Automated Translation: How Does It Work?

Stelios Piperidis
ELRC, ILSP/Athena RC

Simon Krek
Jožef Stefan Institute

Acknowledgements: with adaptations of materials that appeared at MT Marathons, WMTs, etc.

Machine Translation

Agenda:

- Why MT: Volume, Quality and Cost?
- Why is MT hard?
- MT + Human Translators = Quality
- How does modern statistical MT work?
- Its all about Data!
- And the right kind of Data!

Machine Translation, Quality and Cost?

- Europe = Multilinguality
- 24 official languages,
  24+2 CEF languages
- So much to translate!
- Translation costs!?
- Can MT help?
- What about the Quality?

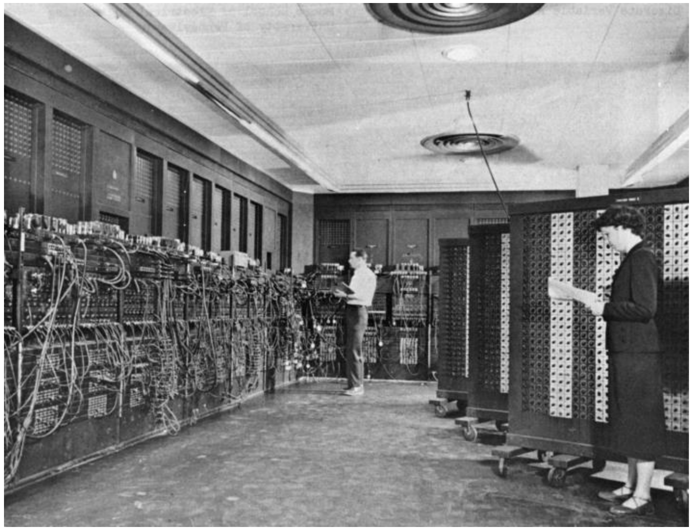


Image: https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg
License: public domain

Why is MT Hard?

- Human languages are:
  - Elegant
  - Efficient
  - Flexible
  - Complex
- One word/sentence may mean many things
- Many ways of saying the same thing
- Meaning depends on context
- Literal and figurative language (metaphor)
- Language and culture (different ways of conceptualising the same thing)
- Word order
- Morphology
- …

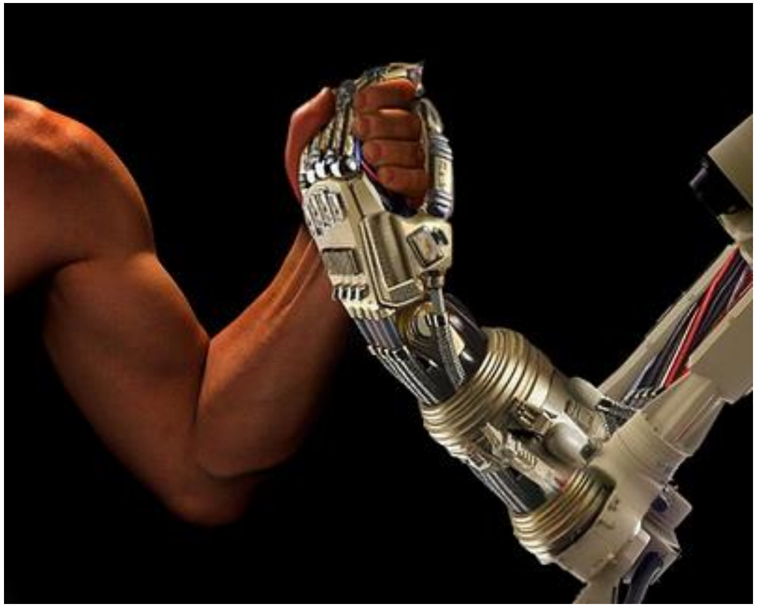
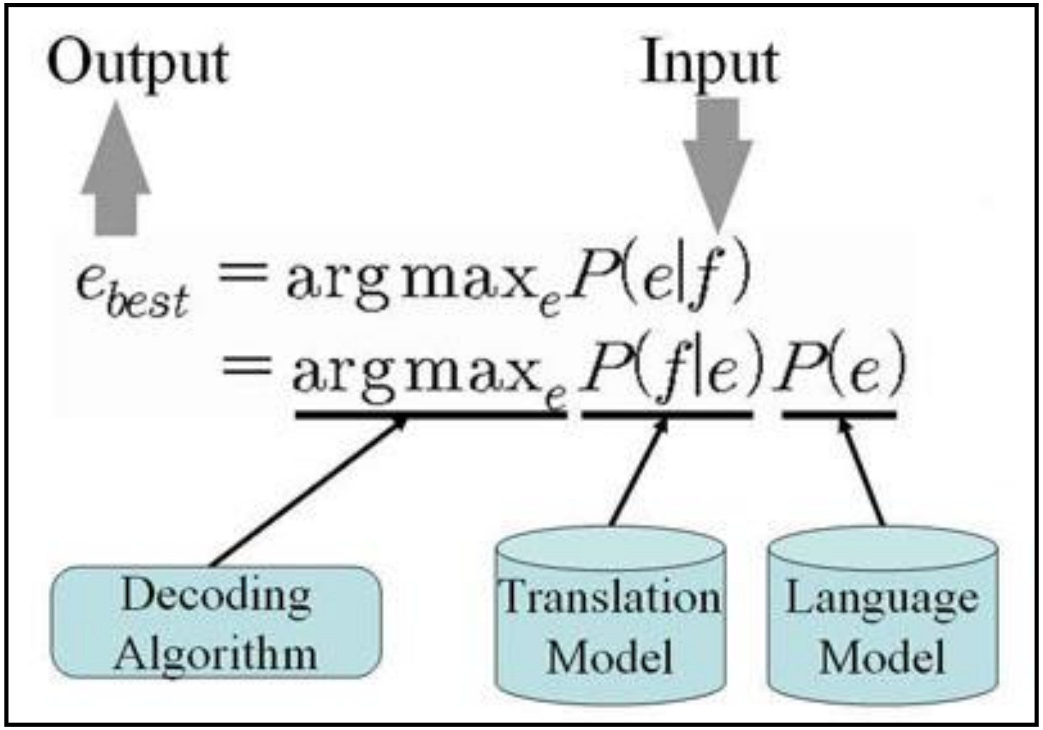


Image: http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg
License: CC BY-NC-SA 3.0

Language and Translation is Complex

- Language/translation is complex
- We cannot compute it exactly
- We tried: rule-based MT and LT …
- What do we do?
- Machine Learning
  - Learns from data ⇒ data is all important
  - Approximate solution ⇒ not perfect, needs help
    - human professional translators
    - Post-editing
    - Automated Translation ≠ Automatic

How does Modern MT Work?


Output
Input
$e_{best} = \arg\max_e P(e|f)$
$= \underbrace{\arg\max_e}\, \underbrace{P(f|e)}\, \underbrace{P(e)}$
Decoding Algorithm
Translation Model
Language Model

- No maths today
- Instead:
- The story of Statistical MT in pictures …
- Its all about **Data** …

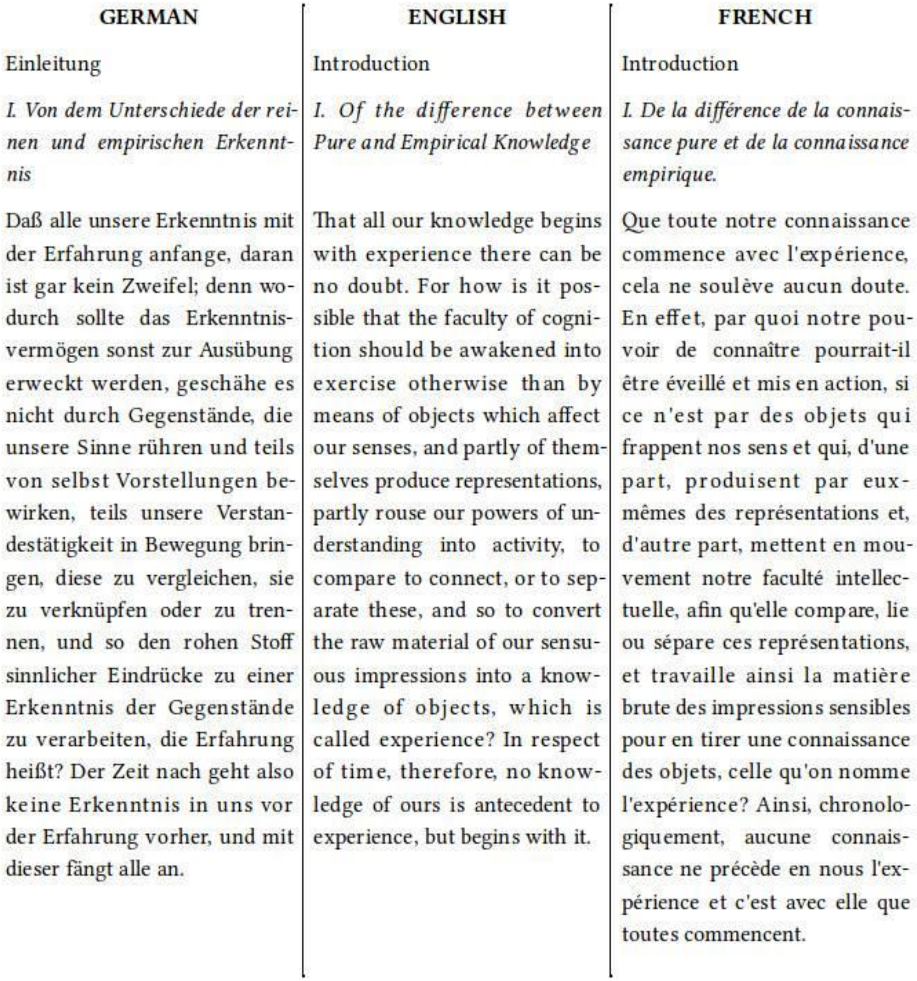
How does Modern MT Work?

Statistical MT learns from data

Two kinds of data:

- Human translations
- Text in the target language

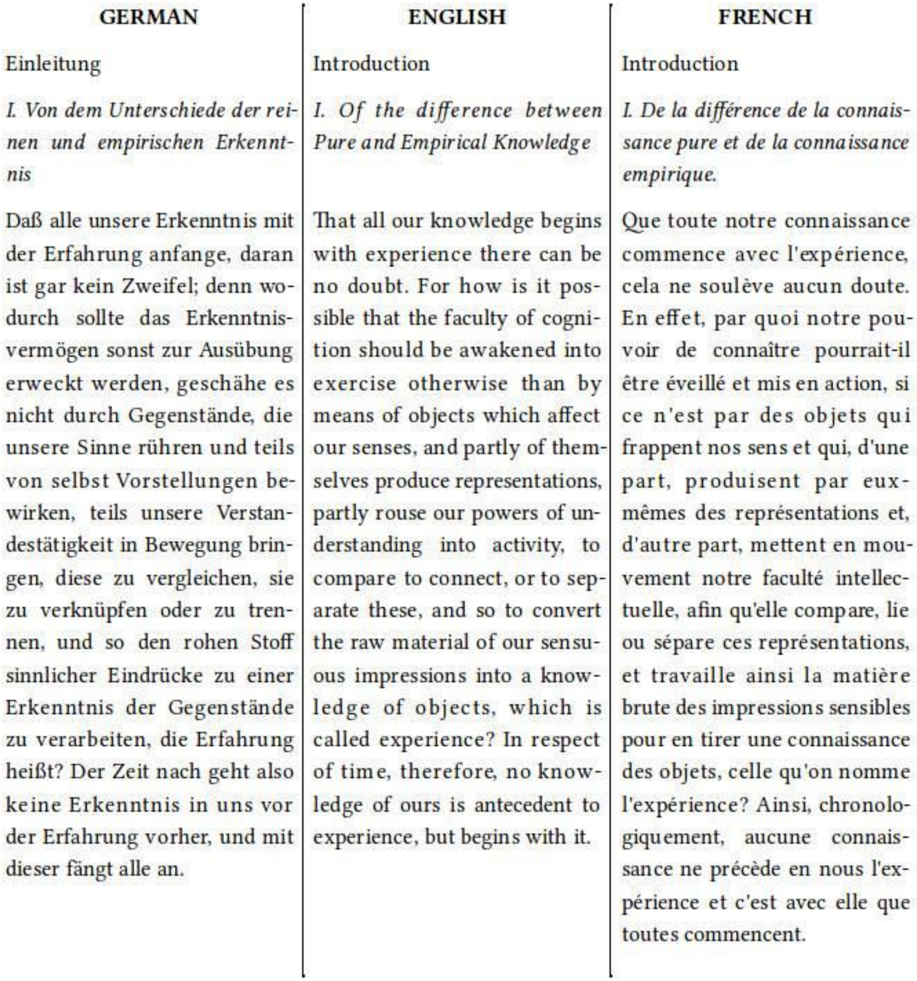
- The more data the better!
- Also: the right kind of data!



GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

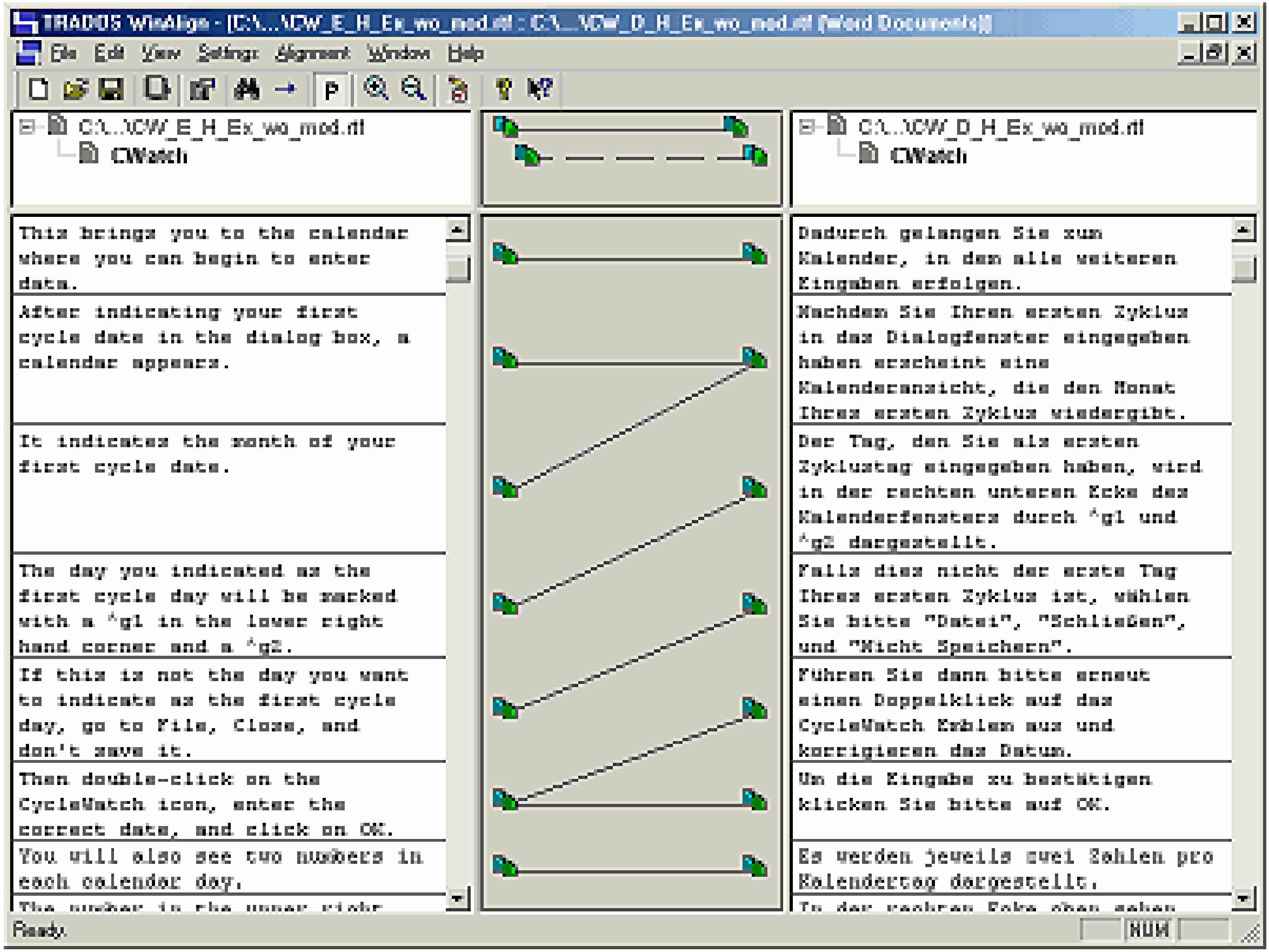
What can/do we Learn from Data?

- Which sentences translate as which: sentence alignment
- Which words translate as which: word alignment + translation probabilities
- What is good target language like: language model



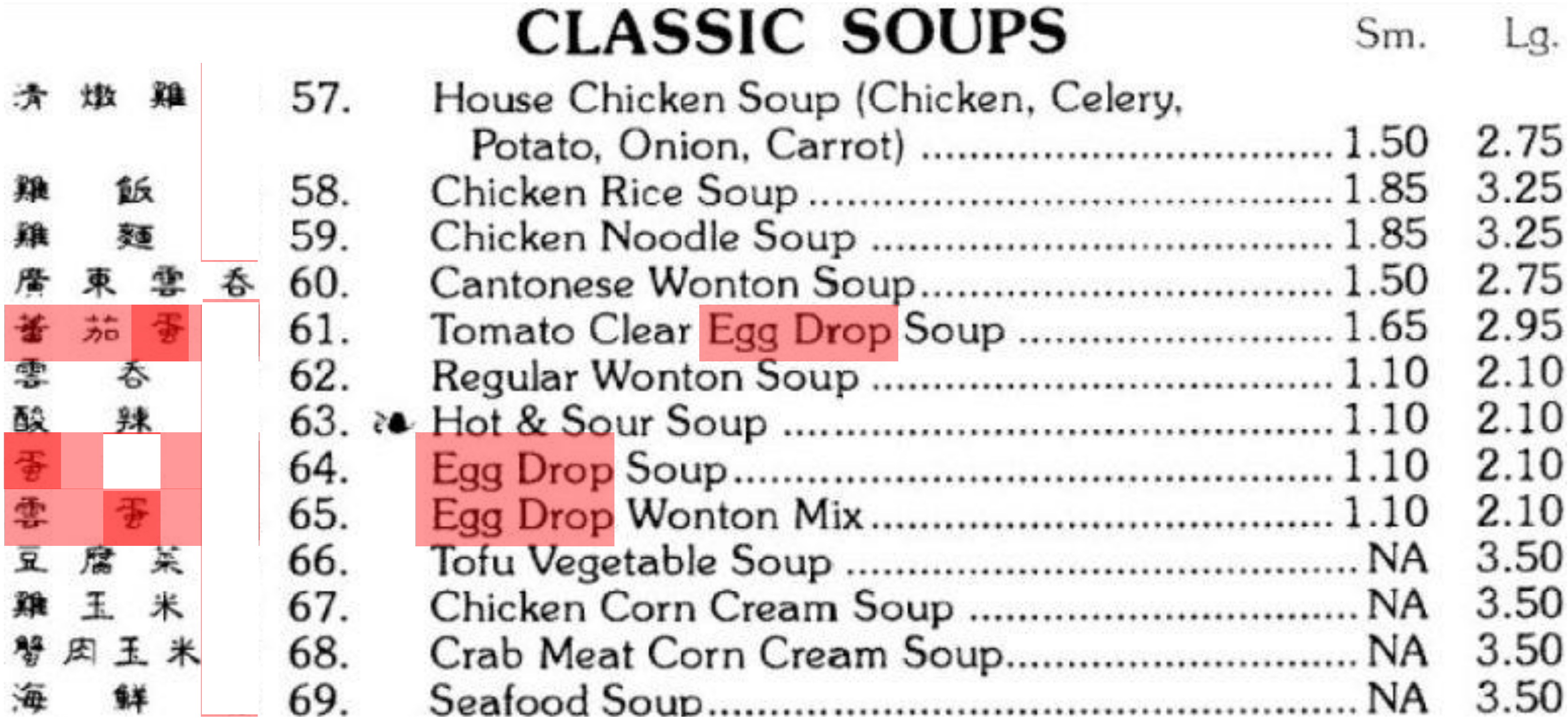
GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations, et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

Sentence Alignment


TRADOS WinAlign - [C:\...\CW_E_H_Ex_wo_mod.rtf : C:\...\CW_D_H_Ex_wo_mod.rtf (Word Documents)]
File Edit View Settings Alignment Window Help
P
C:\...\CW_E_H_Ex_wo_mod.rtf
CWatch
C:\...\CW_D_H_Ex_wo_mod.rtf
CWatch
Ready
NUM

This brings you to the calendar where you can begin to enter data.	Dadurch gelangen Sie zum Kalender, in dem alle weiteren Eingaben erfolgen.
After indicating your first cycle date in the dialog box, a calendar appears.	Nachdem Sie Ihren ersten Zyklus in das Dialogfenster eingegeben haben erscheint eine Kalenderansicht, die den Monat Ihres ersten Zyklus wiedergibt.
It indicates the month of your first cycle date.	Der Tag, den Sie als ersten Zyklustag eingegeben haben, wird in der rechten unteren Ecke des Kalenderfensters durch ^g1 und ^g2 dargestellt.
The day you indicated as the first cycle day will be marked with a ^g1 in the lower right hand corner and a ^g2.	Falls dies nicht der erste Tag Ihres ersten Zyklus ist, wählen Sie bitte "Datei", "Schließen", und "Nicht Speichern".
If this is not the day you want to indicate as the first cycle day, go to File, Close, and don't save it.	Führen Sie dann bitte erneut einen Doppelklick auf das CycleWatch Emblem aus und korrigieren das Datum.
Then double-click on the CycleWatch icon, enter the correct date, and click on OK.	Um die Eingabe zu bestätigen klicken Sie bitte auf OK.
You will also see two numbers in each calendar day.	Es werden jeweils zwei Zahlen pro Kalendertag dargestellt.
The number in the upper right	In der rechten Ecke oben sehen

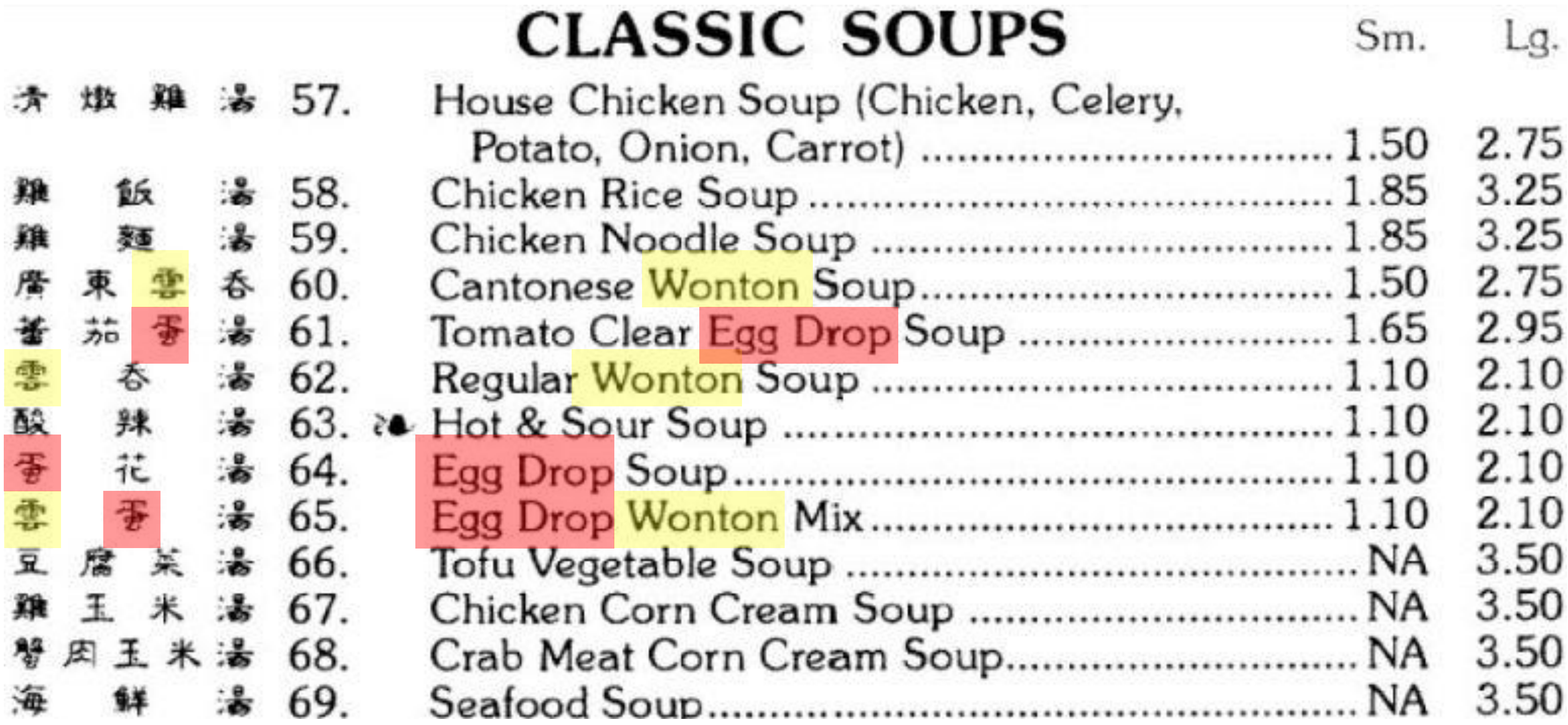
Word Alignment:



CLASSIC SOUPS

			Sm.	Lg.
清 燉 雞	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞 飯	58.	Chicken Rice Soup	1.85	3.25
雞 麵	59.	Chicken Noodle Soup	1.85	3.25
廣 東 雲 吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃 茄 蛋	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲 吞	62.	Regular Wonton Soup	1.10	2.10
酸 辣	63.	Hot & Sour Soup	1.10	2.10
蛋	64.	Egg Drop Soup	1.10	2.10
雲 蛋	65.	Egg Drop Wonton Mix	1.10	2.10
豆 腐 菜	66.	Tofu Vegetable Soup	NA	3.50
雞 玉 米	67.	Chicken Corn Cream Soup	NA	3.50
蟹 肉 玉 米	68.	Crab Meat Corn Cream Soup	NA	3.50
海 鮮	69.	Seafood Soup	NA	3.50

Word Alignment:



CLASSIC SOUPS

			Sm.	Lg.
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄蛋湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63. 🌶	Hot & Sour Soup	1.10	2.10
蛋花湯	64.	Egg Drop Soup	1.10	2.10
雲蛋湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

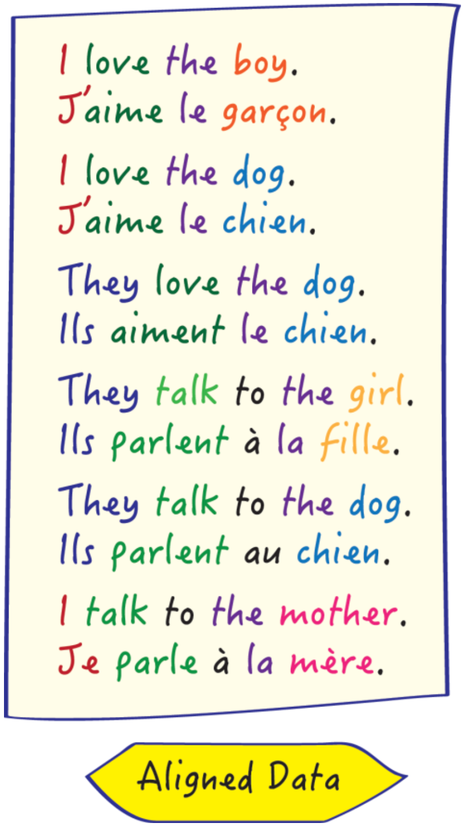
Learning to Translate Words:

- Word alignment mode knows a lot about Chinese soups
- Doesn’t know much else …

- Only knows what it has seen in the training data
- Like people …
- A common theme …

- Given word aligned translation data, can we learn a translation dictionary?
- Yes, really easy …

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

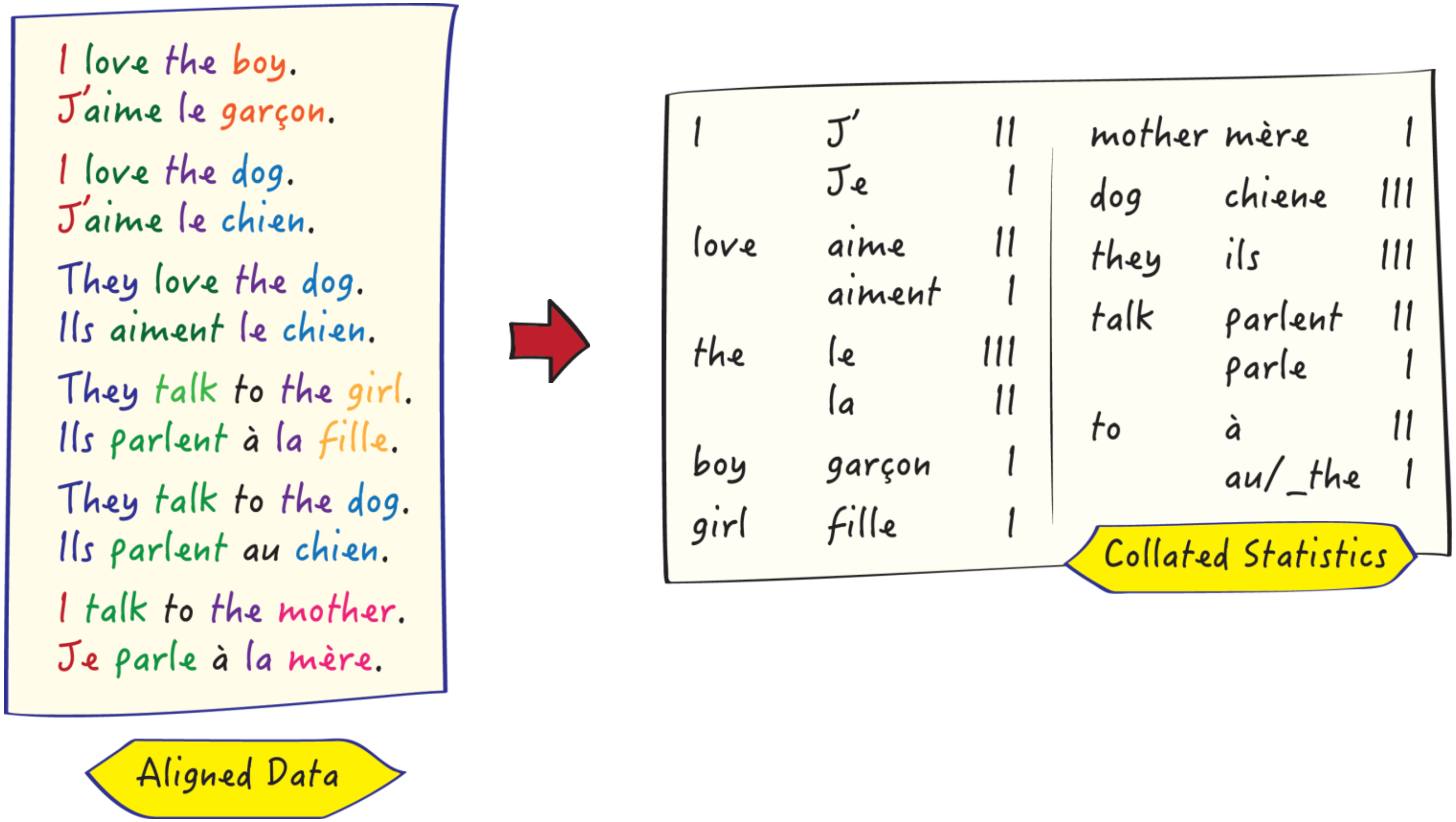
They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

Aligned Data

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

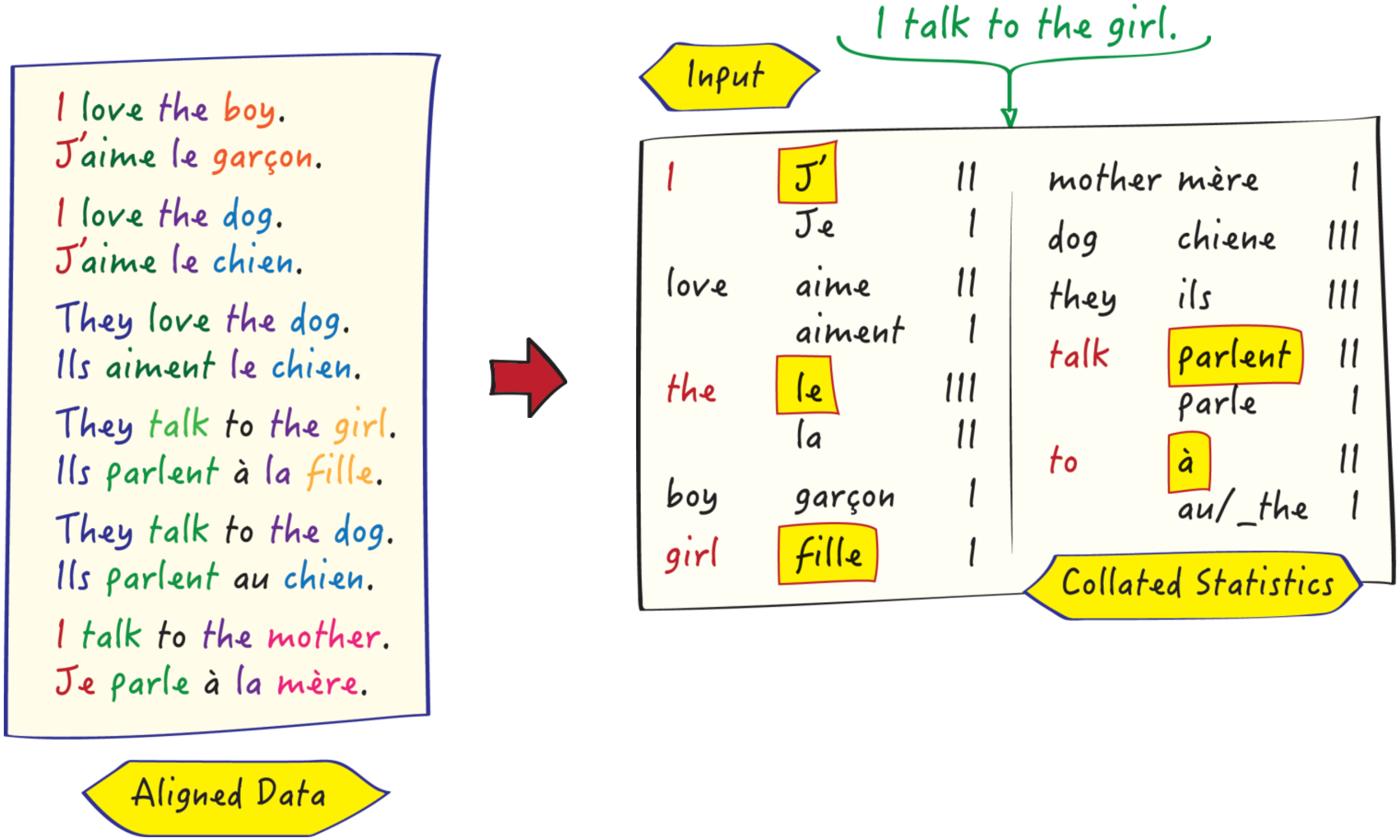
I talk to the mother.
Je parle à la mère.

Aligned Data

I	J'	II
	Je	I
love	aime	II
	aiment	I
the	le	III
	la	II
boy	garçon	I
girl	fille	I
mother	mère	I
dog	chiene	III
they	ils	III
talk	parlent	II
	parle	I
to	à	II
	au/_the	I

Collated Statistics

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

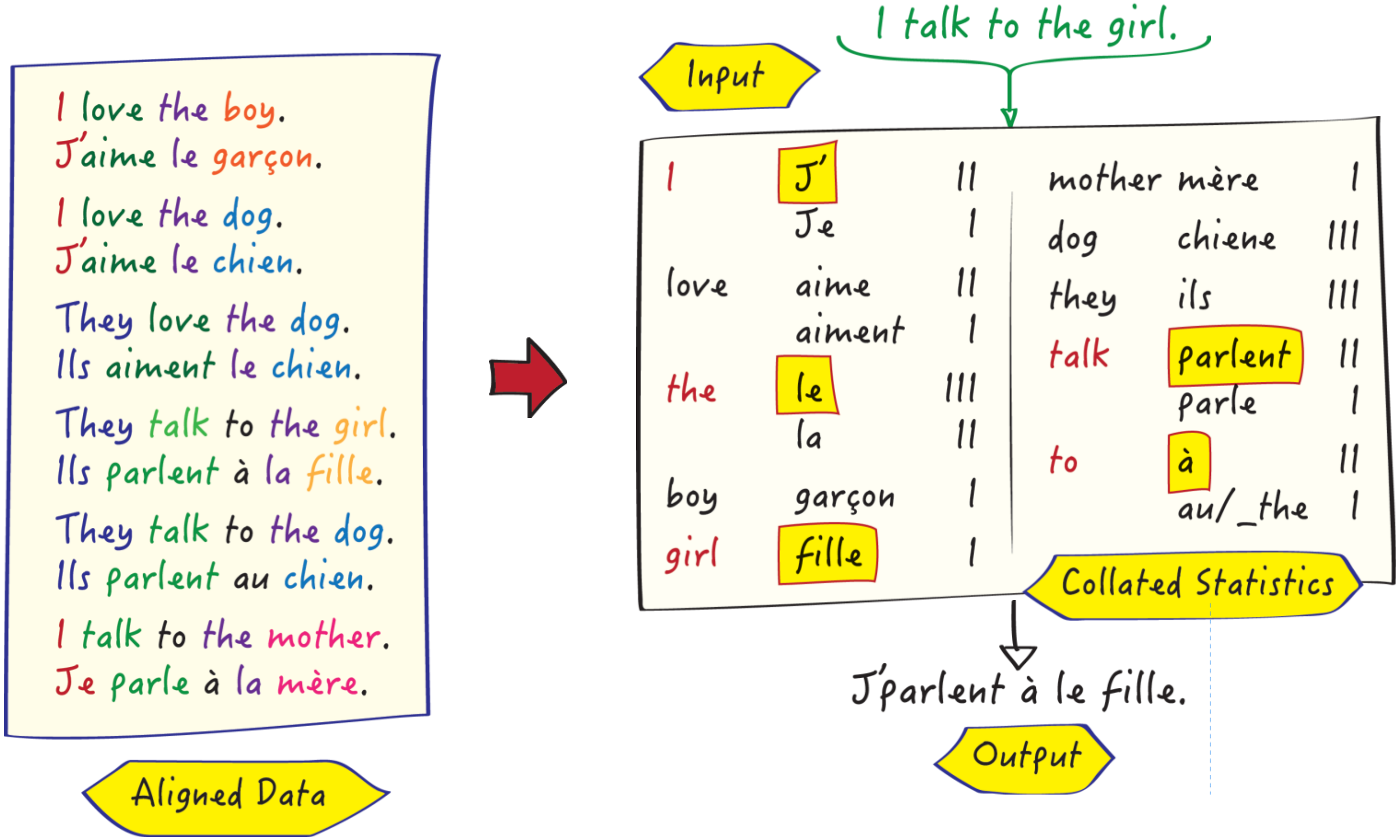
They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

I	J'	II	mother	mère	I
	Je	I	dog	chiene	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	I
	la	II	to	à	II
boy	garçon	I		au/_the	I
girl	fille	I			

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

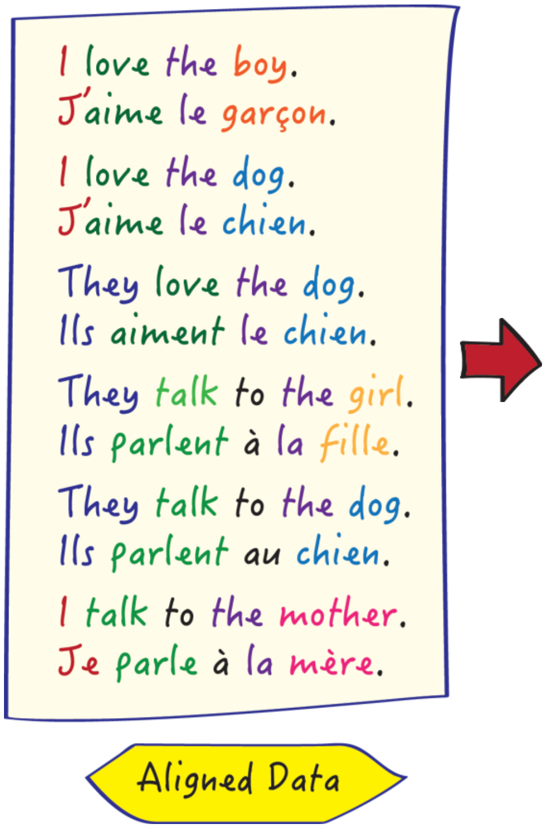
I talk to the mother.
Je parle à la mère.

I talk to the girl.

I	J'	II	mother	mère	I
	Je	I	dog	chiene	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	I
	la	II	to	à	II
boy	garçon	I		au/_the	I
girl	fille	I			

J'parlent à le fille.

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

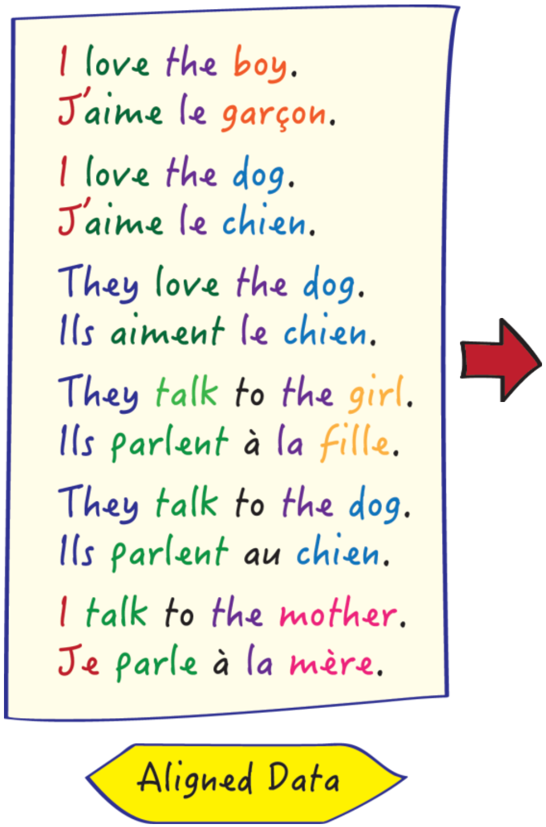
I talk to the mother.
Je parle à la mère.

Aligned Data

I	talk	to	the	girl	
J’	parlent		au	le	fille
2/3	2/3		2/3	3/5	1/1
Je	parle	à	la	fille	
1/3	1/3	1/3	2/5	1/1	

How to choose?

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

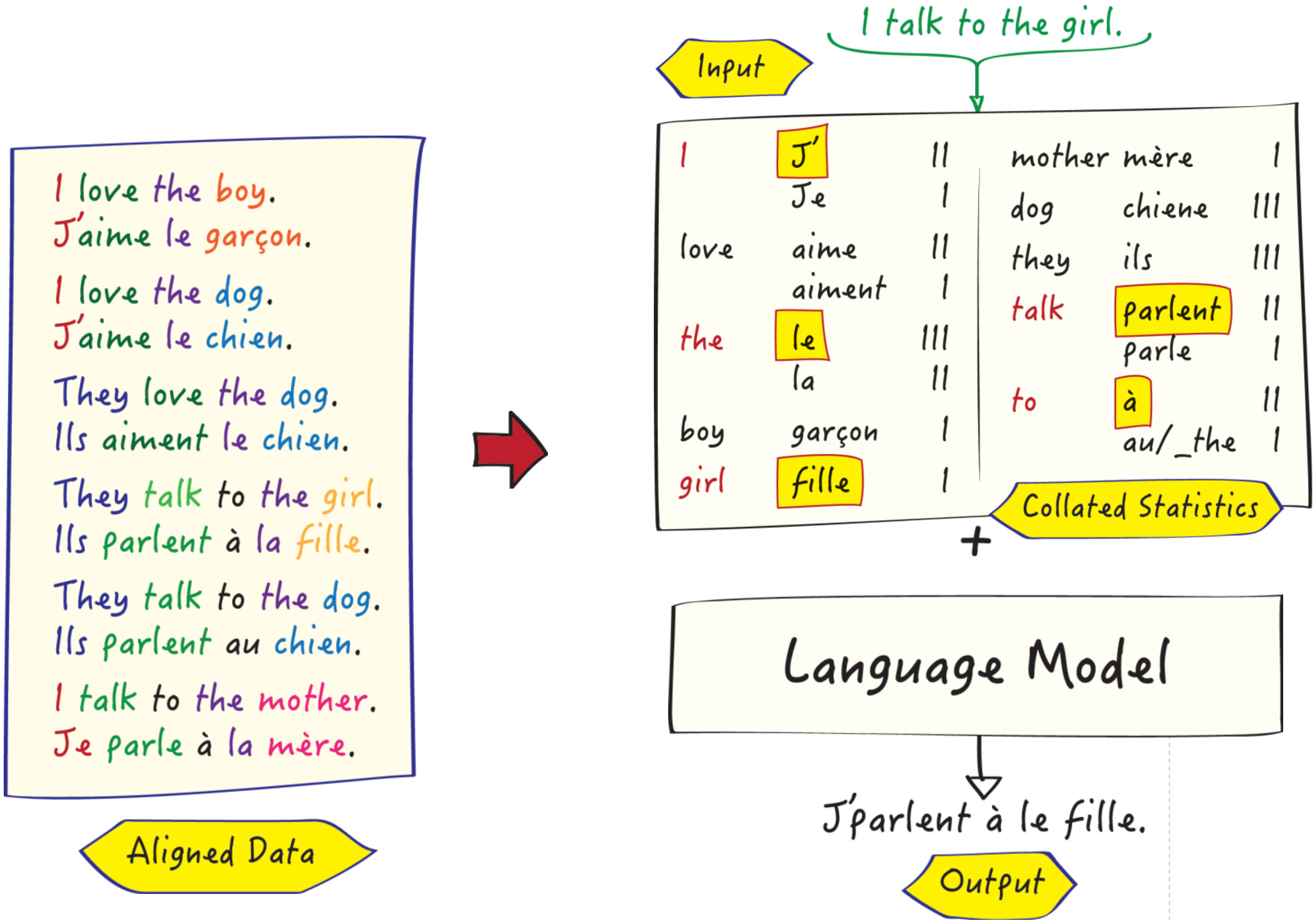
I talk to the mother.
Je parle à la mère.

The Language Model:

- What is good target language?
- Which words can follow which words and which can't … the grammar
- Learnt from the data …
  - *Je parle* is good …
  - *J' parlent* is bad …
  - *la fille* is good …
  - *le fille* is bad …
- *Je parle à la fille >> J' parlent à le fille*

European Language Resource Coordination
Connecting Europe Facility

Statistical Machine Translation



I talk to the girl.

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

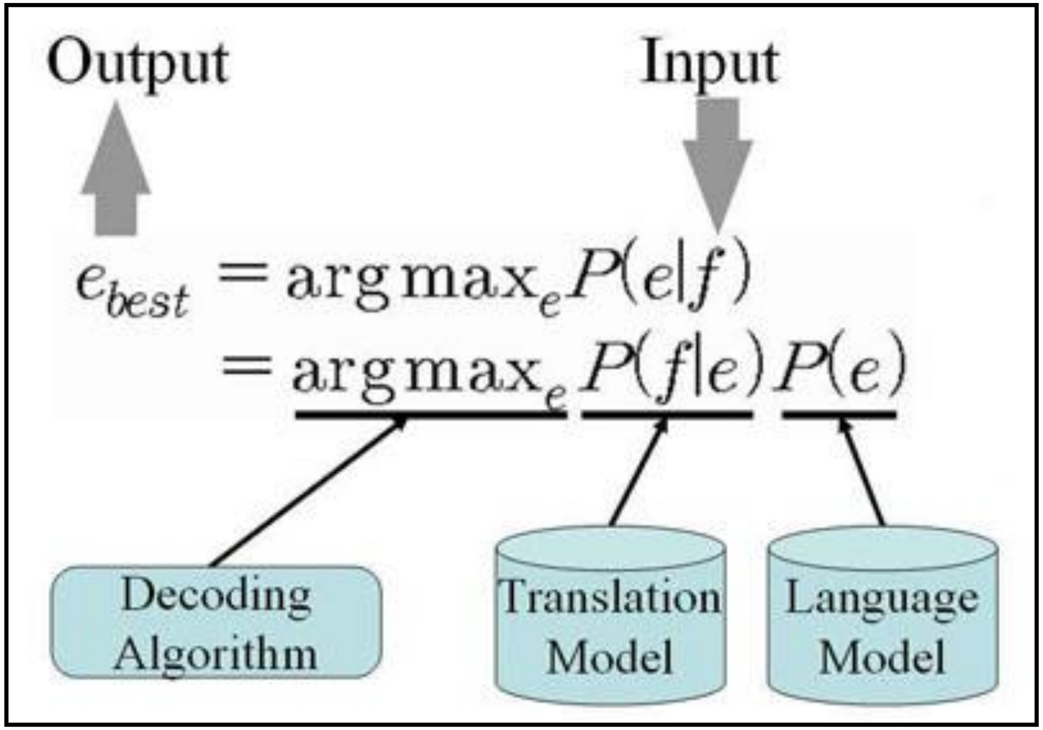
I talk to the mother.
Je parle à la mère.

I	J'	II
	Je	I
love	aime	II
	aiment	I
the	le	III
	la	II
boy	garçon	I
girl	fille	I

mother	mère	I
dog	chiene	III
they	ils	III
talk	parlent	II
	parle	I
to	à	II
	au/_the	I

J'parlent à le fille.

How does Modern MT Work?


Output
Input
$e_{best} = \text{argmax}_e P(e|f)$
$= \underbrace{\text{argmax}_e}\ \underbrace{P(f|e)}\ \underbrace{P(e)}$
Decoding Algorithm
Translation Model
Language Model

- No maths today
- Instead:
- The story of Statistical MT in pictures …
- Its all about **Data** …

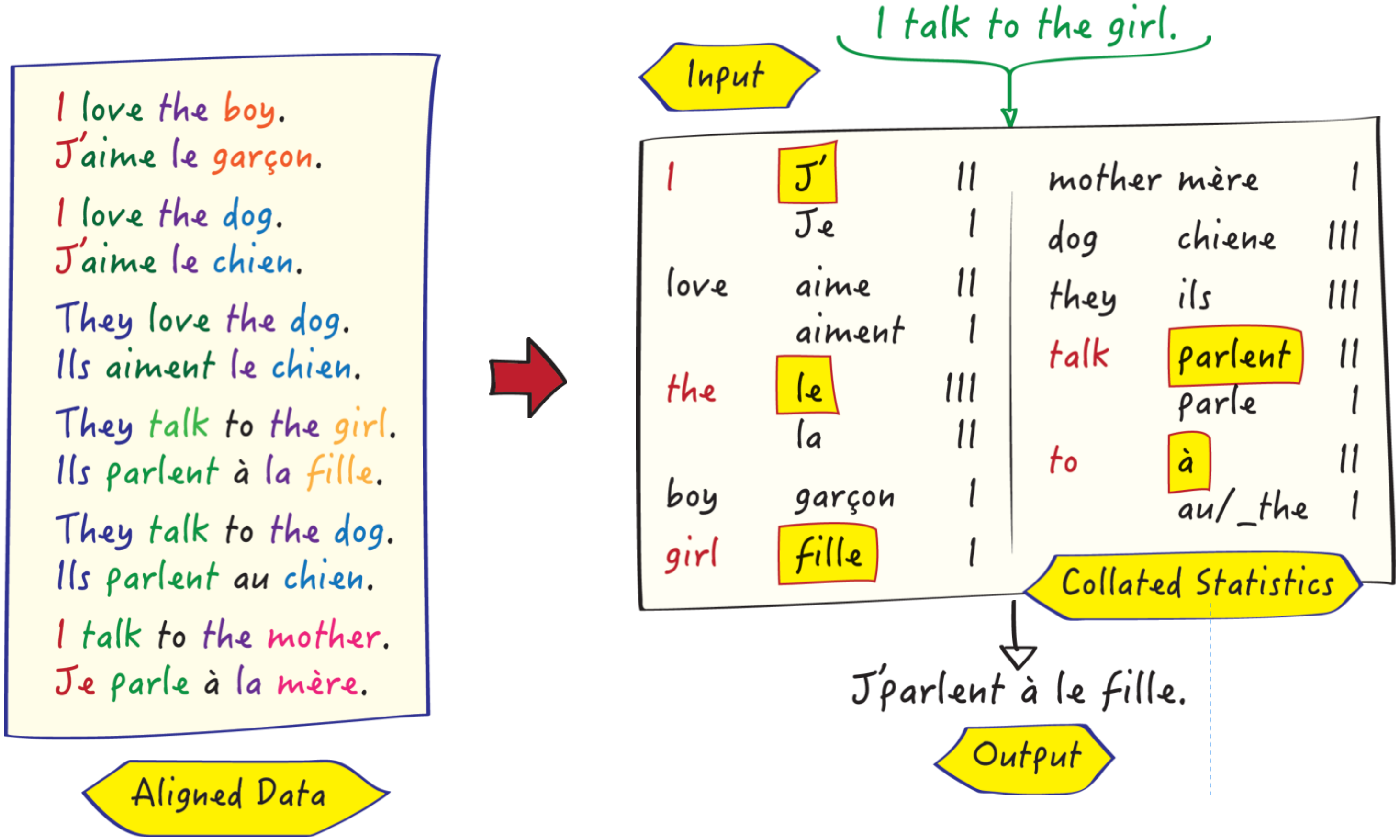
Phrase-Based SMT

- So far: translating single words
- Loses context: such as agreement (*le fille* …) etc.
- To some extent "repaired" by language model

- A better model:
- Not just translations of single words
- But also phrase translations:

– *the girl : la fille*
– *to the girl : a la fille*
– *I talk : Je parle*

Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

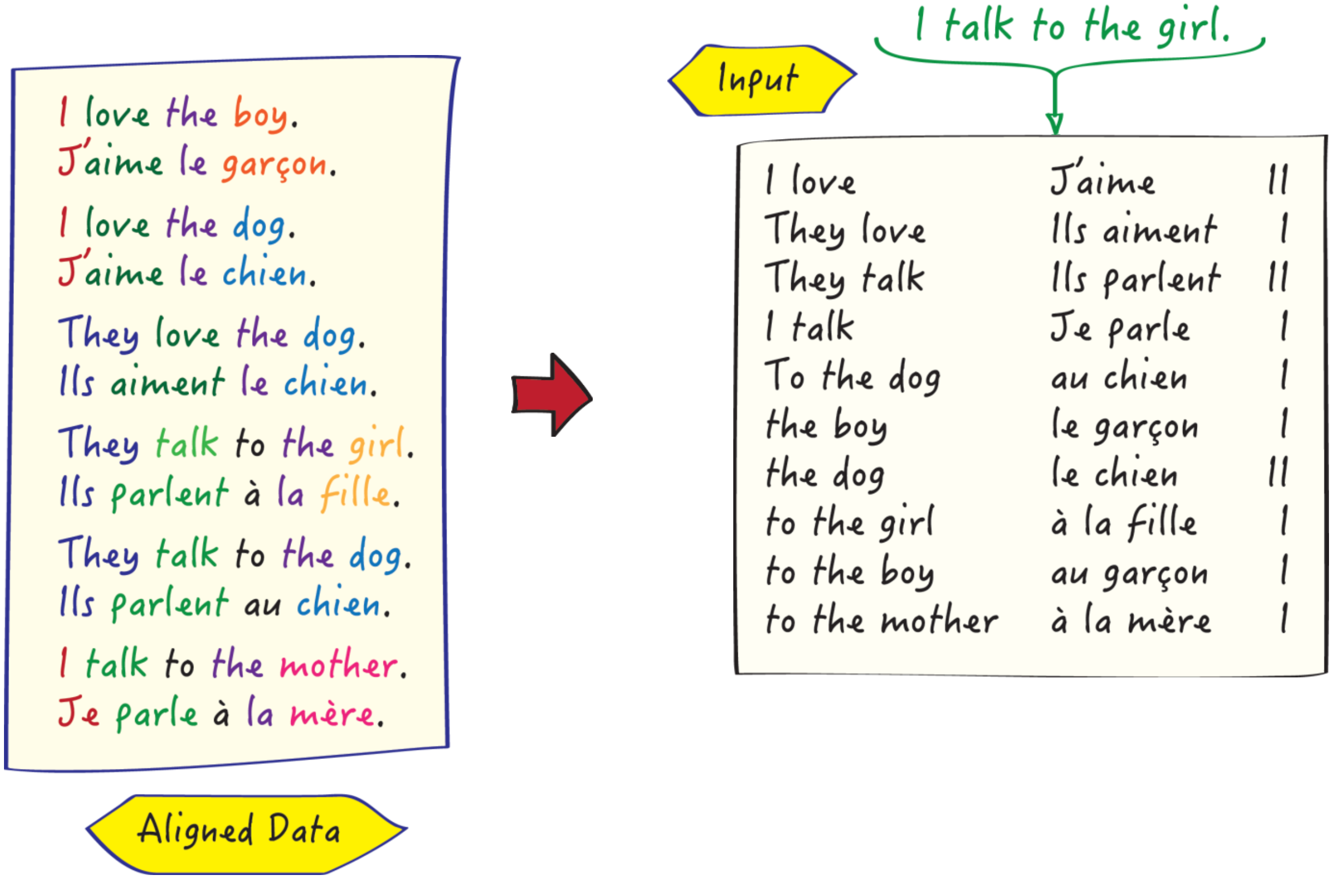
I talk to the mother.
Je parle à la mère.

I talk to the girl.

I	J'	II
	Je	I
love	aime	II
	aiment	I
the	le	III
	la	II
boy	garçon	I
girl	fille	I
mother	mère	I
dog	chiene	III
they	ils	III
talk	parlent	II
	parle	I
to	à	II
	au/_the	I

J'parlent à le fille.

Phrase Based - Statistical Machine Translation



I talk to the girl.

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

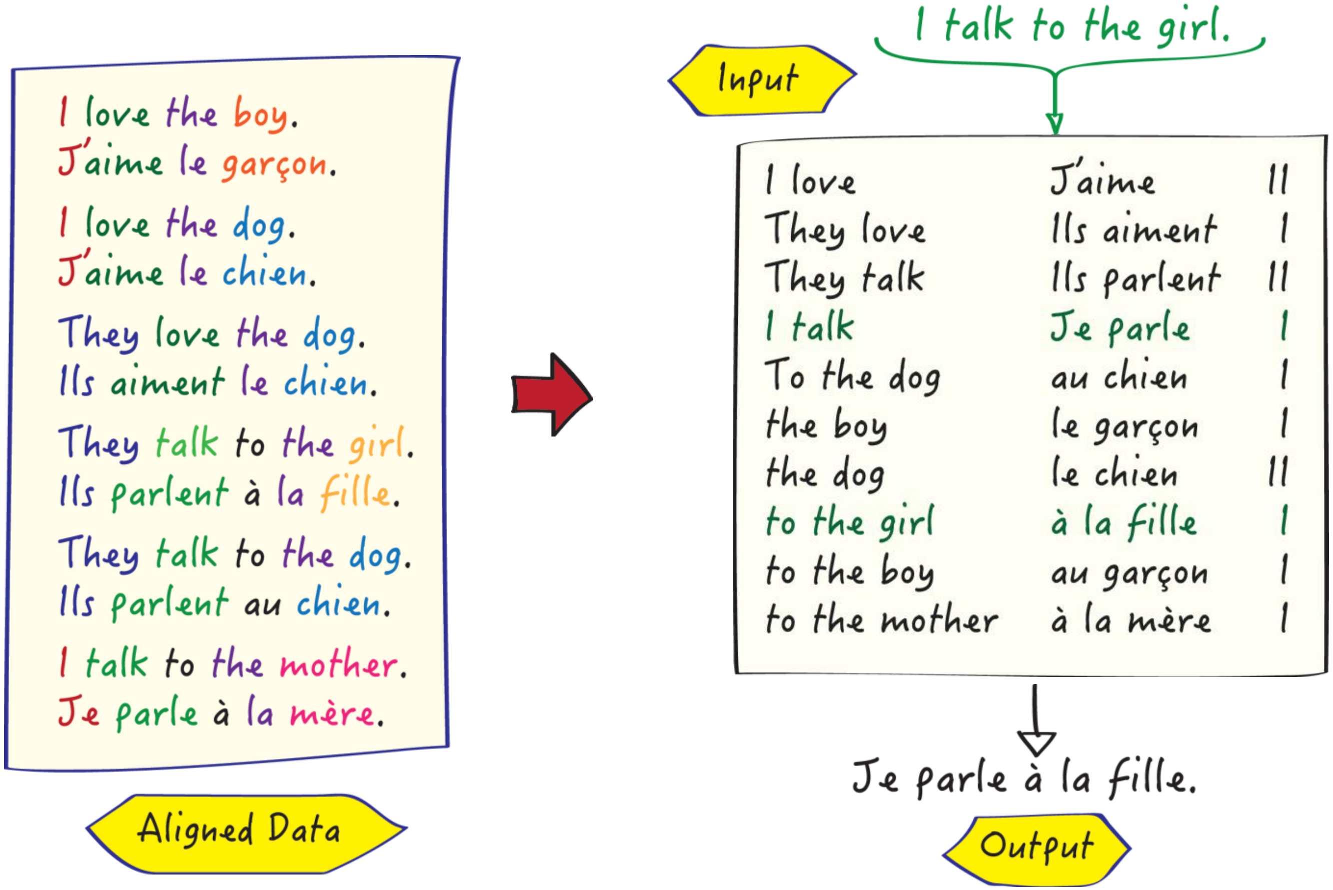
They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

I love	J'aime	II
They love	Ils aiment	I
They talk	Ils parlent	II
I talk	Je parle	I
To the dog	au chien	I
the boy	le garçon	I
the dog	le chien	II
to the girl	à la fille	I
to the boy	au garçon	I
to the mother	à la mère	I

Phrase Based - Statistical Machine Translation



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

I love	J'aime	II
They love	Ils aiment	I
They talk	Ils parlent	II
I talk	Je parle	I
To the dog	au chien	I
the boy	le garçon	I
the dog	le chien	II
to the girl	à la fille	I
to the boy	au garçon	I
to the mother	à la mère	I

Phrase Based - Statistical Machine Translation

- Much better than word-based SMT!
- Standard technology: Google, Microsoft, Baidu, Global Localisation & Translation Industry

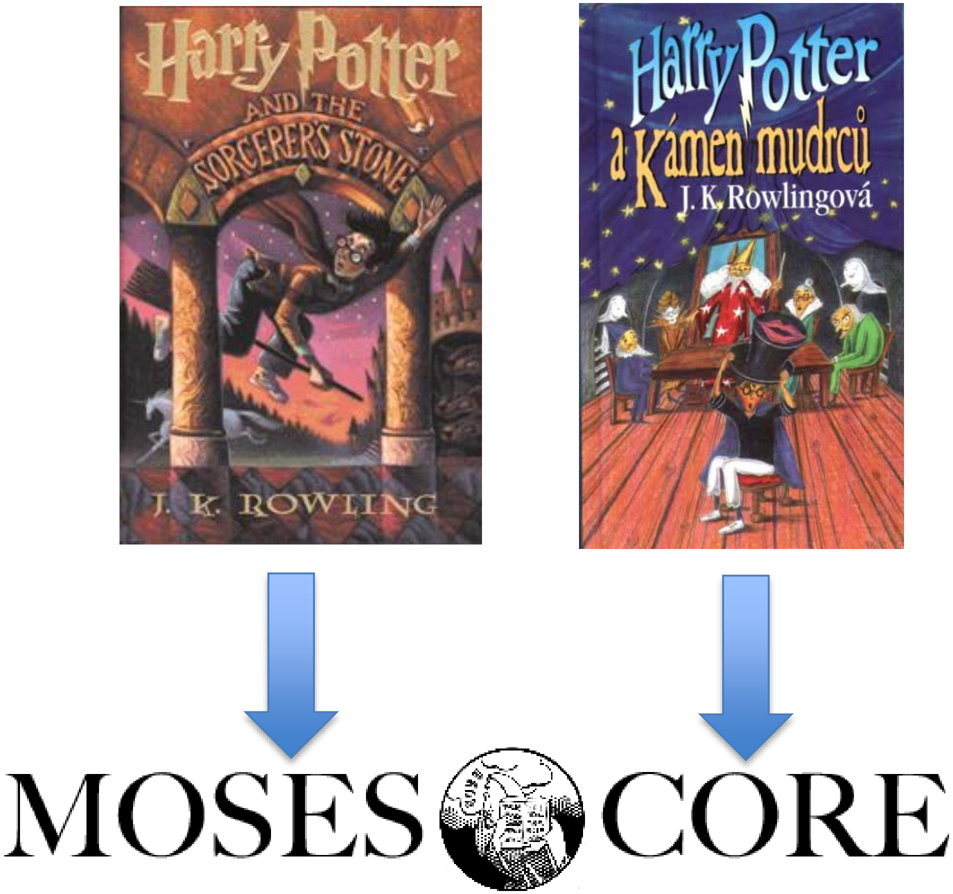
- Moses Open Source PB-SMT
- Most widely used SMT system
- Research funded by EC
- Used by EC DGT’s MT@EC

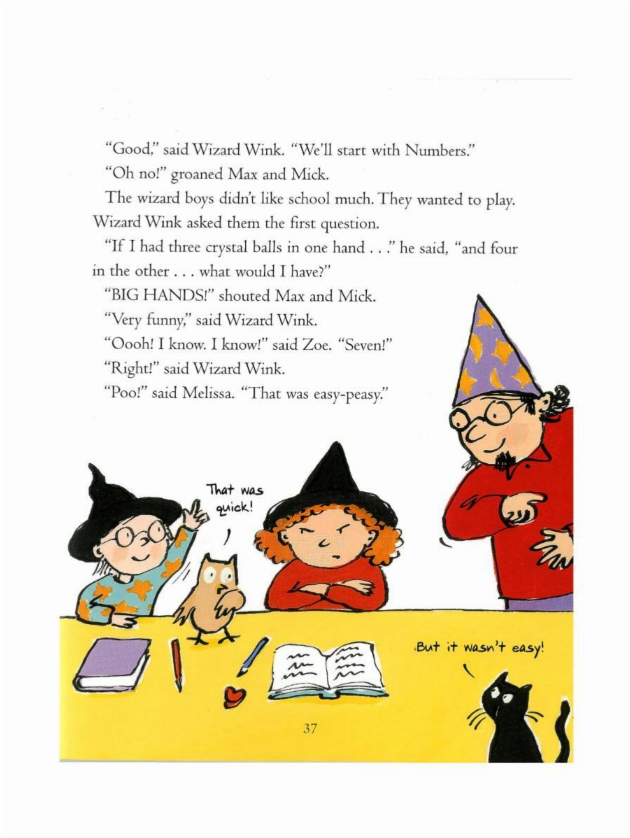
MOSES CORE

Machine Translation and Data

- Statistical Machine Translation is all about data
- SMT learns how to translate from data
- Data
  - translations (bilingual data)
  - Monolingual data (target language text)
  - Dictionaries, terminology, ontologies, named entities
- Like people SMT is good at what it has learned

Machine Translation and Data


Harry Potter AND THE SORCERER'S STONE
J. K. ROWLING
Harry Potter a Kámen mudrců
J. K. Rowlingová
MOSES CORE



"Good," said Wizard Wink. "We'll start with Numbers."

"Oh no!" groaned Max and Mick.

The wizard boys didn't like school much. They wanted to play. Wizard Wink asked them the first question.

"If I had three crystal balls in one hand . . ." he said, "and four in the other . . . what would I have?"

"BIG HANDS!" shouted Max and Mick.

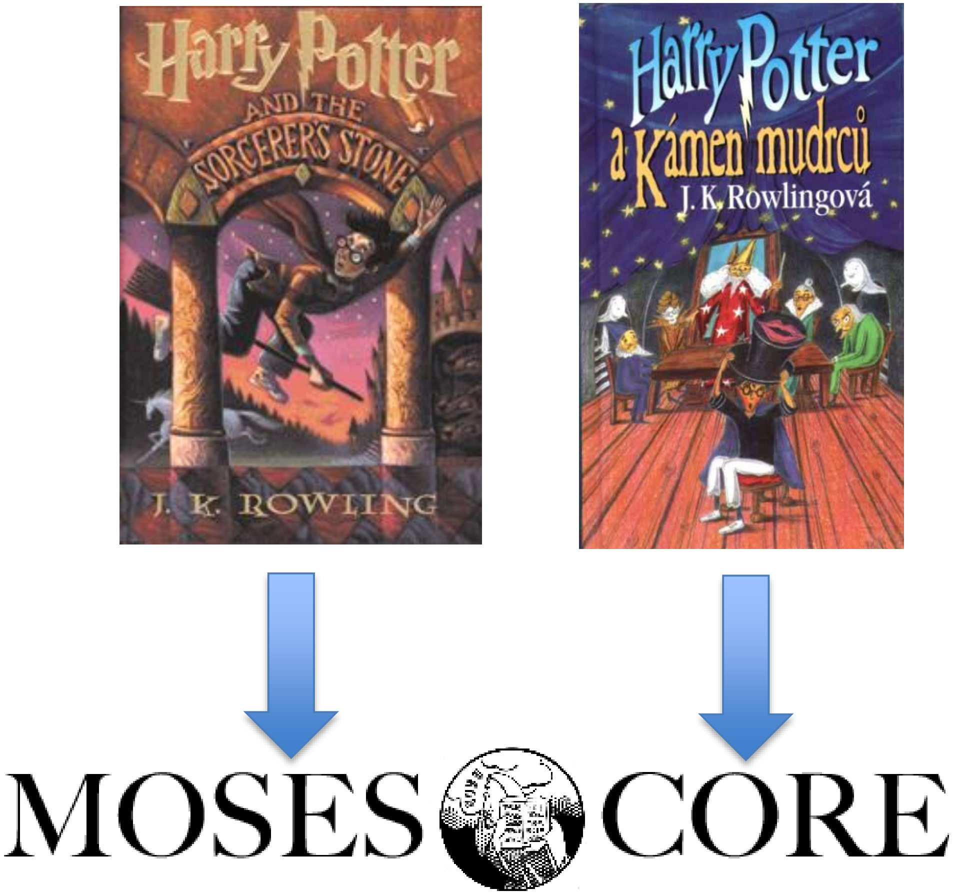
"Very funny," said Wizard Wink.

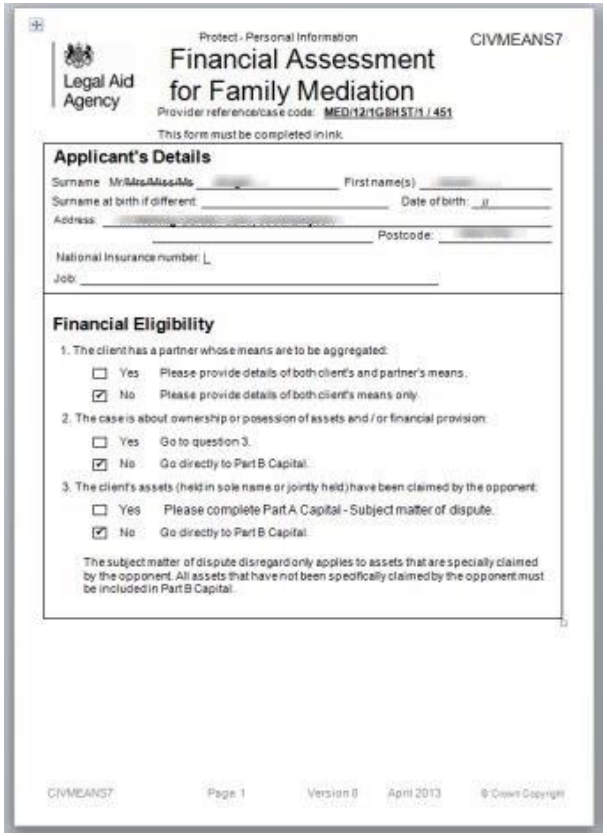
"Oooh! I know. I know!" said Zoe. "Seven!"

"Right!" said Wizard Wink.

"Poo!" said Melissa. "That was easy-peasy."

Machine Translation and Data


Harry Potter AND THE SORCERER'S STONE
J. K. ROWLING
Harry Potter a Kámen mudrců
J. K. Rowlingová
MOSES CORE



Protect - Personal Information

CIVMEANS7

Legal Aid Agency

Financial Assessment for Family Mediation

Provider reference/case code: **MED/12/1GBHST/1 / 451**

This form must be completed in ink.

Applicant's Details

Surname Mr/~~Mrs/Miss/Ms~~ [redacted] First name(s) [redacted]

Surname at birth if different: ______ Date of birth: __//__

Address: [redacted]

Postcode: [redacted]

National Insurance number: L

Job: ______

Financial Eligibility

1. The client has a partner whose means are to be aggregated:
   - □ Yes Please provide details of both client's and partner's means.
   - ☑ No Please provide details of both client's means only.
2. The case is about ownership or posession of assets and / or financial provision:
   - □ Yes Go to question 3.
   - ☑ No Go directly to Part B Capital.
3. The client's assets (held in sole name or jointly held) have been claimed by the opponent:
   - □ Yes Please complete Part A Capital - Subject matter of dispute.
   - ☑ No Go directly to Part B Capital.

The subject matter of dispute disregard only applies to assets that are specially claimed by the opponent. All assets that have not been specifically claimed by the opponent must be included in Part B Capital.

CIVMEANS7 Page 1 Version 8 April 2013 © Crown Copyright

CEF.AT and Data

- CEF.AT needs the right kind of data
- National governments, public administration, public services, NGOs
- CEF provide services for multilingual engagement with national citizens, EU citizens and other customers of public administration

ELRC

- Help us make CEF.AT a success
  - Services for Europe’s citizens
  - Services for you
  - Support multi-linguality
- Help us find the right kind of data
- Supporting our language is supporting Europe and vice versa