European Language Resource Coordination
Connecting Europe Facility

“Wie funktioniert maschinelle Übersetzung?”

Bartholomäus Wloka (Universität Wien)

Basierend auf der Präsentation von Prof. Josef van Genabith
(Deutsches Forschungszentrum für Künstliche Intelligenz)

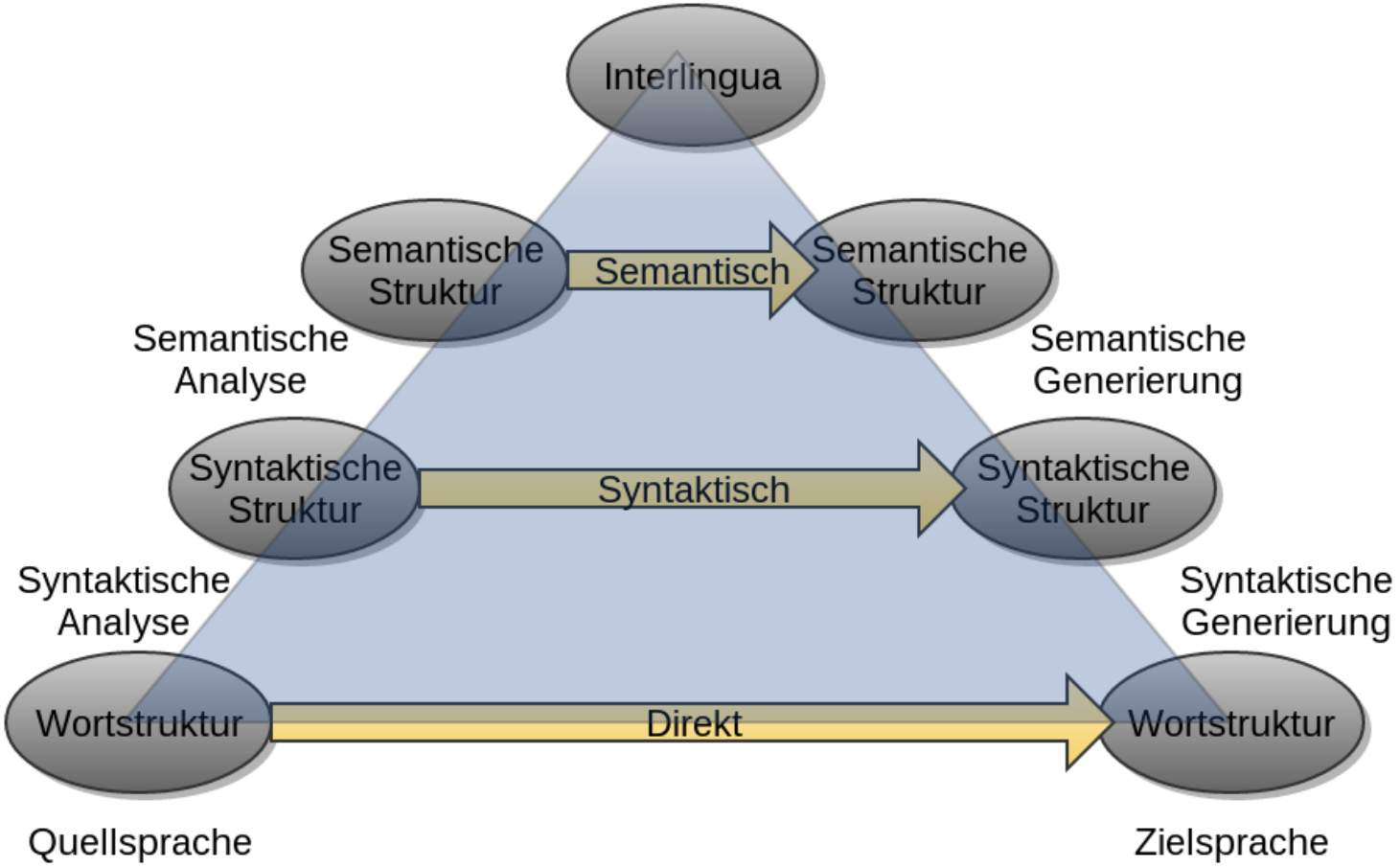
Maschinelle Übersetzung (MÜ)

Überblick:

- Was macht maschinelle Übersetzung so schwierig?
- FAHQMT
- Warum MÜ: Datenmenge, Qualität und Kosten?
- MÜ + professionelle Übersetzer = Qualität
  - CAT Tools
  - MAHT
  - HAMT

European Language Resource Coordination
Connecting Europe Facility

Maschinelle Übersetzung (MÜ)



Maschinelle Übersetzung (MÜ)

Überblick:

- Wie funktioniert die moderne statistisch-basierte MÜ?
- Es geht vor allem um Daten
- Und um die richtige Art von Daten
- Vor allem: Parallele Korpora und Sprachmodelle
- Am besten Domänenspezifisch

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.

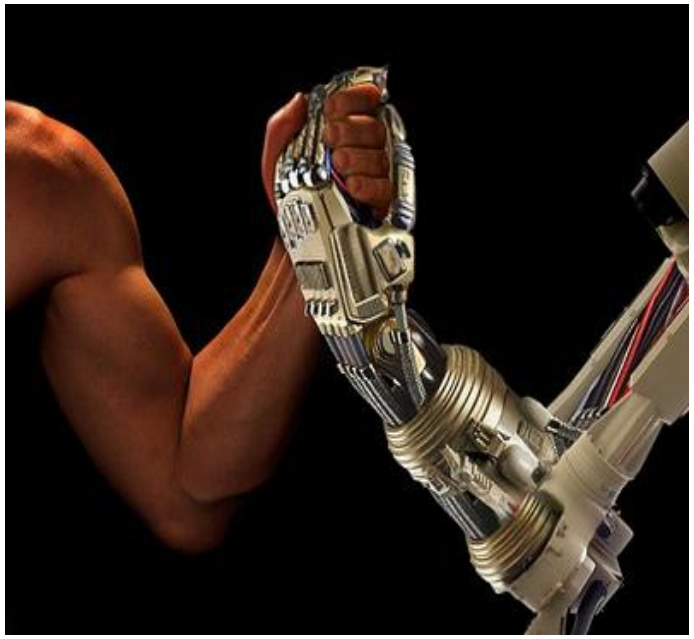
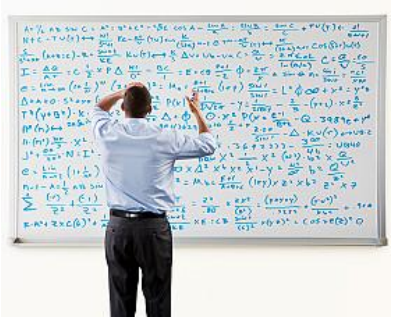


Image: http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg

Hohe Komplexität von Sprache und Übersetzung

- Sprache ist komplex, Übersetzung noch komplexer
- Wir können sie nicht genau berechnen
- Verschiedene MÜ Methoden erforscht
- Hybride Methoden
- Maschinelles Lernen
  - Aus Daten lernen ⇒ zentrale Rolle von Daten
  - Grobe Lösung
    - schafft ersten Überblick, MAHT
    - *Post-editing* durch professionelle Übersetzer



Was macht die MÜ so schwierig?

Schlagzeilen:

1. Minister Accused Of Having 8 Wives In Jail
2. Teacher Strikes Idle Kids
3. Miners refuse to work after death
4. Local High School Dropouts Cut in Half
5. Red Tape Holds Up New Bridges
6. Clinton Wins on Budget, but More Lies Ahead
7. Police: Crack Found in Man's Buttocks

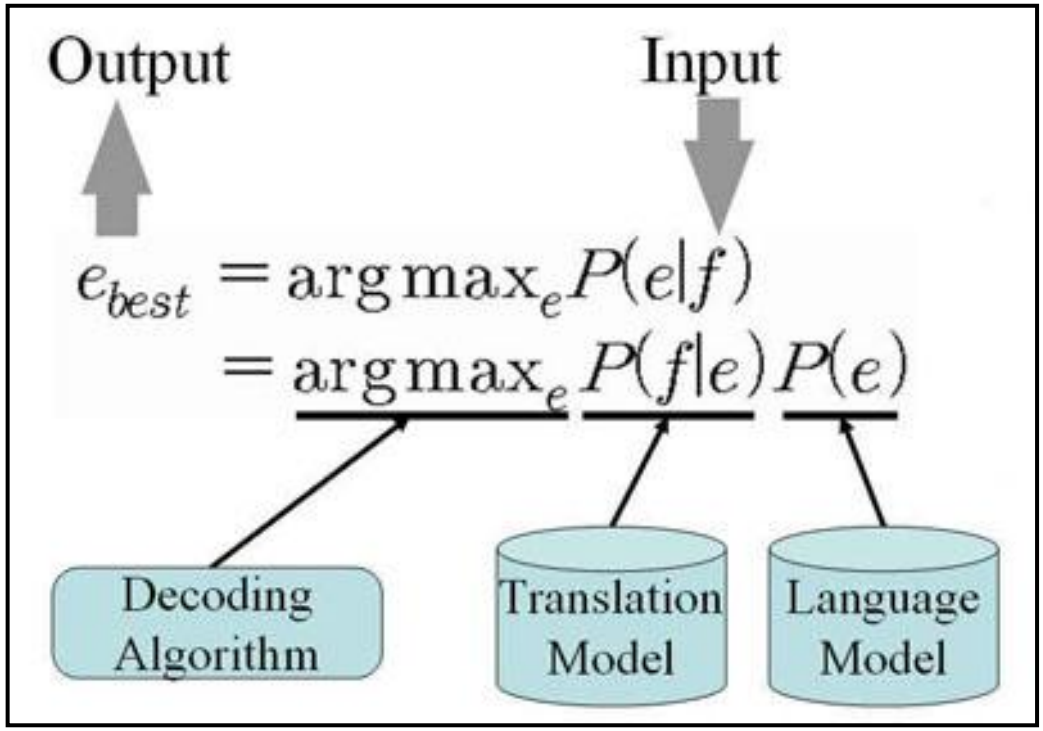
Lecture notes:
Chris Manning

European Language Resource Coordination
Connecting Europe Facility

Was macht die MÜ so schwierig?

- Sprichwörter:
  - The early bird catches the worm
  - Morgenstund hat Gold im Mund
- Satzstellung
  - The German chancellor Angela Merkel will make an announcement on Thursday.
  - Angela Merkel wird am Donnerstag eine Ankündigung machen
- Polysemie
  - Der Angeklagte hat gestanden, jetzt muss er sitzen.

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?

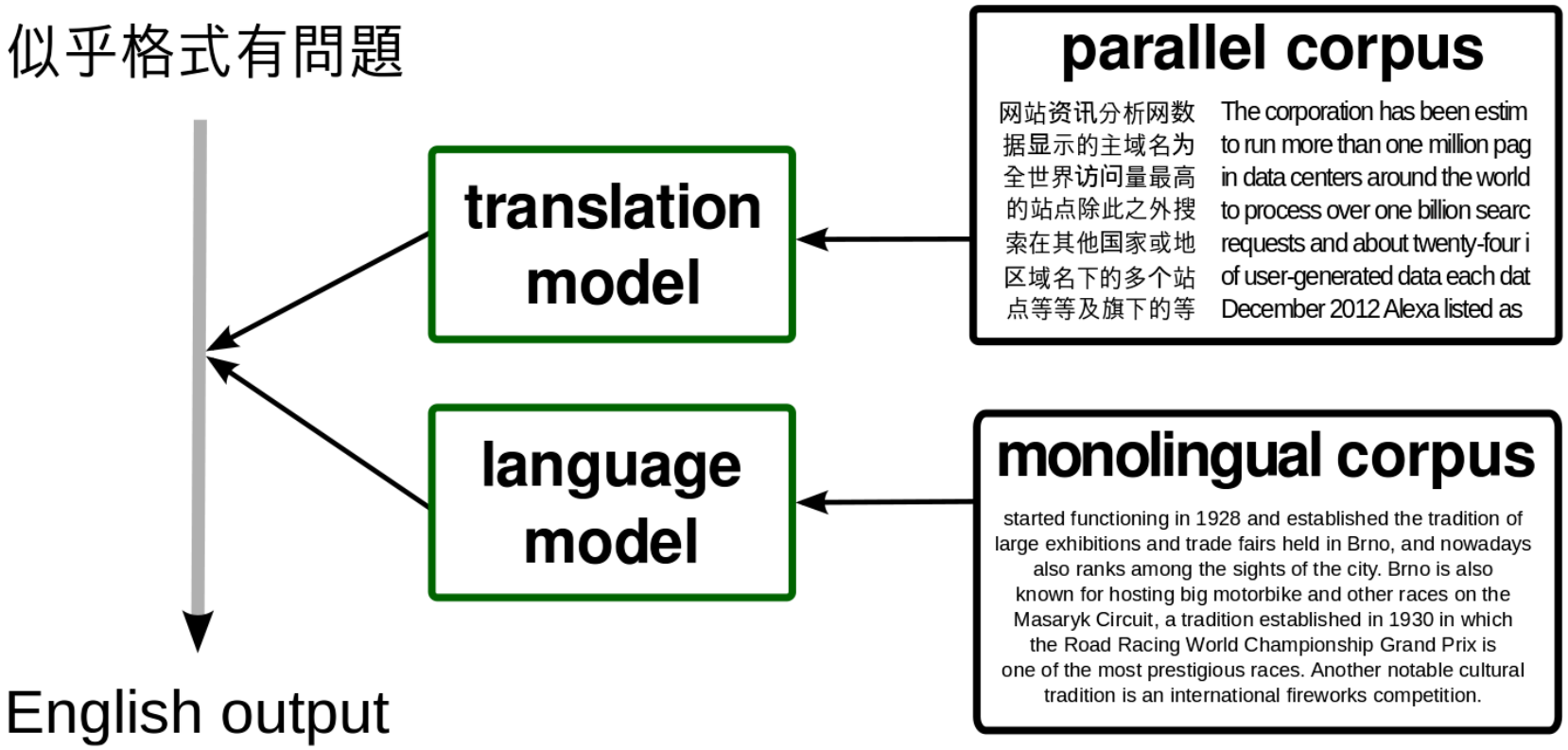
Die statistische MÜ lernt aus zwei Typen von Daten:

- Übersetzungen von Menschen
- Text in der Zielsprache

- So viele **adäquate** Daten wie möglich

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.

Wie funktioniert die moderne MÜ?



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: WSD + Übersetzungswahscheinlich keiten
- 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.

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

Collated Statistics

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.

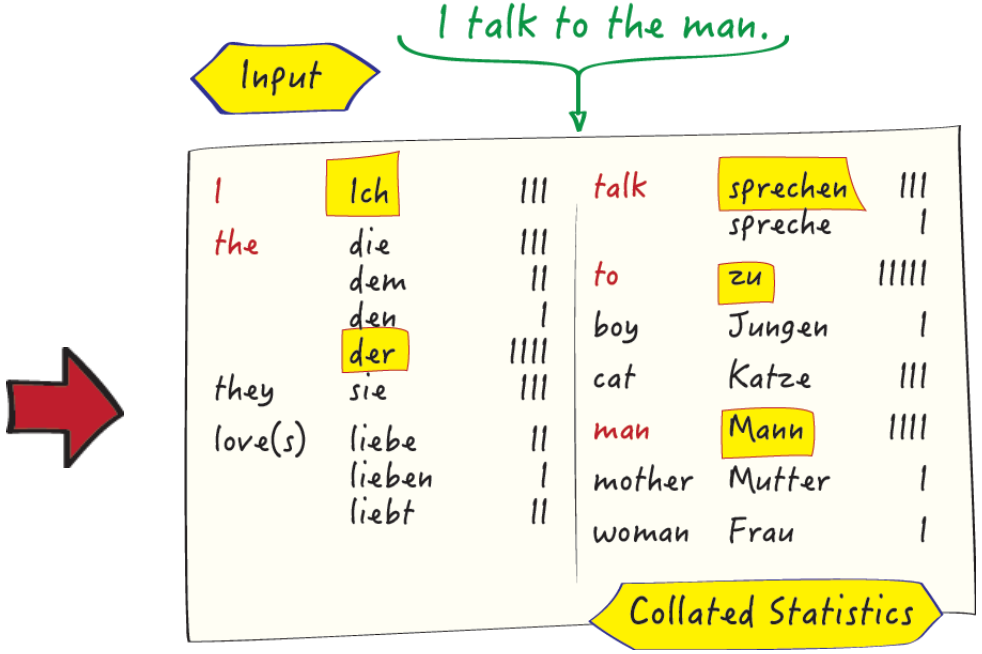
They talk to the cat.
Sie sprechen zu der Katze.

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

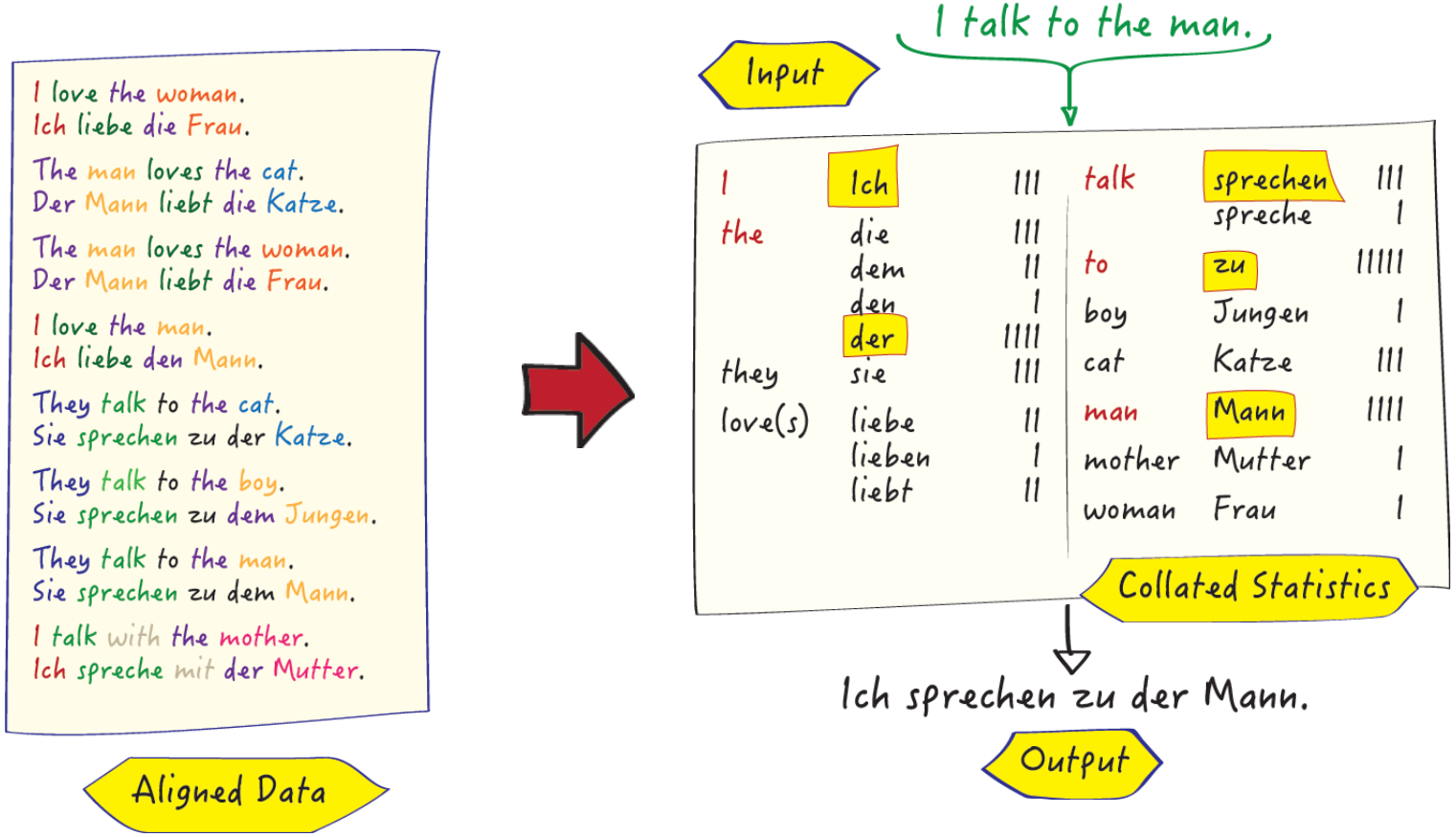
They talk to the man.
Sie sprechen zu dem Mann.

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

Aligned Data



Statistische Maschinelle Übersetzung



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.

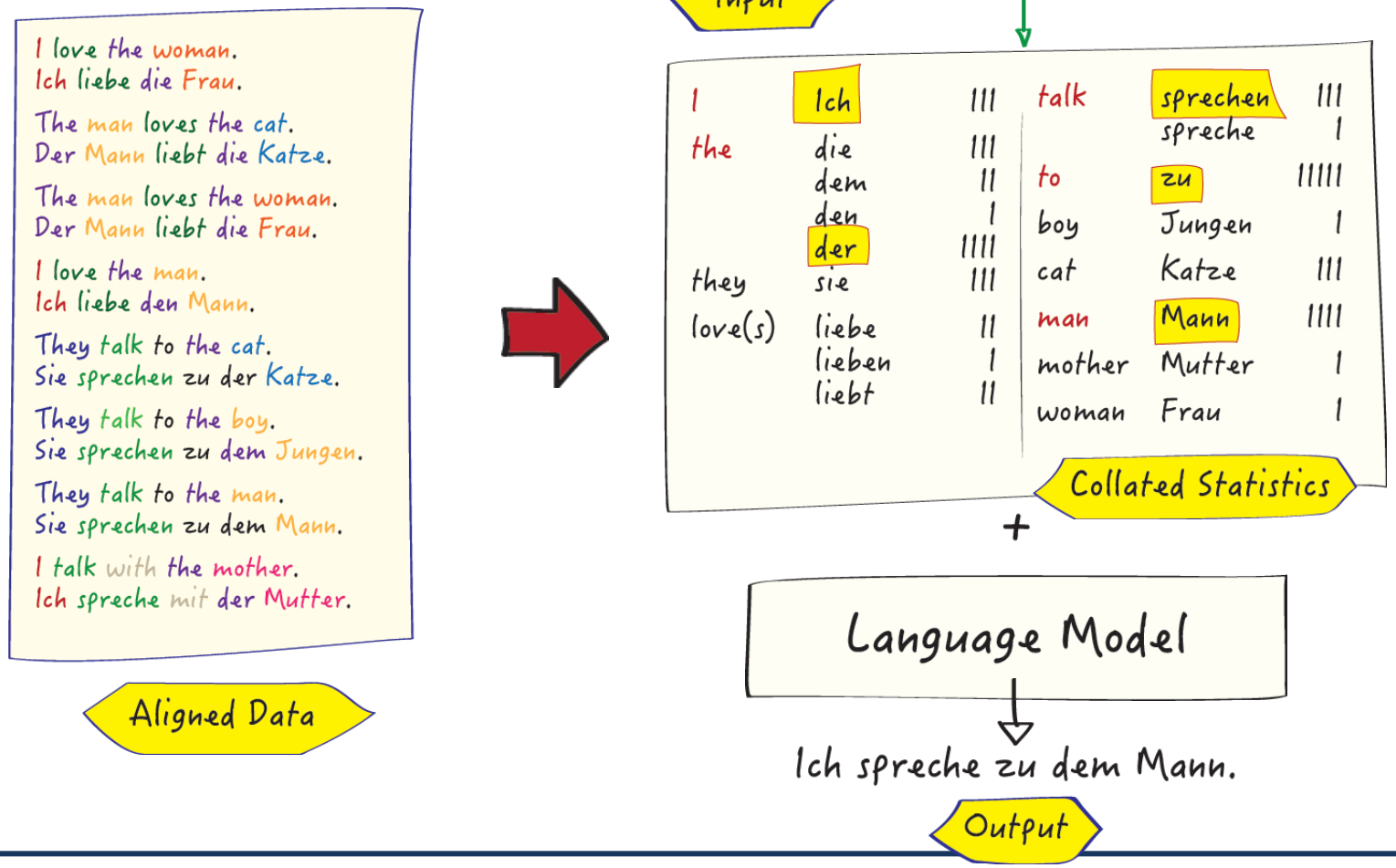
I talk with the mother.
Ich spreche mit der Mutter.

Aligned Data

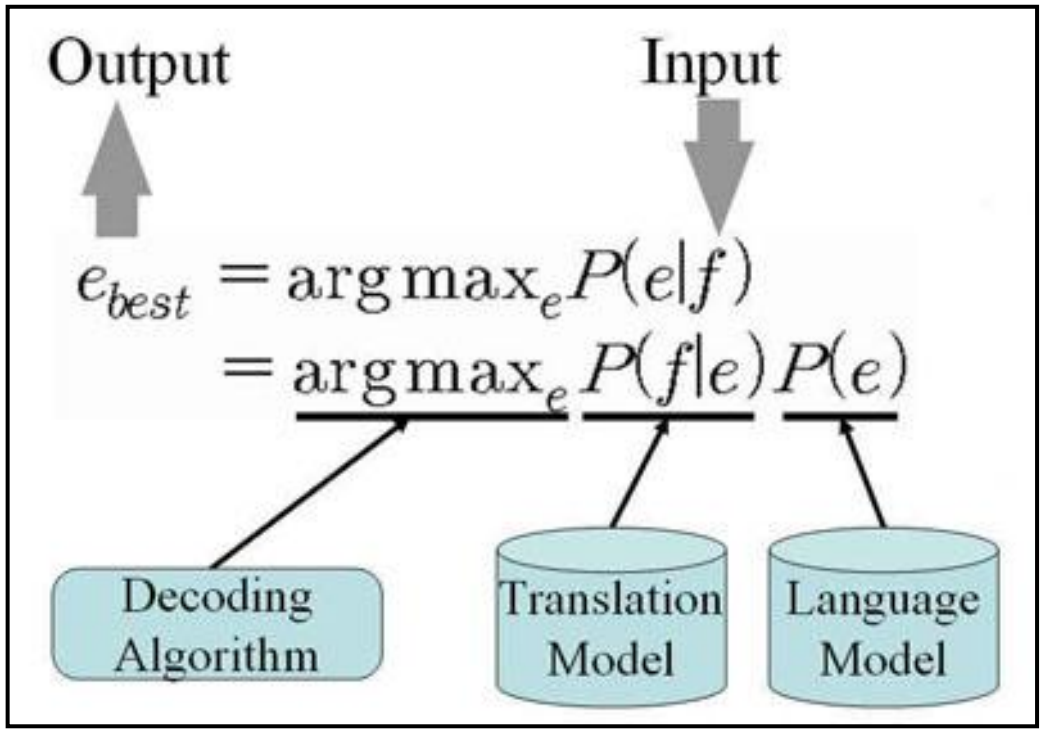
Sprachmodell:

- 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

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

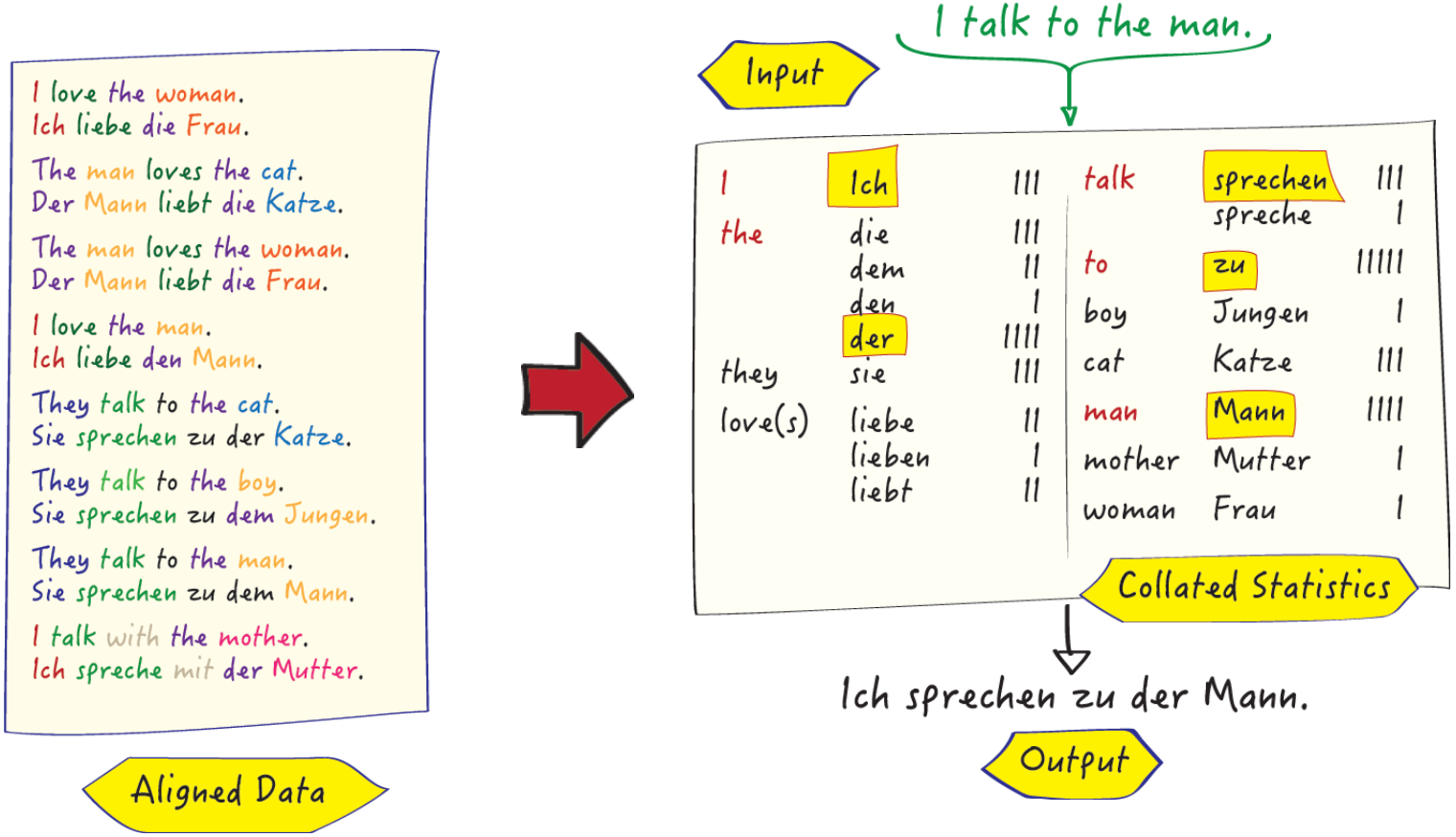
- Bis jetzt: nur einzelne Wörter übersetzt
- Kontext, wie Kongruenz, fehlt (*zu dem Mann* …) usw.
- Bis zu einem gewissen Grad "repariert" mit Hilfe des Sprachmodells

- 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

European Language Resource Coordination
Connecting Europe Facility

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.

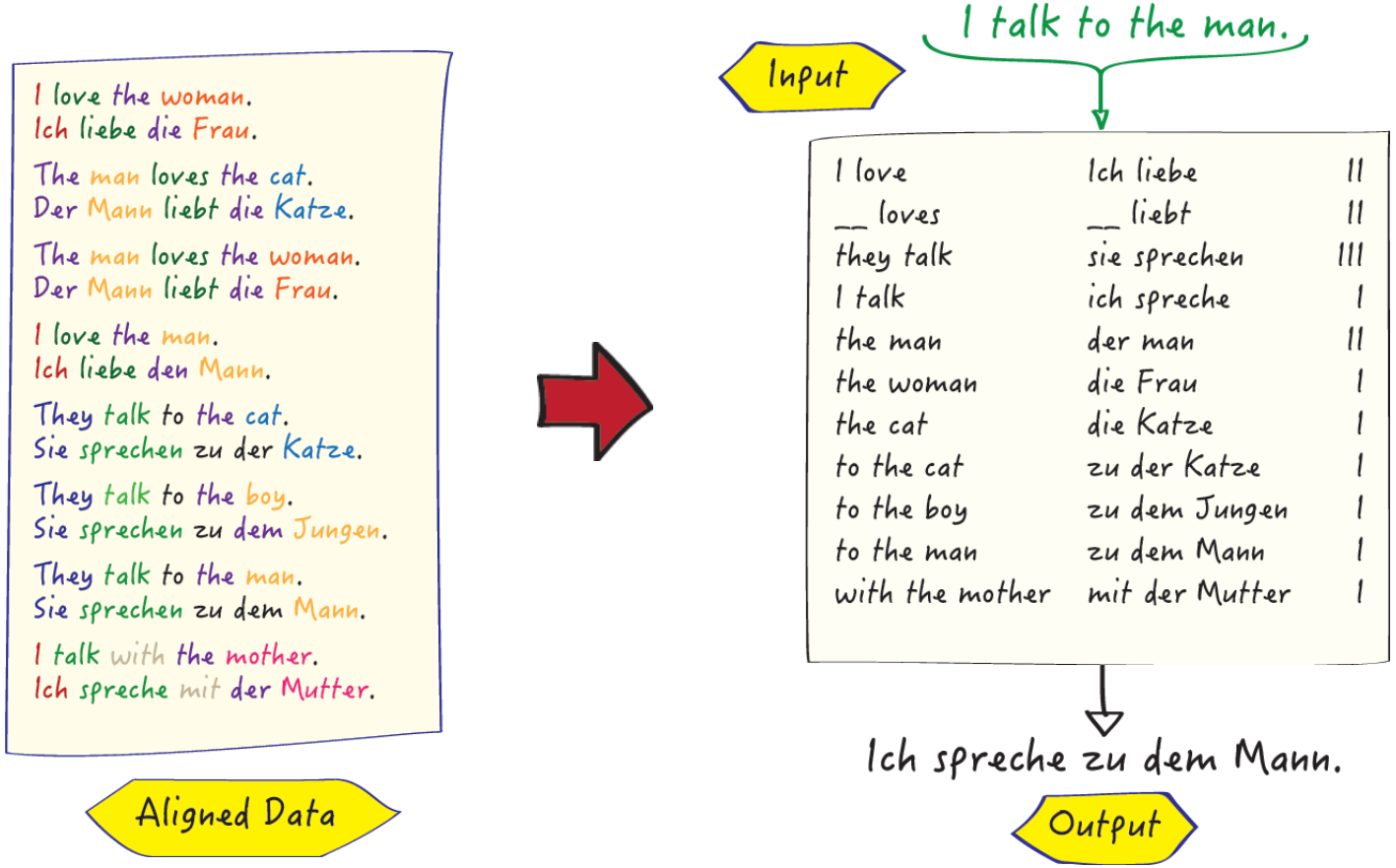
Aligned Data

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

ELRC: Hintergrundinformation

- Ziel: Der “Multilingual Digital Single Market” (mSDM):
  - Keine sprachliche Barriere
  - Freier Verkehr von Leuten, Information, Dienste, Kultur, und Kommerz
- Ziel: CEF.AT:
  - Unterstützung von öffentlichen Diensten, Regierungen, Verwaltung, NGOs in ganz Europa

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)
- Die Qualität der SMÜ hängt vom „Gelernten“ ab
- Nachbearbeitung möglich mit
  - Lexikalische Resourcen, Terminologie, Ontologien, Eigennamen

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 Nutzern 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 mit die richtigen Daten zu finden bzw. zur Verfügung zu stellen