European Language Resource Coordination
Connecting Europe Facility

“Wie funktioniert automatisierte Übersetzung?”

Prof. Josef van Genabith

(Deutsches Forschungszentrum für Künstliche Intelligenz)

Maschinelle Übersetzung (MÜ)

Überblick:

- Warum MÜ: Datenmenge, Qualität und Kosten?
- Was maschinelle Übersetzung so schwierig?
- MÜ + Professionelle Übersetzer = Qualität
- Wie funktioniert die moderne statistisch-basierte MÜ?
- Es geht vor allem um Daten!
- Und um die richtige Art von Daten!

European Language Resource Coordination
Connecting Europe Facility

Maschinelle Übersetzung: Qualität und Kosten?

- Europa = Mehrsprachigkeit
- 24 Amtssprachen
  24+2 CEF Sprachen
- Vieles zu übersetzen!
- Übersetzungskosten!?
- Kann MÜ Hilfe leisten?
- Und in welcher Qualität?

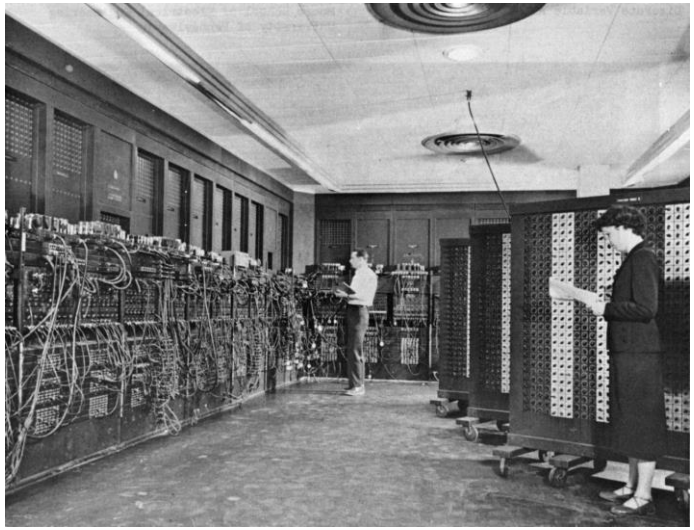


Image: https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg
License: public domain

Wo liegen die Schwierigkeiten bei MÜ?

- Natürliche Sprachen sind:
  - Elegant
  - Effizient
  - Flexibel
  - Komplex
- Ein Wort/Satz kann verschiedenes bedeuten
- Mehrere Möglichkeiten, das Gleiche zu sagen
- Bedeutung hängt von Kontext ab
- Übertragener Sinn (Metapher)
- Sprache und Kultur (unterschiedliche Konzeptualisierungen des gleichen Sachverhalts)
- Wortstellung
- Morphologie u.v.m.

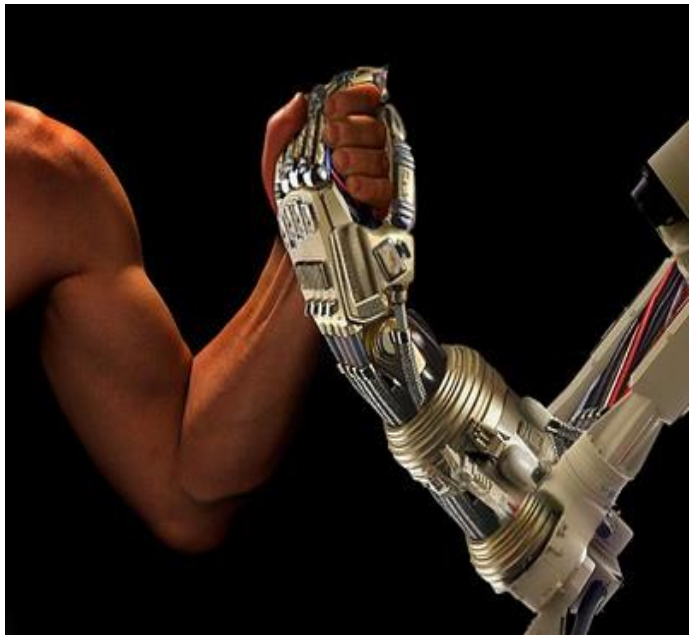
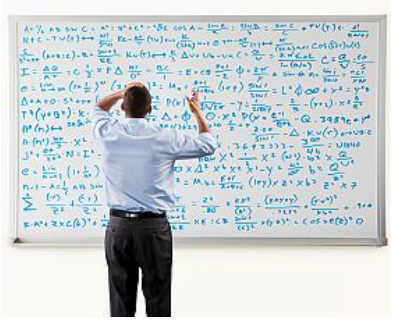


Image: http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg

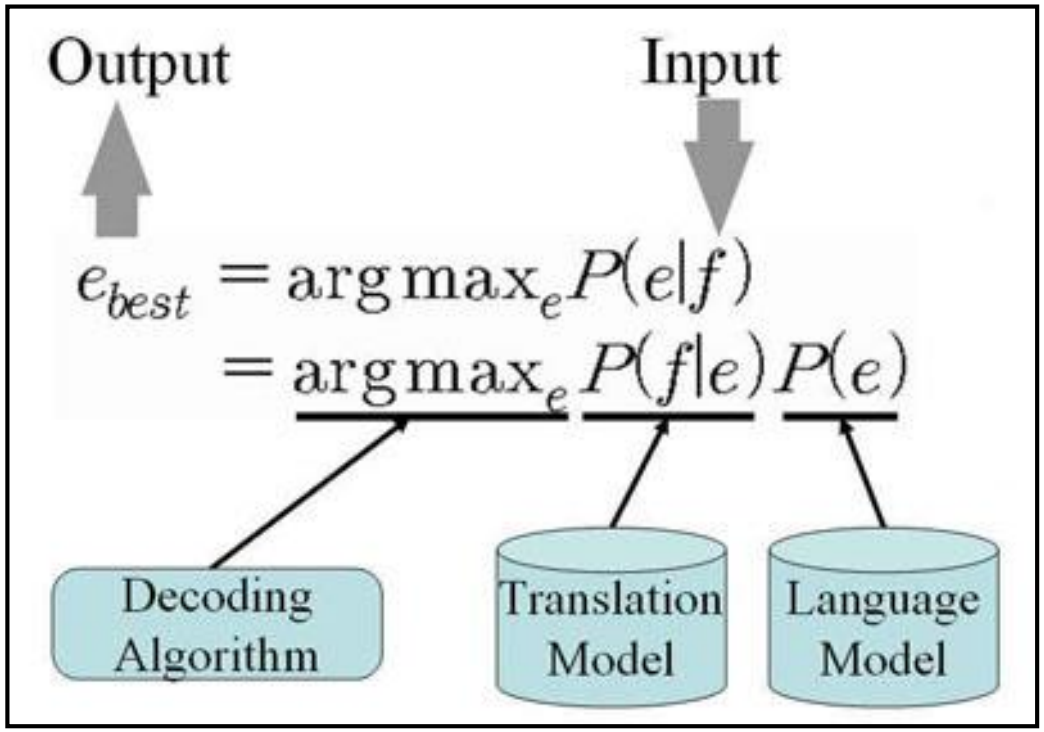
European Language Resource Coordination
Connecting Europe Facility

Hohe Komplexität von Sprache und Übersetzung

- Sprache und Übersetzung sind komplex
- Wir können sie nicht genau berechnen
- Wir haben bereits regel-basierte MÜ und Sprachtechnologie erforscht und eingesetzt
- Was nun?
- Maschinelles Lernen
  - Aus Daten lernen ⇒ zentrale Rolle von Daten
  - Grobe Lösung ⇒ nicht perfekt, Hilfe
    - Professionelle Übersetzer
    - Nach-editieren
  - Automatisierte Übersetzung ≠ automatisch



Wie funktioniert die moderne MÜ?



- Heute kein Mathematikunterricht!
- Sondern:
- Die Geschichte der statistischen MÜ in Bildern …
- Es dreht sich einzig und allein um **Daten** …

Wie funktioniert die moderne MÜ?

Die statistische MÜ lernt aus zwei Typen von Daten:

- Übersetzungen von Menschen
- Text in der Zielsprache

- So viele Daten wie möglich!
- Aber: die richtige Art von Daten!

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.

European Language Resource Coordination
Connecting Europe Facility

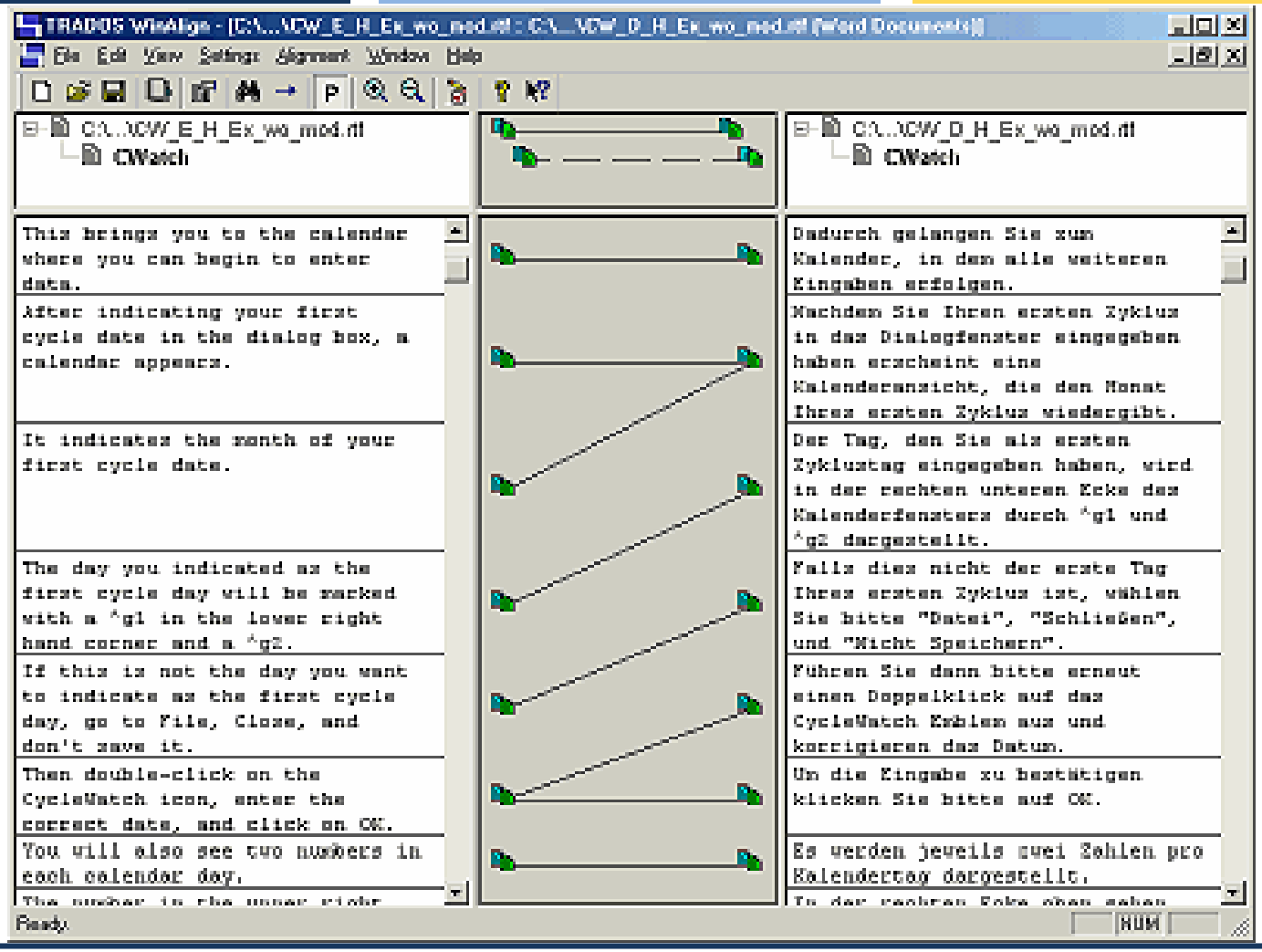
Was kann aus den Daten gelernt werden?

- Welche Sätze wurden wie übersetzt: Satz-Alignierung
- Welche Wörter wurden wie übersetzt: Wort-Alignierung + Übersetzungswahscheinlichkeiten
- Wie sieht eine gute Zielsprache aus: Sprachmodell

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.

European Language Resource Coordination
Connecting Europe Facility

Satz-Alignierung



Wort-Alignierung

CLASSIC SOUPS

		Sm.	Lg.
57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
58.	Chicken Rice Soup	1.85	3.25
59.	Chicken Noodle Soup	1.85	3.25
60.	Cantonese Wonton Soup	1.50	2.75
61.	Tomato Clear Egg Drop Soup	1.65	2.95
62.	Regular Wonton Soup	1.10	2.10
63.	Hot & Sour Soup	1.10	2.10
64.	Egg Drop Soup	1.10	2.10
65.	Egg Drop Wonton Mix	1.10	2.10
66.	Tofu Vegetable Soup	NA	3.50
67.	Chicken Corn Cream Soup	NA	3.50
68.	Crab Meat Corn Cream Soup	NA	3.50
69.	Seafood Soup	NA	3.50

Wort-Alignierung

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
蕃茄雲湯	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
雲花湯	64. Egg Drop Soup	1.10	2.10
雲雲湯	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

Das Lernen von wortbasierten Übersetzungen

- Im Wort-Alignierung Modus vieles ist über chinesische Suppe bekannt
- Aber sonst recht wenig …

- Es wird nur das erkannt, was in den Trainingsdaten gesehen wurde.
- Eigentlich, wie bei Menschen …
- Ein gemeinsames Thema denn …

- Kann auf der Basis von aus wort-alignierten Übersetzungen bestehenden Daten ein Übersetzungslexikon gelernt werden?
- Ja, und zwar recht einfach …

European Language Resource Coordination
Connecting Europe Facility

Statistische Maschinelle Übersetzung

I love the woman.
Ich liebe die Frau.

The man loves the cat.
Der Mann liebt die Katze.

The man loves the woman.
Der Mann liebt die Frau.

I love the man.
Ich liebe den Mann.

They talk to the cat.
Sie sprechen zu der Katze.

They talk to the boy.
Sie sprechen zu dem Jungen.

They talk to the man.
Sie sprechen zu dem Mann.

I talk with the mother.
Ich spreche mit der Mutter.

Aligned Data

Statistische Maschinelle Übersetzung

I love the woman.
Ich liebe die Frau.

The man loves the cat.
Der Mann liebt die Katze.

The man loves the woman.
Der Mann liebt die Frau.

I love the man.
Ich liebe den Mann.

They talk to the cat.
Sie sprechen zu der Katze.

They talk to the boy.
Sie sprechen zu dem Jungen.

They talk to the man.
Sie sprechen zu dem Mann.

I talk with the mother.
Ich spreche mit der Mutter.

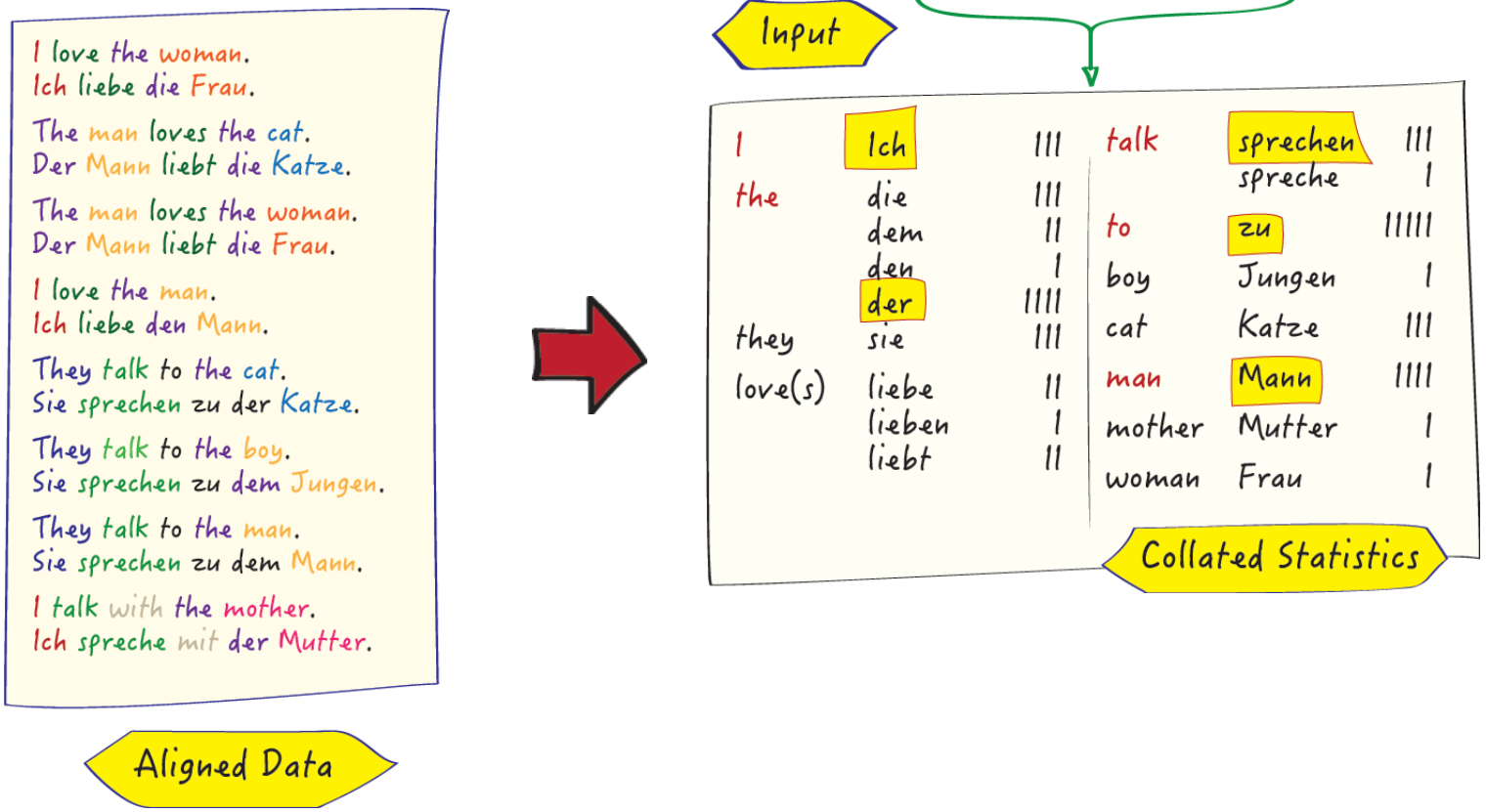
Aligned Data

I	Ich	III	talk	sprechen	III
				spreche	I
the	die	III			
	dem	II	to	zu	IIIII
	den	I	boy	Jungen	I
	der	IIII			
they	sie	III	cat	Katze	III
love(s)	liebe	II	man	Mann	IIII
	lieben	I	mother	Mutter	I
	liebt	II	woman	Frau	I

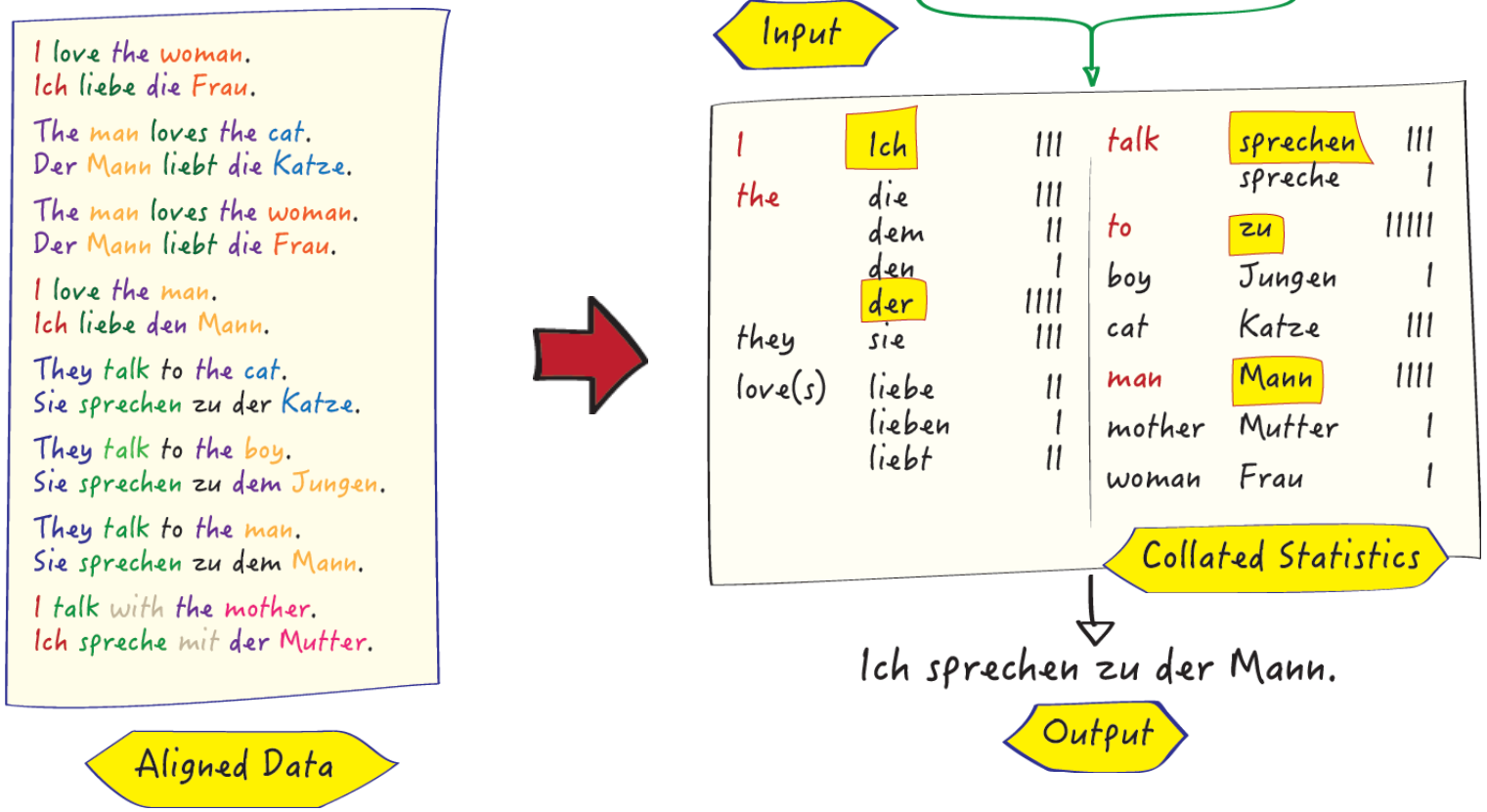
Collated Statistics

European Language Resource Coordination
Connecting Europe Facility

Statistische Maschinelle Übersetzung



Statistische Maschinelle Übersetzung



Statistische Maschinelle Übersetzung

I love the woman.
Ich liebe die Frau.

The man loves the cat.
Der Mann liebt die Katze.

The man loves the woman.
Der Mann liebt die Frau.

I love the man.
Ich liebe den Mann.

They talk to the cat.
Sie sprechen zu der Katze.

They talk to the boy.
Sie sprechen zu dem Jungen.

They talk to the man.
Sie sprechen zu dem Mann.

I talk with the mother.
Ich spreche mit der Mutter.

Aligned Data

I	talk	to	the	man
Ich	sprechen	zu	der	Mann
3/3	3/4	5/5	4/10	4/4
Ich	spreche	zu	dem	Mann
3/3	1/4	5/5	2/10	4/4

Auswahlkriterien?

Statistische Maschinelle Übersetzung

I love the woman.
Ich liebe die Frau.

The man loves the cat.
Der Mann liebt die Katze.

The man loves the woman.
Der Mann liebt die Frau.

I love the man.
Ich liebe den Mann.

They talk to the cat.
Sie sprechen zu der Katze.

They talk to the boy.
Sie sprechen zu dem Jungen.

They talk to the man.
Sie sprechen zu dem Mann.

I talk with the mother.
Ich spreche mit der Mutter.

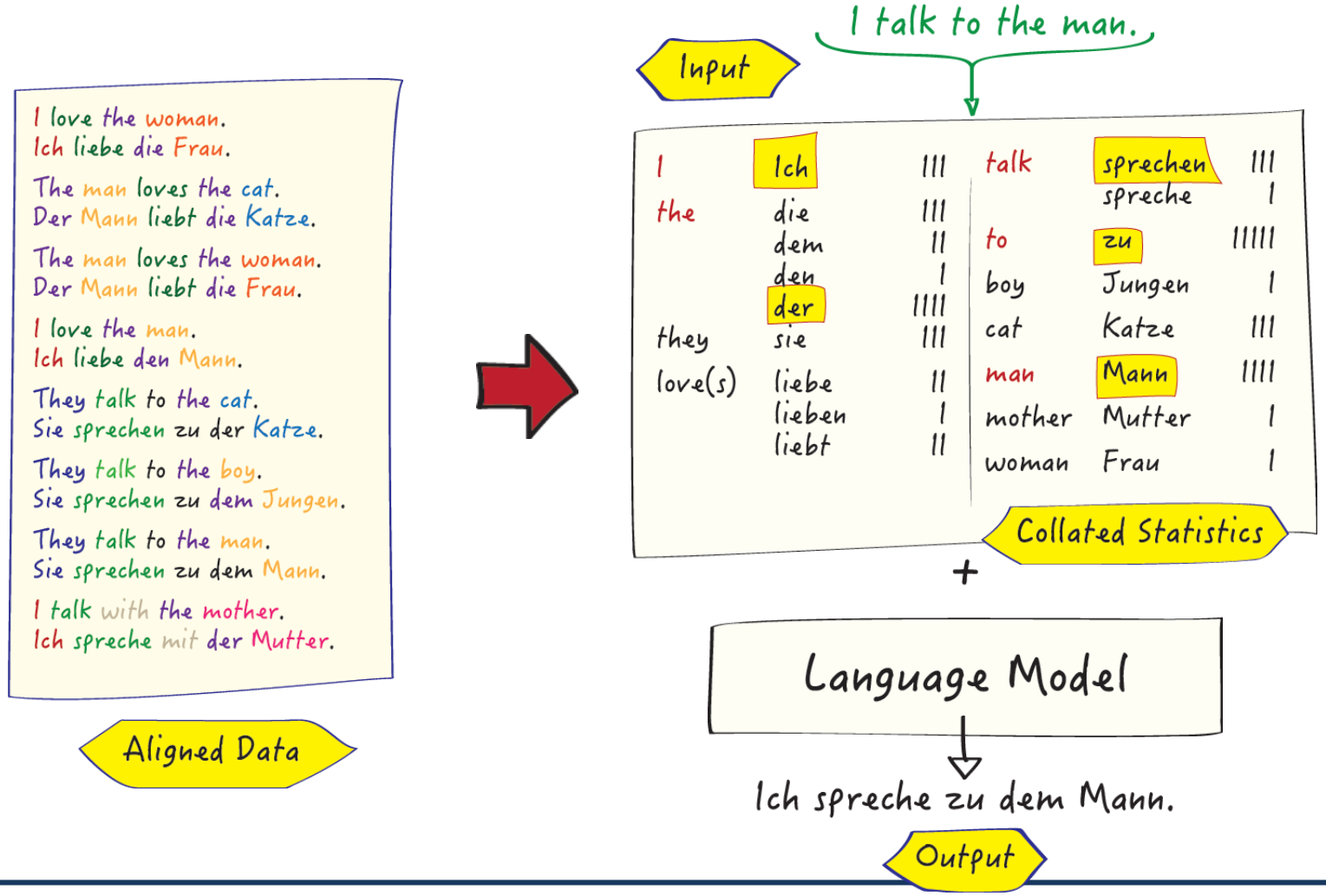
Aligned Data

Sprachmodell:

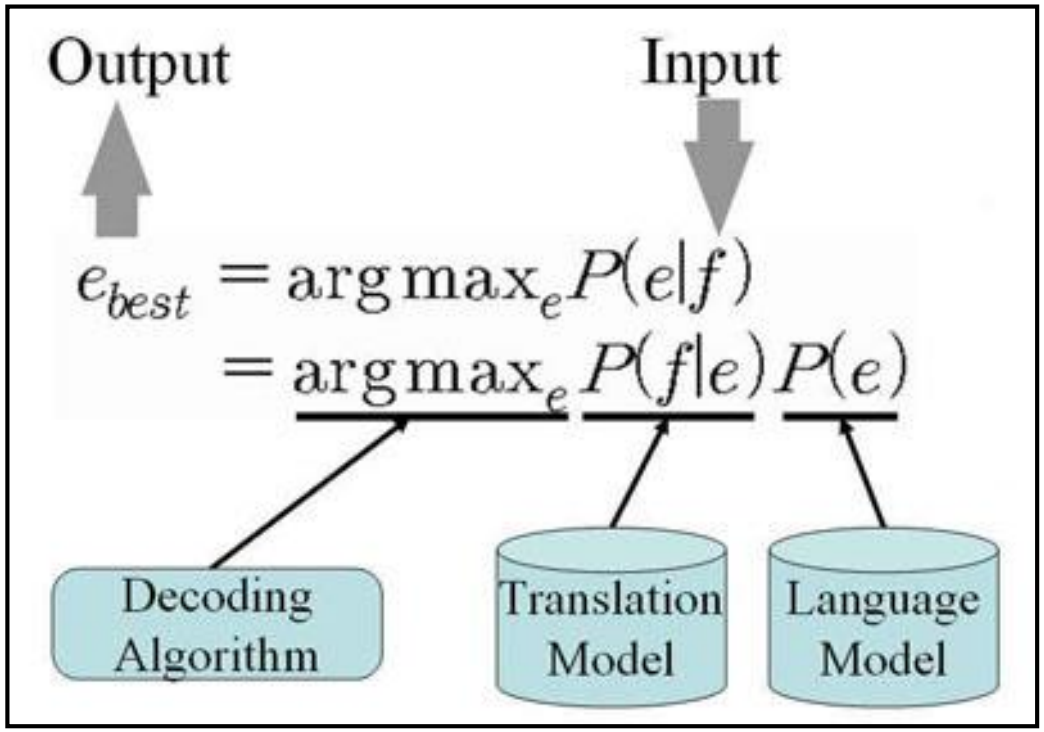
- Was ist eine gute Zielsprache?
- Welche Wörter können aufeinander folgen, und welche nicht…? Die Grammatik
- Aus den Daten lernen …
  - *Ich spreche* is good …
  - *Ich sprechen* is bad …
  - *zu dem Mann* is good …
  - *zu der Mann* is bad …
- *Ich spreche zu dem Mann >>*
  Ich sprechen zu der Mann

European Language Resource Coordination
Connecting Europe Facility

Statistische Maschinelle Übersetzung



Wie funktioniert die moderne MÜ?



- Heute kein Mathematikunterricht!
- Sondern:
- Die Geschichte der statistischen MÜ in Bildern …
- Es dreht sich einzig und allein um **Daten** …

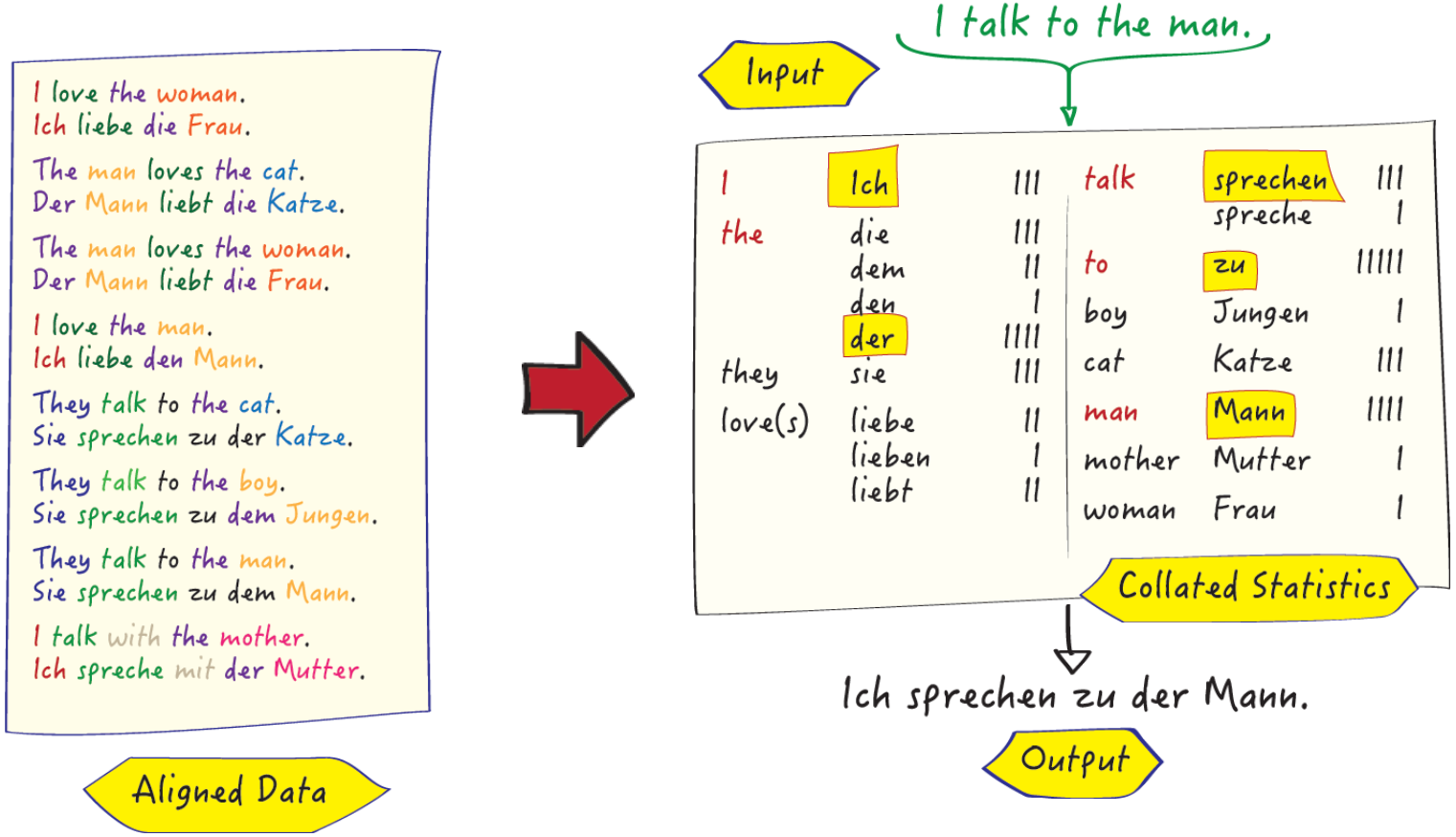
Phrasenbasierte Statistische MÜ

- Bis jetzt: nur einzelne Wörter übersetzt
- Kontext, wie Kongruenz, fehlt (*zu dem Mann* …) usw.
- Bis zu einem gewissen Grad "repariert" mit Hilfe des Sprachmodells

- Ein besserer Ansatz:
- Nicht nur einzelne Wörter, sondern auch Phrasen übersetzen:

the man : der Mann
to the man : zu dem Mann
I talk : Ich spreche

Statistische Maschinelle Übersetzung



European Language Resource Coordination
Connecting Europe Facility

Phrasenbasierte Statistische Maschinelle Übersetzung

I love the woman.
Ich liebe die Frau.

The man loves the cat.
Der Mann liebt die Katze.

The man loves the woman.
Der Mann liebt die Frau.

I love the man.
Ich liebe den Mann.

They talk to the cat.
Sie sprechen zu der Katze.

They talk to the boy.
Sie sprechen zu dem Jungen.

They talk to the man.
Sie sprechen zu dem Mann.

I talk with the mother.
Ich spreche mit der Mutter.

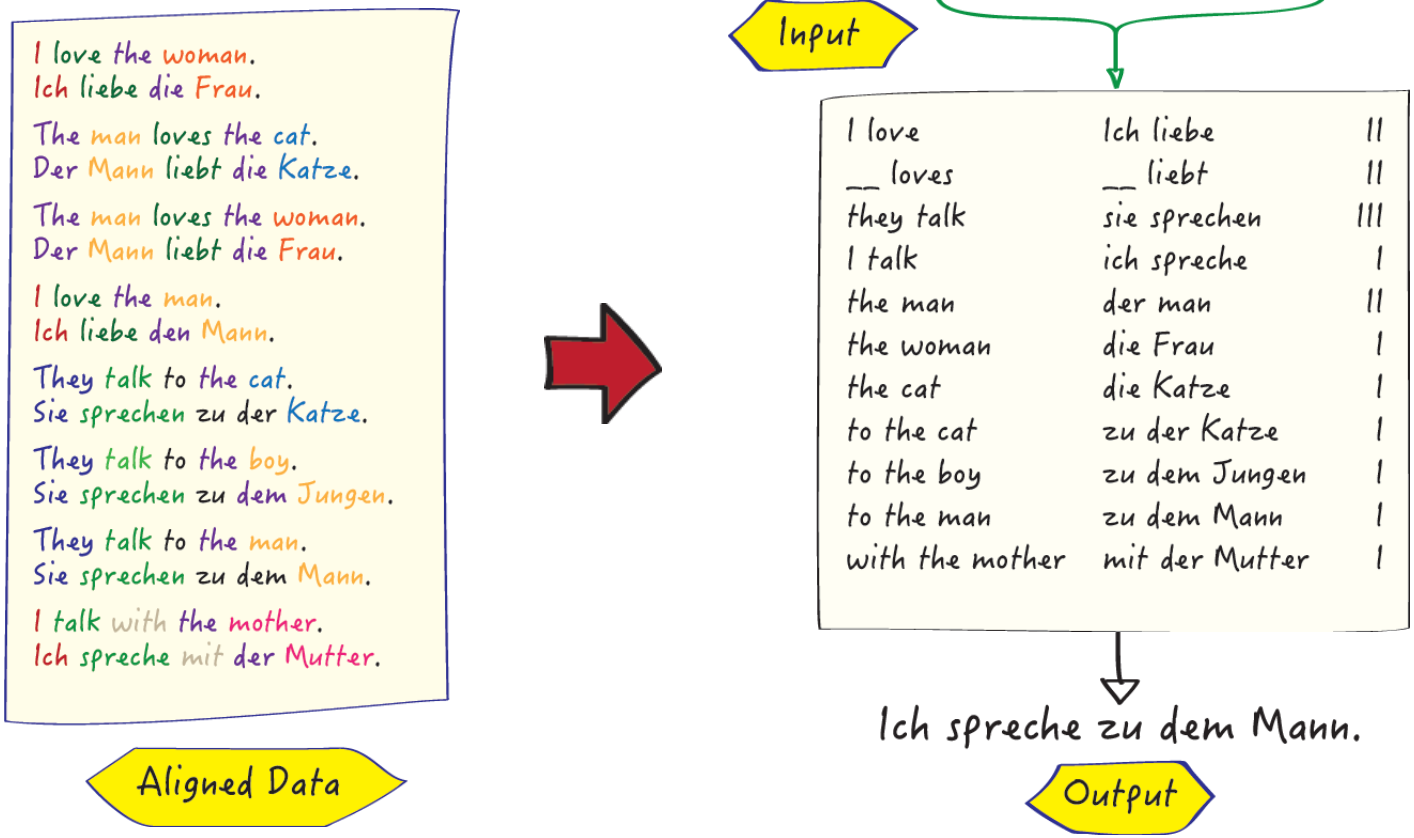
Aligned Data

Input: I talk to the man.

I love	Ich liebe	II
__ loves	__ liebt	II
they talk	sie sprechen	III
I talk	ich spreche	I
the man	der man	II
the woman	die Frau	I
the cat	die Katze	I
to the cat	zu der Katze	I
to the boy	zu dem Jungen	I
to the man	zu dem Mann	I
with the mother	mit der Mutter	I

European Language Resource Coordination
Connecting Europe Facility

Phrasenbasierte Statistische Maschinelle Übersetzung



Phrasenbasiert - Statistische Maschinelle Übersetzung

- Viel besser als wortbasierte SMÜ!
- Standard Technologie: Google, Microsoft, Baidu, globale Lokalisierungs- und Übersetzungsindustrie

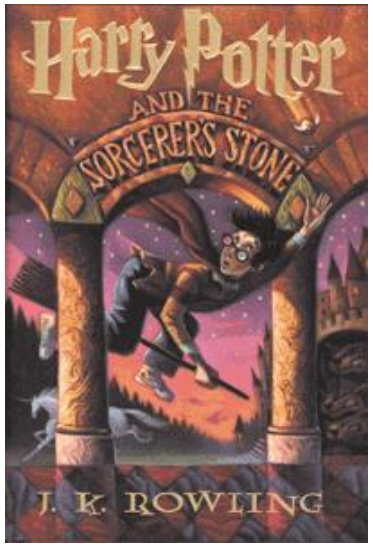
- Moses Open Source PB-SMÜ
- Am meisten verwendetes System für SMÜ
- Forschung auch von der EC finanziert
- Eingesetzt bei dem Direktorat EC DGT's MT@EC

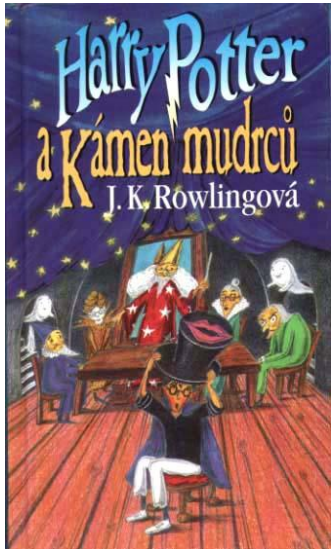
Maschinelle Übersetzung und Daten

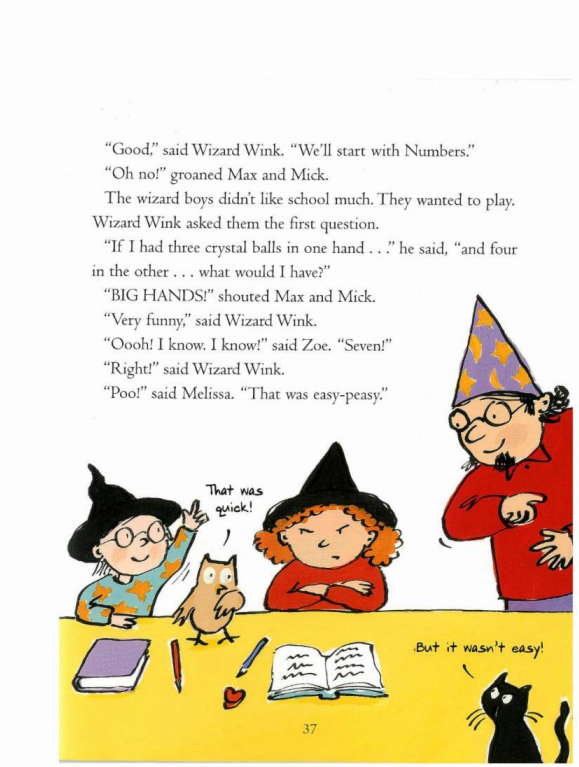
- Bei der Statistischen Maschinellen Übersetzung dreht sich alles um Daten
- SMÜ lernt das Übersetzen aus den Daten
- Daten
  - Übersetzungen (zweisprachige Daten)
  - Einzelsprachliche Daten (Text in der Zielsprache)
  - Wörterbücher, Terminologie, Ontologien, Eingennamen
- Genauso wie bei den Menschen, die SMÜ ist gut bei dem was sie gelernt hat.

European Language Resource Coordination
Connecting Europe Facility

Maschinelle Übersetzung und Daten



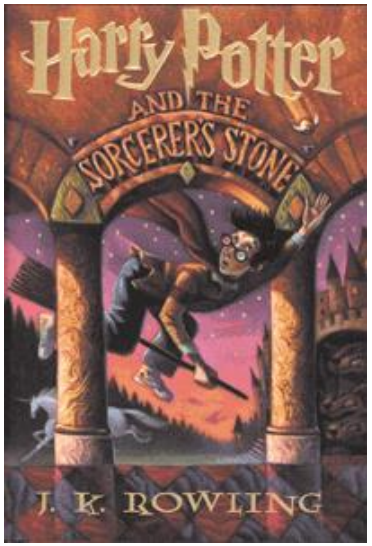


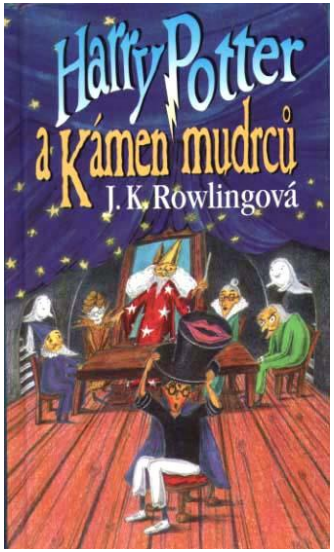

"Good," said Wizard Wink. "We'll start with Numbers."
"Oh no!" groaned Max and Mick.
The wizard boys didn't like school much. They wanted to play. Wizard Wink asked them the first question.
"If I had three crystal balls in one hand . . ." he said, "and four in the other . . . what would I have?"
"BIG 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."

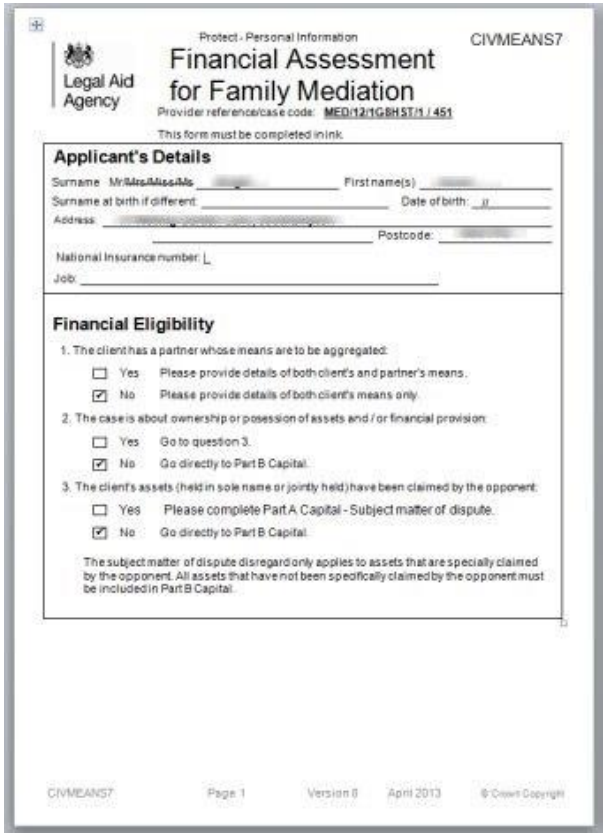
MOSES CORE

European Language Resource Coordination
Connecting Europe Facility

Maschinelle Übersetzung und Daten






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~~ First name(s):

Surname at birth if different: Date of birth:

Address: Postcode:

National Insurance number:

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

MOSES CORE

CEF.AT und Daten

- CEF.AT braucht die richtigen Daten
- Nationale Regierungen, öffentliche Verwaltungen, öffentliche Dienste, NRO/NGOs
- CEF bietet Diensten für multilinguale Interaktion mit den nationalen Bürgern, EU Bürgern und anderen Nützern von öffentlichen Verwaltungen.

ELRC

- Helfen Sie uns, CEF.AT zum Erfolg zu führen
  - Dienste für Europäische Bürger
  - Dienste für Sie
  - Unterstützung von Mehrsprachigkeit
- Helfen Sie, die richtigen Daten zu finden
- Unsere Sprachen zu unterstützen heißt Europa zu unterstützen, und umgekehrt