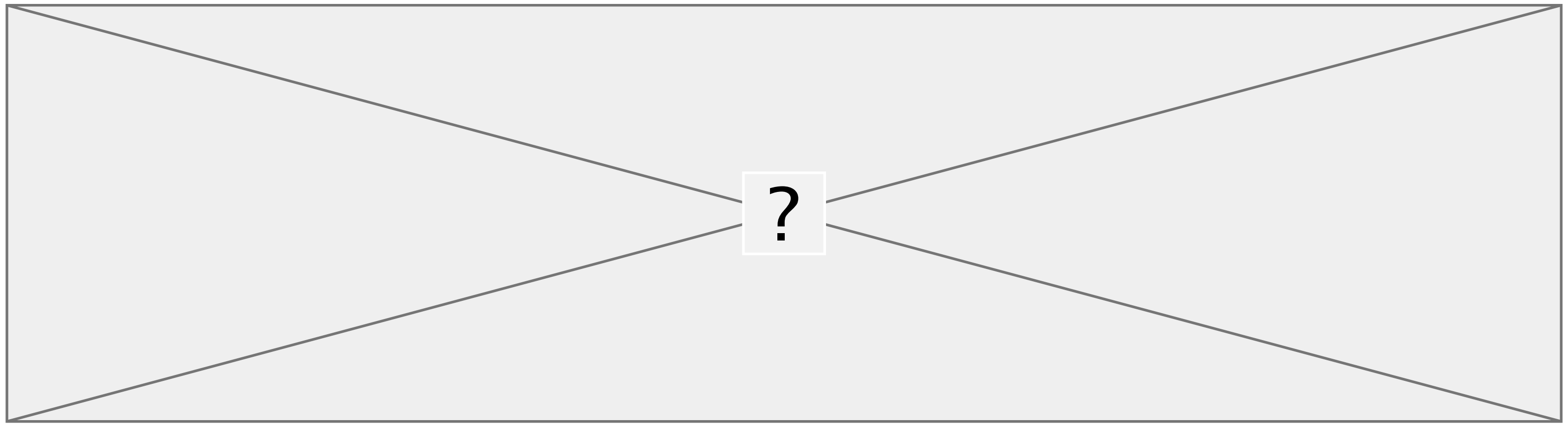
$$\exp -||q - e||^2/2\sigma^2$$

Gaussian proximity kernel

Question Answering

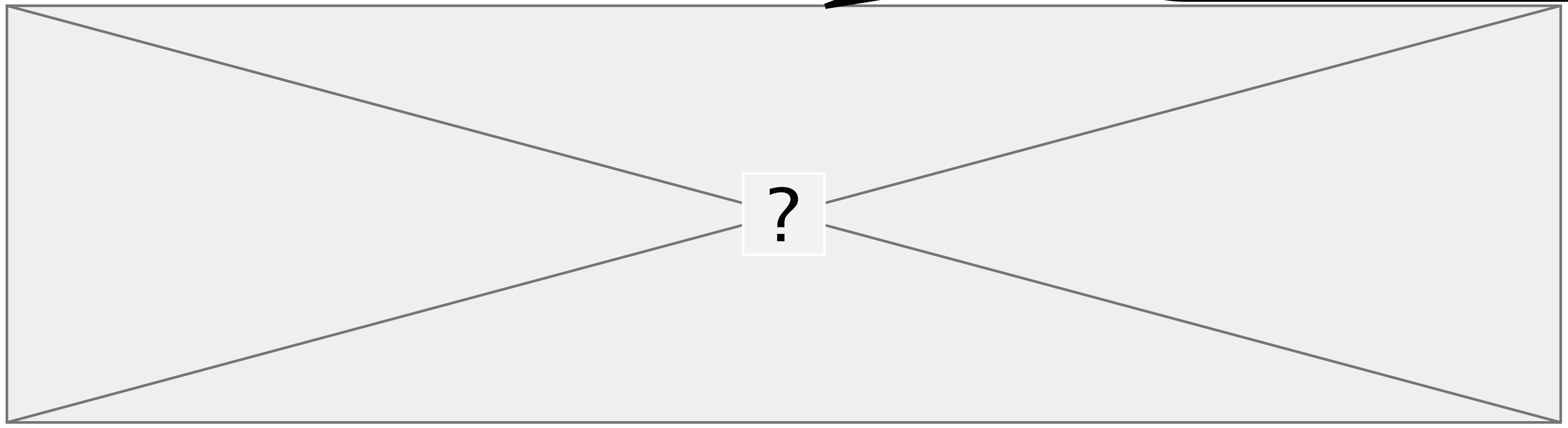
- Human vs. machine: Compare search engine queries with, e.g., StackOverflow
- Current QA systems perform simple "factoid" retrieval
  - Who invented the paper clip?
  - Where is the Valley of the Kings?
  - When was the last major eruption of Mt. St. Helens?

Question Answering

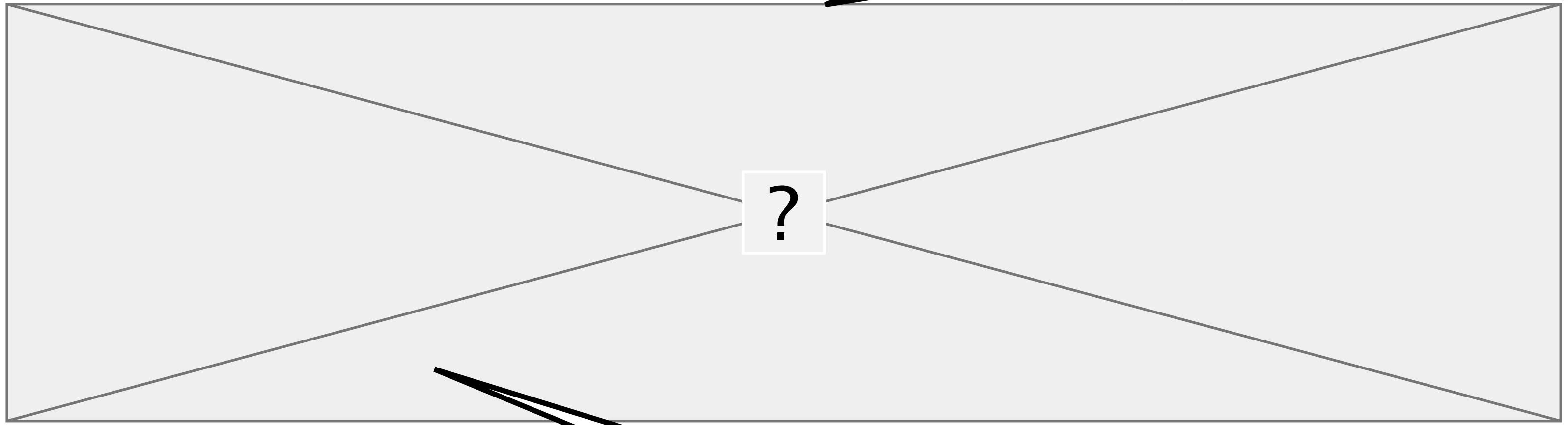


Question Answering

Example: Indri system used as first stage of IBM Watson



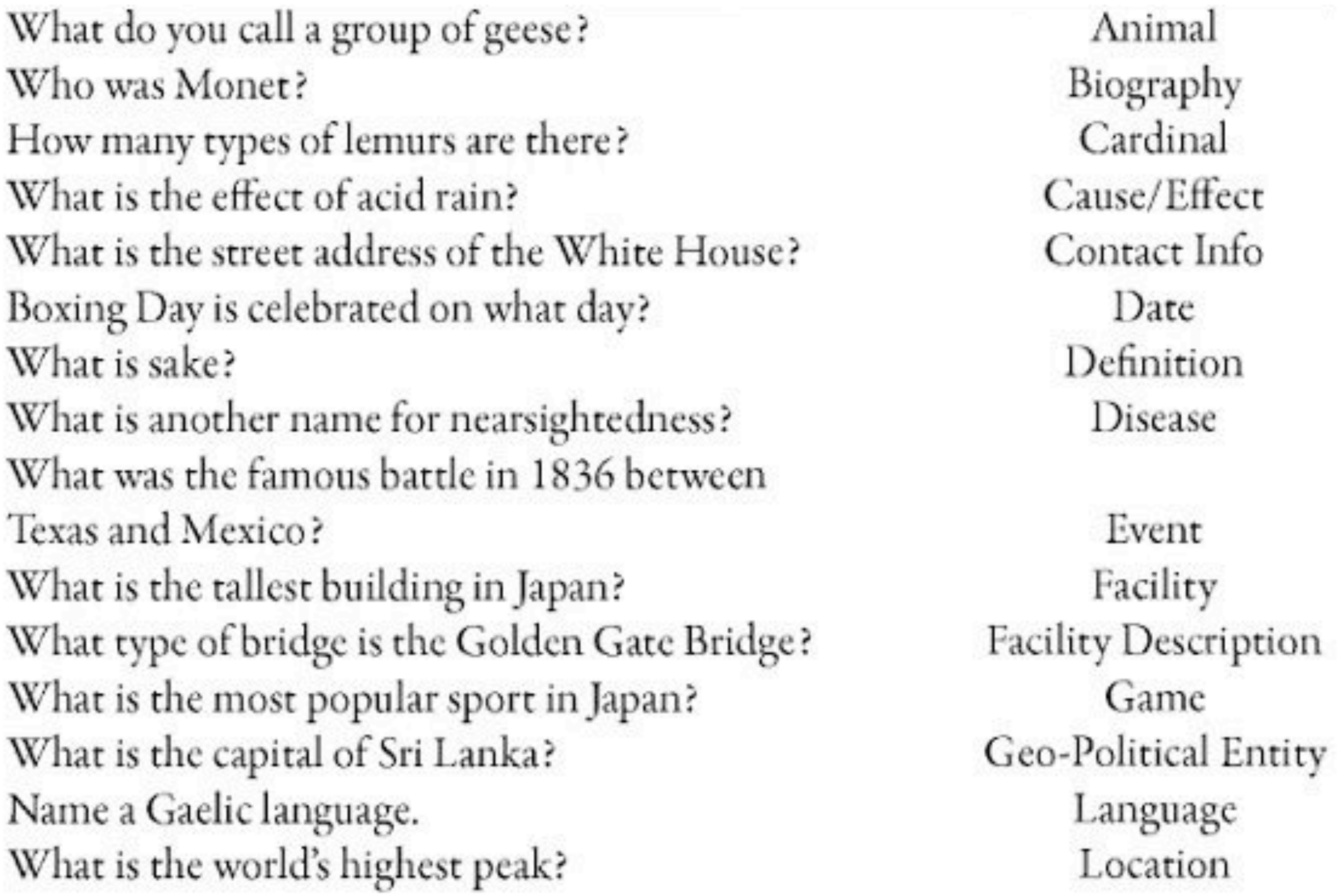
Question Answering



Example: Indri system used as first stage of IBM Watson

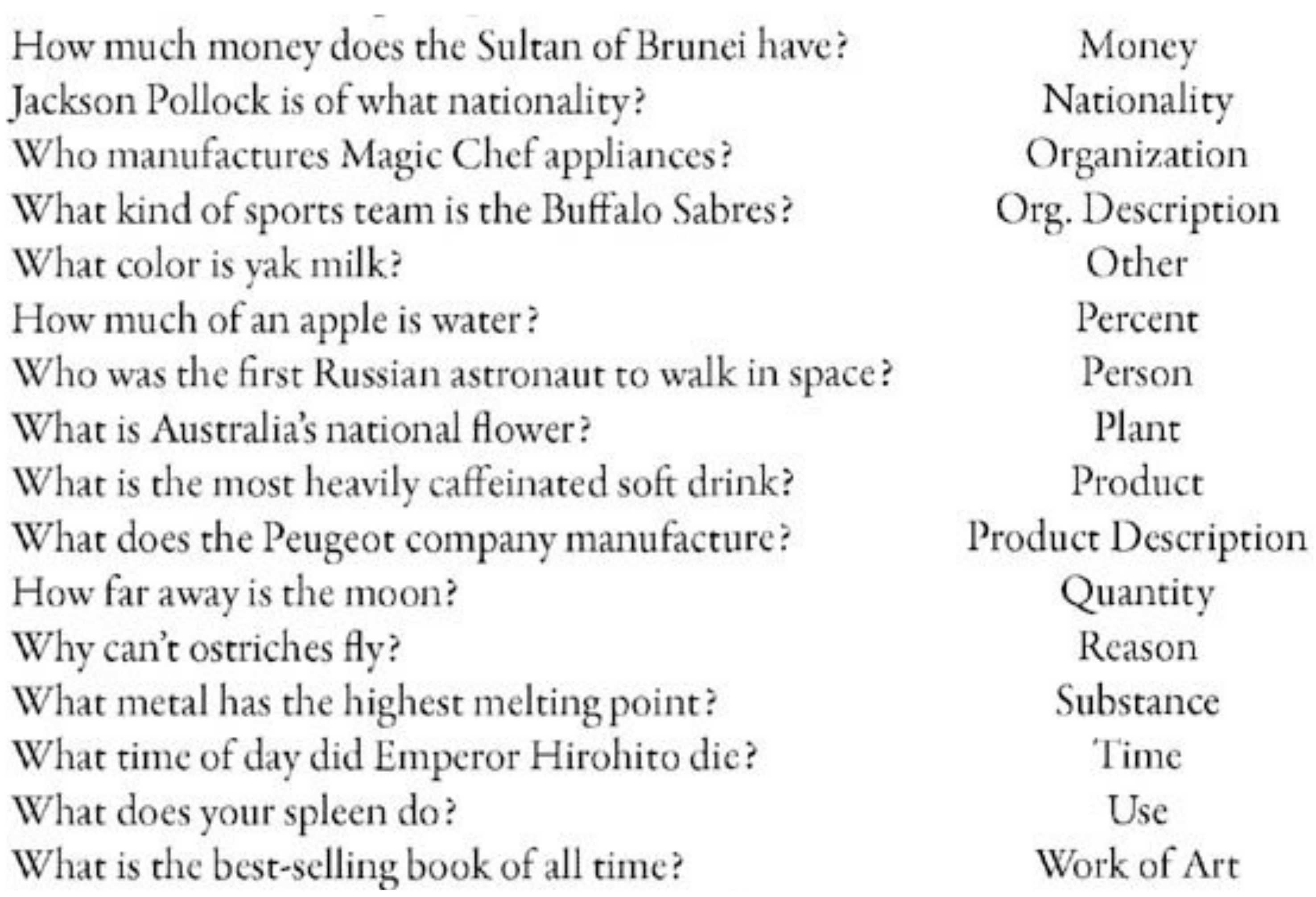
What kind of answer is the user expecting?

Answer Categorization



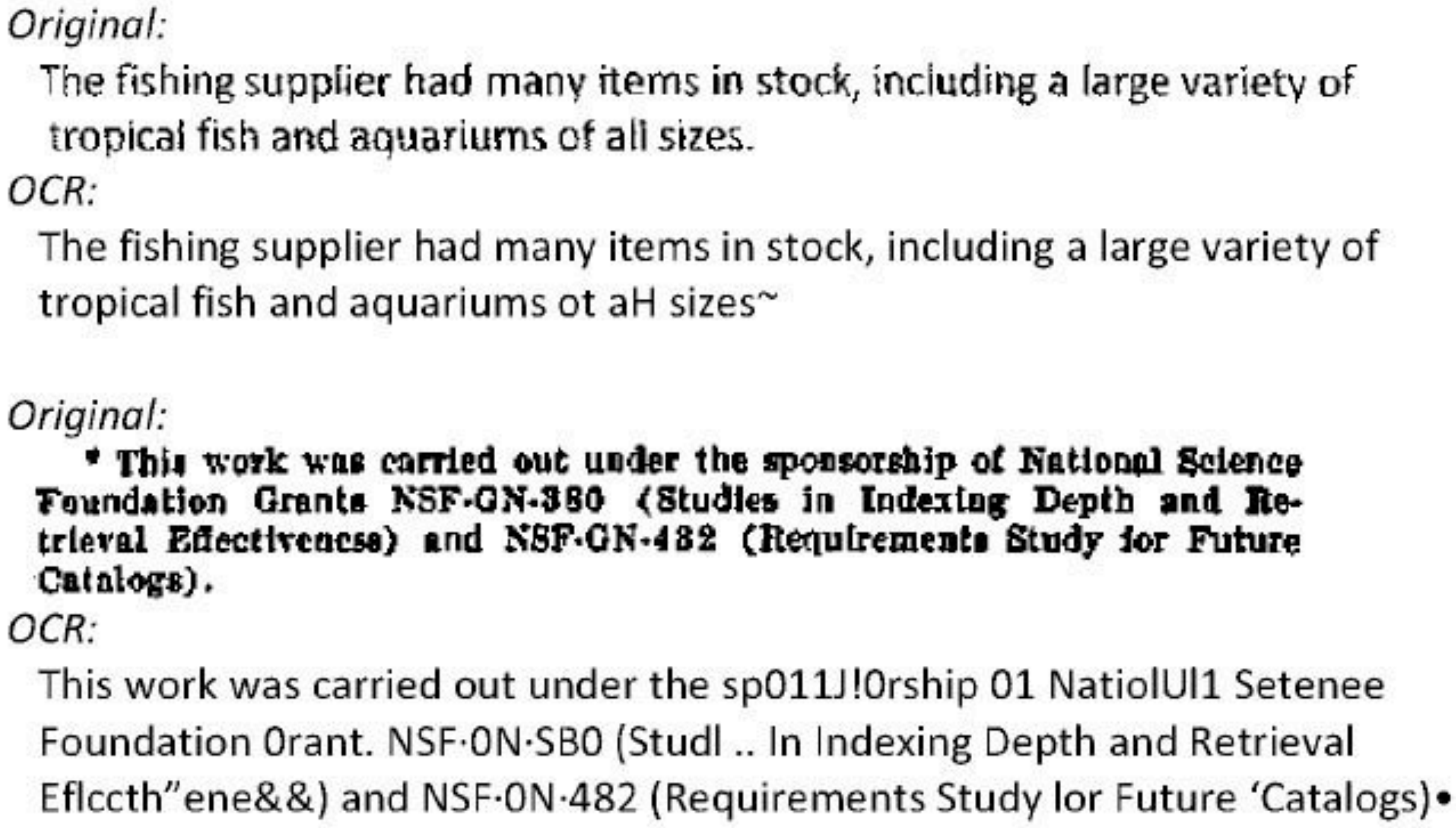
What do you call a group of geese?	Animal
Who was Monet?	Biography
How many types of lemurs are there?	Cardinal
What is the effect of acid rain?	Cause/Effect
What is the street address of the White House?	Contact Info
Boxing Day is celebrated on what day?	Date
What is sake?	Definition
What is another name for nearsightedness?	Disease
What was the famous battle in 1836 between Texas and Mexico?	Event
What is the tallest building in Japan?	Facility
What type of bridge is the Golden Gate Bridge?	Facility Description
What is the most popular sport in Japan?	Game
What is the capital of Sri Lanka?	Geo-Political Entity
Name a Gaelic language.	Language
What is the world's highest peak?	Location

Answer Categorization



How much money does the Sultan of Brunei have?	Money
Jackson Pollock is of what nationality?	Nationality
Who manufactures Magic Chef appliances?	Organization
What kind of sports team is the Buffalo Sabres?	Org. Description
What color is yak milk?	Other
How much of an apple is water?	Percent
Who was the first Russian astronaut to walk in space?	Person
What is Australia's national flower?	Plant
What is the most heavily caffeinated soft drink?	Product
What does the Peugeot company manufacture?	Product Description
How far away is the moon?	Quantity
Why can't ostriches fly?	Reason
What metal has the highest melting point?	Substance
What time of day did Emperor Hirohito die?	Time
What does your spleen do?	Use
What is the best-selling book of all time?	Work of Art

OCR Transcripts



Original:

The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums of all sizes.

OCR:

The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums ot aH sizes~

Original:

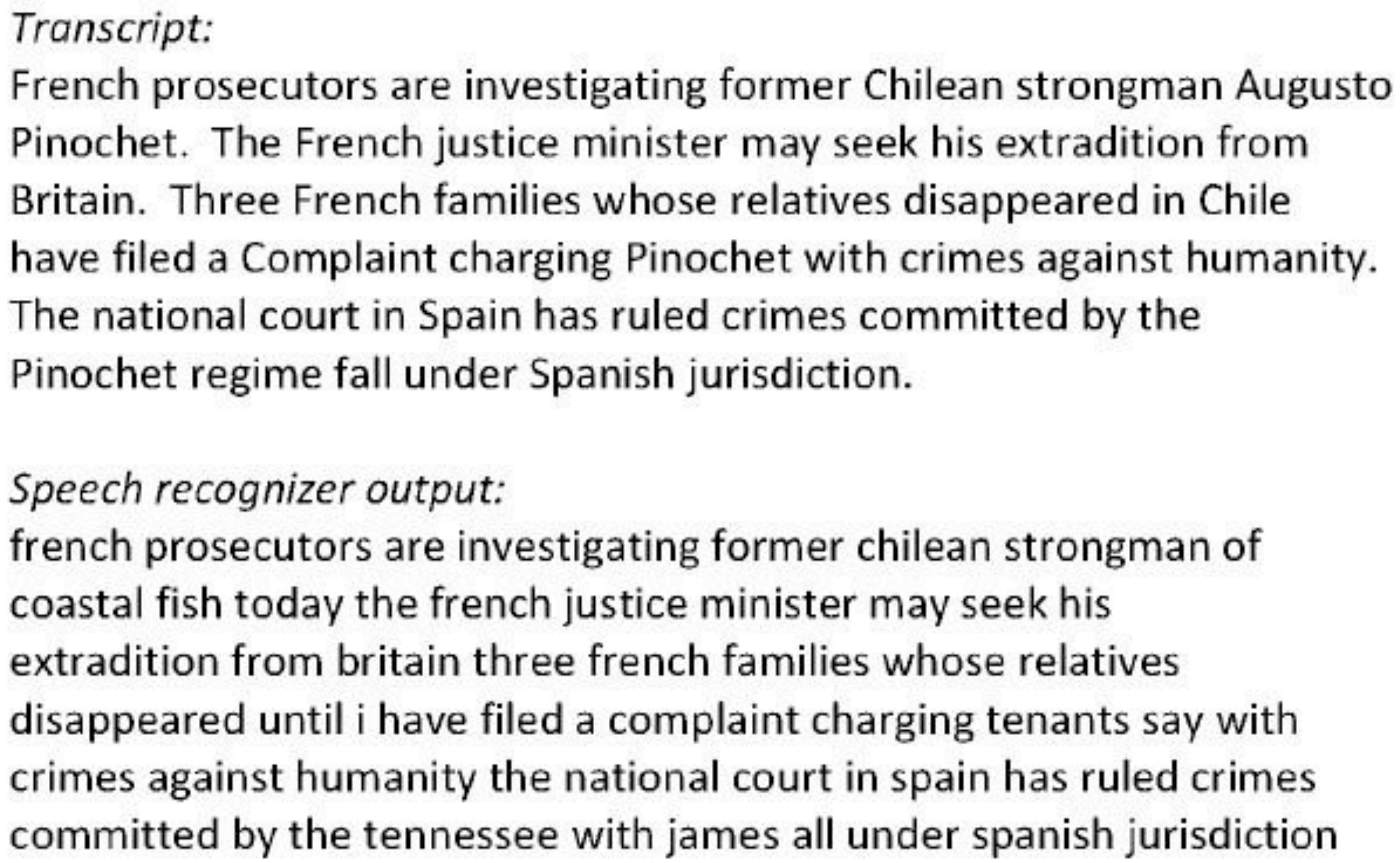
*** This work was carried out under the sponsorship of National Science Foundation Grants NSF-GN-380 (Studies in Indexing Depth and Retrieval Effectiveness) and NSF-GN-482 (Requirements Study for Future Catalogs).**

OCR:

This work was carried out under the sp011J!0rship 01 NatiolUl1 Setenee Foundation 0rant. NSF·0N·SB0 (Studl .. In Indexing Depth and Retrieval Eflccth"ene&&) and NSF·0N·482 (Requirements Study lor Future 'Catalogs)•

Optical character recognition

ASR Transcripts



Transcript:

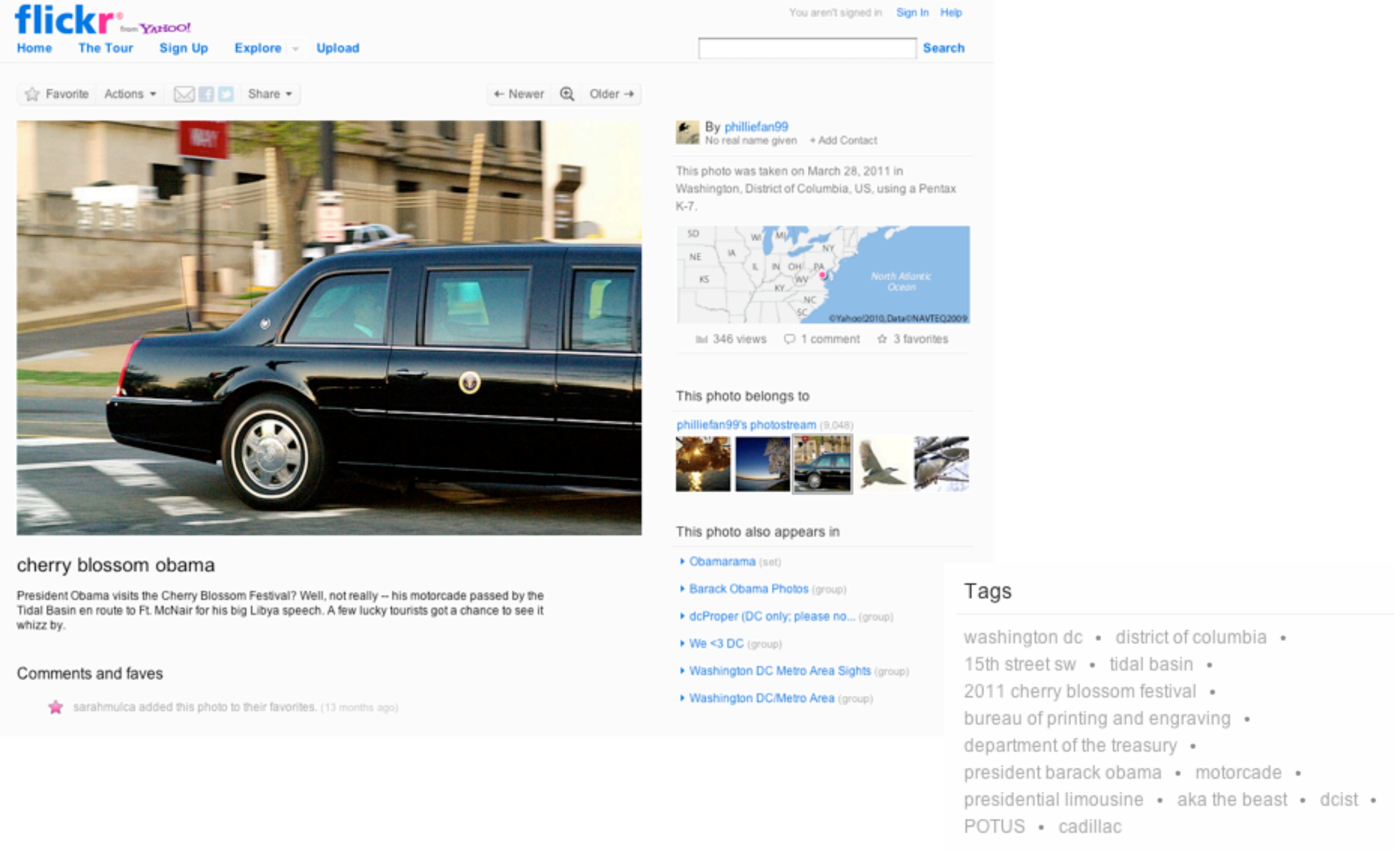
French prosecutors are investigating former Chilean strongman Augusto Pinochet. The French justice minister may seek his extradition from Britain. Three French families whose relatives disappeared in Chile have filed a Complaint charging Pinochet with crimes against humanity. The national court in Spain has ruled crimes committed by the Pinochet regime fall under Spanish jurisdiction.

Speech recognizer output:

french prosecutors are investigating former chilean strongman of coastal fish today the french justice minister may seek his extradition from britain three french families whose relatives disappeared until i have filed a complaint charging tenants say with crimes against humanity the national court in spain has ruled crimes committed by the tennessee with james all under spanish jurisdiction

Automatic speech recognition

Image Tagging



By philliefan99
No real name given + Add Contact

This photo was taken on March 28, 2011 in Washington, District of Columbia, US, using a Pentax K-7.

346 views 1 comment 3 favorites

This photo belongs to

philliefan99's photostream (9,048)

This photo also appears in

- Obamarama (set)
- Barack Obama Photos (group)
- dcProper (DC only; please no... (group)
- We <3 DC (group)
- Washington DC Metro Area Sights (group)
- Washington DC/Metro Area (group)

cherry blossom obama

President Obama visits the Cherry Blossom Festival? Well, not really -- his motorcade passed by the Tidal Basin en route to Ft. McNair for his big Libya speech. A few lucky tourists got a chance to see it whizz by.

Comments and faves

sarahmulca added this photo to their favorites. (13 months ago)

Tags

washington dc • district of columbia • 15th street sw • tidal basin • 2011 cherry blossom festival • bureau of printing and engraving • department of the treasury • president barack obama • motorcade • presidential limousine • aka the beast • dcist • POTUS • cadillac

Image Annotation

- Offline annotation
- Online pseudo-relevance feedback
  - Retrieve images by text annotation
  - Estimate text/image relevance model
  - Rerank more images

people, pool,
swimmers, water

cars, formula,
tracks, wall

clouds, jet,
plane, sky

fox, forest,
river, water