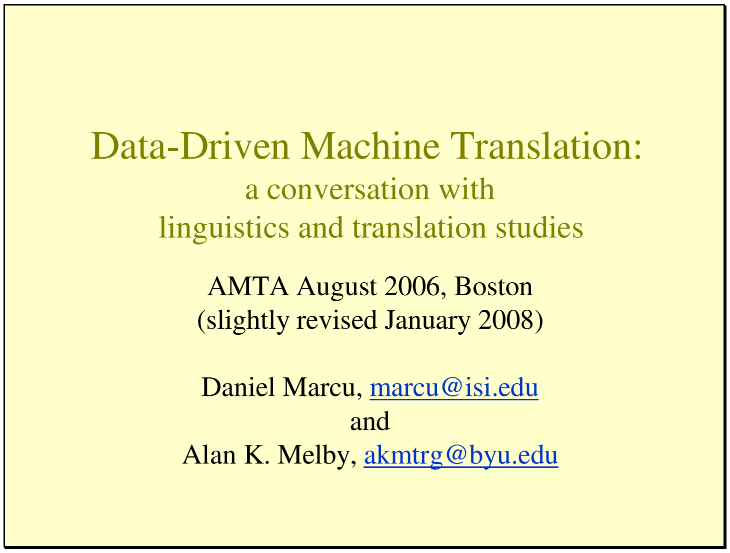
Slide 1



Data-Driven Machine Translation:

a conversation with linguistics and translation studies

AMTA August 2006, Boston
(slightly revised January 2008)

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

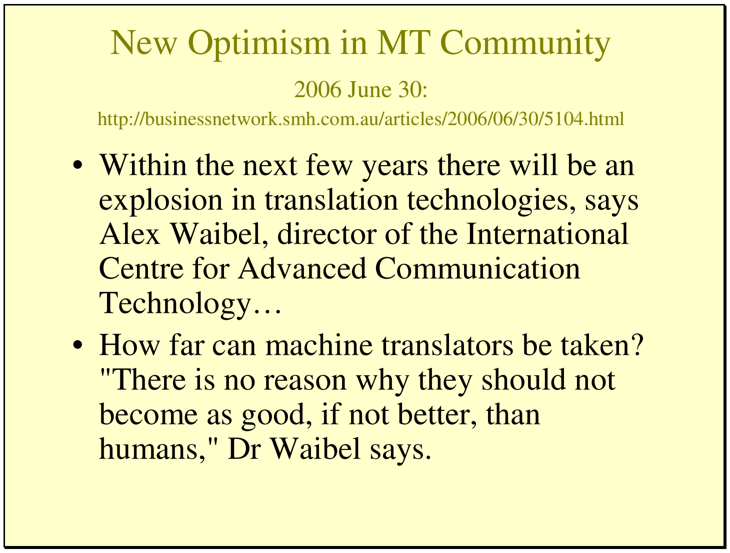
Daniel Marcu is a professor at the University of Southern California (http://www.isi.edu/~marcu/) and Founder of Language Weaver Inc. (www.languageweaver.com).

Alan Melby is a professor at Brigham Young University (http://www.ttt.org/akm-cv.html).

This debate is about the future of data-driven machine translation (MT), in particular about hurdles to overcome if it is to produce translations indistinguishable from those of professional human translators.

Data-driven machine translation uses large bitext corpora (primarily of human translations) as the basis for producing translations, as opposed to traditional machine translation, which uses morphological, syntactic, and semantic analysis, transfer and generation, based on dictionary lookup and rules.

Slide 2



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.

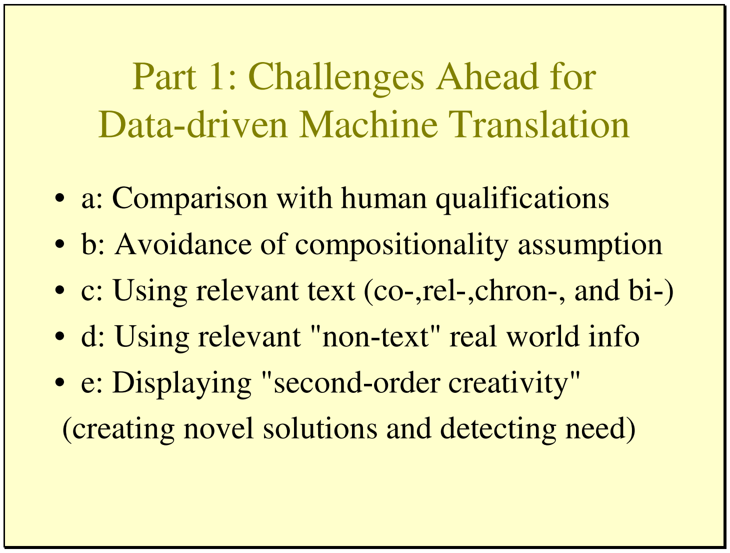
There was an initial period of optimism in the 1950s, when machine translation research and development began.

Then, there was a dark period after the publication of the ALPAC report (http://en.wikipedia.org/wiki/ALPAC) in the mid 1960s.

Then, there was another period of optimism in the 1980s, followed by a period of realism in the 1990s. In the 1990s, few if any MT developers claimed that MT would compete directly with human translators in the foreseeable future. Instead, proponents of MT looked for appropriate applications of MT, especially (a) domain-specific, controlled-language systems and (b) information-gathering (as opposed to publication-quality) systems where highly accurate and readable translation is not necessary.

Now, in the first decade of the new century, there is new optimism among MT developers, based primarily on the hope that a data-driven approach will overcome the obstacles encountered by those who have focused on a rule-based approach.

Slide 3



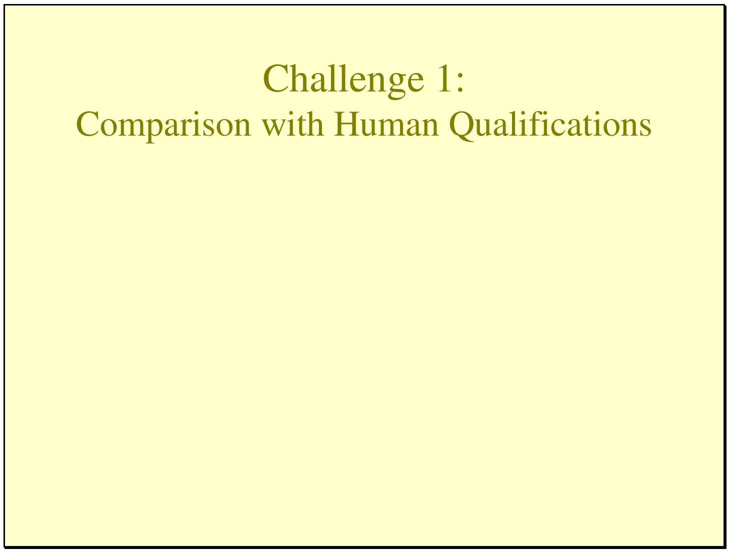
Part 1: Challenges Ahead for Data-driven Machine Translation

- a: Comparison with human qualifications
- b: Avoidance of compositionality assumption
- c: Using relevant text (co-,rel-,chron-, and bi-)
- d: Using relevant "non-text" real world info
- e: Displaying "second-order creativity"

(creating novel solutions and detecting need)

Melby suggested that there are at least these five types of challenges ahead for data-driven MT as it attempts to go beyond the traditional applications of MT and compete directly with professional human translators. Each is further described in subsequent slides.

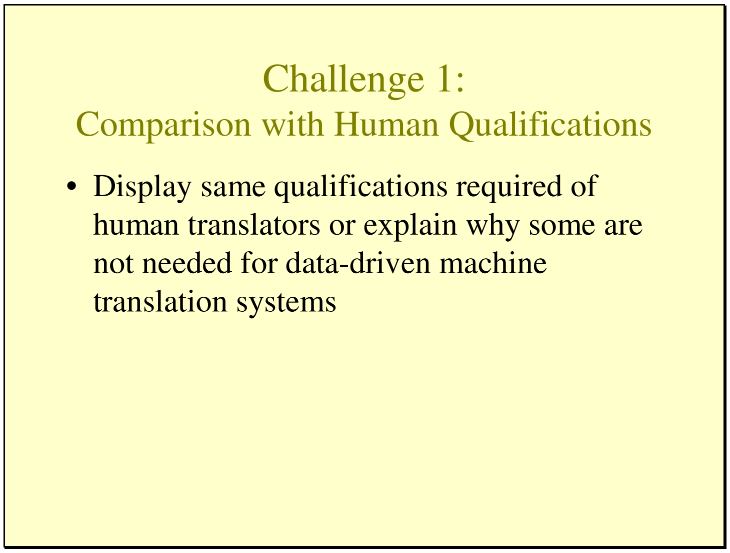
Slide 4



Challenge 1:
Comparison with Human Qualifications

Next five slides describe the process of commercial translation involving human translators and lists some of their qualifications.

Slide 5

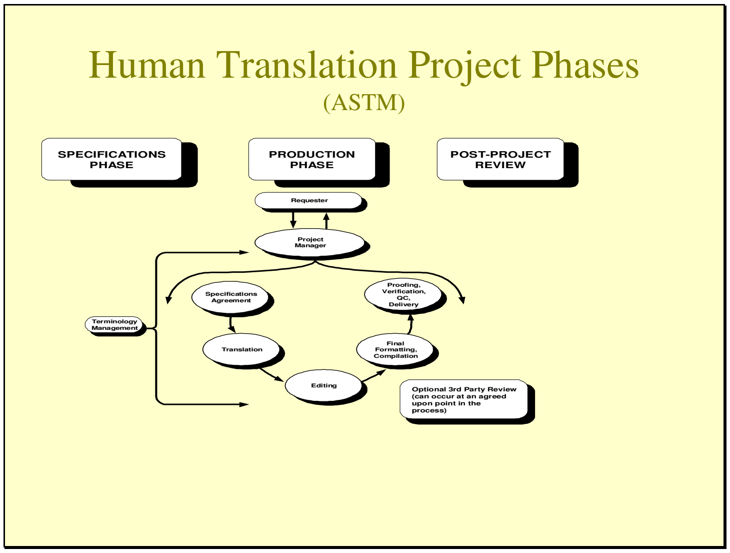


Challenge 1:
Comparison with Human Qualifications

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

Melby suggested that a machine translation system that performs as well as a professional human translator would have to exhibit the same qualifications required of human translators. If machine translation proponents want to challenge this, they should be able to explain why some of those qualifications required of humans should not be required of machines.

Slide 6



Human Translation Project Phases

(ASTM)

SPECIFICATIONS PHASE

PRODUCTION PHASE

POST-PROJECT REVIEW

Requester

Project Manager

Specifications Agreement

Proofing, Verification, QC, Delivery

Terminology Management

Translation

Final Formatting, Compilation

Editing

Optional 3rd Party Review (can occur at an agreed upon point in the process)

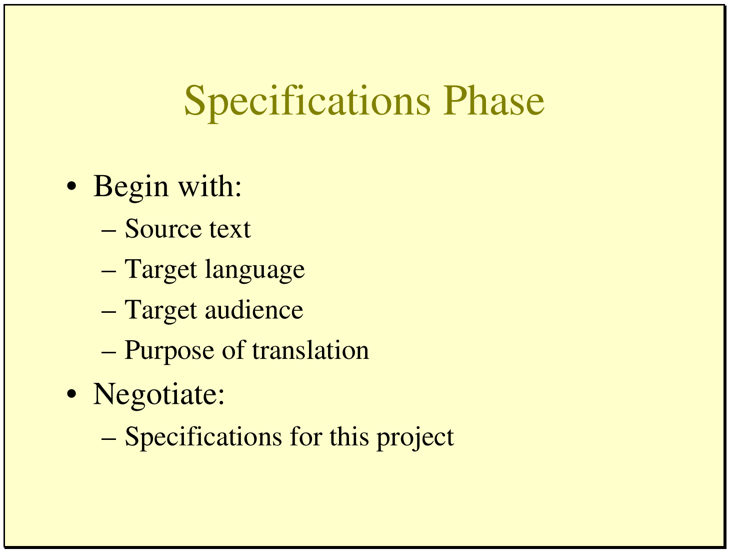
This diagram is from a draft of the ASTM translation standard, showing the phases of a typical translation project:

The specifications phase

The production phase

The post-project review phase

Slide 7

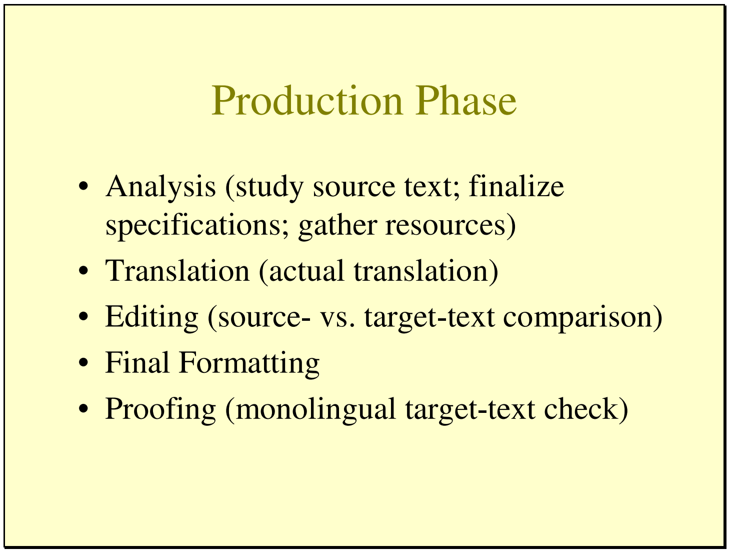


Specifications Phase

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

The specifications phase involves a dialogue between the requester (the client) and the provider (the translator or translation project manager) about the nature of the source text, the exact variation of the target language that is to be produced (e.g. Canadian French vs. Belgian French), the audience for the translation, and the purpose of the translation, to name a just a few translation parameters. A more extensive list of translation parameters relevant to producing specifications for a translation project are available at the www.ttt.org/specs webpage.

Slide 8

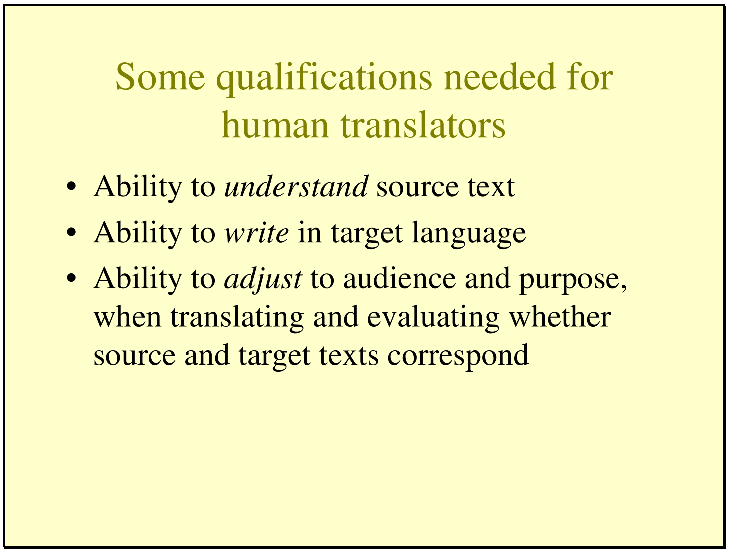


Production Phase

- Analysis (study source text; finalize specifications; gather resources)
- Translation (actual translation)
- Editing (source- vs. target-text comparison)
- Final Formatting
- Proofing (monolingual target-text check)

This slide shows the major steps that are followed during an actual translation project.

Slide 9

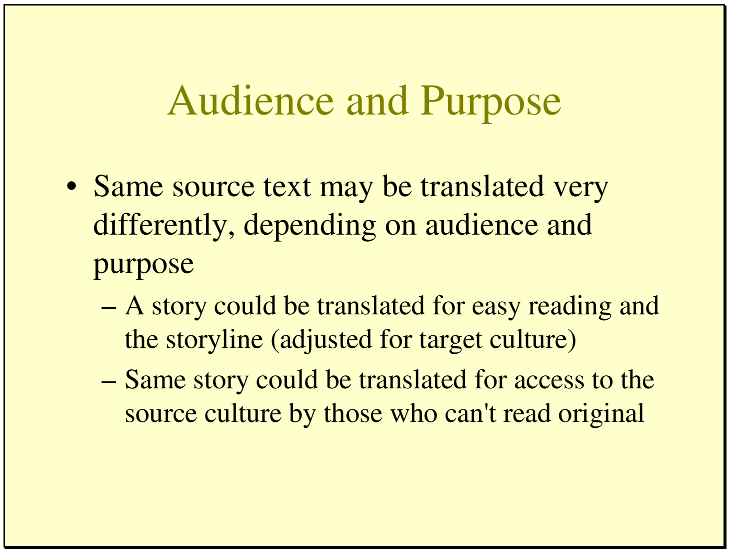


Some qualifications needed for human translators

- 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

In this slide, Melby points out that human translators are expected to understand the source text, be able to write well in the target language and be able to adjust to the audience and purpose (so that the same source text may have different translations depending on audience and purpose). These requirements may seem totally obvious to a professional human translator but they may not be so obvious to a person who believes in data-driven machine translation.

Slide 10

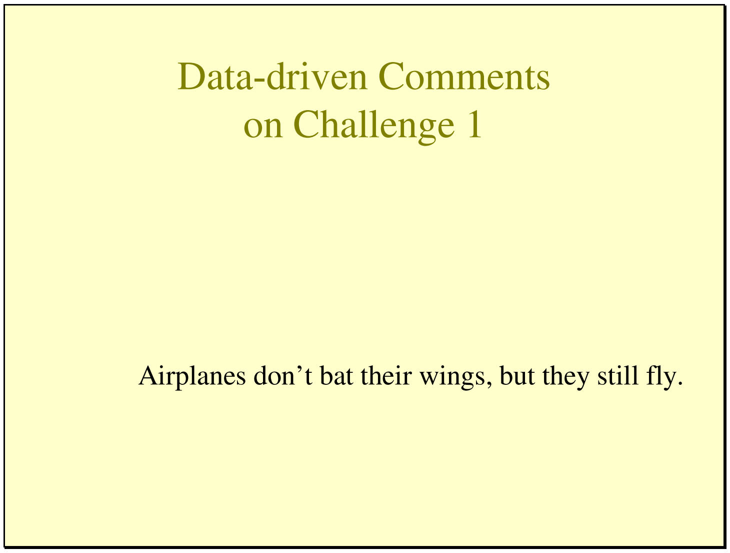


Audience and Purpose

- 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

How will data-driven MT adjust its output according to the audience and purpose of the translation?

Slide 11

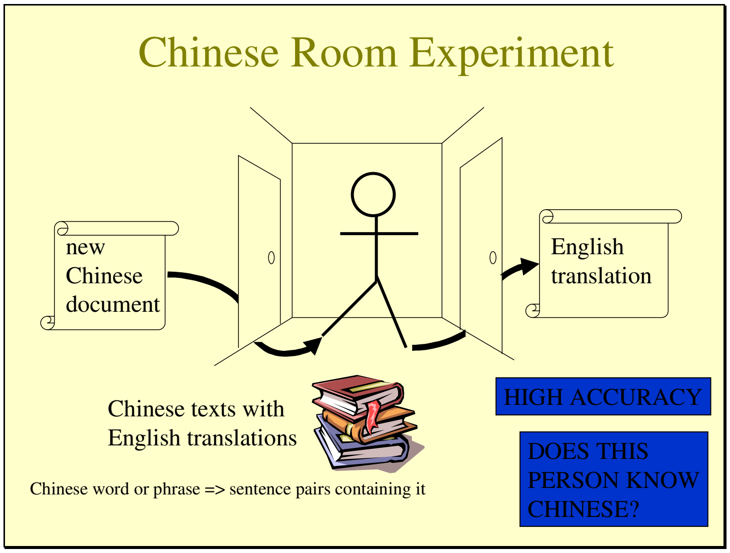


Data-driven Comments on Challenge 1

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

In this slide and the next four, Marcu counters Melby by pointing out that airplanes do not bat their wings, implying that a machine translation system need not have the same qualifications as a human translator.

Slide 12



Chinese Room Experiment

new Chinese document

English translation

Chinese texts with English translations

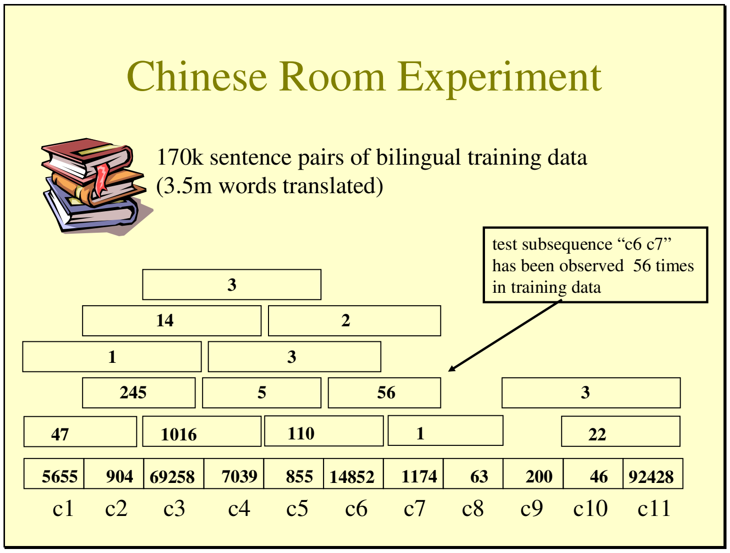
Chinese word or phrase => sentence pairs containing it

HIGH ACCURACY

DOES THIS PERSON KNOW CHINESE?

The Chinese Room experiment was originally proposed by the philosopher Searle (http://en.wikipedia.org/wiki/Chinese_room). In this experiment, a human who does not speak Chinese is asked to translate a Chinese document into English and is given detailed instructions on how to do it without understanding the source text.

Slide 13



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

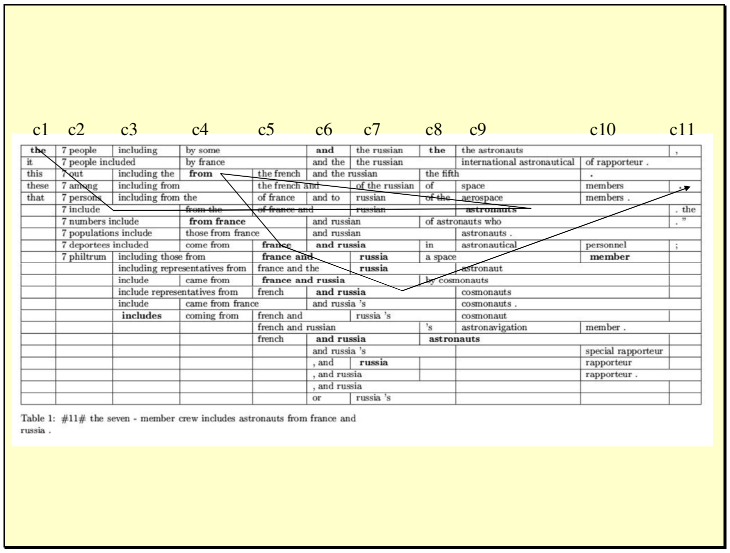
245 5 56 3

47 1016 110 1 22

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

Marcu describes a Chinese room experiment recently conducted in which humans were able to translate Chinese into English without understanding it.

Slide 14

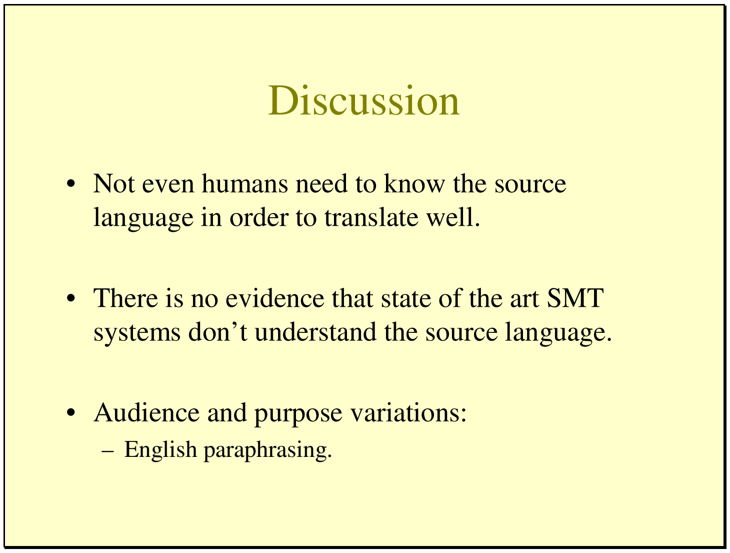


c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	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 .

This diagram from Marcu suggests how the English translation was produced.

Slide 15



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.

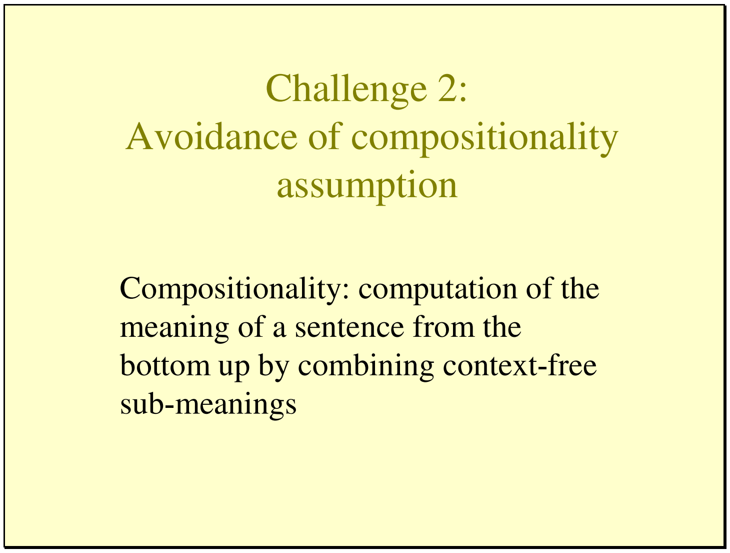
Marcu claims that his Chinese Room experiment demonstrates that humans need not understand the source language in order to translate well.

Marcu also claims that we have no way of knowing whether SMT (statistical MT – a type of data-driven MT) systems understand the source language or not.

Melby notes that Marcu's response does not address the proposed need for adjusting to audience and purpose (have the machine "paraphrase the output") is vague and does not demonstrate that it would actually work. Furthermore, Melby suggests that the Chinese Room experiment should be replicated with a variety of source texts and target languages. Furthermore, Melby points out that even if this experiment does succeed, it does not prove that data-driven MT will succeed, since the humans involved do understand the target language and are able to write in it, while the machine may not.

A fundamental question is left open: is understanding of the source text necessary for a high-quality translation? For Melby, the answer is obviously "yes". For Marcu, the answer is obviously "no".

Slide 16

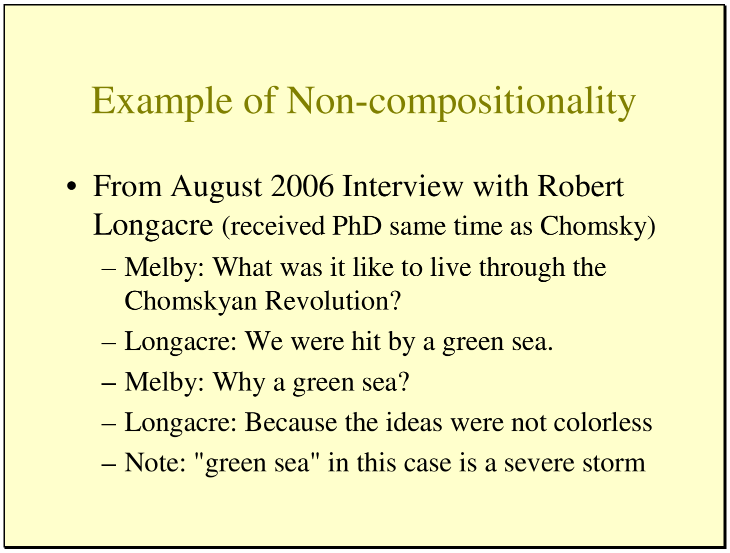


Challenge 2:
Avoidance of compositionality assumption

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

Melby points out that mainstream linguistics assumes that the meaning of a sentence (or at least the possible meanings) can be computed by combining context-independent units of meaning at the word level into larger and larger units.

Slide 17



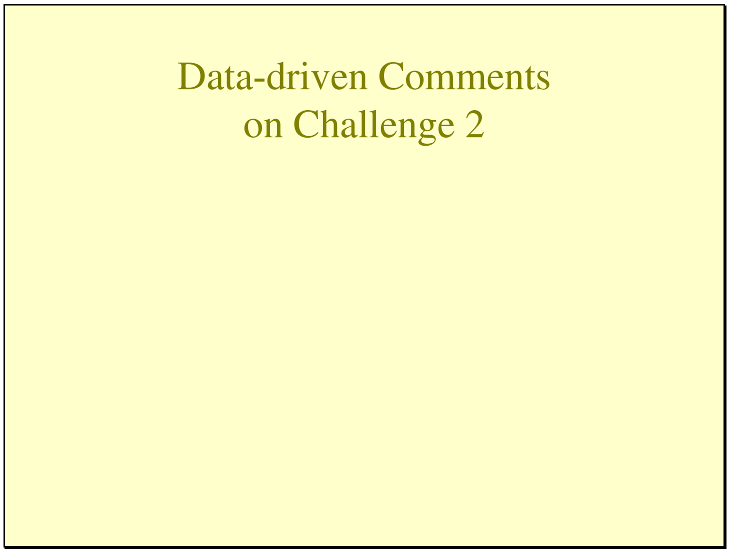
Example of Non-compositionality

- 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

Melby gives an example of real language where the meaning of the sentences is not easily derived from context-free meanings of the individual words. Longacre’s sentence refers to a sentence in Chomsky’s *Syntactic Structures* (“Colorless green ideas sleep furiously”).

Melby claims that much of the meaning of language is non-compositional.

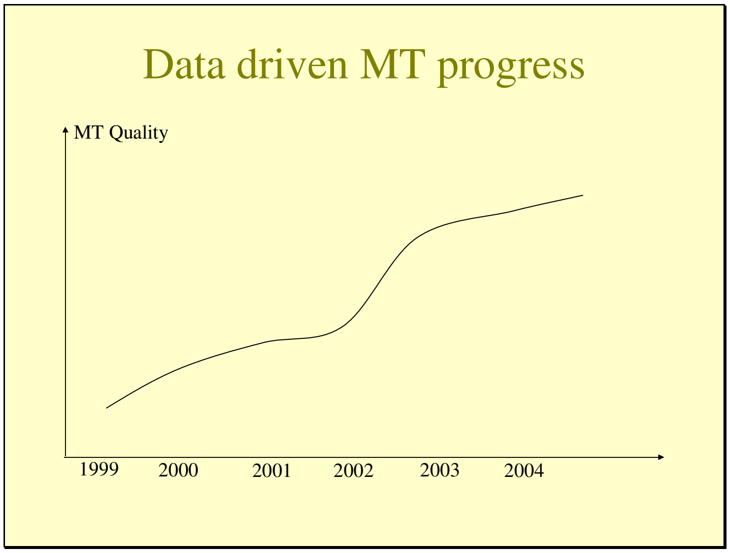
Slide 18



Data-driven Comments on Challenge 2

Marcu presents a series of slides in response to the second challenge.

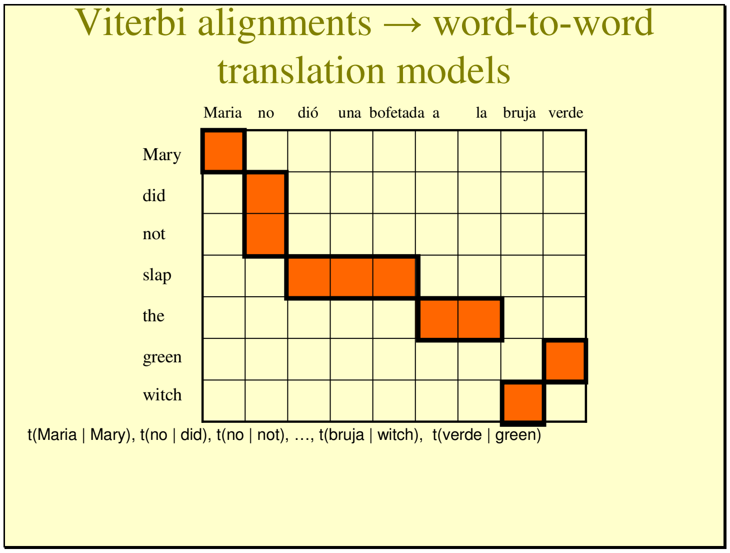
Slide 19



Data driven MT progress

Maru pointed out that data-driven MT has been improving in quality and therefore will likely to continue improving in quality until it meets or exceeds human quality translation.

Slide 20



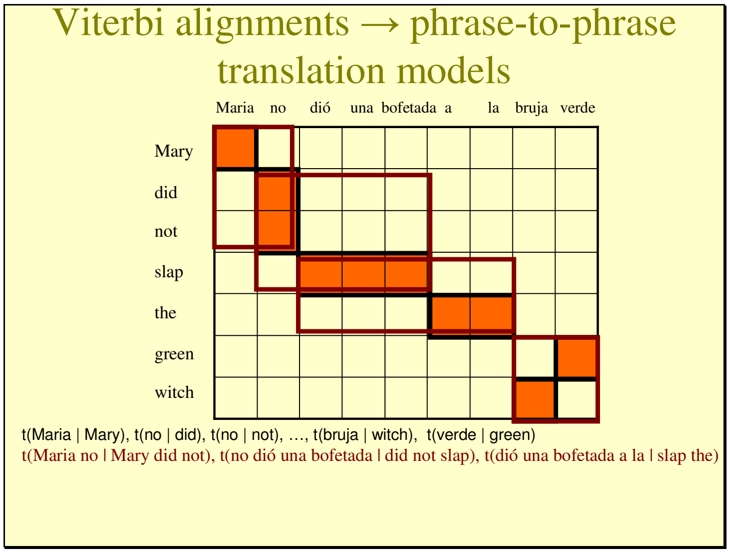
Viterbi alignments → word-to-word translation models

	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)

Marcu shows what can be done with a word-for-word translation approach (see text at bottom of slide).

Slide 21

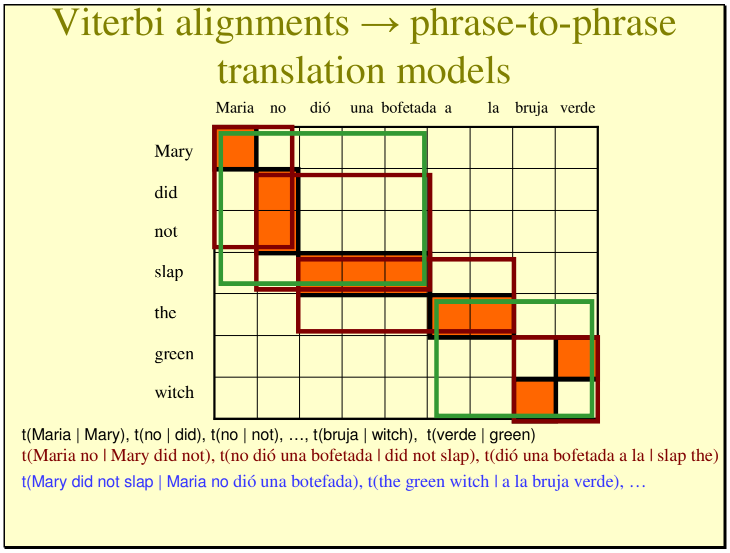


Viterbi alignments → phrase-to-phrase translation models

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)

Marcu shows what more can be done with a phrase-for-phrase approach. Note that this approach does not attempt to take into account meaning. It only matches sequences of words.

Slide 22



Viterbi alignments → phrase-to-phrase translation models

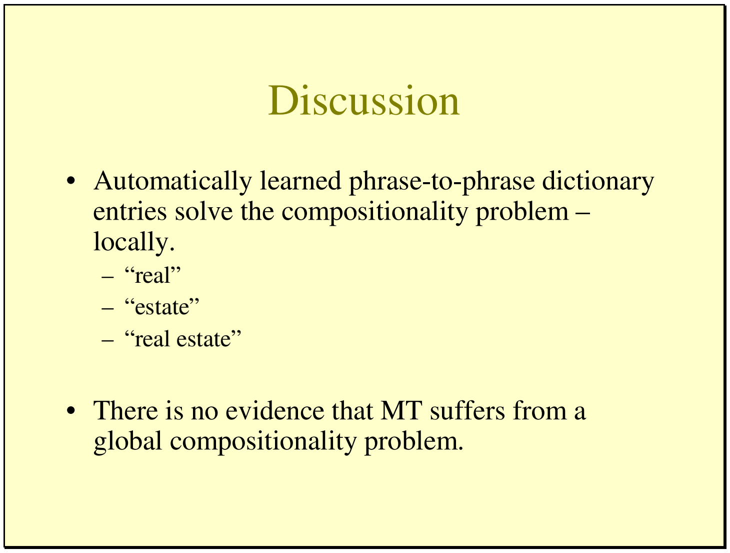
t(Maria | Mary), t(no | did), t(no | not), …, t(bruja | witch), t(verde | green)

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

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

Marcu replies initially by pointing out that data-driven MT is working better and better. He does not directly address the issue raised by Melby. What is not explicit in these slides is the claim by Melby that data-driven MT currently assumes that translations can be put together from pieces of text from a bitext without taking larger context into account.

Slide 23



Discussion

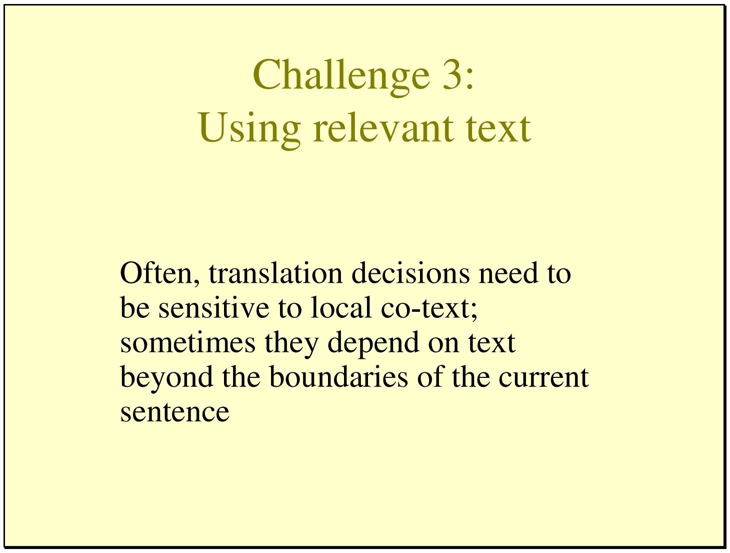
- Automatically learned phrase-to-phrase dictionary entries solve the compositionality problem – locally.
  - “real”
  - “estate”
  - “real estate”
- There is no evidence that MT suffers from a global compositionality problem.

Marcu is saying that data-driven MT has not reached its limits and that phrase dictionaries (for terms such as "real estate" that do not have the meaning of the combination of the two words of which they are composed) may be sufficient to produce MT output indistinguishable from professional human translation. Marcu also believes that compositionality is a troublesome concept only as long as it is confined to constructions derived from atomic, word-based meanings. If one is open to operate with longer phrases and abstract concepts, Marcu predicts that data-driven systems should evolve to the level where such concepts can be learned from data.

Melby points out that there is no proof that data-driven MT will not eventually plateau in quality, unless it abandons the compositionality assumption. Producing phrase dictionaries only pushes the compositionality question up a level: is meaning compositional based on combinations of phrases?

This issue is left open for further debate in the future.

Slide 24

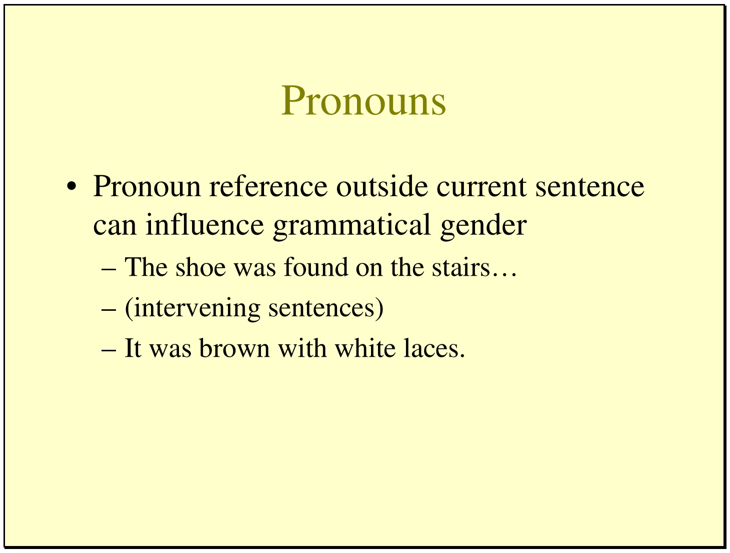


Challenge 3:
Using relevant text

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

Melby points out that translation sometimes depends on information beyond the current sentence and gives examples. For a discussion of various aspects of context, see the paper downloadable from the http://www.ttt.org/context/ webpage.

Slide 25

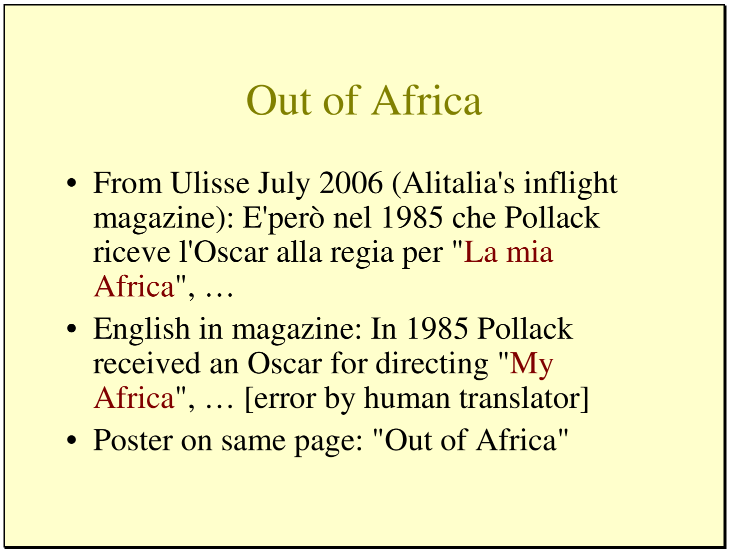


Pronouns

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

Sometimes the relevant co-text is far from the current sentence. How will data-driven MT take this into account?

Slide 26

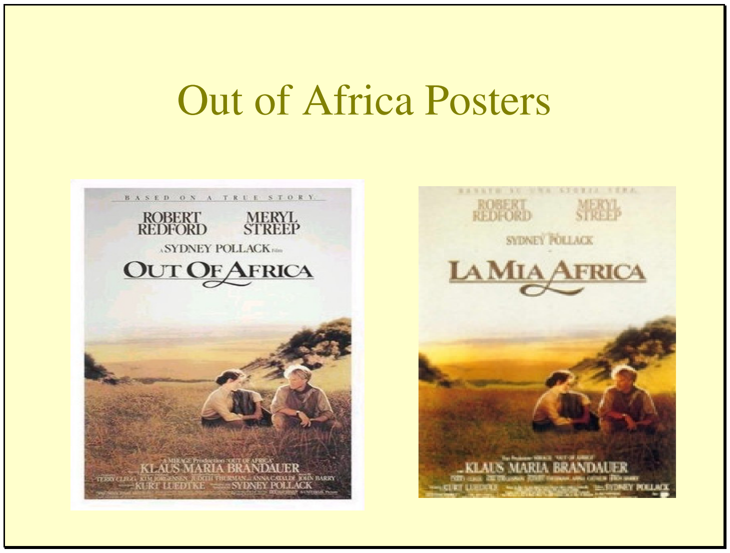


Out of Africa

- 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"

Even human translators can make mistakes when they do not search relevant texts and images. How will data-driven MT do better?

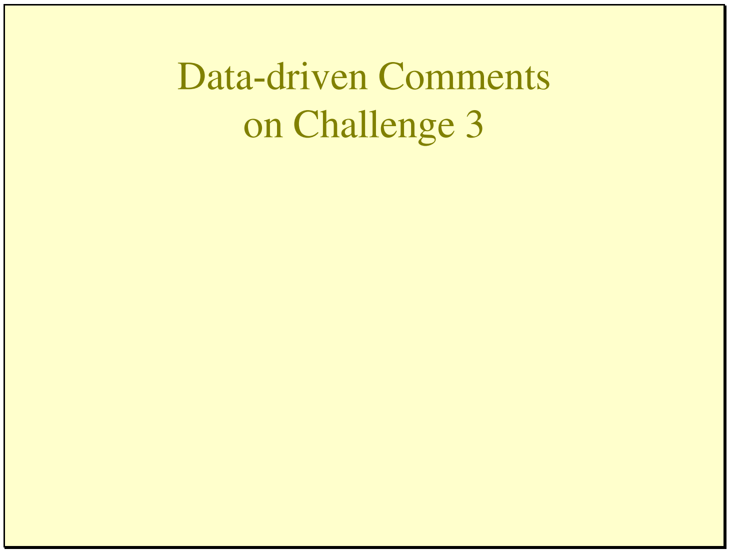
Slide 27



Out of Africa Posters

These posters show that the Italian movie title should be translated as "Out of Africa", not "My Africa".

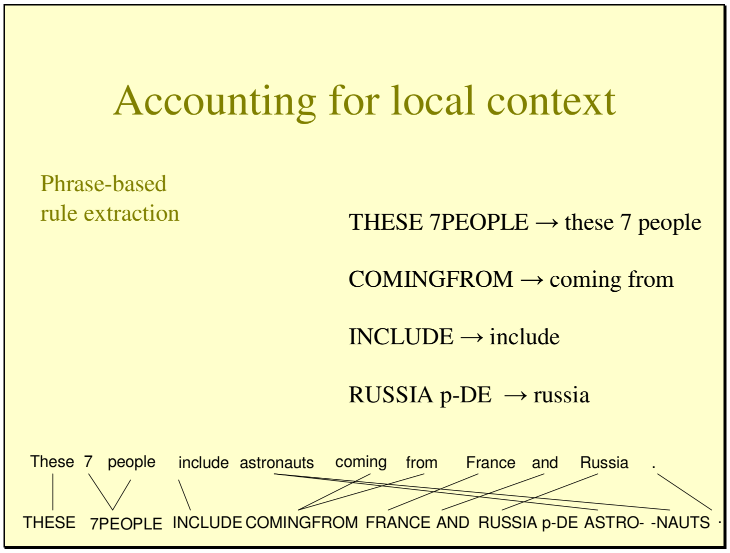
Slide 28



Data-driven Comments
on Challenge 3

Marcu responds to the challenge of using relevant text that is not in the immediate vicinity, say within the sentence, of the text being translated.

Slide 29



Accounting for local context

Phrase-based rule extraction

THESE 7PEOPLE → these 7 people

COMINGFROM → coming from

INCLUDE → include

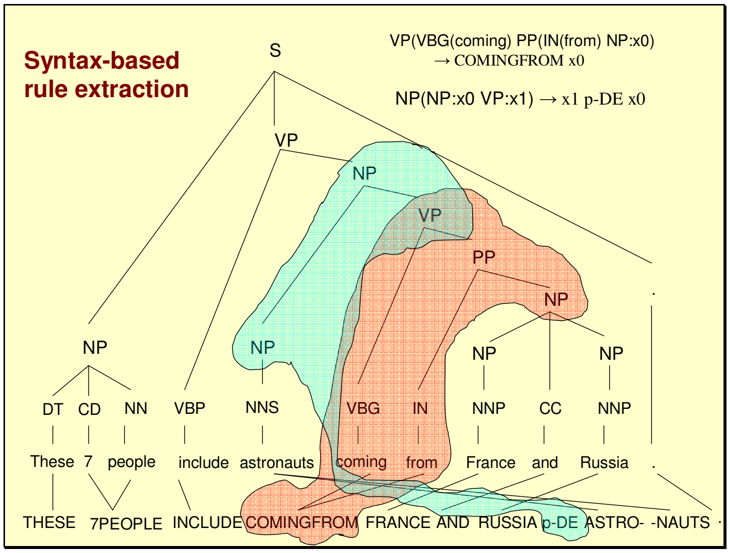
RUSSIA p-DE → russia

These 7 people include astronauts coming from France and Russia .

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

Marcu shows how local (adjacent word) context can be taken into account. Melby does not challenge that data-driven MT can do this to a degree.

Slide 30



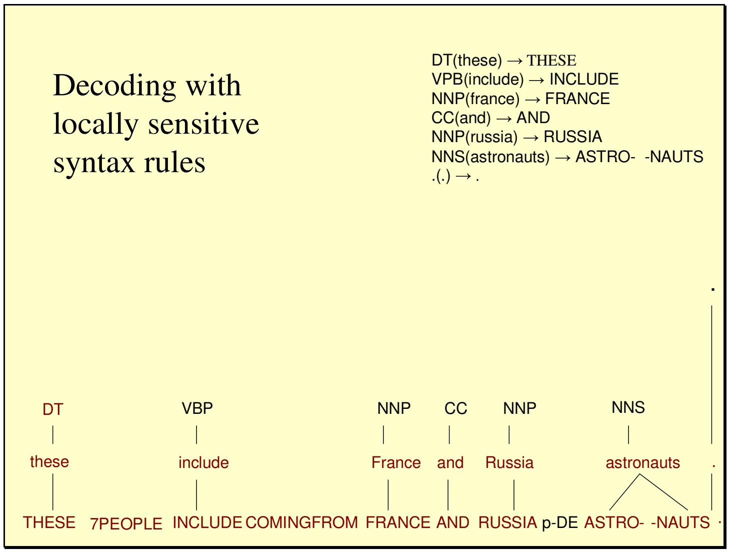
Syntax-based rule extraction

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

NP(NP:x0 VP:x1) → x1 p-DE x0

S
VP
NP
VP
PP
NP
NP
NP
NP
NP
DT CD NN VBP NNS VBG IN NNP CC NNP
These 7 people include astronauts coming from France and Russia .
.
THESE 7PEOPLE INCLUDE COMINGFROM FRANCE AND RUSSIA p-DE ASTRO- -NAUTS .

Slide 31



Decoding with locally sensitive syntax rules

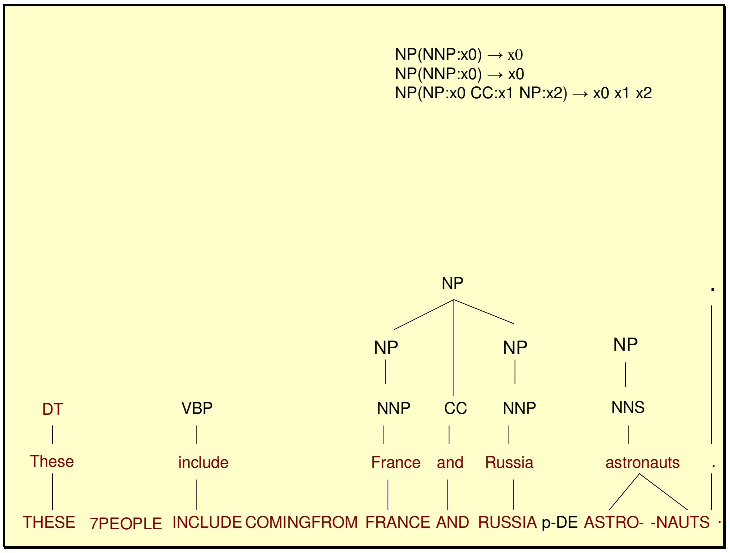
DT(these) → THESE
VPB(include) → INCLUDE
NNP(france) → FRANCE
CC(and) → AND
NNP(russia) → RUSSIA
NNS(astronauts) → ASTRO- -NAUTS
.(.) → .

DT VBP NNP CC NNP NNS .

these include France and Russia astronauts .

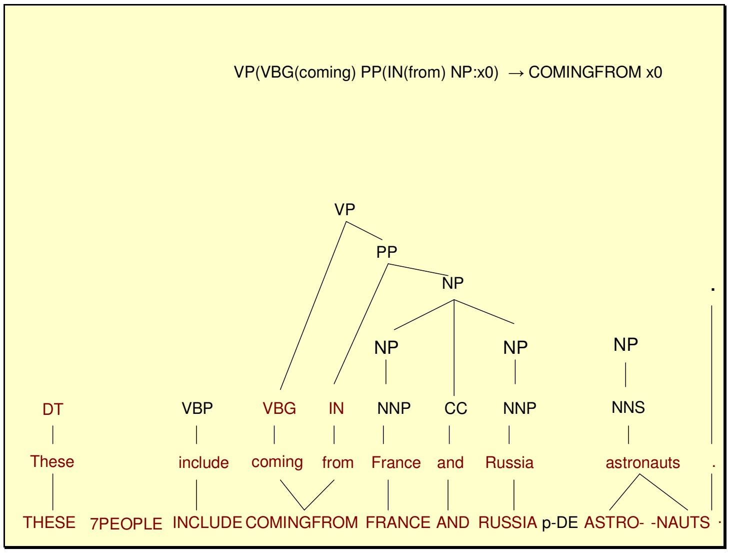
THESE 7PEOPLE INCLUDE COMINGFROM FRANCE AND RUSSIA p-DE ASTRO- -NAUTS .

Slide 32



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

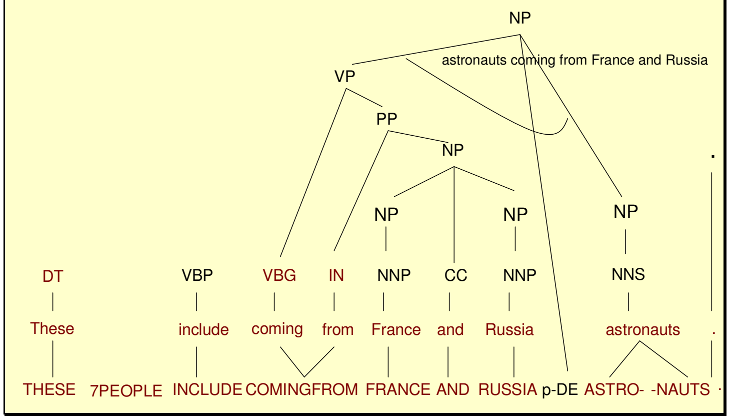
Slide 33



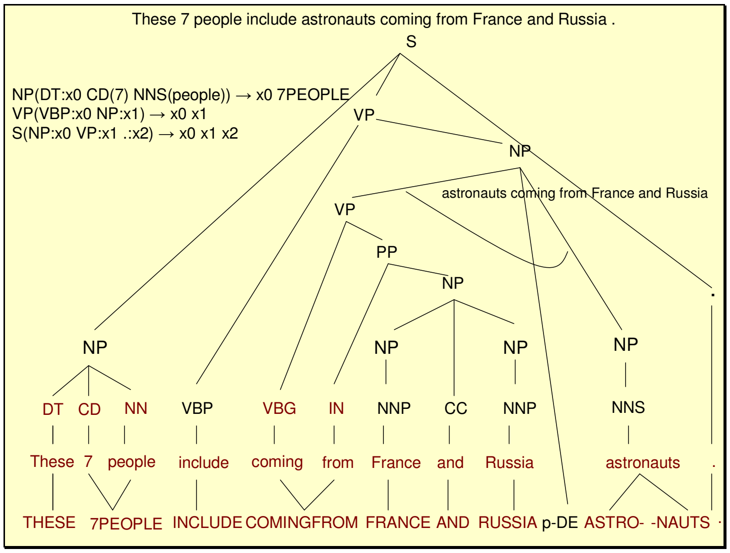
VP(VBG(coming) PP(IN(from) NP:x0) → COMINGFROM x0

Slide 34

NP(NP:x0 p-DE VP:x1) → x1 x0



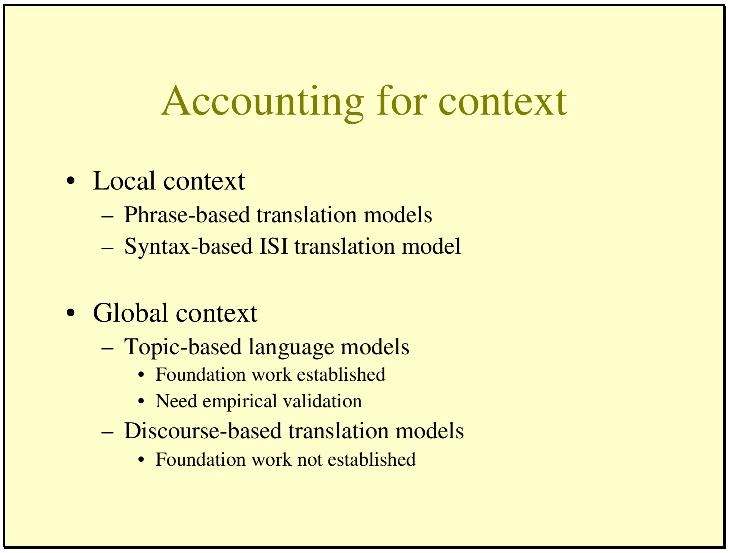
Slide 35



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

Slide 36



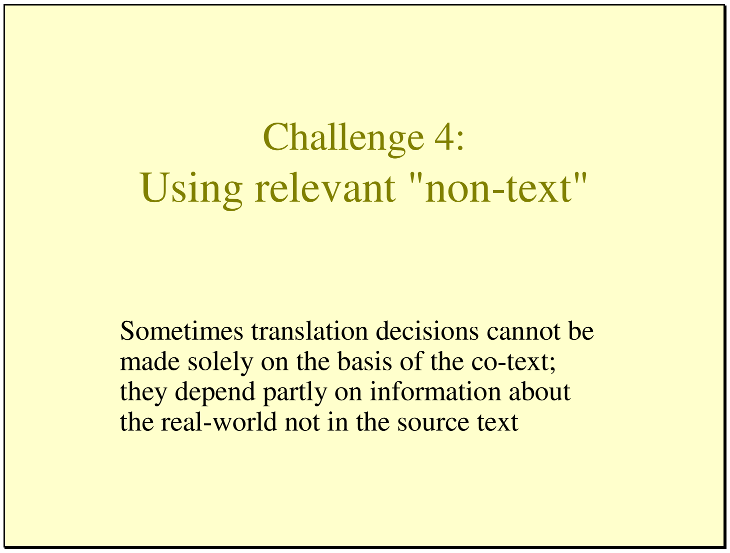
Accounting for context

- 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

Note that most data-driven MT systems do not currently include full syntactic analysis, which would be needed for the processing suggested by Marcu's local context slides. In this slide, Marcu admits that data-driven MT has not yet addressed the question of taking into account extra-sentential co-text.

The question of relevant text that is not local is left open for future discussion.

Slide 37

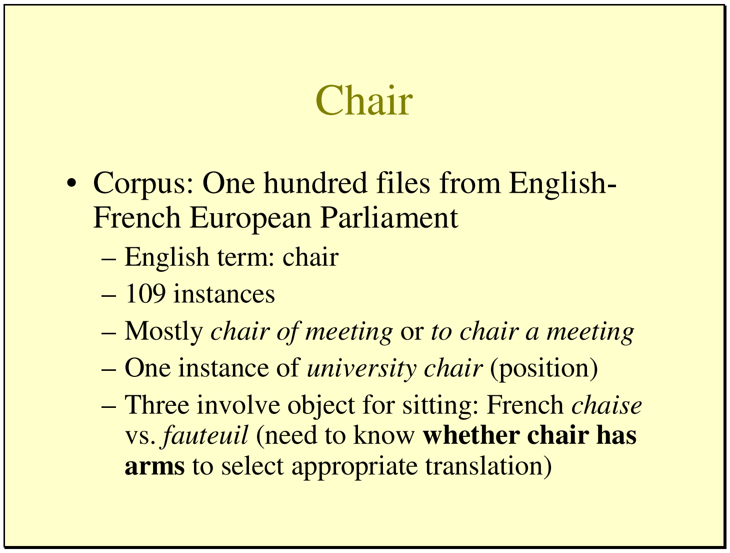


Challenge 4:
Using relevant "non-text"

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

Melby points out that sometimes translation requires even more than text, namely “non-text” (which could also be called “extra-text” or "un-text").

Slide 38

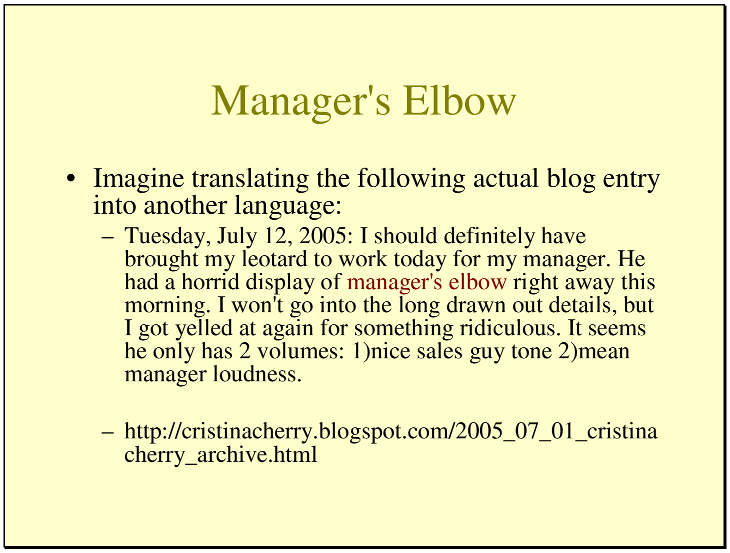


Chair

- 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)

The text surrounding the word "chair" may indicate that it is furniture as opposed to being a person, but may not be sufficient to determine whether the chair has arms or not. And this non-textual information is needed to translate correctly into French.

Slide 39

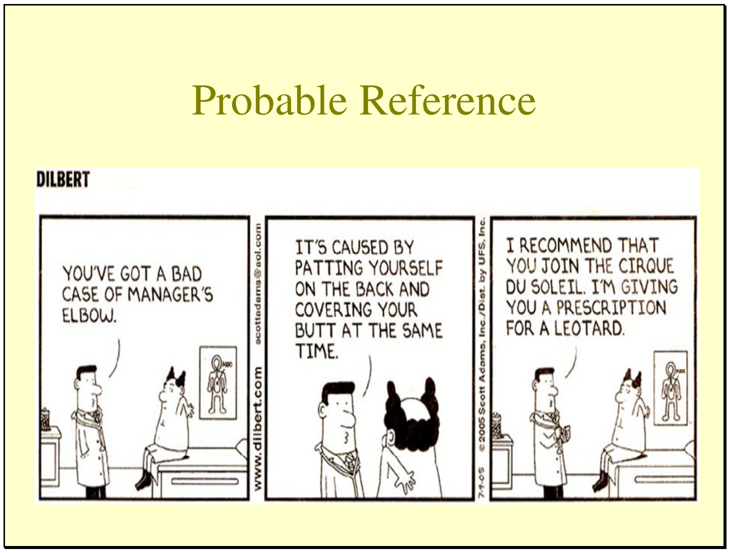


Manager's Elbow

- 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 sales guy tone 2)mean manager loudness.
  - http://cristinacherry.blogspot.com/2005_07_01_cristinacherry_archive.html

What if the term "manager's elbow" has not yet been translated into German or Hungarian? How would a data-driven MT system translate this term correctly without non-textual information about the real world (e.g. what the term means)? This challenge is related to Melby's claim that high quality translation requires an understanding of the source text. Without information beyond the source text, how will the system know that "manager's elbow" should not be translated literally?

Slide 40

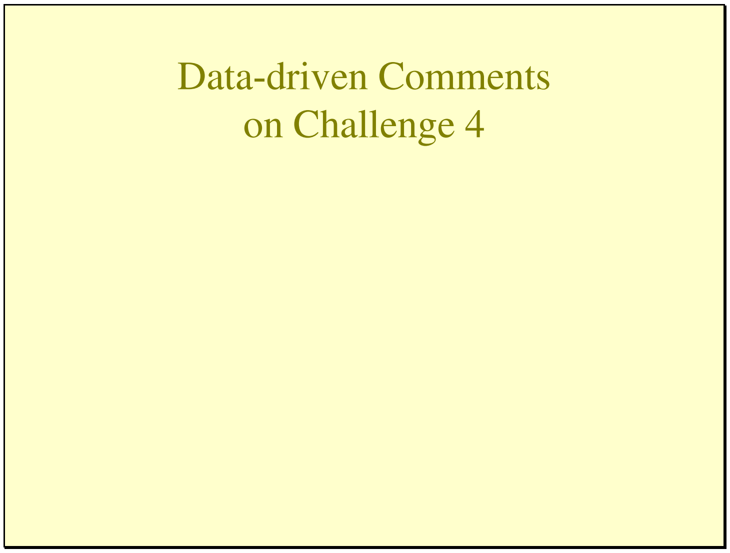


Probable Reference

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.

These cartoons (where the letters are probably not available as searchable text) is the probable real-world reference of the blog entry. On the previous slide.

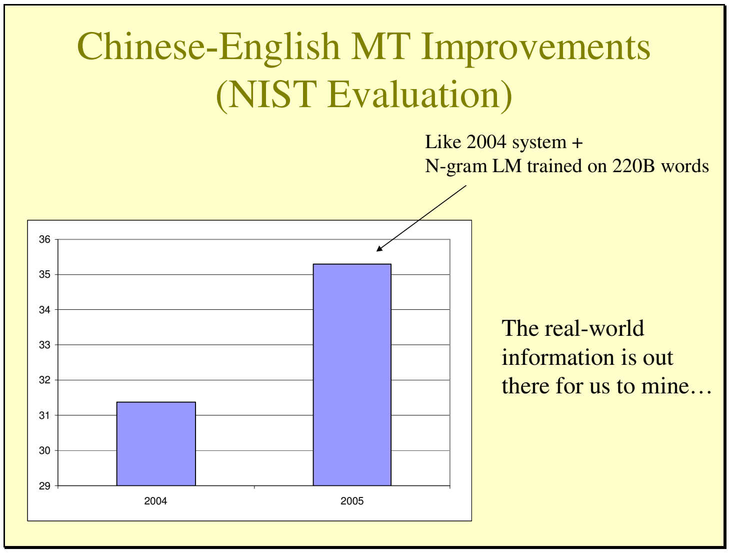
Slide 41



Data-driven Comments on Challenge 4

Marcu responds to Melby's claim that sometimes non-textual information is needed to produce a high-quality translation.

Slide 42



Chinese-English MT Improvements (NIST Evaluation)

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

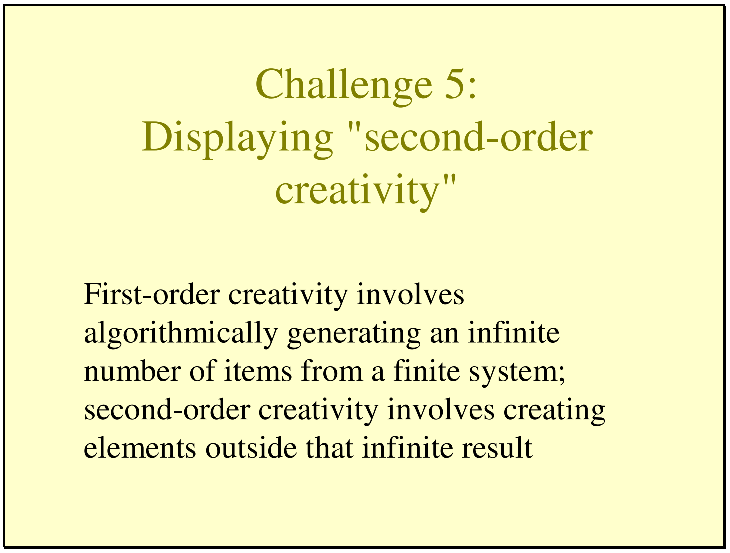
The real-world information is out there for us to mine…

Marcu claims that there is no need to use information that is not explicitly in text. Lots of information is available in text explicitly in long n-grams, for example. And even more is available implicitly in abstractions that are not learned by the current technology. Marcu believes that as the field matures, computer programs will be able to learn all the required information (explicit and implicit) from large bilingual and monolingual corpora.

Melby replies: Time will tell.

This issue is unresolved. The claim that non-textual information is needed constitutes a direct challenge to the data-driven approach. Marcu has not shown how Melby's examples, "chair" and "manager's elbow" would be handled by the data-driven approach. But even if he does solve these problems, there is an infinite number of other similar problems out there. This is a philosophical issue: is everything needed to produce high-quality translations available somewhere in a textual form?

Slide 43

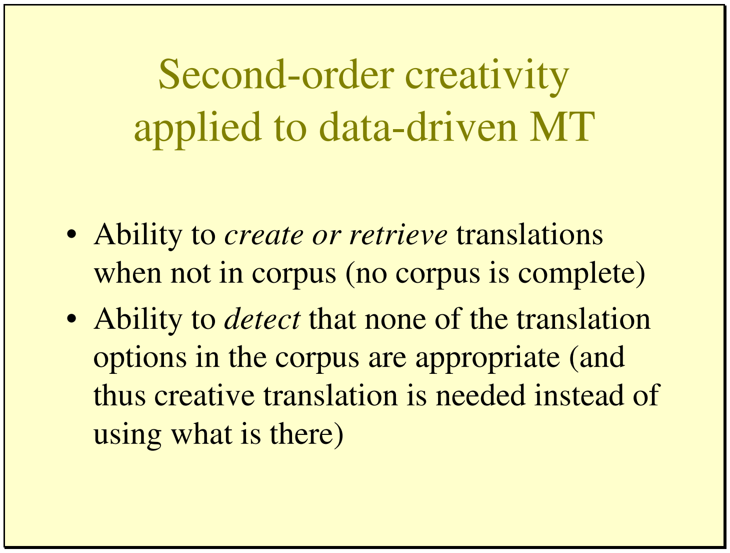


Challenge 5: Displaying "second-order creativity"

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

Melby defines "second order creativity". Note that this definition does not correspond directly to "second-order creativity" as used by John P. Miller in *The Holistic Curriculum*.

Slide 44

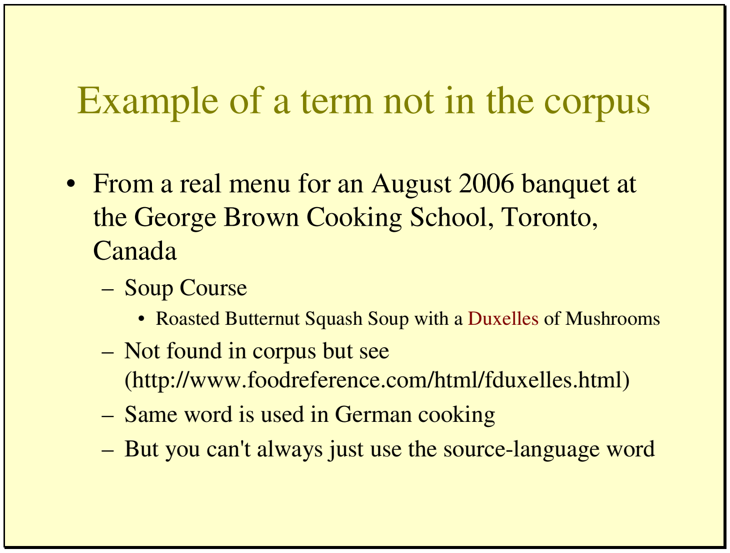


Second-order creativity applied to data-driven MT

- 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)

Melby points out that sometimes the answer is not in the corpus. (Afterwards, he wished he had further emphasized that that even if the appropriate solution is in the corpus, it may not be easy for the machine to detect which is the appropriate one for this particular source text.)

Slide 45

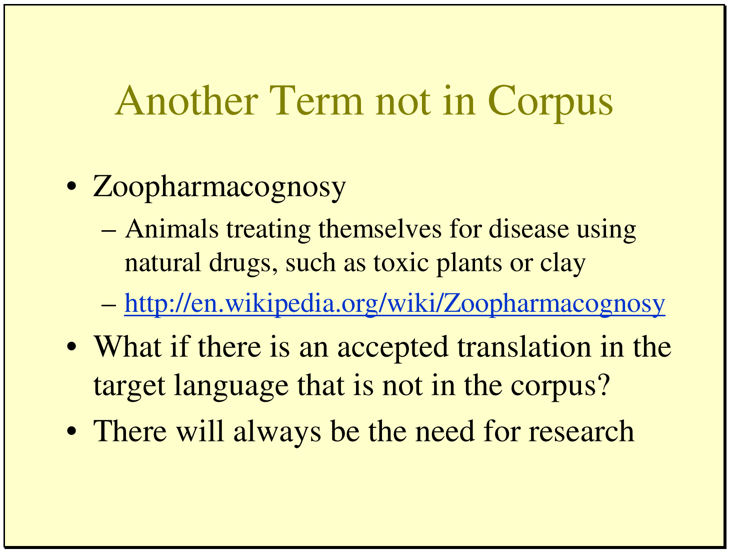


Example of a term not in the corpus

- 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

How would "a duxelles of mushrooms" be translated in Greek, Urdu, and Japanese? What if there is no segment of text in the corpus that contains this phrase and is already translated into the target language?

Slide 46

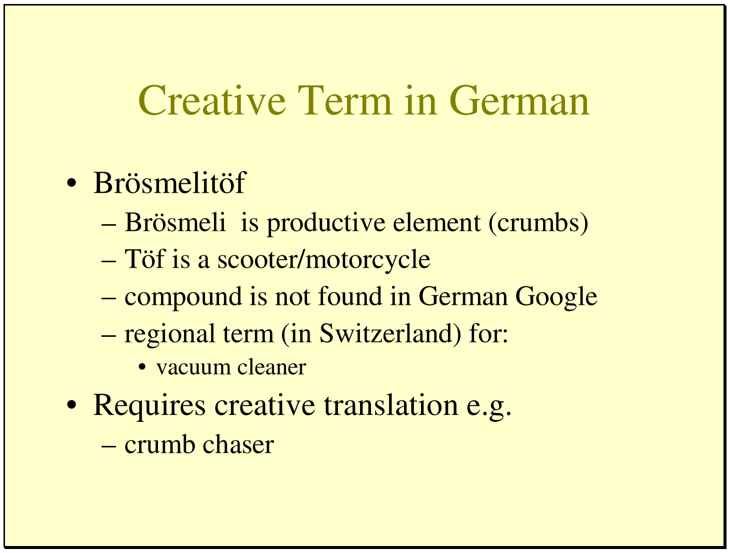


Another Term not in Corpus

- 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

What if Zoopharmacognosy is not your corpus?

Slide 47

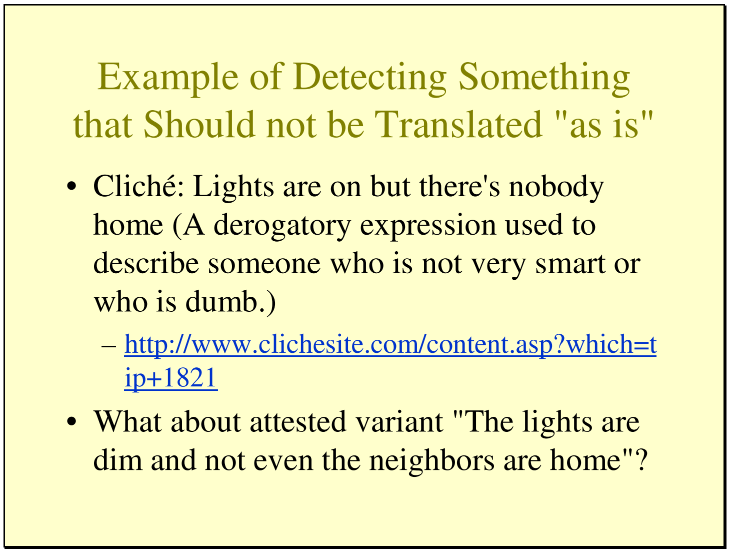


Creative Term in German

- 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

The word Brösmelitöf, a synonym for vacuum cleaner, was made up by a relative of Melby living in Switzerland.

Slide 48

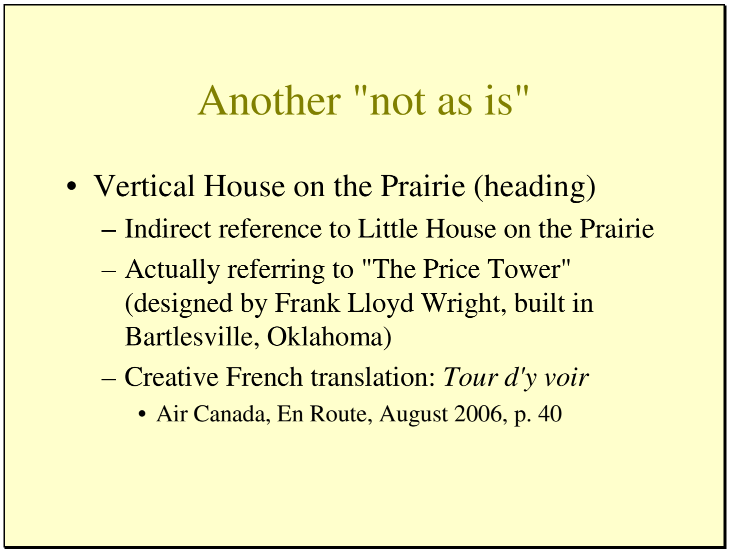


Example of Detecting Something that Should not be Translated "as is"

- 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=tip+1821
- What about attested variant "The lights are dim and not even the neighbors are home"?

Even in the phrase "Lights are on but there's nobody home" is in the corpus, how would a data-driven MT system know that "The lights are dim and not even the neighbors are home" is not literal but instead a variant of the idiomatic expression?

Slide 49

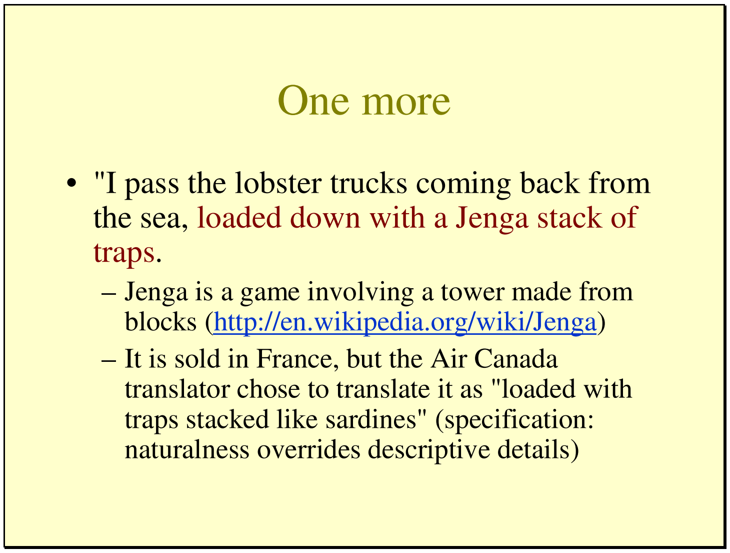


Another "not as is"

- 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

How would a data-driven MT system know that "Vertical House on the Prairie" is based on the name of an old TV series and needs an equally creative translation into French, related to the literal referent (The Price Tower).

Slide 50

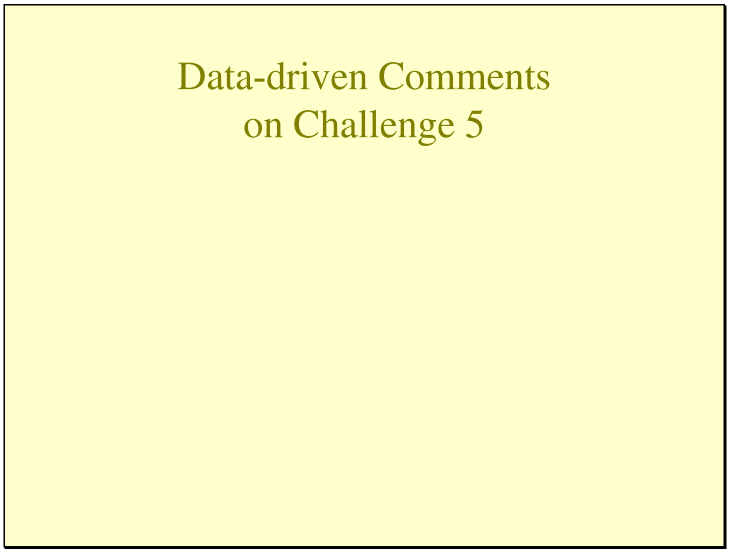


One more

- "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)

Sometimes a specification of naturalness for a translation requires considerable creativity to avoid a translation that uses a term (such as "Jenga") that is very uncommon in the target language. No one translation will work in all contexts, so a corpus search may not be sufficient.

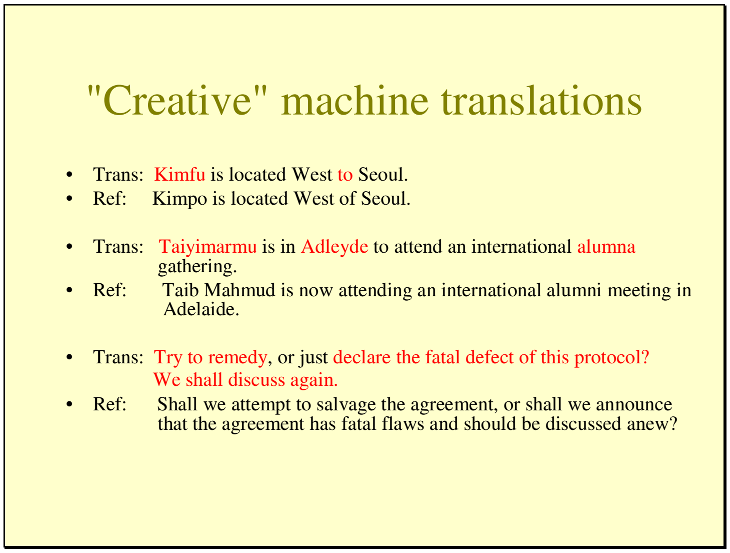
Slide 51



Data-driven Comments
on Challenge 5

Now Marcu responds to the need for second-order creativity in some cases.

Slide 52



"Creative" machine translations

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

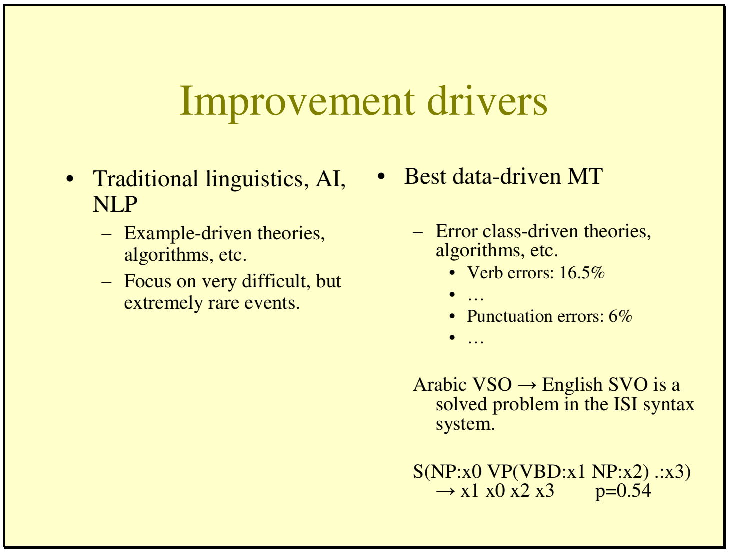
- Trans: Taiyimarmu is in Adleyde to attend an international alumna gathering.
- Ref: Taib Mahmud is now attending an international alumni meeting in Adelaide.

- 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?

Marcu gives examples of human translations that are different from other human translations. Marcu makes the point that humans make some of the types of mistakes that Melby is complaining about machines making. Marcu believes that machines will get as good as humans at the task of translating not because they will produce perfect outputs, but because not even humans produce perfect outputs.

Melby believes that this answer sidesteps the issue. The fact that humans make mistakes does not guarantee that machines will exhibit second-order creativity.

Slide 53



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.

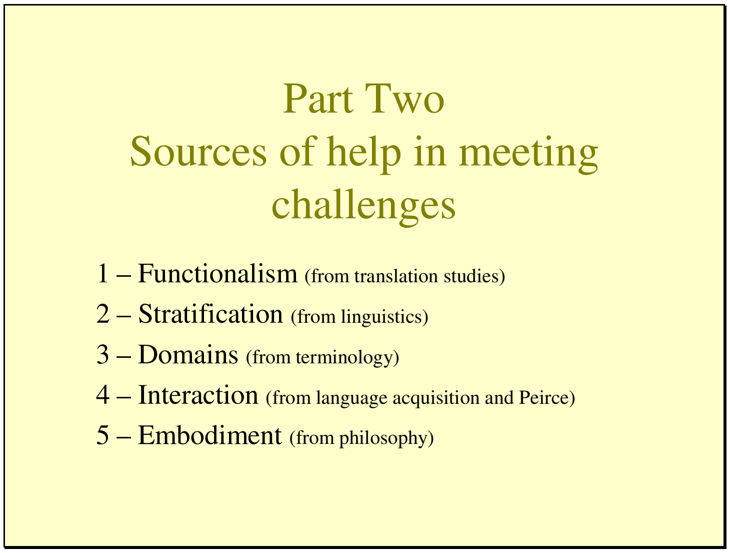
S(NP:x0 VP(VBD:x1 NP:x2) .:x3)
→ x1 x0 x2 x3 p=0.54

Marcu hopes that data-driven MT will succeed. Time will tell.

The question of whether data-driven MT can handle situations where second-order creativity is needed is open to future discussion.

There was a tentative agreement to re-stage this debate after five years (in 2011) to see how data-driven MT has done relative to the challenges pointed out by Melby.

Slide 54



Part Two
Sources of help in meeting challenges

1 – Functionalism (from translation studies)

2 – Stratification (from linguistics)

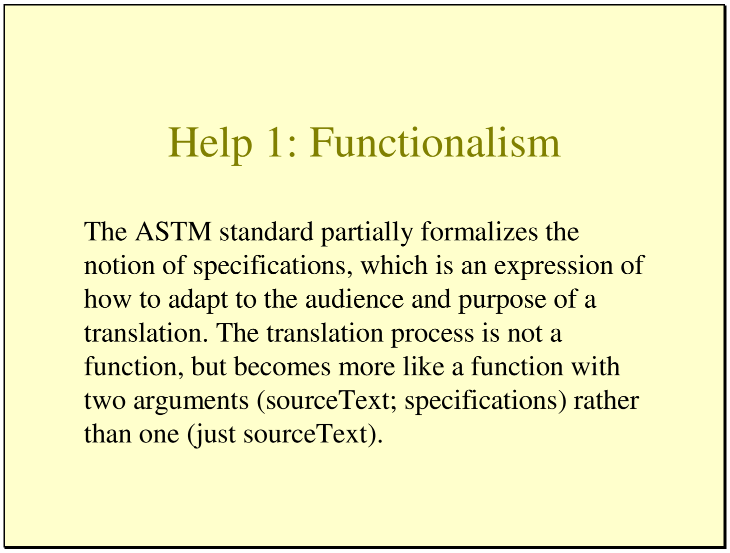
3 – Domains (from terminology)

4 – Interaction (from language acquisition and Peirce)

5 – Embodiment (from philosophy)

Melby lists some areas in various disciplines that might help data-driven MT advance.

Slide 55

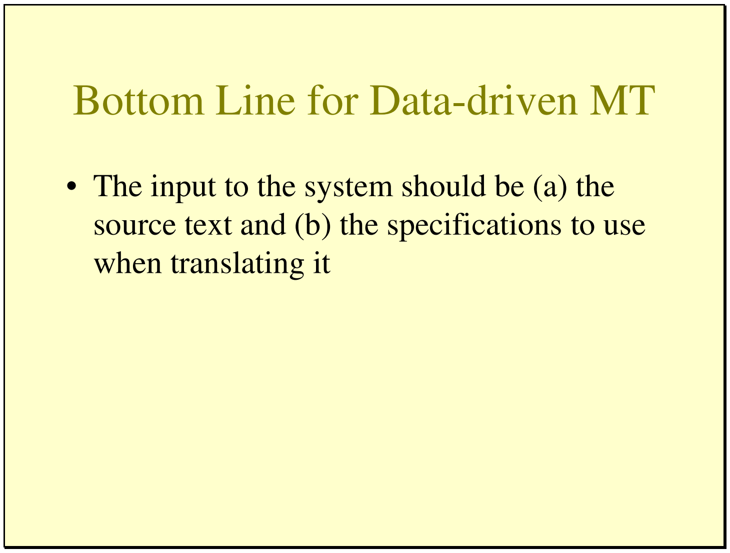


Help 1: Functionalism

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 (just sourceText).

Functionalism is a trend in translation studies. Melby suggests that data-driven MT developers would do well to study Functionalism.

Slide 56

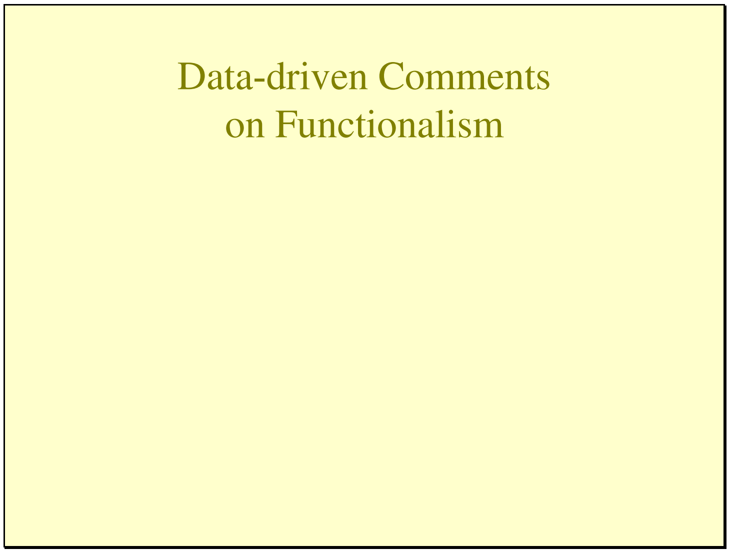


Bottom Line for Data-driven MT

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

Melby suggests that data-driven MT cannot reach human quality unless the input is not just a source text but also a set of specifications.

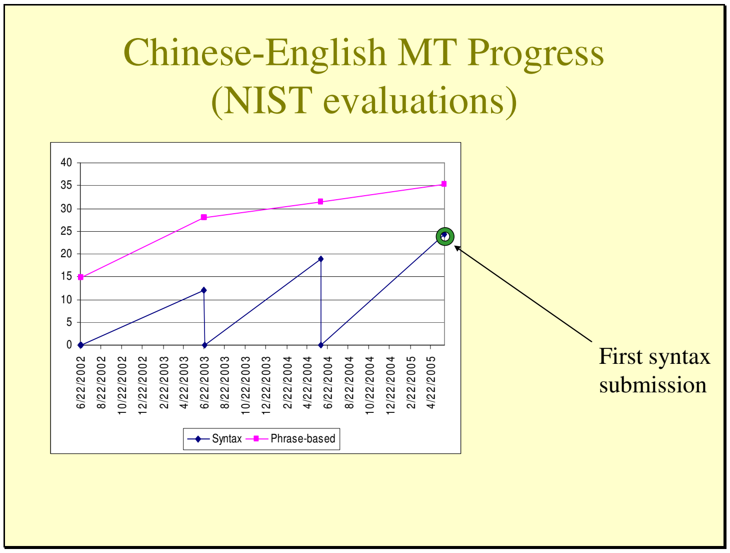
Slide 57



Data-driven Comments on Functionalism

Marcu responds.

Slide 58

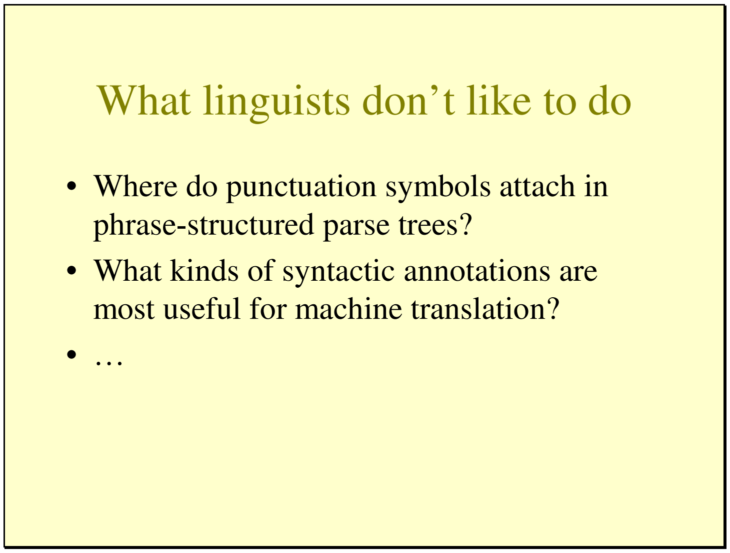


Chinese-English MT Progress
(NIST evaluations)

First syntax submission

Marcu notes that MT quality is improving by one measure (but, Melby points out, that does not include specifications).

Slide 59

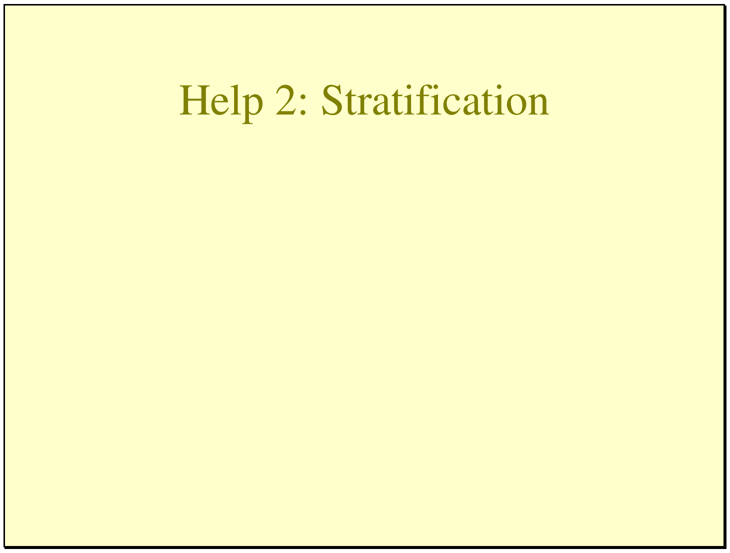


What linguists don’t like to do

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

Marcu asks linguists to help MT with the problems that MT developers point out, not the ones linguistics are interested in.

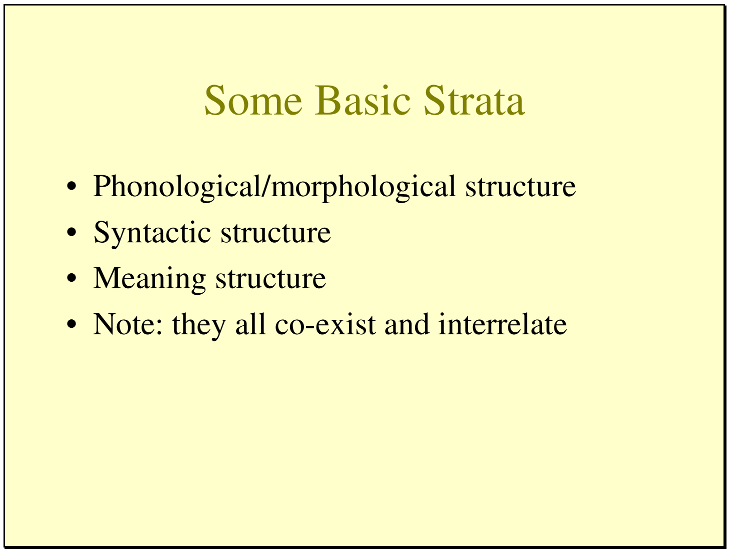
Slide 60



Help 2: Stratification

Melby discusses stratification.

Slide 61

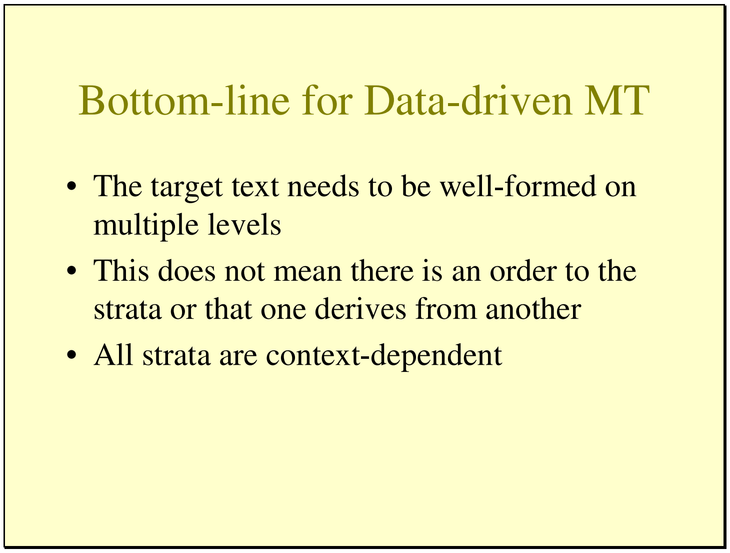


Some Basic Strata

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

Most linguistics recognize various strata.

Slide 62

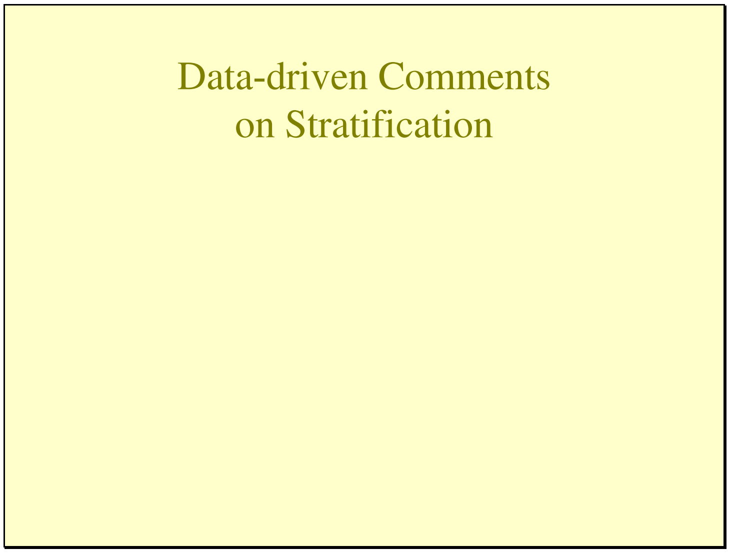


Bottom-line for Data-driven MT

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

Melby suggests that a data-driven MT system must take strata into account and check for well-formedness on multiple levels (syntactically, semantically, etc).

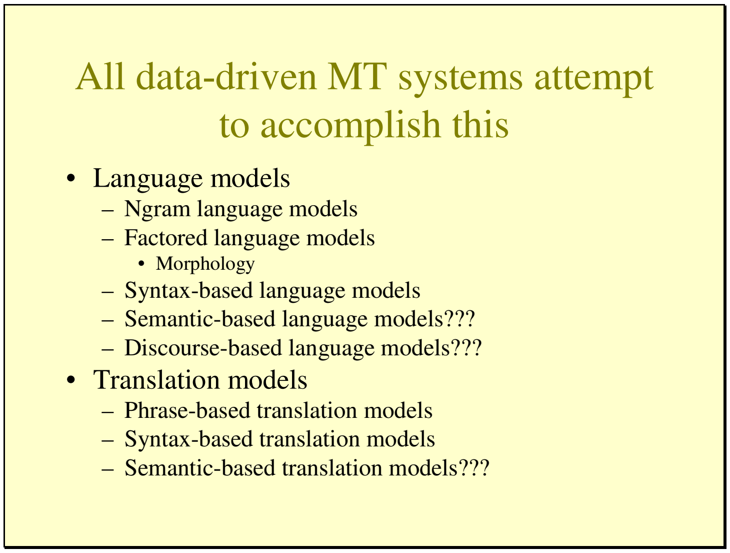
Slide 63



Data-driven Comments
on Stratification

Marcu responds.

Slide 64

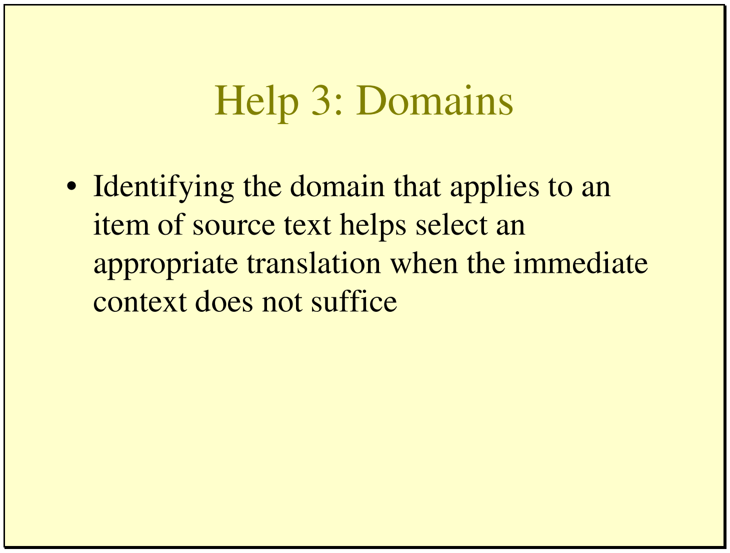


All data-driven MT systems attempt to accomplish this

- 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???

Marcu points out that data-driven MT does recognize multiple strata. The question is whether it will reach the higher strata.

Slide 65



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

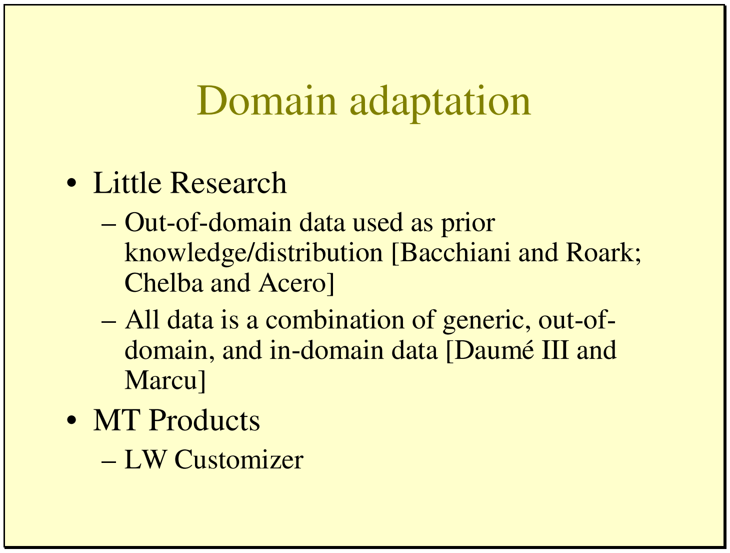
Melby suggests that identifying the domain that a source text item (word, phrase, etc.) belongs to would help improve MT accuracy.

Slide 66

Data-driven Comments
on Domains

Marcu responds.

Slide 67

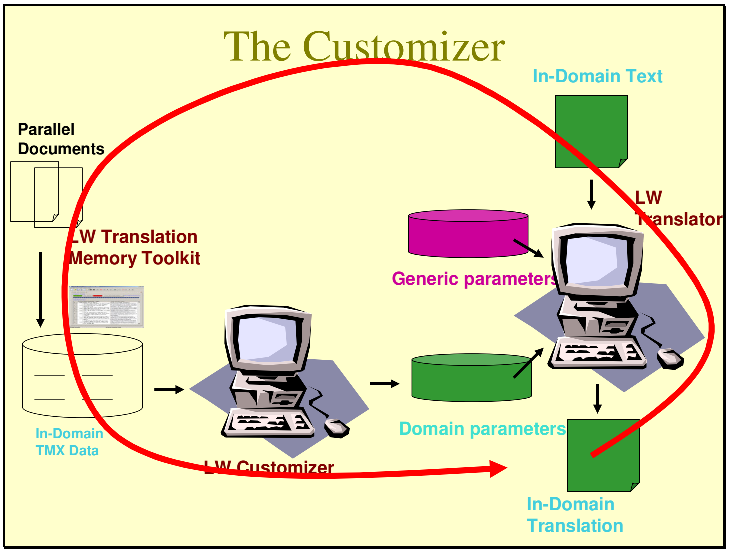


Domain adaptation

- 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

Marcu admits there is relatively little research so far on domains in MT and suggests that the LW (Language Weaver) Customizer uses domains.

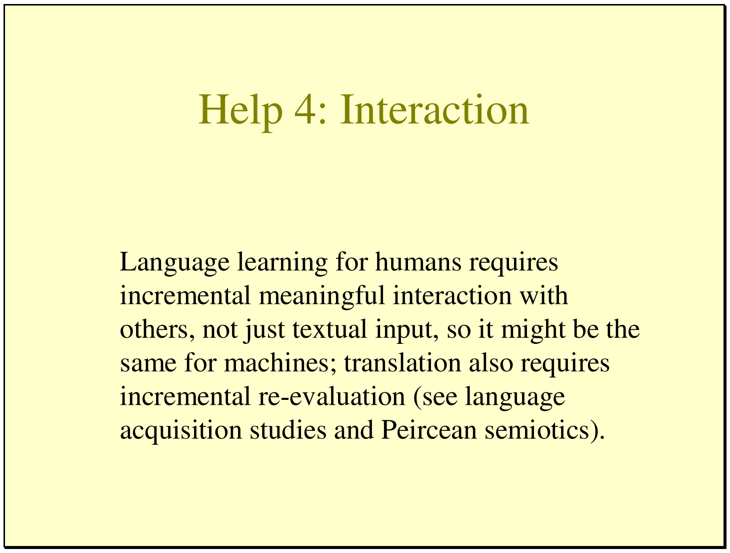
Slide 68



The LW Customizer customizes a data-driven MT system to texts that are part of a particular domain, in order to increase accuracy.

Melby pointed out later that there is a tension between the size of the bitext corpus used for data-driven MT and the focus of the corpus to the relevant domain. You can't have it both ways: either a corpus is really big and somewhat unfocused or smaller and more focused. So the argument that the main answer to obtaining better quality in data-driven MT is a bigger corpus is flawed because of this tension between size and focus. There may eventually be huge bitext corpora in particular domains for particular language combinations, but not for all languages in all domains. So here may be an inherent limitation to the data-driven MT approach. Professional human translators like having domain-specific resources to draw on but often have to function without them. They use intuition to combine domain-specific and general language information. How will data-driven MT decide when it is ok to use information from outside the subset of the corpus that is attached to the current domain? And how will it even identify the relevant domain?

Slide 69

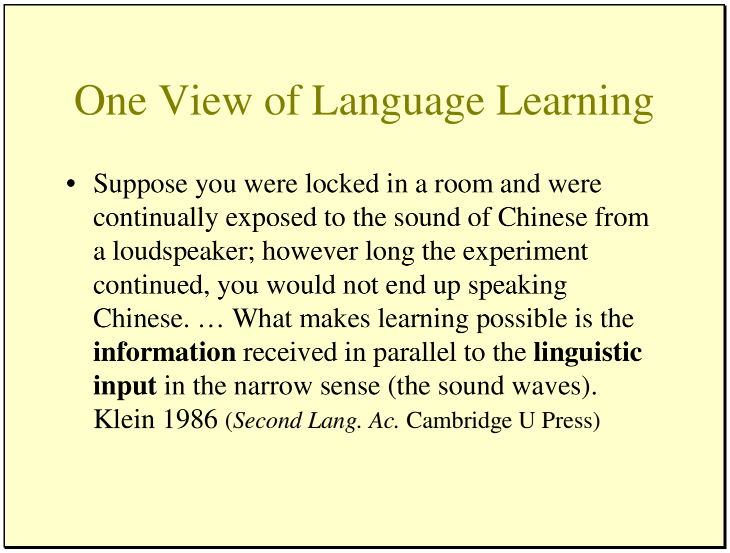


Help 4: Interaction

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).

Melby suggests that human-to-human interaction is essential to the development of professional translation proficiency.

Slide 70

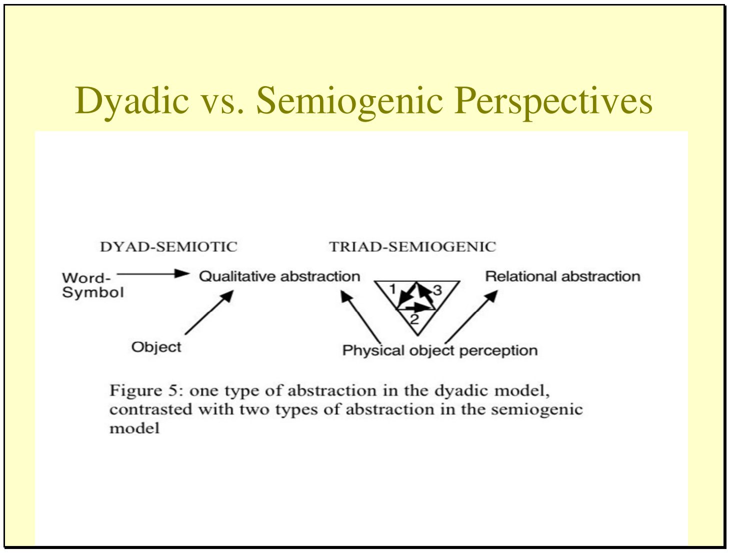


One View of Language Learning

- 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)

Melby claims that language learning is more than just hearing language. At least partial understanding is also needed, or the language is not much different from noise.

Slide 71

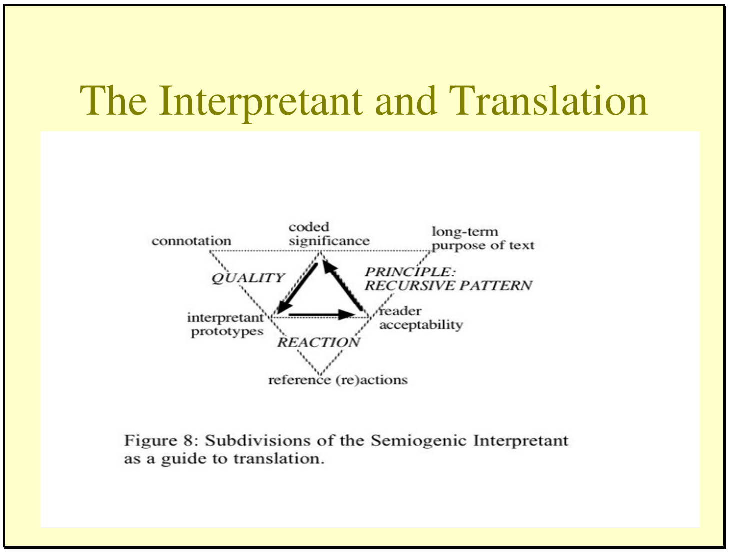


Dyadic vs. Semiogenic Perspectives

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

Melby suggests that the Peircean semiogenic model of language is relevant to translation and that the dyadic model is insufficient. This idea is discussed at length in the article "Quality in translation: a lesson for the study of meaning" published in the journal *Linguistics and the Human Sciences*, Volume 1, number 3 (copyright 2005, appeared in 2006).

Slide 72

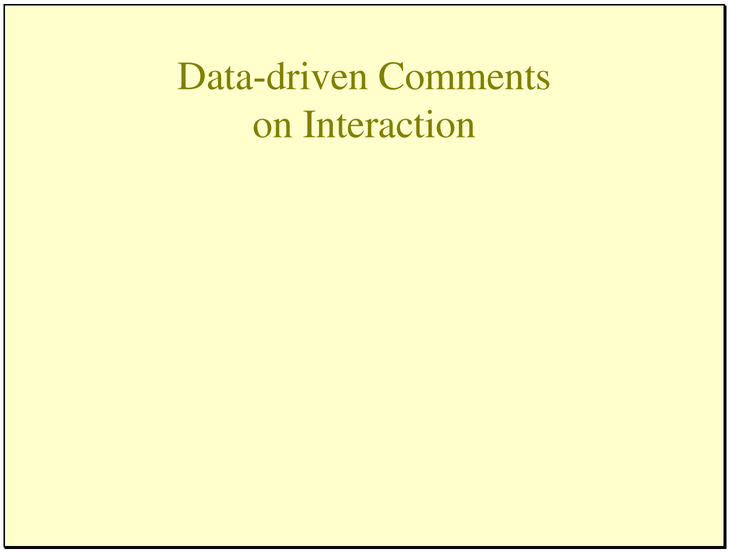


The Interpretant and Translation

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

This diagram from the article shows how the Peircean triangle supports the notion of interaction in translation. Tentative translations are evaluated in terms of such factors as possibly connotations (desirable and less so), possible referential ambiguities, purpose of the translation, and predicted reader response.

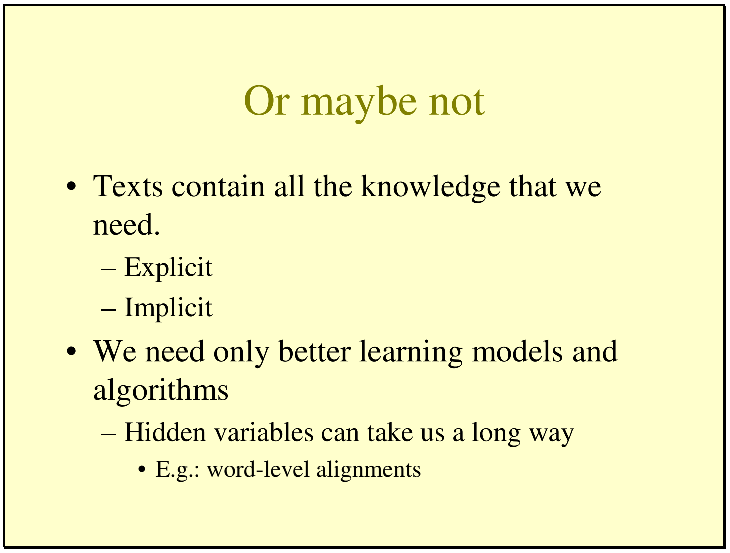
Slide 73



Data-driven Comments on Interaction

Marcus responds.

Slide 74

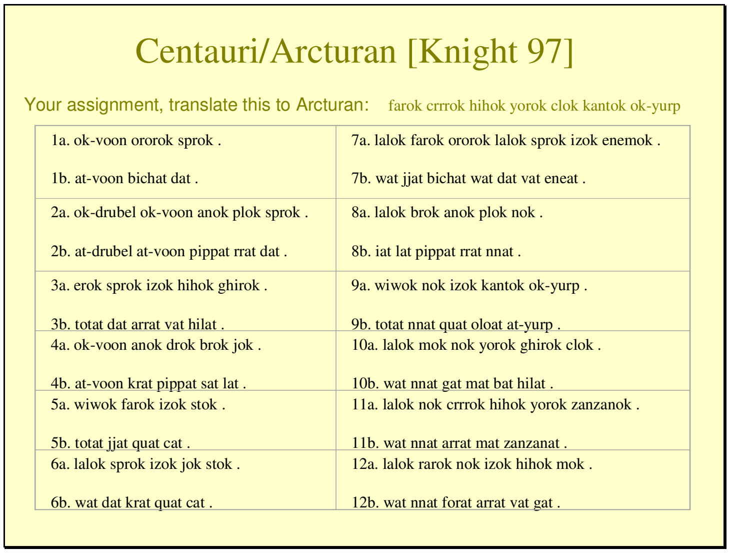


Or maybe not

- 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

Marcu suggests that interaction is not needed and that everything needed for translation is in texts.

Slide 75



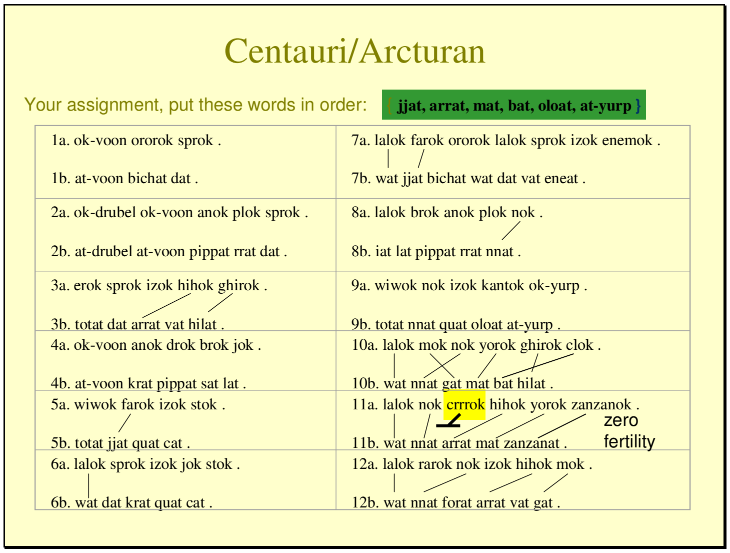
Centauri/Arcturan [Knight 97]

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

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Marcu gives an example of how humans can figure out languages they do not know.

Slide 76



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 .

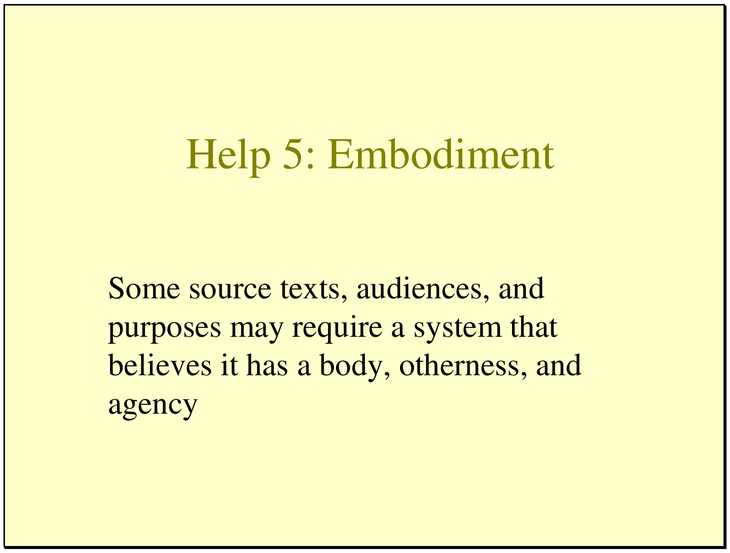
zero fertility

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Example continues.

Slide 77

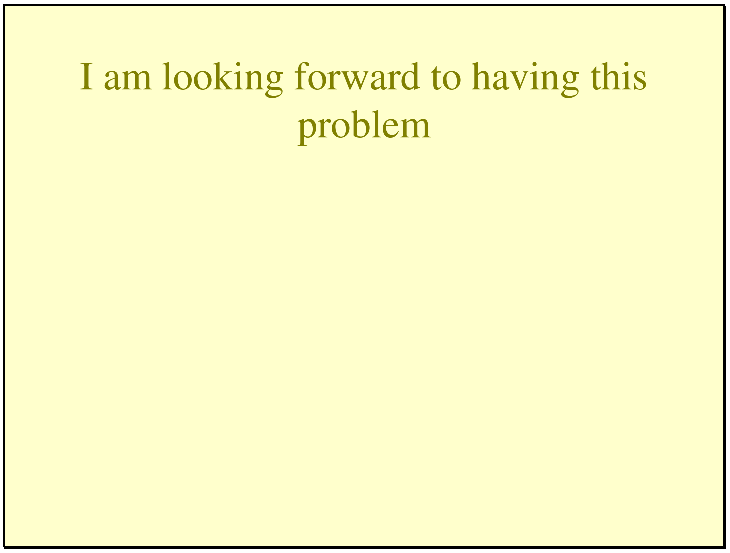


Help 5: Embodiment

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

Melby suggests that a machine translation system that does really well compared to a professional human translator will need to be conscious and believe that it has a body, that other conscious being exist, and that it has free will (agency).

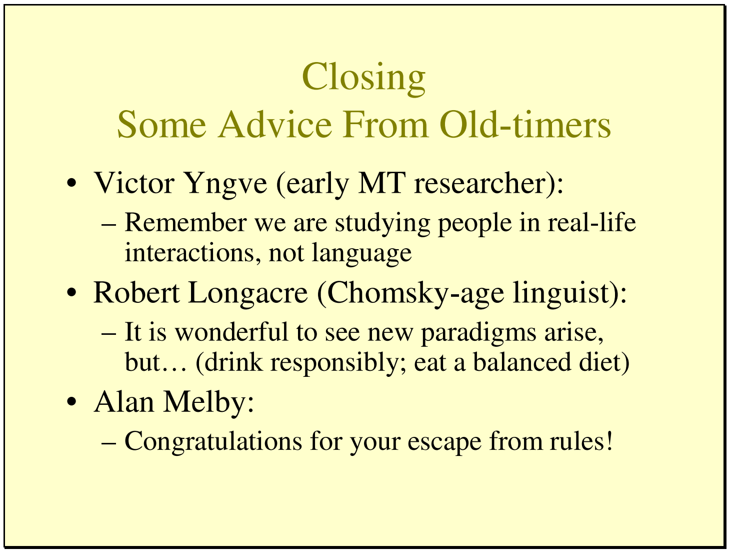
Slide 78



I am looking forward to having this problem

Marcu indicates that there is plenty to work on in MT that does not require embodiment.

Slide 79

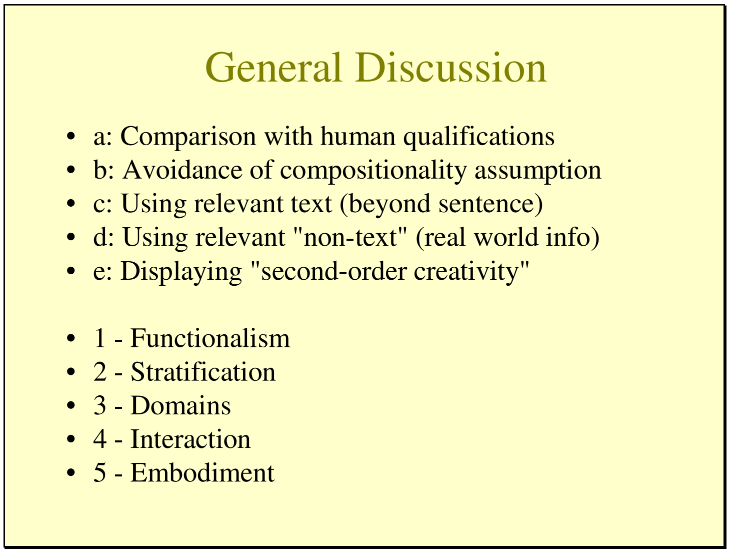


Closing
Some Advice From Old-timers

- 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!

Melby asked two old-timers (contemporaries of Noam Chomsky) for advice to the data-driven MT community and congratulates the data-driven MT community for escaping from the assumptions of the previous generation of rule-based MT systems.

Slide 80



General Discussion

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

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

The bottom line is that it is too soon to tell whether data-driven MT will succeed in producing translations indistinguishable from professional human translations. We should hold another debate every five years and see what progress has been made in data-driven MT relative to the five challenges identified by Melby and the five areas of translation studies, linguistics, and philosophy that might need to be taken into account in future MT systems.

Melby and Marcu ended the debate as friends who disagree on almost everything except that the debate is far from over.