

Presentation Outline for ELRC Workshop, Section 7a:

“Automated Translation: How Does It Work?”

Author: Josef van Genabith (DFKI)

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



- Europe = Multilinguality
- 24 official languages,
- 24x23 language pairs
- So much to translate!
- Translation costs!?
- Can MT help?
- What about the Quality?

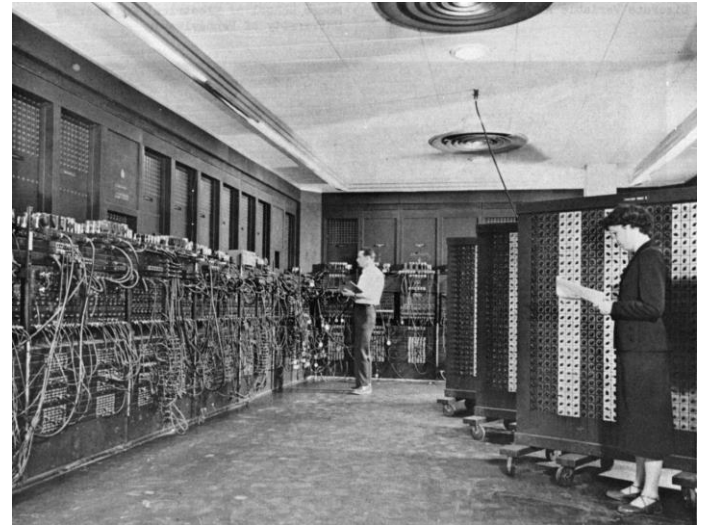


Image: <https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg>
License: public domain

- 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)



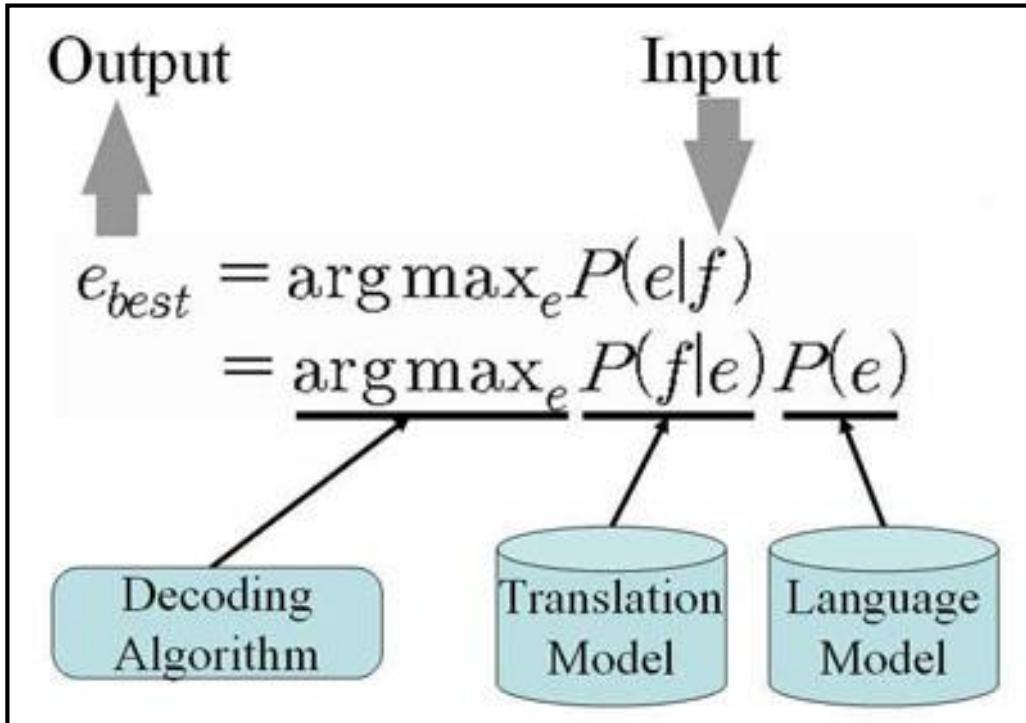
Image: <http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg>
License: CC BY-NC-SA 3.0



- 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** \Rightarrow data is all important
 - Approximate solution \Rightarrow not perfect, needs help
 - human professional translators
 - Post-editing
 - Automated Translation \neq Automatic



How does Modern MT Work?



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



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

Einleitung

I. Von dem Unterschiede der reinen und empirischen Erkenntnis

Daß alle unsere Erkenntnis mit der Erfahrung anfangt, 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 Verstandstätigkeit 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.

ENGLISH

Introduction

I. Of the difference between Pure and Empirical Knowledge

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 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.

FRENCH

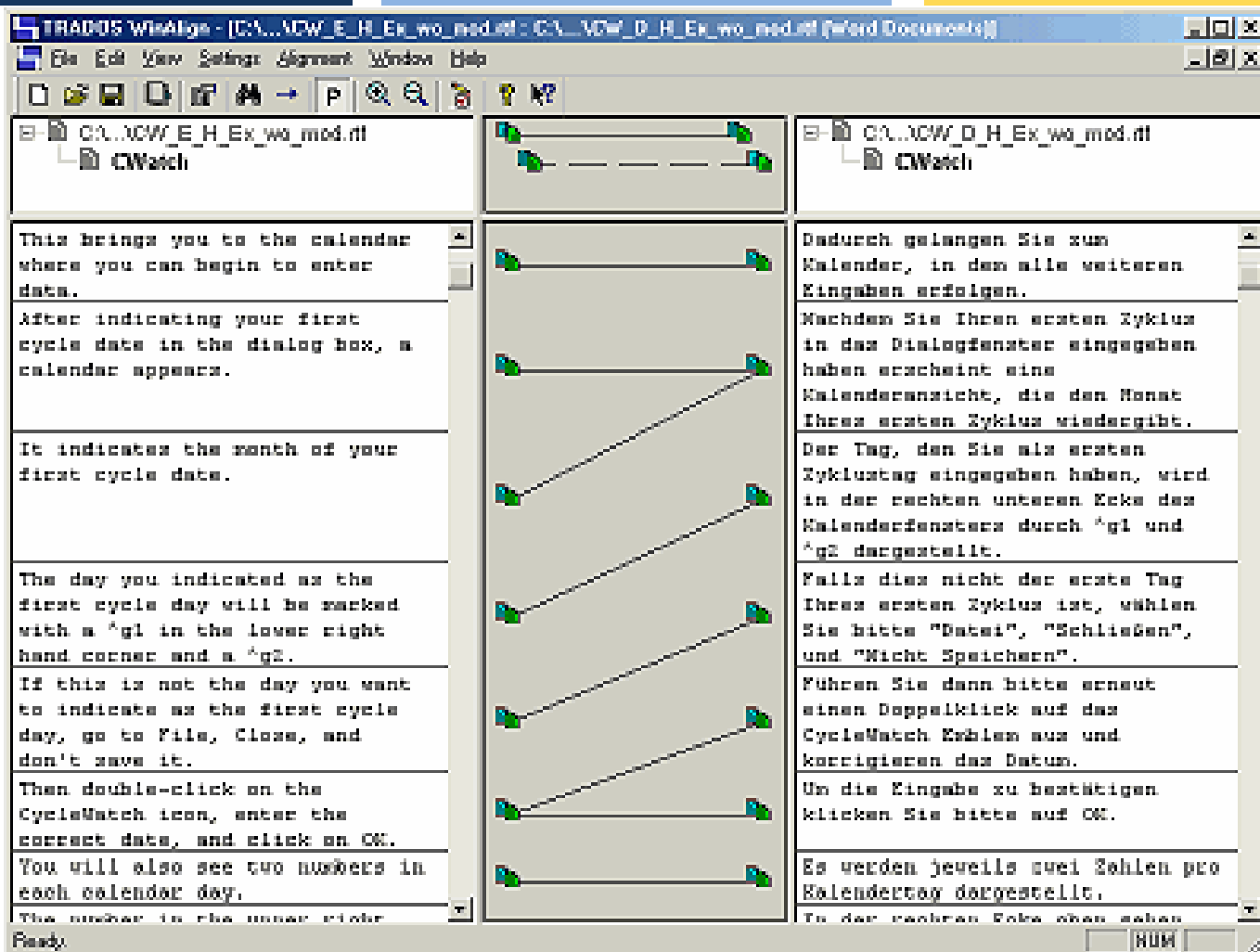
Introduction

I. De la différence de la connaissance pure et de la connaissance empirique.

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.

- 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
<p>Einleitung</p> <p><i>I. Von dem Unterschiede der reinen und empirischen Erkenntnis</i></p> <p>Daß alle unsere Erkenntnis mit der Erfahrung anfangt, 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 Verstandstätigkeit 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.</p>	<p>Introduction</p> <p><i>I. Of the difference between Pure and Empirical Knowledge</i></p> <p>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 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.</p>	<p>Introduction</p> <p><i>I. De la différence de la connaissance pure et de la connaissance empirique.</i></p> <p>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.</p>



The screenshot shows the TRADOS WinAlign application window. The title bar reads "TRADOS WinAlign - [C:\...NOW_E_H_Ex_wo_mod.tif : C:\...NOW_D_H_Ex_wo_mod.tif (Inford Documents)]". The menu bar includes "File", "Edit", "View", "Settings", "Segment", "Window", and "Help". The toolbar contains various icons for file operations and editing. The main workspace is divided into three panes:

- Left Pane:** Shows the file path "C:\...NOW_E_H_Ex_wo_mod.tif" and a sub-folder "CWatch". Below this, there is a list of English text segments with scrollbars.
- Center Pane:** Displays a visual alignment interface. It shows two horizontal lines representing the source and target sentences. Small green and blue icons are placed along these lines, connected by arrows, indicating the alignment of words between the two languages.
- Right Pane:** Shows the file path "C:\...NOW_D_H_Ex_wo_mod.tif" and the sub-folder "CWatch". Below this, there is a list of German text segments with scrollbars.

The status bar at the bottom of the window shows the text "Ready." and a "NUM" button.



		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	
雲吞	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	68. Crab Meat Corn Cream Soup.....	NA	3.50
海	鮮		69. Seafood Soup.....	NA	3.50



		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	61.	Tomato Clear Egg Drop Soup	1.65	2.95		
雲吞	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		
雲吞	吞	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		



- 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 ...

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 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



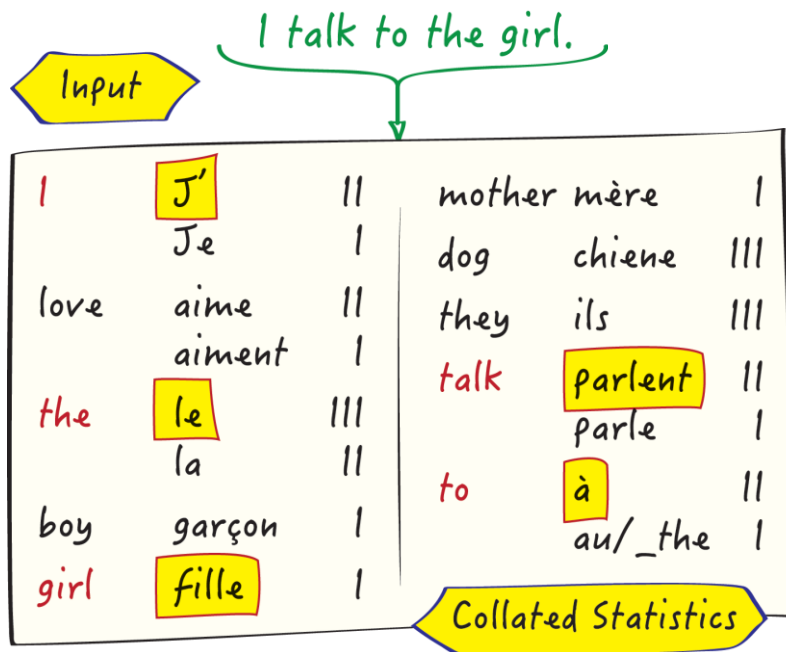
I	J'		mother	mère	
	Je		dog	chiene	
love	aime		they	ils	
	aiment		talk	parlent	
the	le			parle	
	la		to	à	
boy	garçon			au/_the	
girl	fille				

Collated Statistics



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.

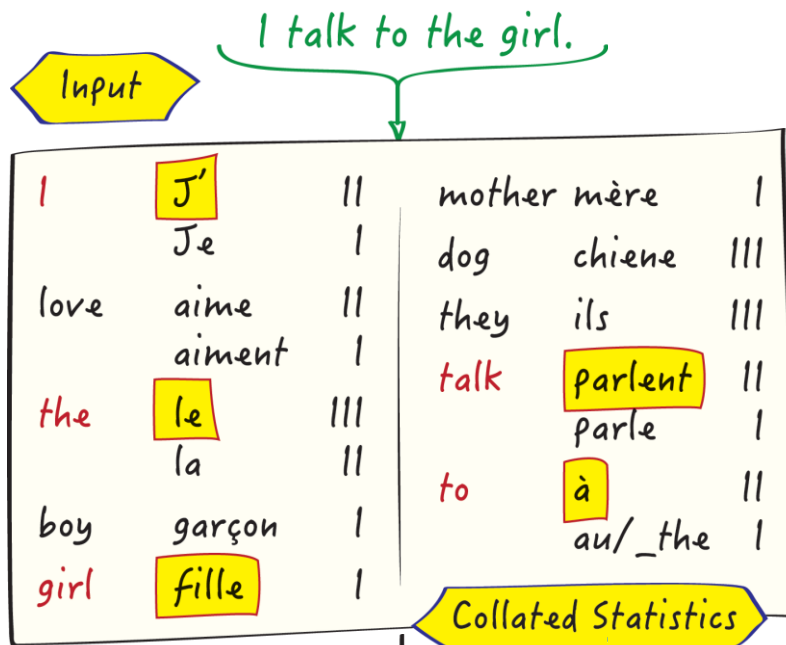
Aligned Data





I 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



J'parlent à le fille.

Output



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?



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

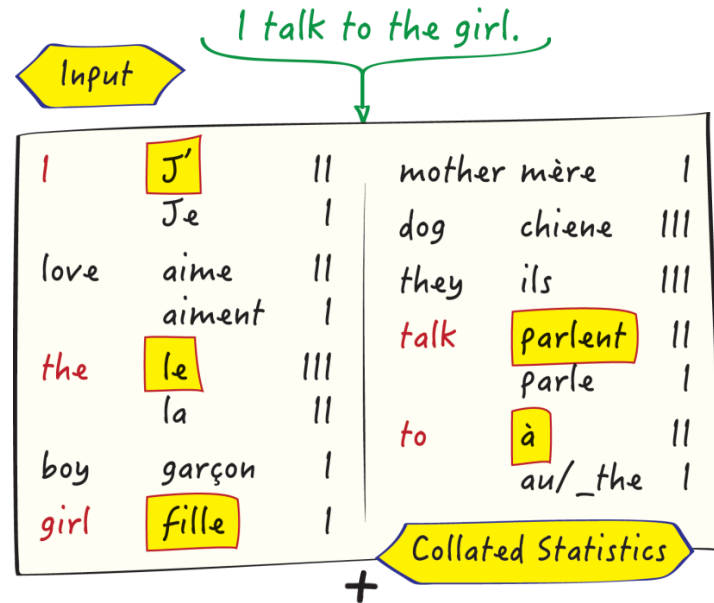
The Language Model:

- 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*



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

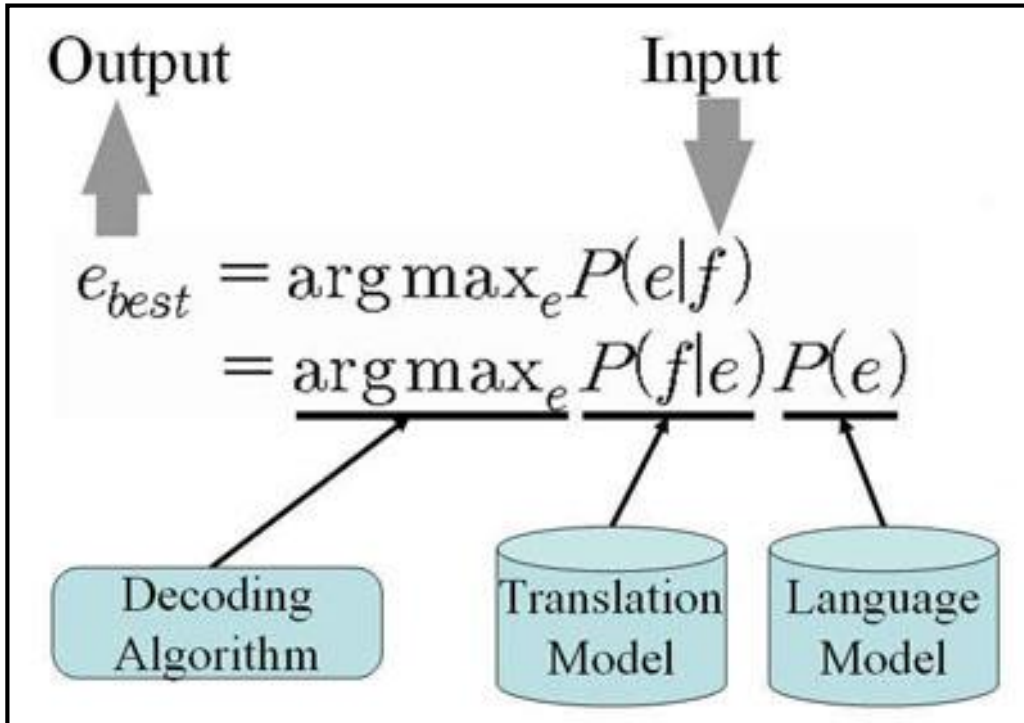


Language Model

J'parlent à le fille.

Output

How does Modern MT Work?



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



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

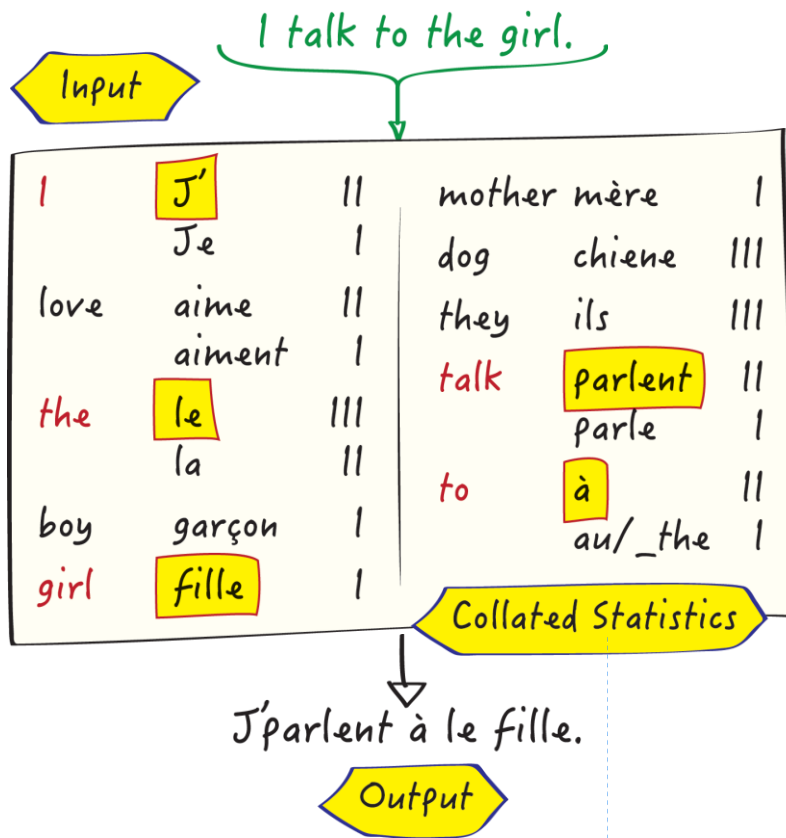
- A better model:
- Not just translations of single words
- But also phrase translations:

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



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 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

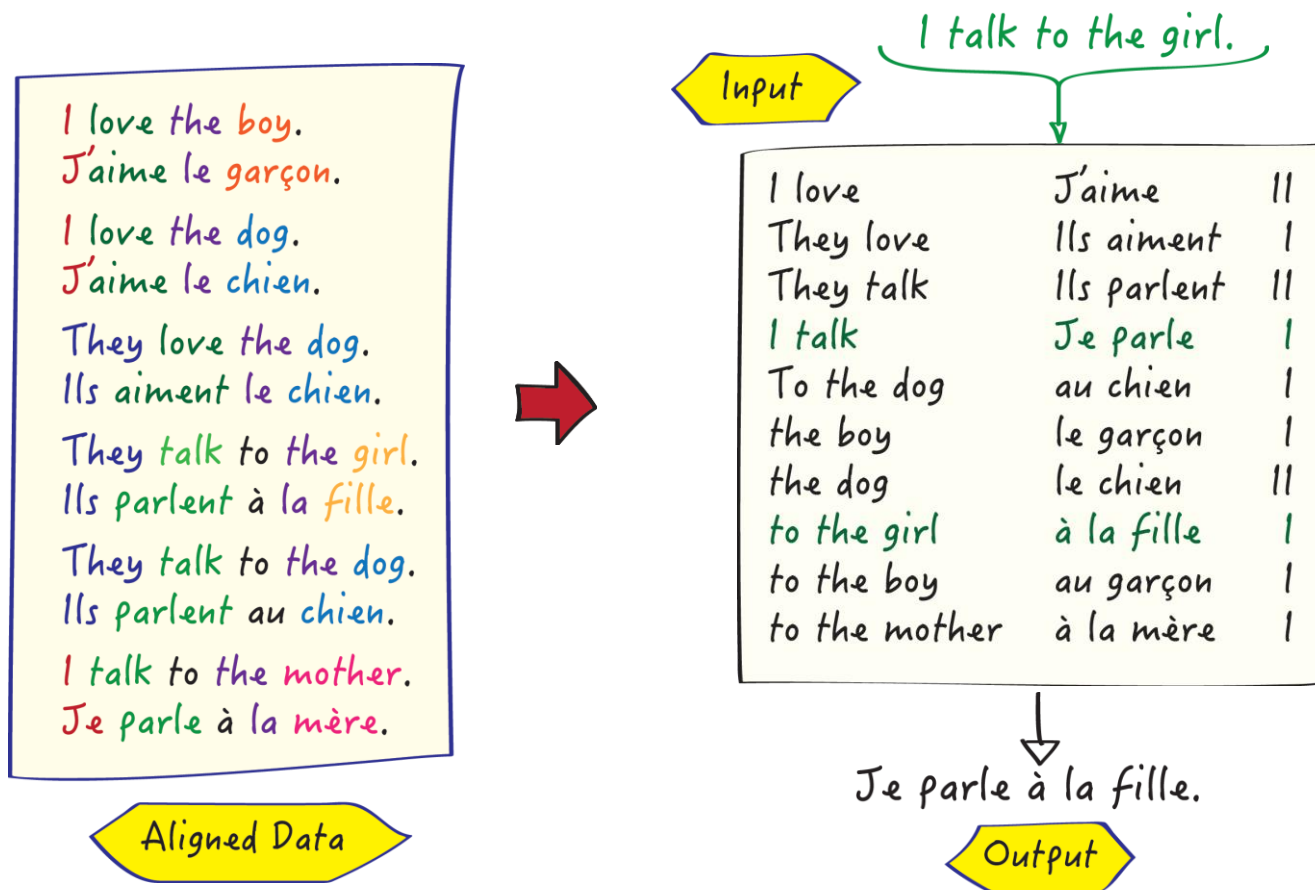


Input

I talk to the girl.



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



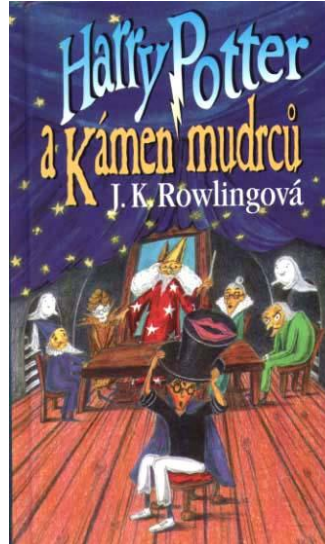
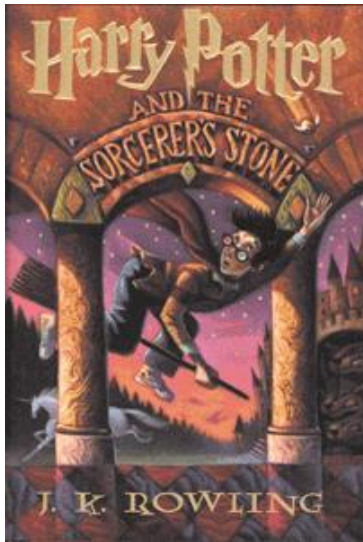


- Much better than word-based SMT!
- Standard technology: Google, Microsoft, Baidu, Global Localisation & Translation Industry
- Moses Open Source PB-SMT
- Most widely used SMT system
- Research funded by EC
- Used by EC DGT's MT@EC

MOSES  CORE

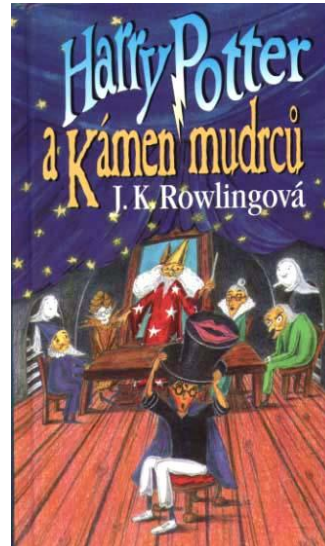
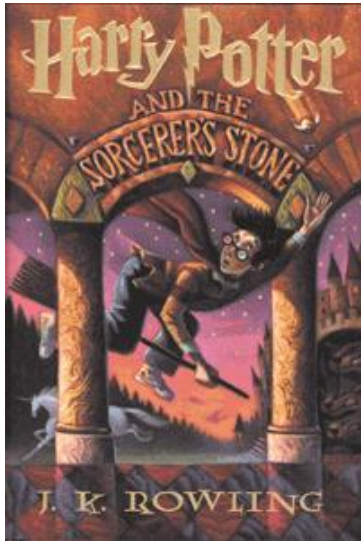


- Statistical Machine Translation is all about data
- SMT learns how to translate from data
- Data
 - translations (bilingual data)
 - Monolingual data (target language text)
 - Dictionaries, terminology, ontologies, named entities
- Like people SMT is good at what it has learned



MOSES  CORE





MOSES  CORE

Protect - Personal Information CIVMEANS7

Legal Aid Agency
Financial Assessment for Family Mediation
Provider reference/case code: MED12/1GBHST/1/451

This form must be completed in ink.

Applicant's Details

Surname: Mr/Ms/Miss/Ms _____ First name(s) _____
Surname at birth if different: _____ Date of birth: JJ _____
Address: _____ Postcode: _____
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 possession 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 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

- 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