Data-Driven Machine Translation:

a conversation with linguistics and translation studies

AMTA 2006, Boston

Daniel Marcu, marcu@isi.edu
and
Alan K. Melby, akmtrg@byu.edu

New Optimism in MT Community

2006 June 30:

http://businessnetwork.smh.com.au/articles/2006/06/30/5104.html

- Within the next few years there will be an explosion in translation technologies, says Alex Waibel, director of the International Centre for Advanced Communication Technology…
- How far can machine translators be taken? "There is no reason why they should not become as good, if not better, than humans," Dr Waibel says.

Part 1: Challenges Ahead for Data-driven Machine Translation

- a: Comparison with human qualifications
- b: Avoidance of compositionality assumption
- c: Using relevant co-text (beyond sentence)
- d: Using relevant "extra-text" (real world info)
- e: Displaying "second-order creativity"

(creating novel solutions and detecting need)

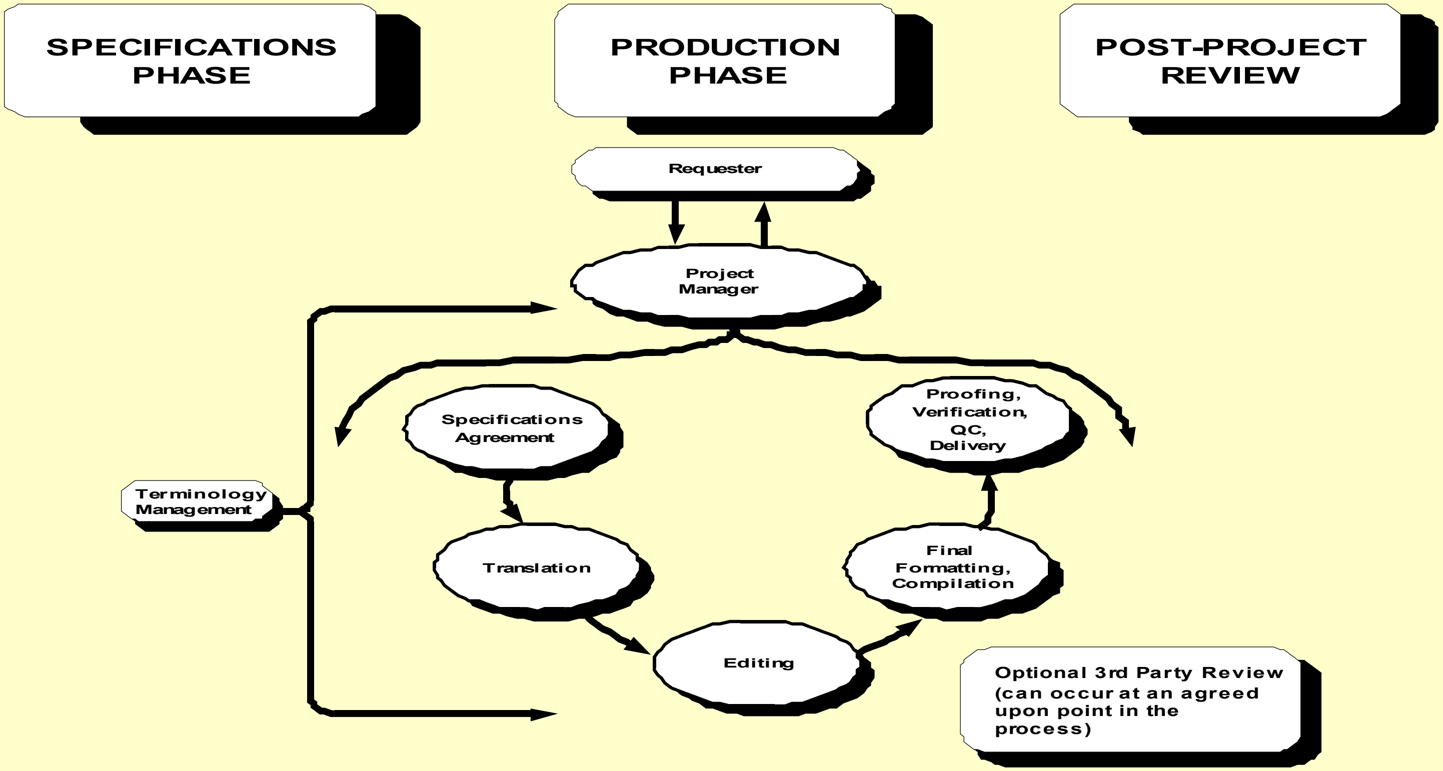
Challenge 1:
Comparison with Human Qualifications

Challenge 1:
Comparison with Human Qualifications

- Display same qualifications required of human translators or explain why some are not needed for data-driven machine translation systems

Human Translation Project Phases

(ASTM)



Specifications Phase

- Begin with:
  - Source test
  - Target language
  - Target audience
  - Purpose of translation
- Negotiate:
  - Specifications for this project

Production Phase

- Specifications Agreement (mode adjustment)
- Translation (actual translation)
- Editing (source- vs. target-text comparison)
- Formatting (e.g. integrate source format)
- Proofing (monolingual target-text check)

Some qualifications needed for human translators

- Ability to *understand* source text
- Ability to *write* in target language
- Ability to *adjust* to audience and purpose, when translating and evaluating whether source and target texts correspond

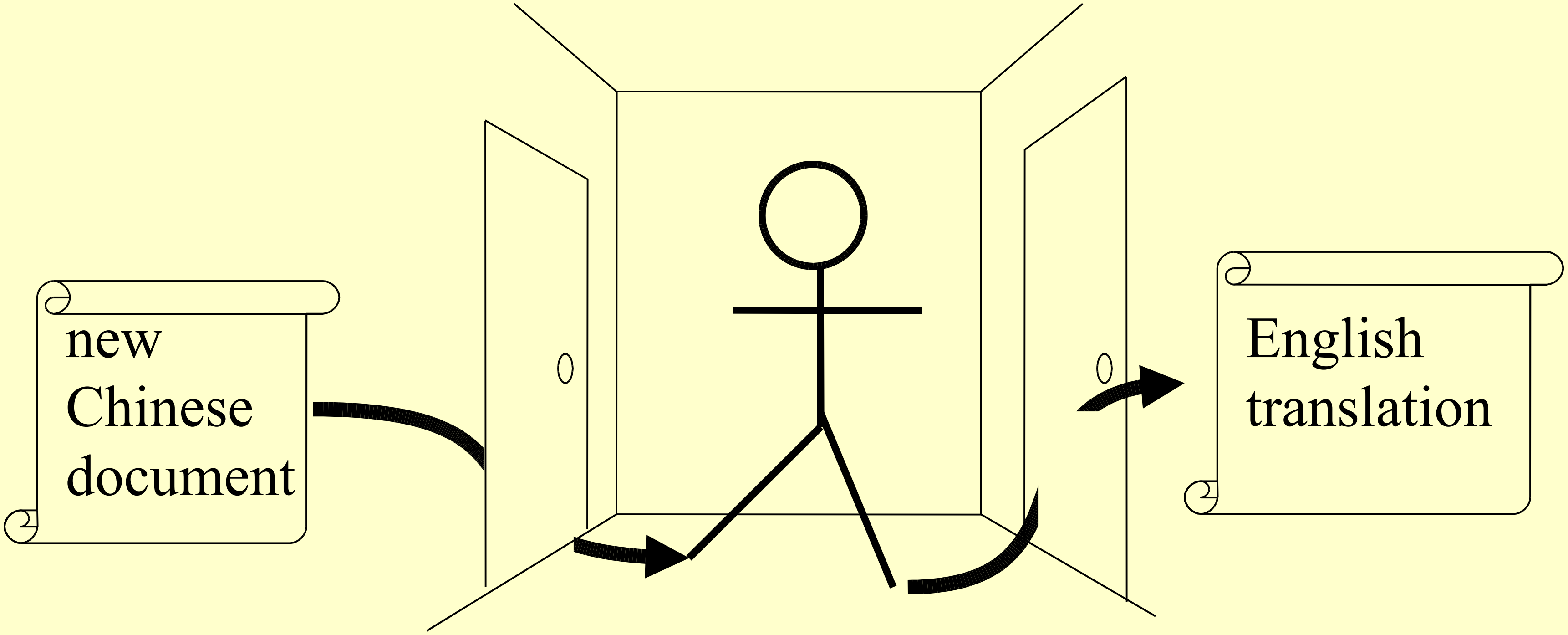
Audience and Purpose

- Same source text may be translated very differently, depending on audience and purpose
  - A story could be translated for easy reading and the storyline (adjusted for target culture)
  - Same story could be translated for access to the source culture by those who can't read original

Data-driven Comments on Challenge 1

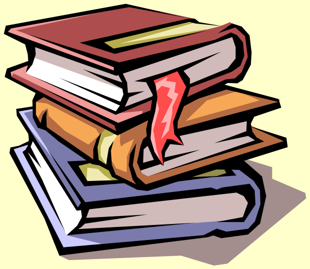
Airplanes don’t bat their wings, but they still fly.

Chinese Room Experiment



new Chinese document

English translation



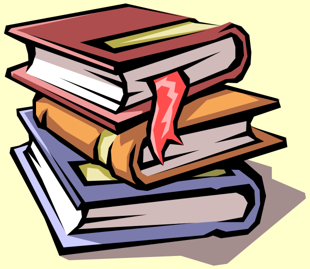
Chinese texts with English translations

Chinese word or phrase => sentence pairs containing it

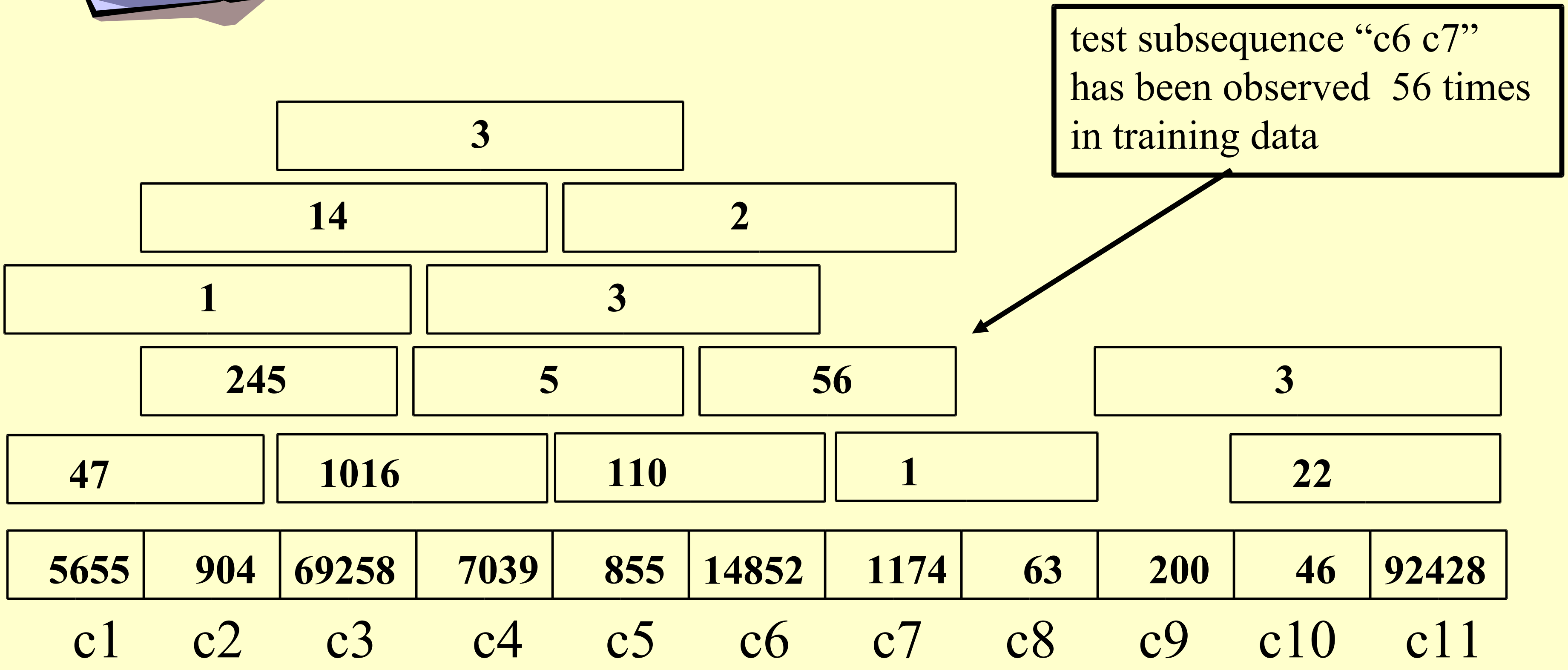
HIGH ACCURACY

DOES THIS PERSON KNOW CHINESE?

Chinese Room Experiment



170k sentence pairs of bilingual training data
(3.5m words translated)



test subsequence “c6 c7” has been observed 56 times in training data

3

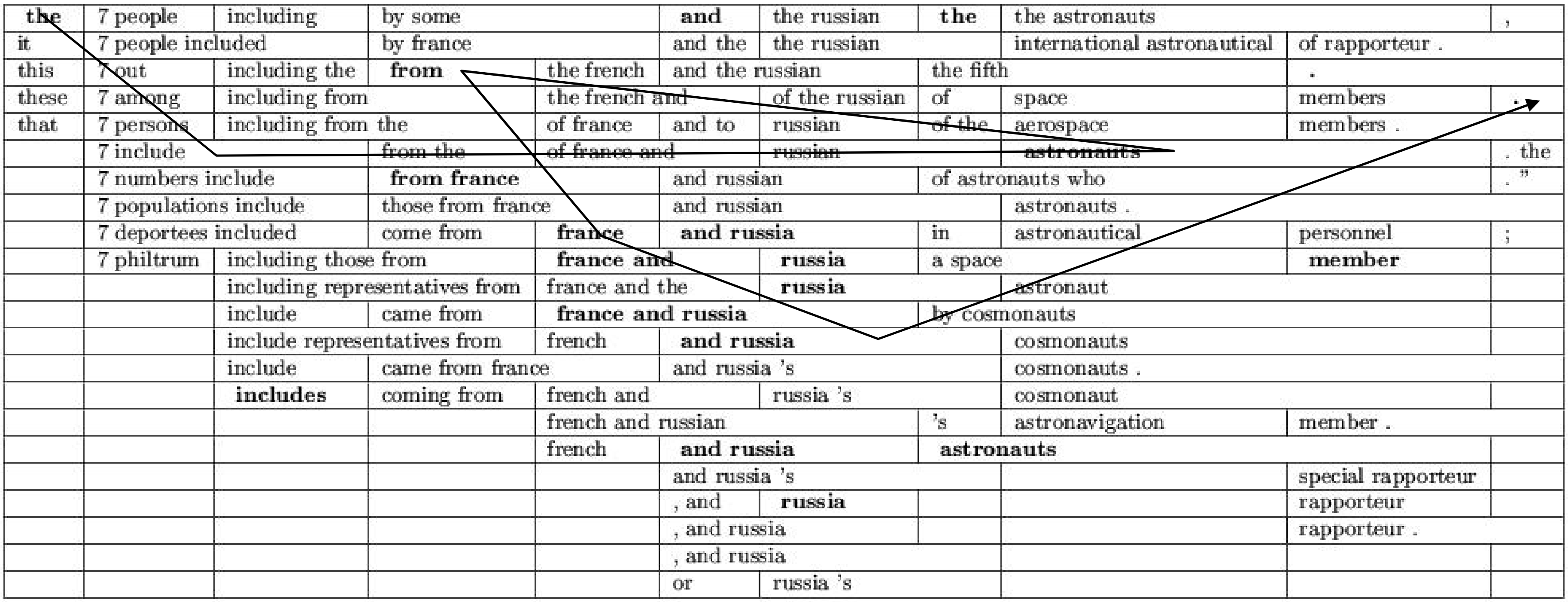
14 2

1 3

245 5 56 3

47 1016 110 1 22

5655	904	69258	7039	855	14852	1174	63	200	46	92428
c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11



c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		**.**	
these	7 among	including from		the french and		of the russian	of	space	members	**.**
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Discussion

- Not even humans need to know the source language in order to translate well.

- There is no evidence that state of the art SMT systems don’t understand the source language.

- Audience and purpose variations:
  - English paraphrasing.

Challenge 2:
Avoidance of compositionality assumption

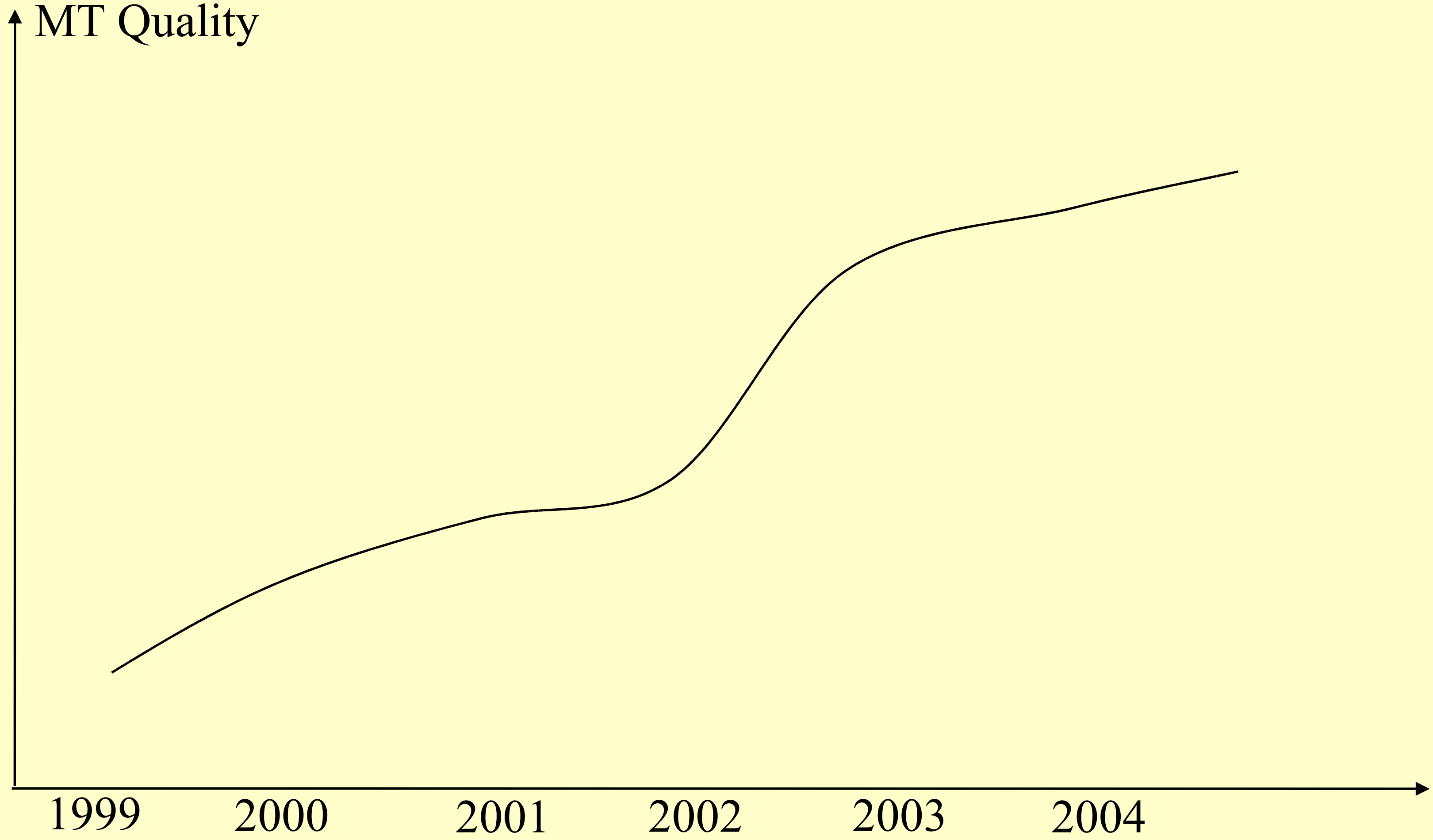
Compositionality: computation of the meaning of a sentence from the bottom up by combining context-free sub-meanings

Example of Non-compositionality

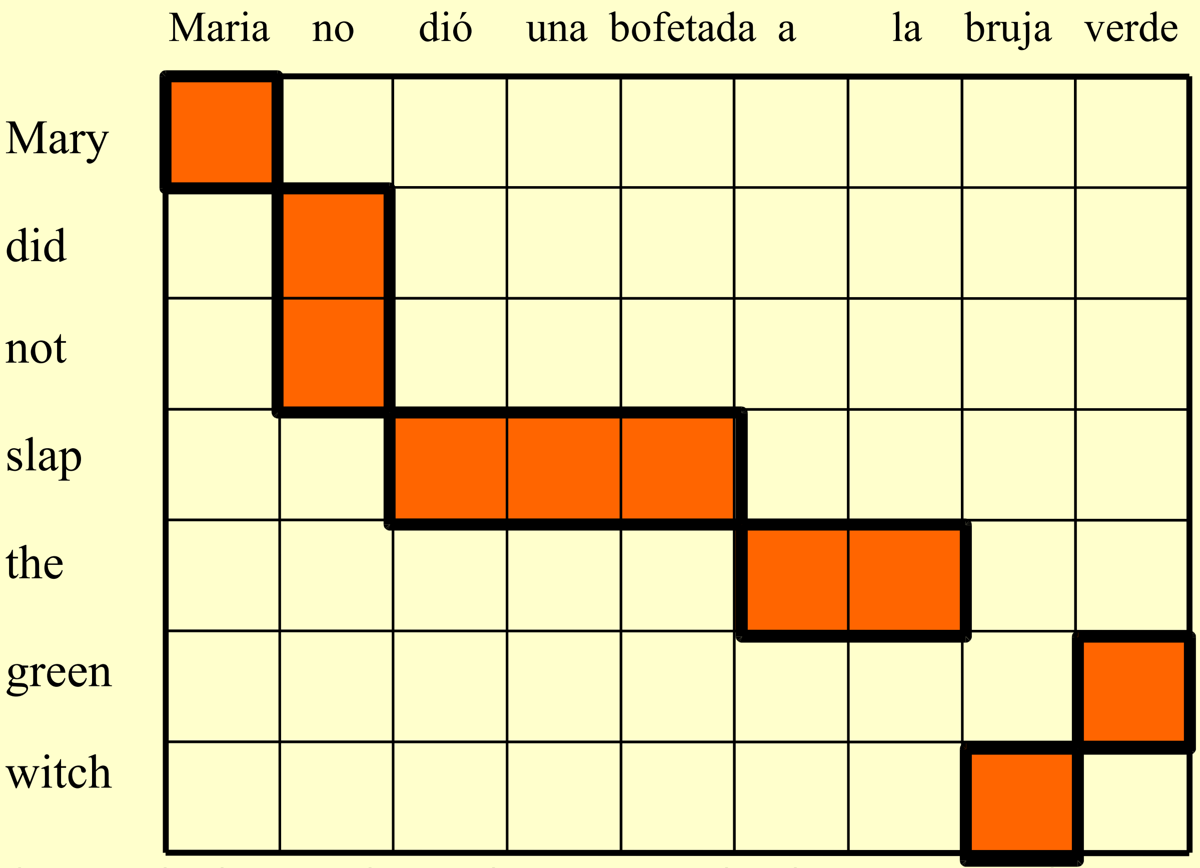
- From August 2006 Interview with Robert Longacre (received PhD same time as Chomsky)
  - Melby: What was it like to live through the Chomskyan Revolution?
  - Longacre: We were hit by a green sea.
  - Melby: Why a green sea?
  - Longacre: Because the ideas were not colorless
  - Note: "green sea" in this case is a severe storm

Data-driven Comments on Challenge 2

Data driven MT progress



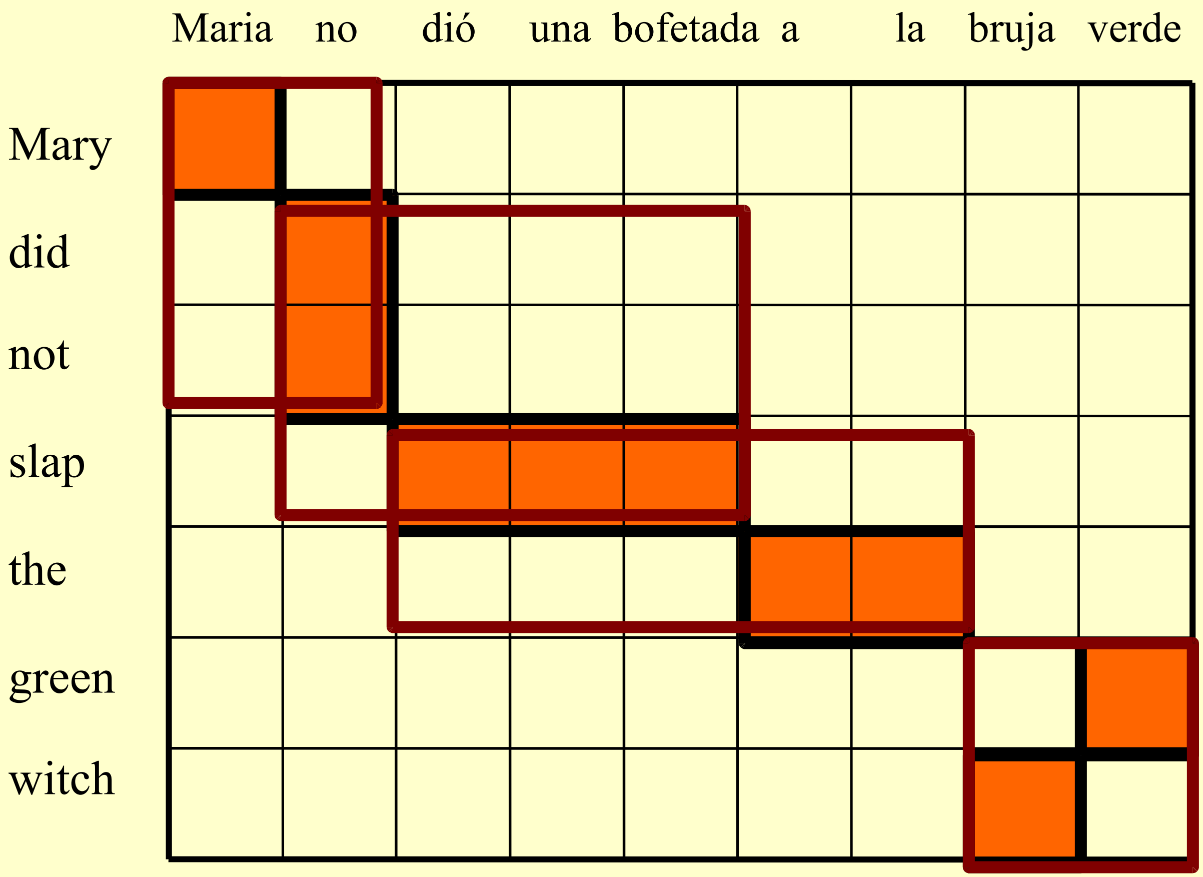
Viterbi alignments → word-to-word translation models



	Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap			■	■	■				
the						■	■		
green									■
witch								■	

t(Maria | Mary), t(no | did), t(no | not), …, t(bruja | witch), t(verde | green)

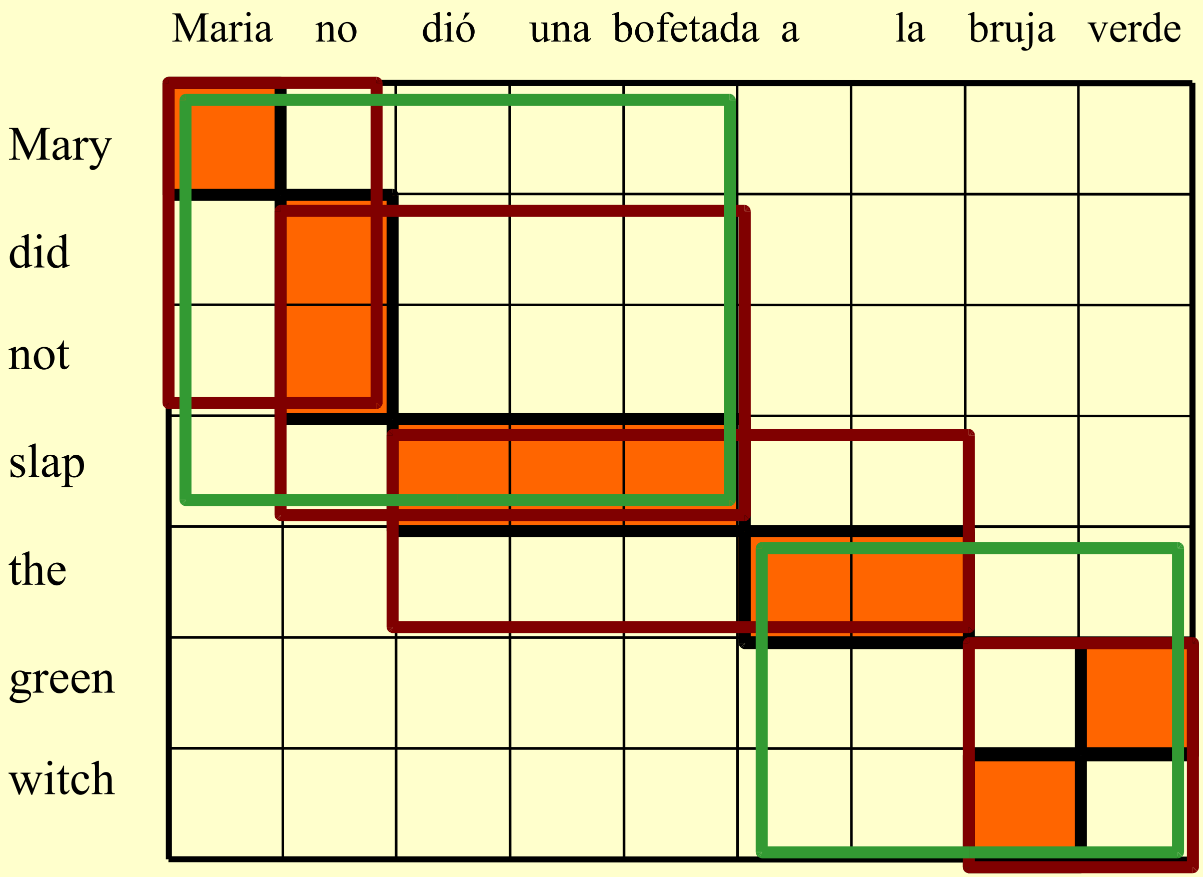
Viterbi alignments → phrase-to-phrase translation models



t(Maria | Mary), t(no | did), t(no | not), …, t(bruja | witch), t(verde | green)

t(Maria no | Mary did not), t(no dió una bofetada | did not slap), t(dió una bofetada a la | slap the)

Viterbi alignments → phrase-to-phrase translation models



	Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap			■	■	■				
the						■	■		
green									■
witch								■	

t(Maria | Mary), t(no | did), t(no | not), …, t(bruja | witch), t(verde | green)

t(Maria no | Mary did not), t(no dió una bofetada | did not slap), t(dió una bofetada a la | slap the)

t(Mary did not slap | Maria no dió una botefada), t(the green witch | a la bruja verde), …

Discussion

- Automatically learned phrase-to-phrase dictionary entries solve the compositionality problem – locally.
  - “real”
  - “estate”
  - “real estate”

- There is no evidence that MT suffers from a global compositionality problem.

Challenge 3:
Using relevant co-text

Often, translation decisions need to be sensitive to local context; sometimes they depend on co-text beyond the boundaries of the current sentence

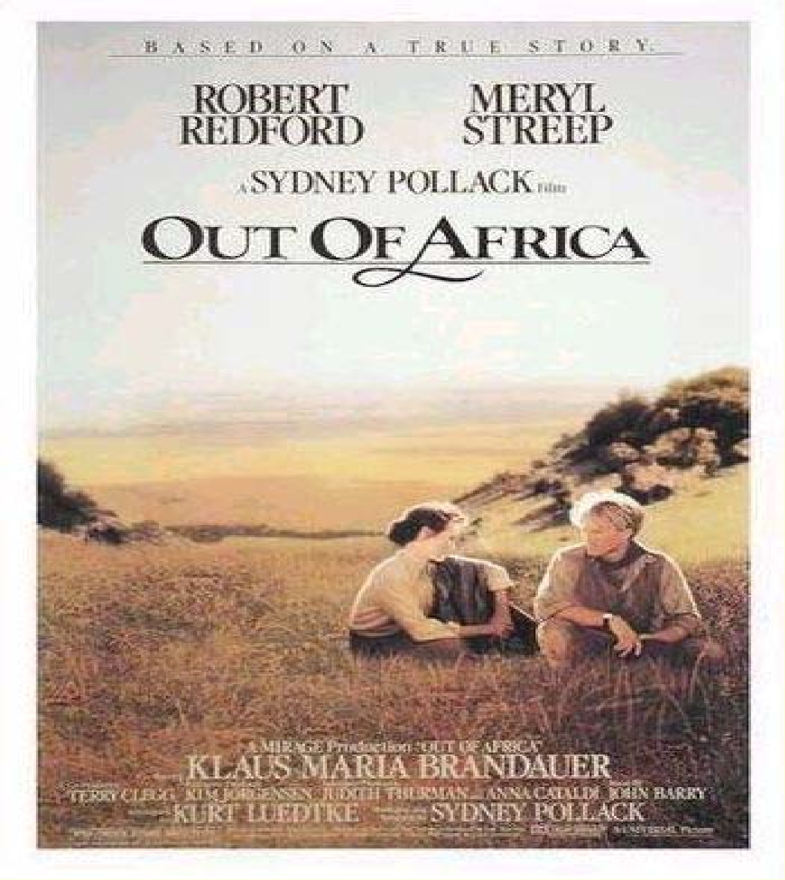
Pronouns

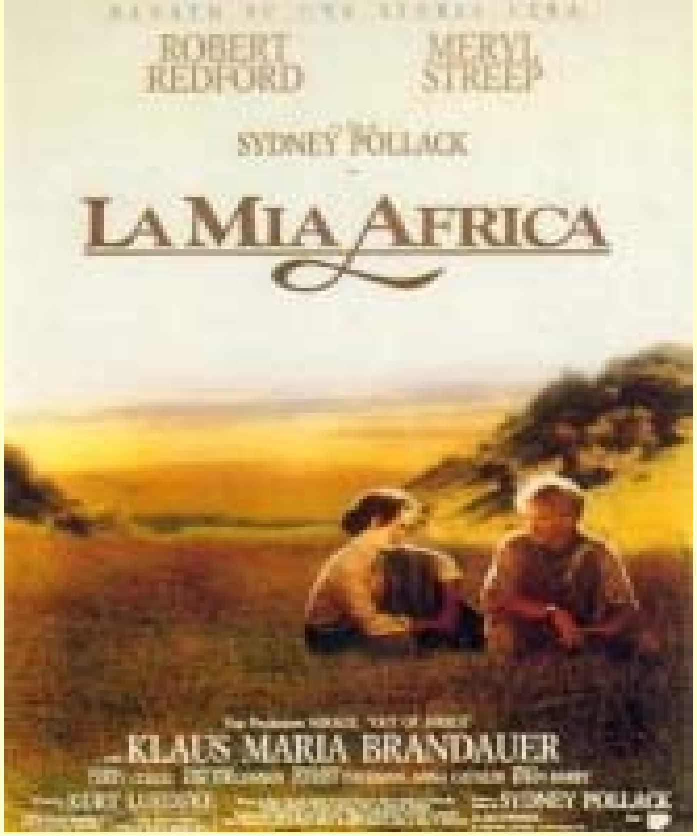
- Pronoun reference outside current sentence can influence grammatical gender
  - The shoe was found on the stairs…
  - (intervening sentences)
  - It was brown with white laces.

Out of Africa

- From Ulisse July 2006 (Alitalia's inflight magazine): E'però nel 1985 che Pollack riceve l'Oscar alla regia per "La mia Africa", …
- English in magazine: In 1985 Pollack received an Oscar for directing "My Africa", … [error by human translator]
- Poster on same page: "Out of Africa"

Out of Africa Posters





Data-driven Comments on Challenge 3

Accounting for local context

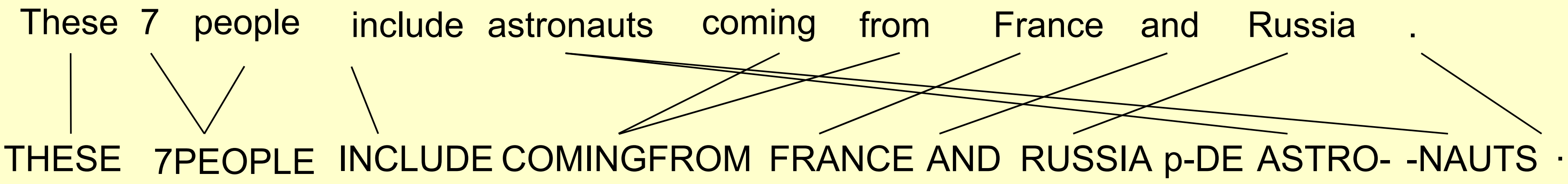
Phrase-based
rule extraction

THESE 7PEOPLE → these 7 people

COMINGFROM → coming from

INCLUDE → include

RUSSIA p-DE → russia



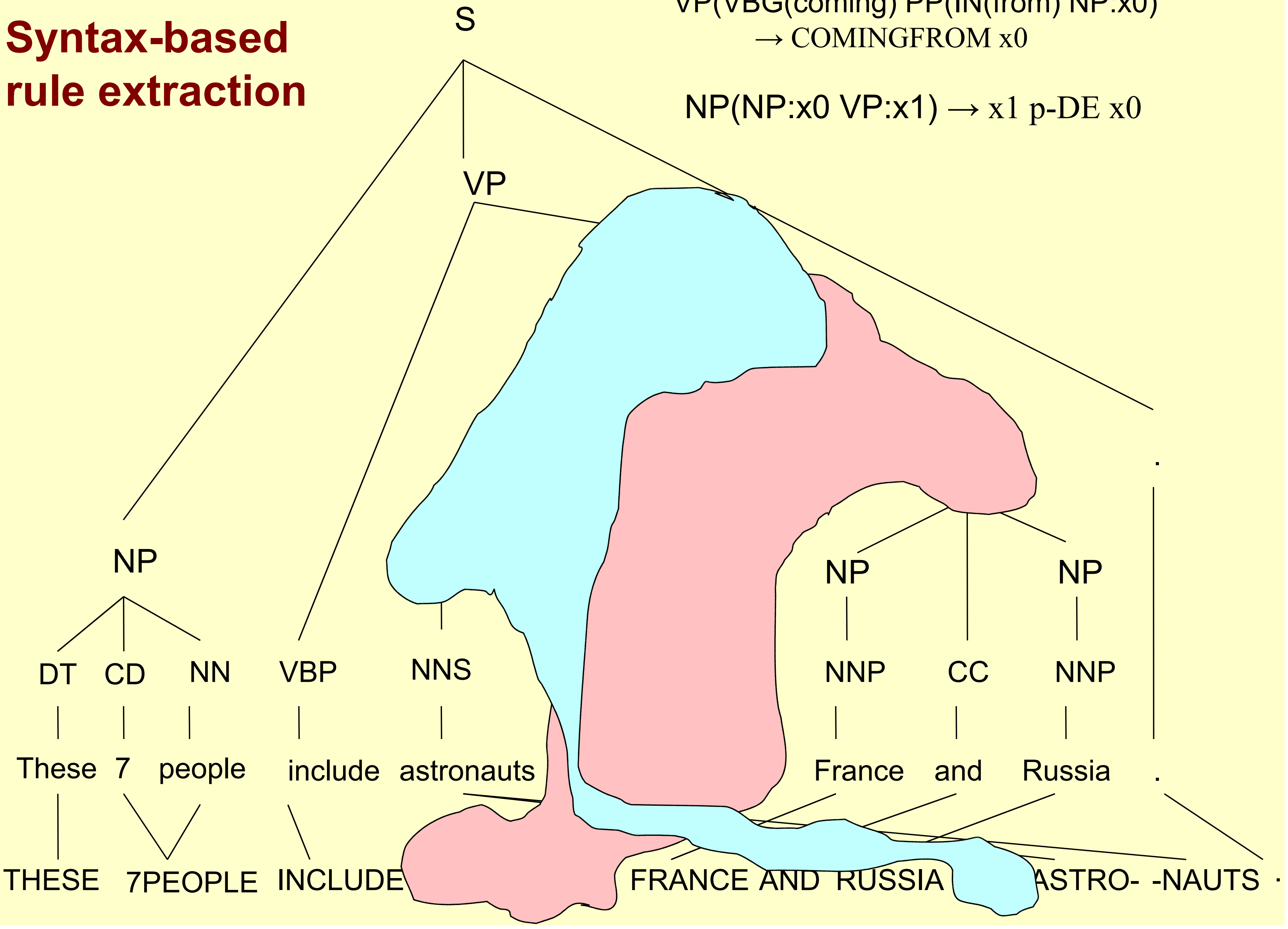
These 7 people include astronauts coming from France and Russia .

THESE 7PEOPLE INCLUDE COMINGFROM FRANCE AND RUSSIA p-DE ASTRO- -NAUTS .

Syntax-based rule extraction

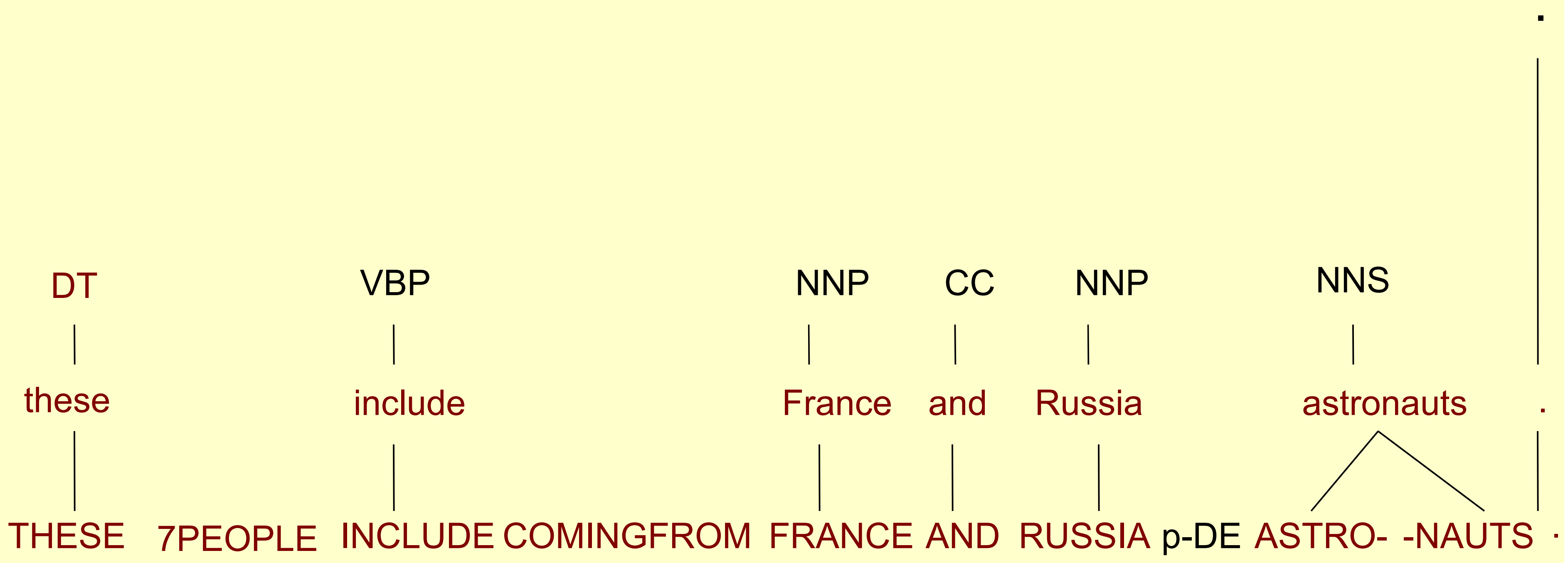
VP(VBG(coming) PP(IN(from) NP:x0)
→ COMINGFROM x0

NP(NP:x0 VP:x1) → x1 p-DE x0



Decoding with locally sensitive syntax rules

DT(these) → THESE
VPB(include) → INCLUDE
NNP(france) → FRANCE
CC(and) → AND
NNP(russia) → RUSSIA
NNS(astronauts) → ASTRO- -NAUTS
.(.) → .



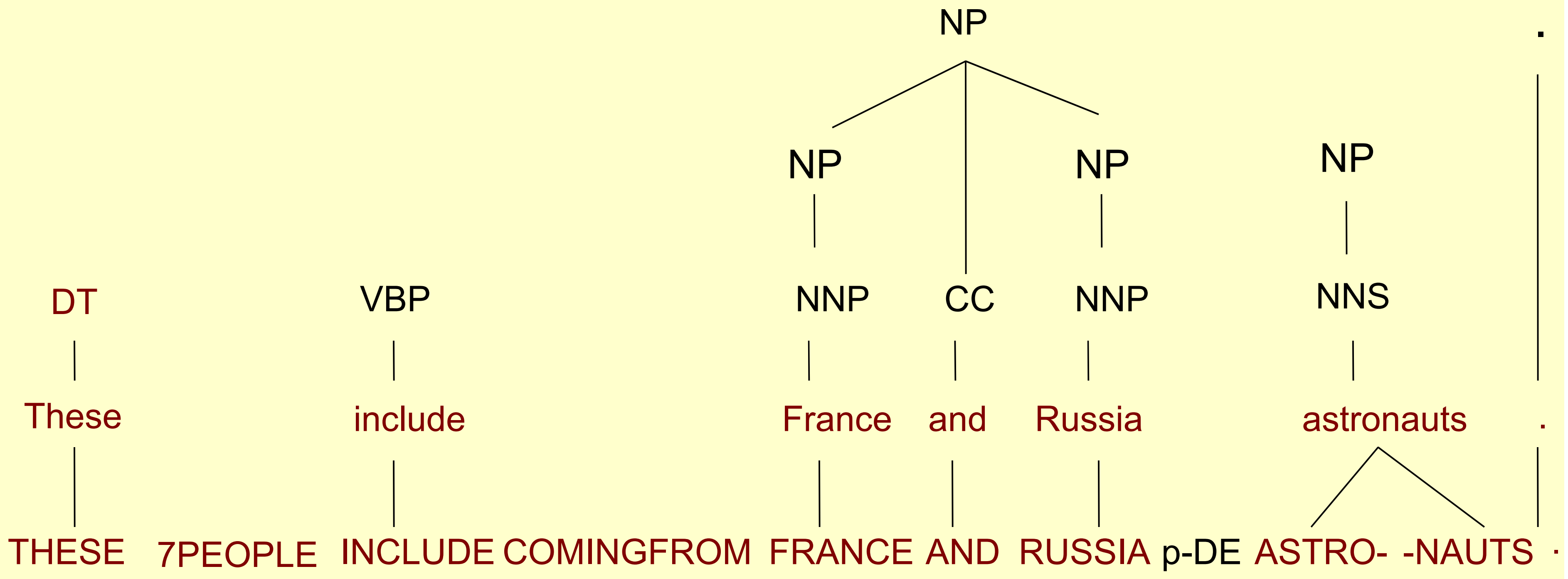
.

DT VBP NNP CC NNP NNS

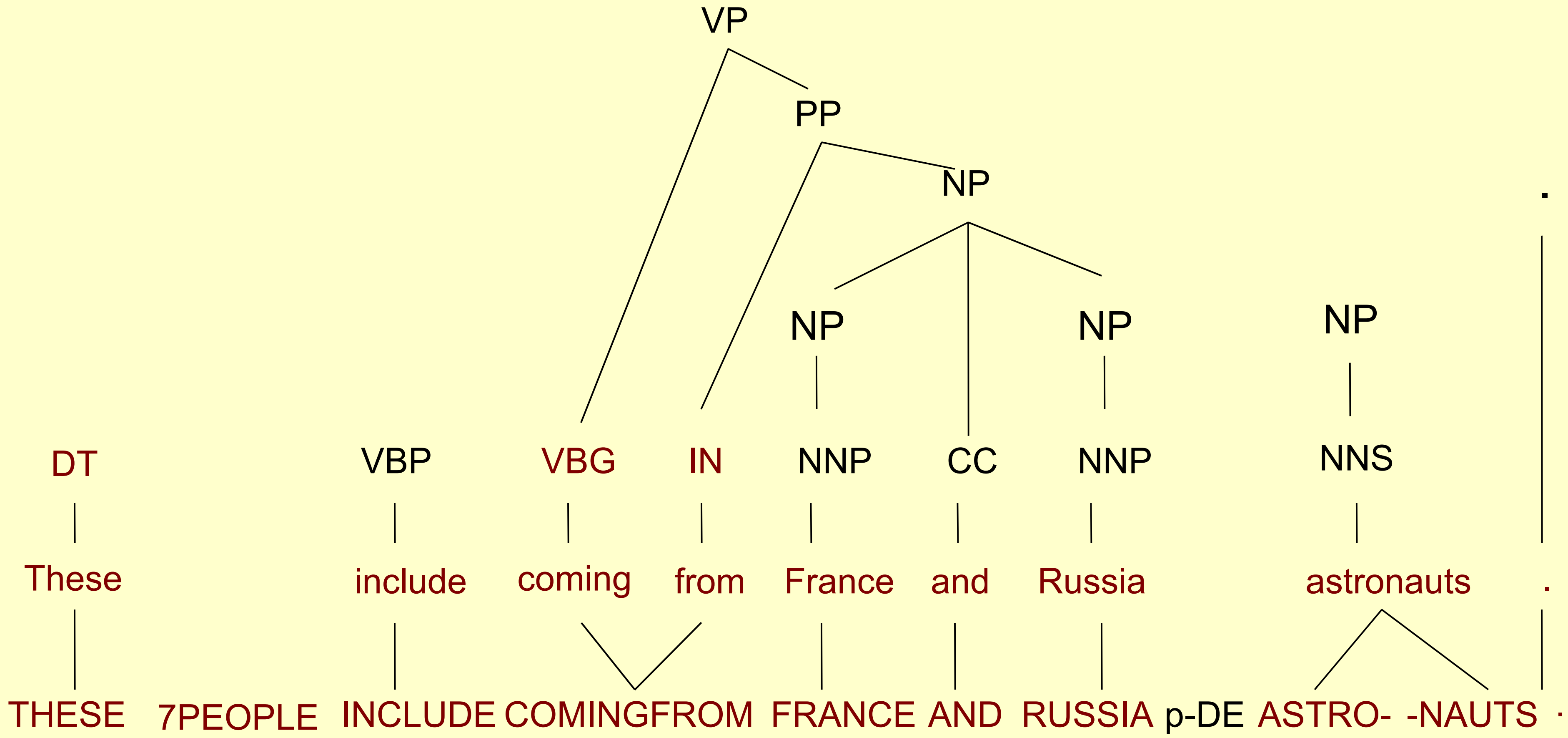
these include France and Russia astronauts .

THESE 7PEOPLE INCLUDE COMINGFROM FRANCE AND RUSSIA p-DE ASTRO- -NAUTS .

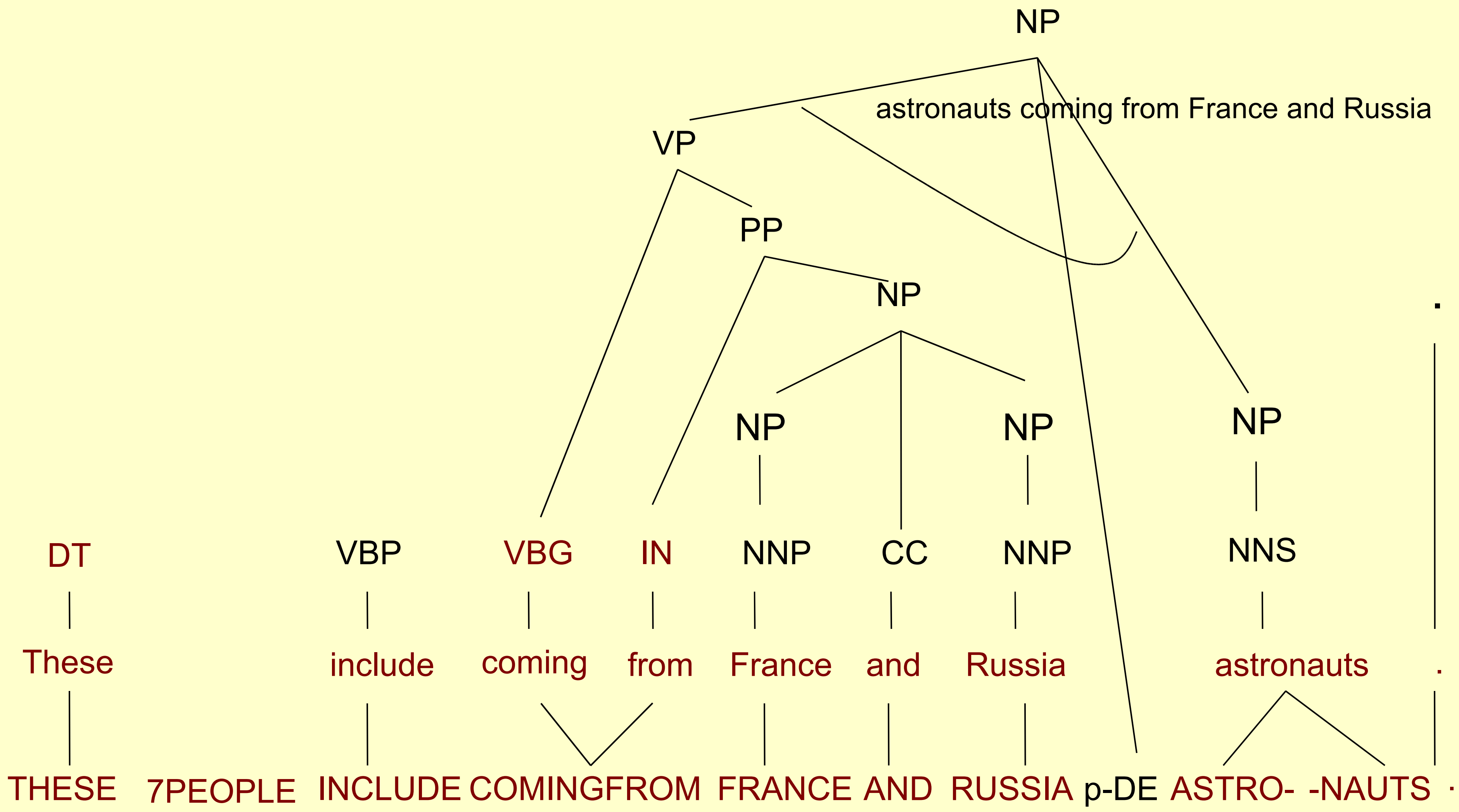
NP(NNP:x0) → x0
NP(NNP:x0) → x0
NP(NP:x0 CC:x1 NP:x2) → x0 x1 x2



VP(VBG(coming) PP(IN(from) NP:x0) → COMINGFROM x0

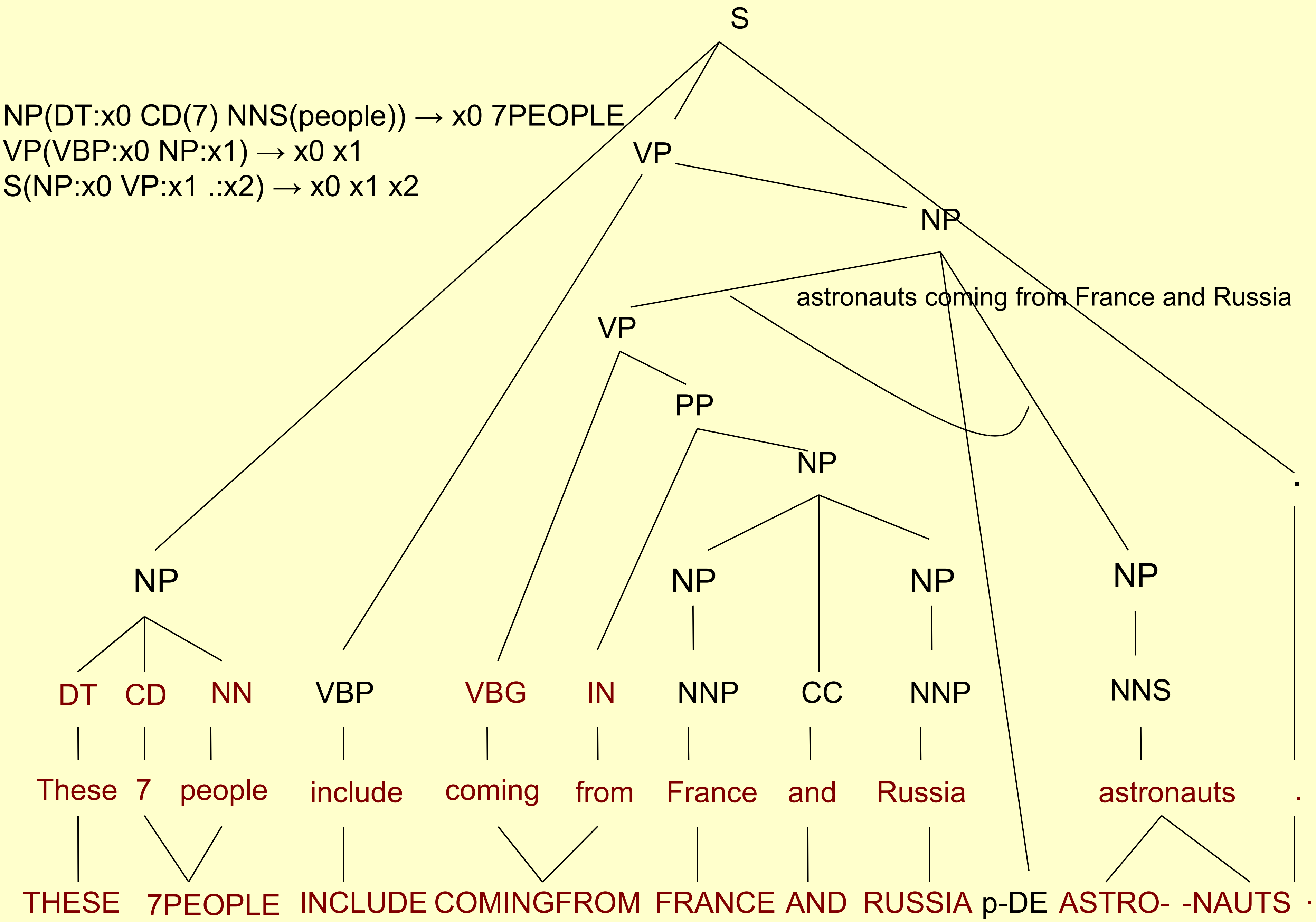


NP(NP:x0 p-DE VP:x1) → x1 x0



These 7 people include astronauts coming from France and Russia .

NP(DT:x0 CD(7) NNS(people)) → x0 7PEOPLE
VP(VBP:x0 NP:x1) → x0 x1
S(NP:x0 VP:x1 .:x2) → x0 x1 x2



Accounting for context

- Local context
  - Phrase-based translation models
  - Syntax-based ISI translation model
- Global context
  - Topic-based language models
    - Foundation work established
    - Need empirical validation
  - Discourse-based translation models
    - Foundation work not established

Challenge 4:
Using relevant "extra-text"

Sometimes translation decisions cannot be made solely on the basis of the co-text; they depend partly on information about the real-world not in the source text

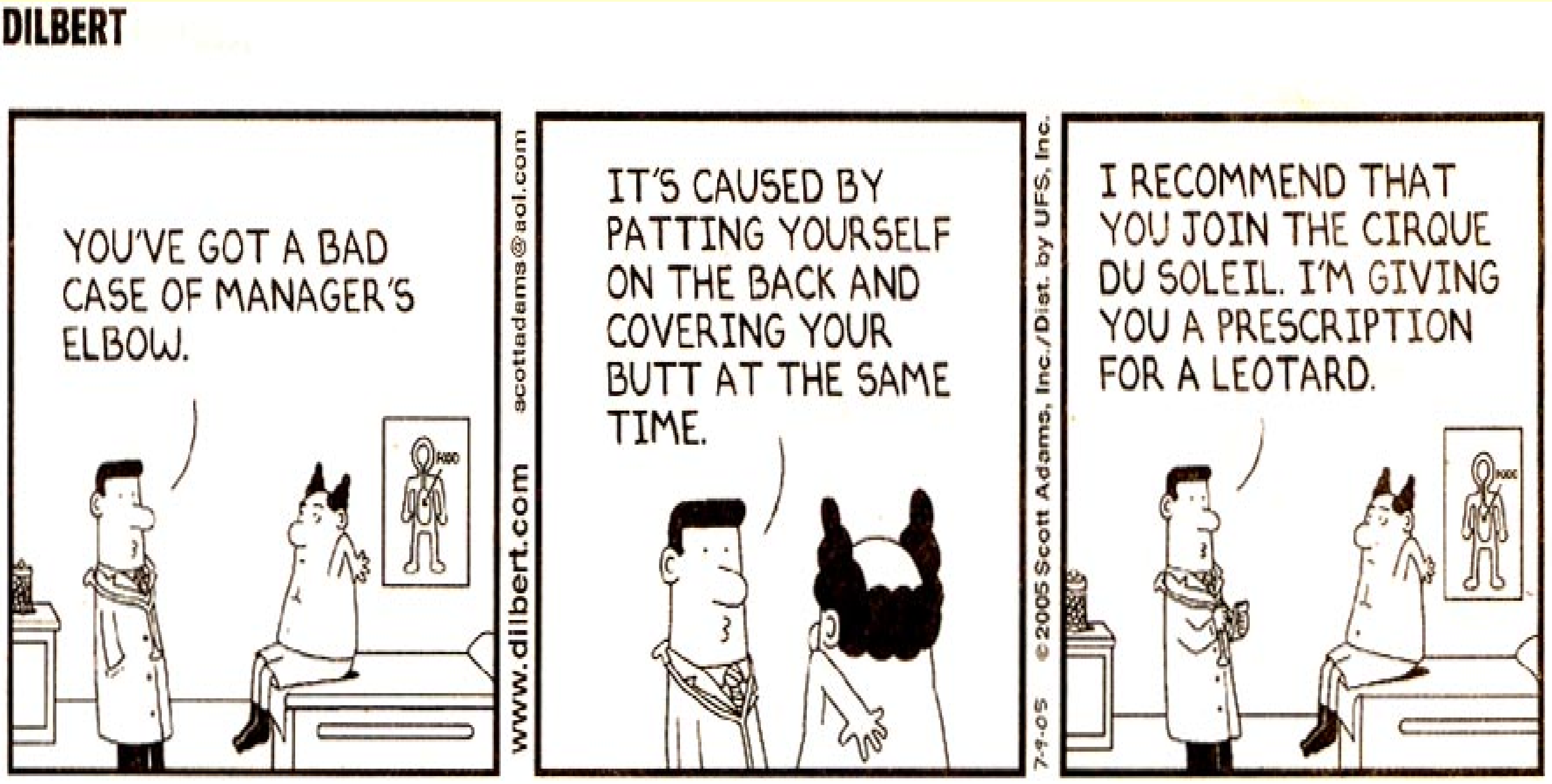
Chair

- Corpus: One hundred files from English-French European Parliament
  - English term: chair
  - 109 instances
  - Mostly *chair of meeting* or *to chair a meeting*
  - One instance of *university chair* (position)
  - Three involve object for sitting: French *chaise* vs. *fauteuil* (need to know **whether chair has arms** to select appropriate translation)

Manager's Elbow

- Imagine translating the following actual blog entry into another language:
  - Tuesday, July 12, 2005: I should definitely have brought my leotard to work today for my manager. He had a horrid display of manager's elbow right away this morning. I won't go into the long drawn out details, but I got yelled at again for something ridiculous. It seems he only has 2 volumes: 1)nice salesguy tone 2)mean manager loudness.
  - http://cristinacherry.blogspot.com/2005_07_01_cristinacherry_archive.html

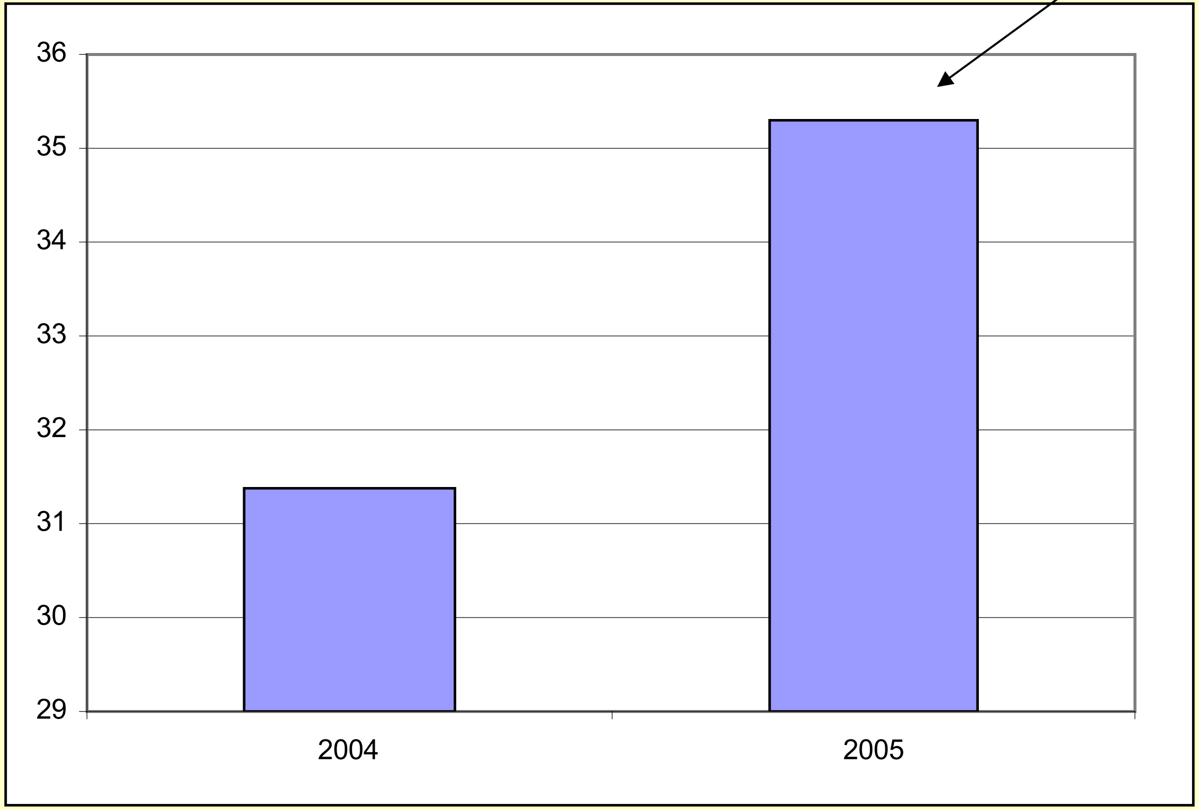
Probable Reference


DILBERT
YOU'VE GOT A BAD CASE OF MANAGER'S ELBOW.
scottadams@aol.com
www.dilbert.com
IT'S CAUSED BY PATTING YOURSELF ON THE BACK AND COVERING YOUR BUTT AT THE SAME TIME.
© 2005 Scott Adams, Inc./Dist. by UFS, Inc.
7-9-05
I RECOMMEND THAT YOU JOIN THE CIRQUE DU SOLEIL. I'M GIVING YOU A PRESCRIPTION FOR A LEOTARD.

Data-driven Comments on Challenge 4

Chinese-English MT Improvements (NIST Evaluation)

Like 2004 system +
N-gram LM trained on 220B words



The real-world information is out there for us to mine…

Challenge 5: Displaying "second-order creativity"

First-order creativity involves algorithmically generating an infinite number of items from a finite system; second-order creativity involves creating elements outside that infinite result

Second-order creativity applied to data-driven MT

- Ability to *create or retrieve* translations when not in corpus (no corpus is complete)
- Ability to *detect* that none of the translation options in the corpus are appropriate (and thus creative translation is needed instead of using what is there)

Example of a term not in the corpus

- From a real menu for an August 2006 banquet at the George Brown Cooking School, Toronto, Canada
  - Soup Course
    - Roasted Butternut Squash Soup with a Duxelles of Mushrooms
  - Not found in corpus but see (http://www.foodreference.com/html/fduxelles.html)
  - Same word is used in German cooking
  - But you can't always just use the source-language word

Another Term not in Corpus

- Zoopharmacognosy
  - Animals treating themselves for disease using natural drugs, such as toxic plants or clay
  - http://en.wikipedia.org/wiki/Zoopharmacognosy
- What if there is an accepted translation in the target language that is not in the corpus?
- There will always be the need for research

Creative Term in German

- Brösmelitöf
  - Brösmeli is productive element (crumbs)
  - Töf is a scooter/motorcycle
  - compound is not found in German Google
  - regional term (in Switzerland) for:
    - vacuum cleaner
- Requires creative translation e.g.
  - crumb chaser

Example of Detecting Something that Should not be Translated "as is"

- Cliché: Lights are on but there's nobody home (A derogatory expression used to describe someone who is not very smart or who is dumb.)
  - http://www.clichesite.com/content.asp?which=ti
- What about attested variant "The lights are dim and not even the neighbors are home"?

Another "not as is"

- Vertical House on the Prairie (heading)
  - Indirect reference to Little House on the Prairie
  - Actually referring to "The Price Tower" (designed by Frank Lloyd Wright, built in Bartlesville, Oklahoma)
  - Creative French translation: *Tour d'y voir*
    - Air Canada, En Route, August 2006, p. 40

One more

- "I pass the lobster trucks coming back from the sea, loaded down with a Jenga stack of traps.
  - Jenga is a game involving a tower made from blocks (http://en.wikipedia.org/wiki/Jenga)
  - It is sold in France, but the Air Canada translator chose to translate it as "loaded with traps stacked like sardines" (specification: naturalness overrides descriptive details)

Data-driven Comments on Challenge 5

"Creative" machine translations

- Trans: Kimfu is located West to Seoul.
- Ref: Kimpo is located West of Seoul.

- Trans: Taiyimarmu is in Adleyde to attend an international alumna gathering.
- Ref: Taib Mahmud is now attending an international alumni meeting in Adelaide.

- Trans: Try to remedy, or just declare the fatal defect of this protocol? We shall discuss again.
- Ref: Shall we attempt to salvage the agreement, or shall we announce that the agreement has fatal flaws and should be discussed anew?

Improvement drivers

- Traditional linguistics, AI, NLP
  - Example-driven theories, algorithms, etc.
  - Focus on very difficult, but extremely rare events.

- Best data-driven MT
  - Error class-driven theories, algorithms, etc.
    - Verb errors: 16.5%
    - …
    - Punctuation errors: 6%
    - …

Arabic VSO → English SVO is a solved problem in the ISI syntax system.

S(NP:x0 VP(VBD:x1 NP:x2) .:x3) → x1 x0 x2 x3 p=0.54

Part Two
Sources of help in meeting challenges

1 – Functionalism (from translation studies)

2 – Stratification (from linguistics)

3 – Domains (from terminology)

4 – Interaction (from language acquisition and Peirce)

5 – Embodiment (from philosophy)

Help 1: Functionalism

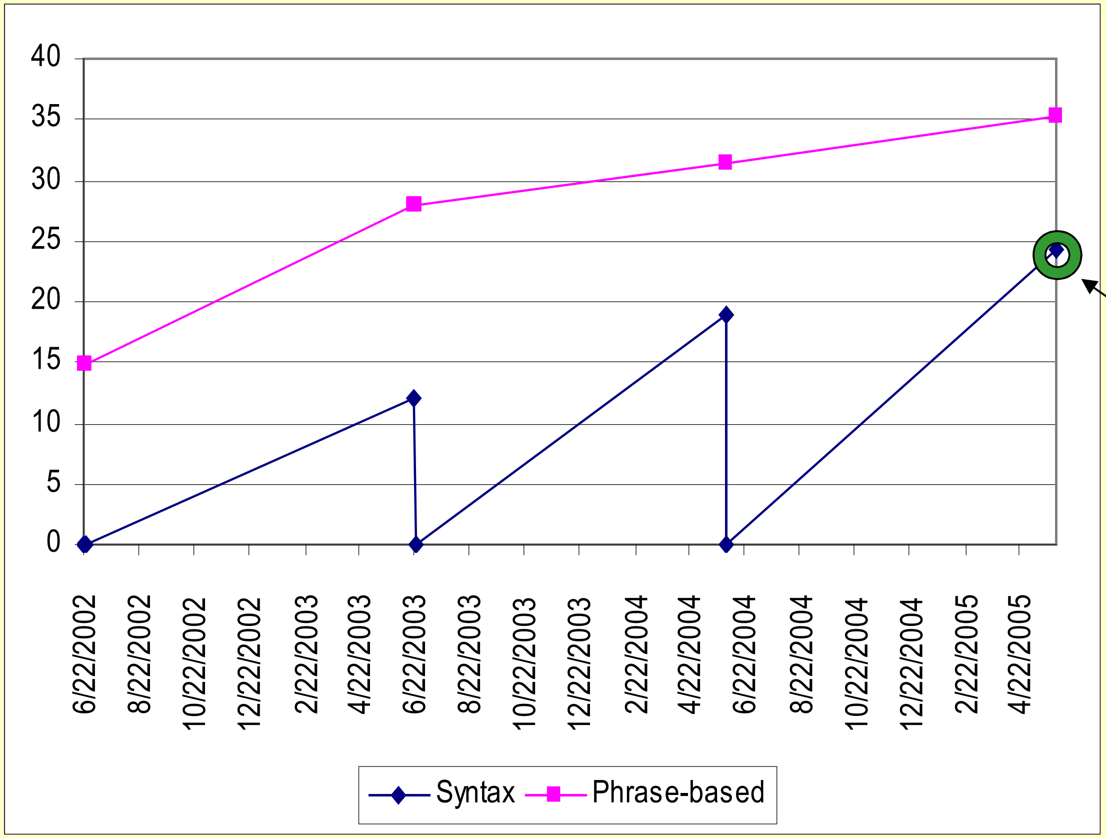
The ASTM standard partially formalizes the notion of specifications, which is an expression of how to adapt to the audience and purpose of a translation. The translation process is not a function, but becomes more like a function with two arguments (sourceText, specifications) rather than one (sourceText).

Bottom Line for Data-driven MT

- The input to the system should be (a) the source text and (b) the specifications to use when translating it

Data-driven Comments on Functionalism

Chinese-English MT Progress (NIST evaluations)



First syntax submission

What linguists don’t like to do

- Where do punctuation symbols attach in phrase-structured parse trees?
- What kinds of syntactic annotations are most useful for machine translation?
- …

Help 2: Stratification

Some Basic Strata

- Phonological/morphological structure
- Syntactic structure
- Meaning structure
- Note: they all co-exist and interrelate

Bottom-line for Data-driven MT

- The target text needs to be well-formed on multiple strata
- This does not mean there is an order to the strata or that one derives from another
- All strata are context-dependent

Data-driven Comments on Stratification

All data-driven MT systems attempt to accomplish this

- Language models
  - Ngram language models
  - Factored language models
    - Morphology
  - Syntax-based language models
  - Semantic-based language models???
  - Discourse-based language models???
- Translation models
  - Phrase-based translation models
  - Syntax-based translation models
  - Semantic-based translation models???

Help 3: Domains

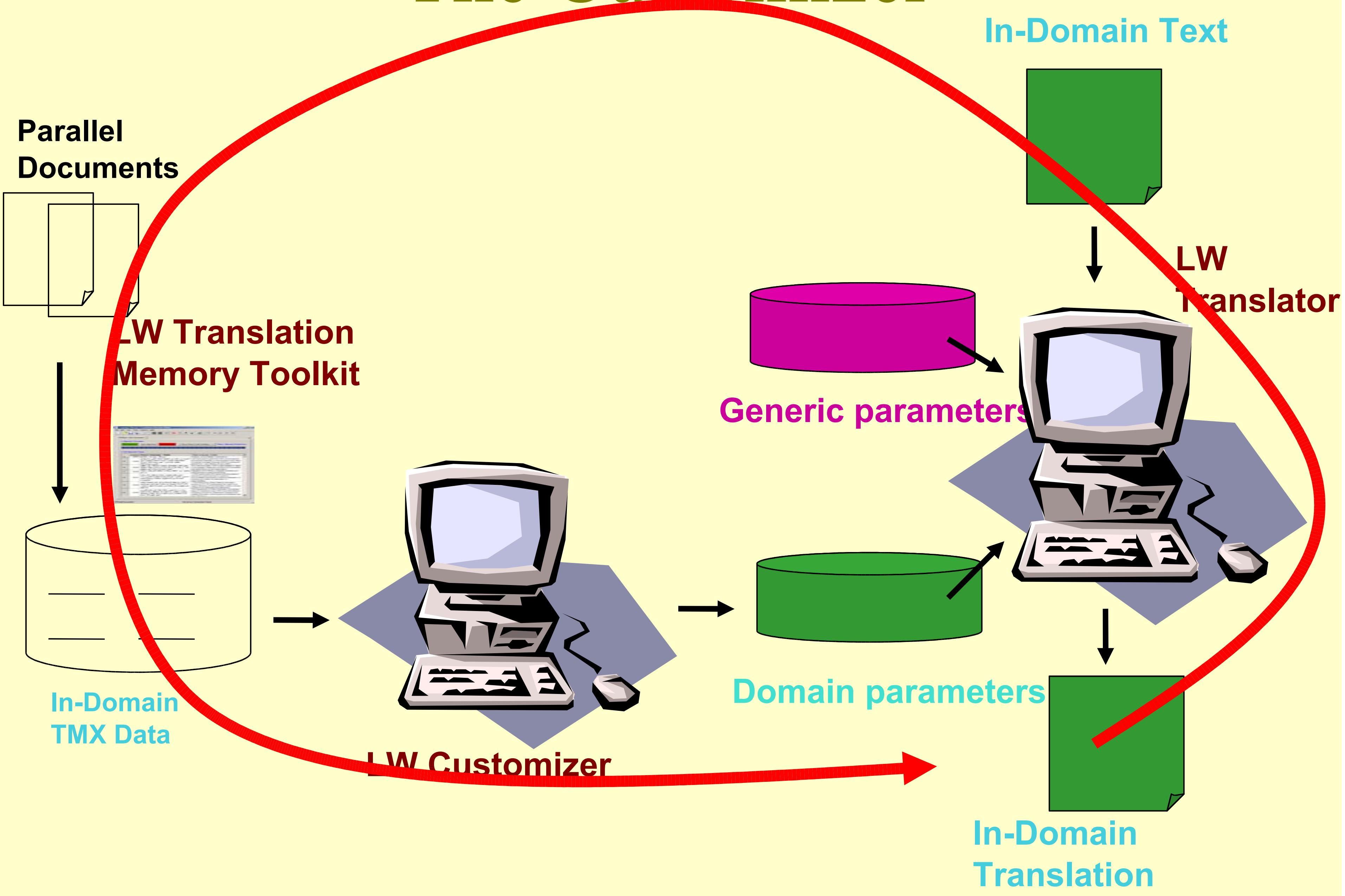
- Identifying the domain that applies to an item of source text helps select an appropriate translation when the immediate context does not suffice

Data-driven Comments on Domains

Domain adaptation

- Little Research
  - Out-of-domain data used as prior knowledge/distribution [Bacchiani and Roark; Chelba and Acero]
  - All data is a combination of generic, out-of-domain, and in-domain data [Daumé III and Marcu]
- MT Products
  - LW Customizer

The Customizer



Parallel Documents

LW Translation Memory Toolkit

In-Domain TMX Data

LW Customizer

Domain parameters

Generic parameters

In-Domain Text

LW Translator

In-Domain Translation

Help 4: Interaction

Language learning for humans requires incremental meaningful interaction with others, not just textual input, so it might be the same for machines; translation also requires incremental re-evaluation (see language acquisition studies and Peircean semiotics).

One View of Language Learning

- Suppose you were locked in a room and were continually exposed to the sound of Chinese from a loudspeaker; however long the experiment continued, you would not end up speaking Chinese. … What makes learning possible is the **information** received in parallel to the **linguistic input** in the narrow sense (the sound waves). Klein 1986 (*Second Lang. Ac.* Cambridge U Press)

Dyadic vs. Semiogenic Perspectives

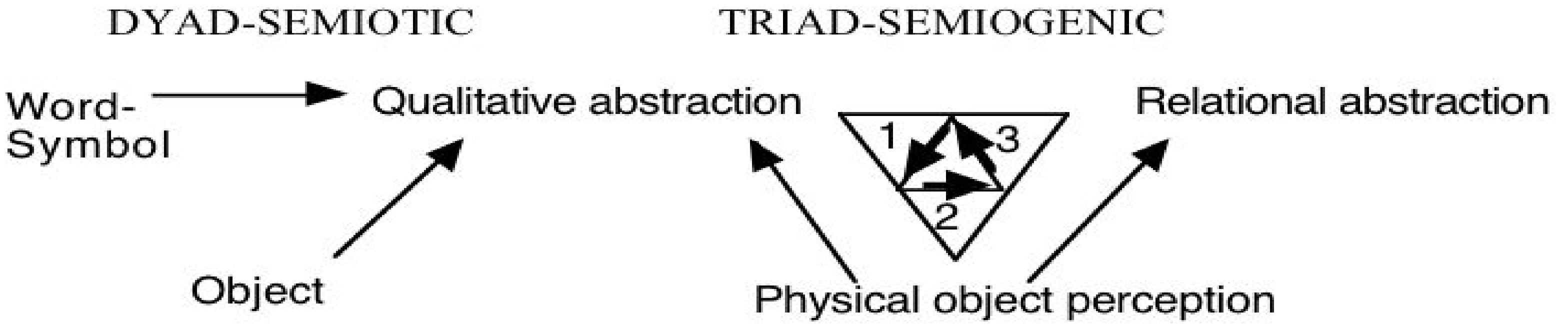


Figure 5: one type of abstraction in the dyadic model, contrasted with two types of abstraction in the semiogenic model

The Interpretant and Translation

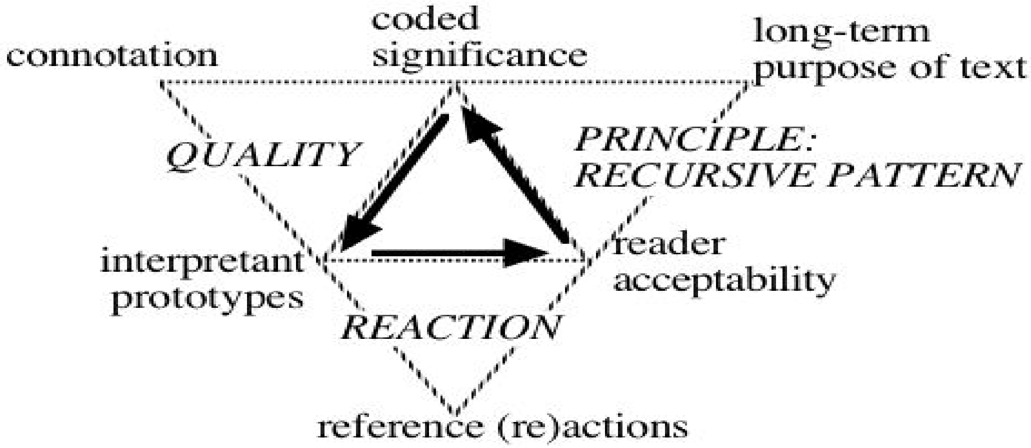


Figure 8: Subdivisions of the Semiogenic Interpretant as a guide to translation.

Data-driven Comments on Interaction

Or maybe not

- Texts contain all the knowledge that we need.
  - Explicit
  - Implicit
- We need only better learning models and algorithms
  - Hidden variables can take us a long way
    - E.g.: word-level alignments

Centauri/Arcturan [Knight 97]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

zero fertility

Help 5: Embodiment

Some source texts, audiences, and purposes may require a system that believes it has a body, otherness, and agency

I am looking forward to having this problem

Closing
Some Advice From Old-timers

- Victor Yngve (early MT researcher):
  - Remember we are studying people in real-life interactions, not language
- Robert Longacre (Chomsky-age linguist):
  - It is wonderful to see new paradigms arise, but… (drink responsibly; eat a balanced diet)
- Alan Melby:
  - Congratulations for your escape from rules!

General Discussion

- a: Comparison with human qualifications
- b: Avoidance of compositionality assumption
- c: Using relevant co-text (beyond sentence)
- d: Using relevant "extra-text" (real world info)
- e: Displaying "second-order creativity"

- 1 - Functionalism
- 2 - Stratification
- 3 - Domains
- 4 - Interaction
- 5 - Embodiment