# Statistical Machine Translation

**Prof. Andy Way**

ADAPT Centre,
School of Computing,
Dublin City University,
Dublin 9, Ireland

**andy.way@adaptcentre.ie**

**http://www.computing.dcu.ie/~away/**

# Overview

- Why MT?
- How are people using MT?
- What’s the Role of the Human Translator?
- Why Corpus-Based MT?
- How might you go about translating languages you have no knowledge of?
- The Importance of Data

TAUS
Enabling Better Translation

Email | Password | Login
Create an Account | Reset Password | Blog

MY TAUS | INSIGHTS | DATA | EVALUATION | DIRECTORIES | LABS

About | Membership | Resources | Events | Standards

About ▸ Newsroom ▸ Press releases ▸ Size Machine Translation Market is $250 Million – TAUS Publishes New Market Report

# SIZE MACHINE TRANSLATION MARKET IS $250 MILLION – TAUS PUBLISHES NEW MARKET REPORT

*August 26, 2014, Amsterdam* - TAUS estimates the size of the machine translation market at $250 million in its newl published machine translation market report. The 60 page report offers a detailed overview of all facets of the marke machine translation with sections on the different types of offerings, the players, open-source systems, challenges, opportunities and trends.

“*The size of the MT market is relatively small compared to its innovation power and impact*”, says Jaap van der Mee one of the co-authors of the report. “*MT technology is a key enabler and a force multiplier for new services and growth. MT technology finds a high adoption rate among language service providers. Innovative companies in information technology and other sectors are converging MT technology in new applications and products or they use MT to enhance their existing products.*”

For this market report TAUS has identified 65 different MT operators. More than 80 companies responded to the surveys and the TAUS team interviewed 37 users and developers of MT. The largest MT providers in alphabetical order are: CSLi, Google, IBM, LionBridge, Microsoft, PROMT, Raytheon BBN, SDL, Smart Communications, SYSTRAN. The MT market is a vibrant sector with new companies entering the market place and long-term players being acquired. Around 20 of the 65 identified MT players started business in the last five years. At the start of this year SYSTRAN was acquired by CSLi from South Korea and AppTek was acquired by eBay.

“*The dynamics in the MT market have changed dramatically in the last five years*”, says Achim Ruopp, product development manager at TAUS and co-author of the report. “*The increased availability of easy to use and integrate MT with sufficient quality has ignited the emergence of new business models. This has been promoted by many new MT suppliers that base their offering on the open source statistical MT system Moses. A bigger impact still has come from some of the internet giants like Google, Microsoft and Yandex that offer free or very cheap MT services.*”

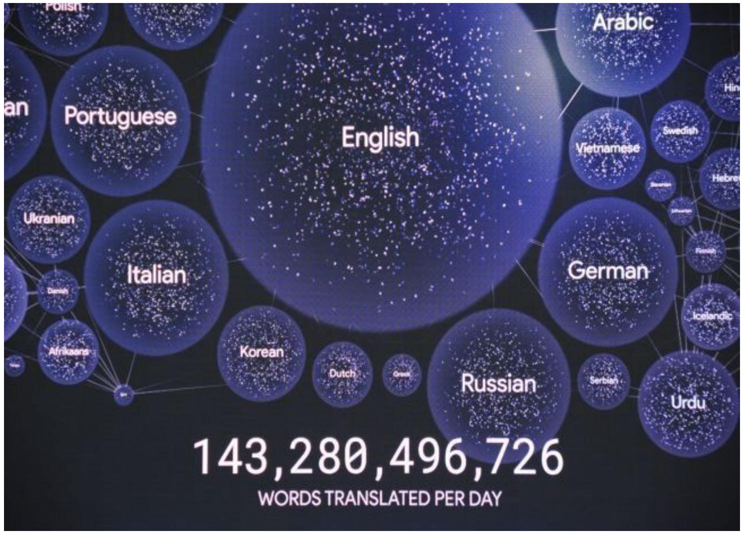
The TAUS Machine Translation Market Report is available on the TAUS website.

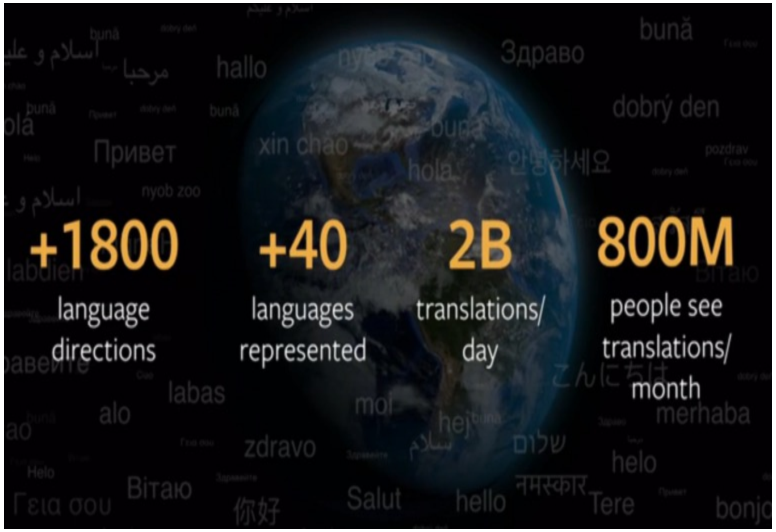
Mission
History
Services
People
Team
Advisors
Representatives
Partners
Newsroom
Press releases
Newsletters
Working at TAUS
Contact

# Highlights from TAUS Claims

- Size of the market: €250,000,000
- “MT technology is a key enabler and a force multiplier for new services”
- “Innovative companies in IT and other sectors are converging MT technology in new applications and products or they use MT to enhance their existing products”
- “The increased availability of easy to use and integrate MT with sufficient quality has ignited the emergence of new business models. This has been promoted by many new MT suppliers that base their offering on the open source statistical MT system Moses”

# MT is being used every day ...





- Google Translate provides a billion translations a day for 200 million users
- Amount of text translated daily is more than what's in a million books
- Surpasses what professional translators handle in a year

# Client-customised engines

- Improve productivity,
- Translate content previously not feasible due to time or cost constraints,
- Reduce time to market, and
- Reduce translation costs.

# Lots of successful case studies

- Adobe & ProMT
- Church of Jesus Christ of Latter-day Saints & Microsoft Translator Hub
- Dell & Safaba/welocalize
- DuDu & CapitaTI
- Ford & Systran/SAIC
- Sajan & Asia Online
- text&form & LucySoft

# The time for MT is now!

- At *MT Summit XIV* in France, for the first time the number of commercial attendees exceeded those from academia.
- Ruopp (2013): for the first time in a TAUS survey, largest group of respondents was LSPs & translation agencies, not research institutes.
- Trends likely to continue, with more large multinational companies, LSPs and MT developers attending such events ...

# Not everyone agrees ...

# Why Corpus-Based MT?

- the (relative) failure of rule-based approaches
- the increasing availability of machine-readable text
- the increase in capability of hardware (CPU, memory, disk space) with decrease in cost

# 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
- …

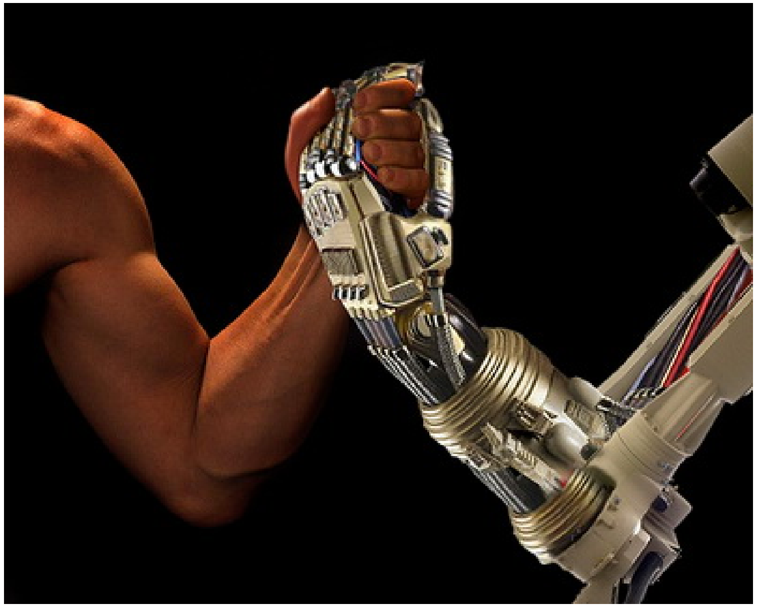


Image: http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg
License: CC BY-NC-SA 3.0

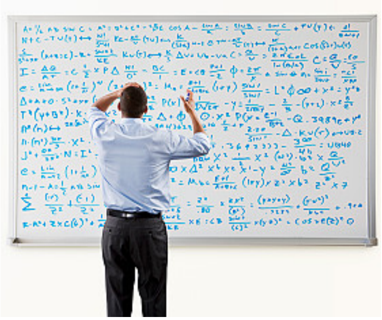
# Why is MT Hard?

## Newspaper Headlines:

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

*Thanks to Chris Manning*

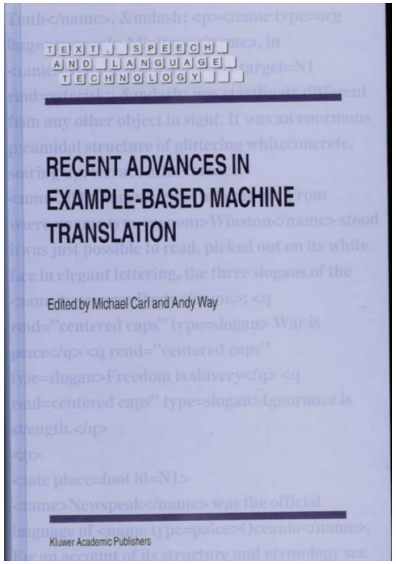
# Language & Translation is Complex



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

# Types of Corpus-based MT

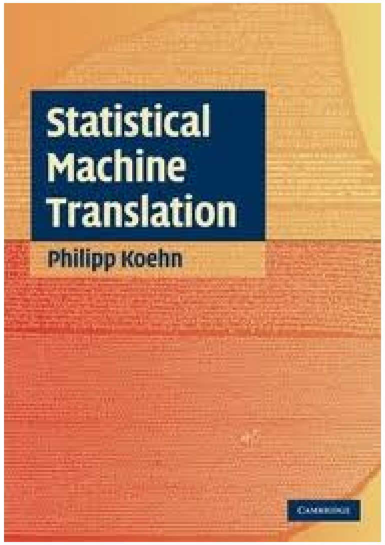
- Example-Based MT (Nagao, 1984)
- Statistical MT
  - 1988: word-based (IBM)
  - 2002—now: phrase-based (Moses)
  - 2005—now: tree-based (Hiero)
- Neural MT (1997 … and 2013 onwards)


TEXT, SPEECH AND LANGUAGE TECHNOLOGY

RECENT ADVANCES IN EXAMPLE-BASED MACHINE TRANSLATION

Edited by Michael Carl and Andy Way

Kluwer Academic Publishers


Statistical Machine Translation

Philipp Koehn

CAMBRIDGE

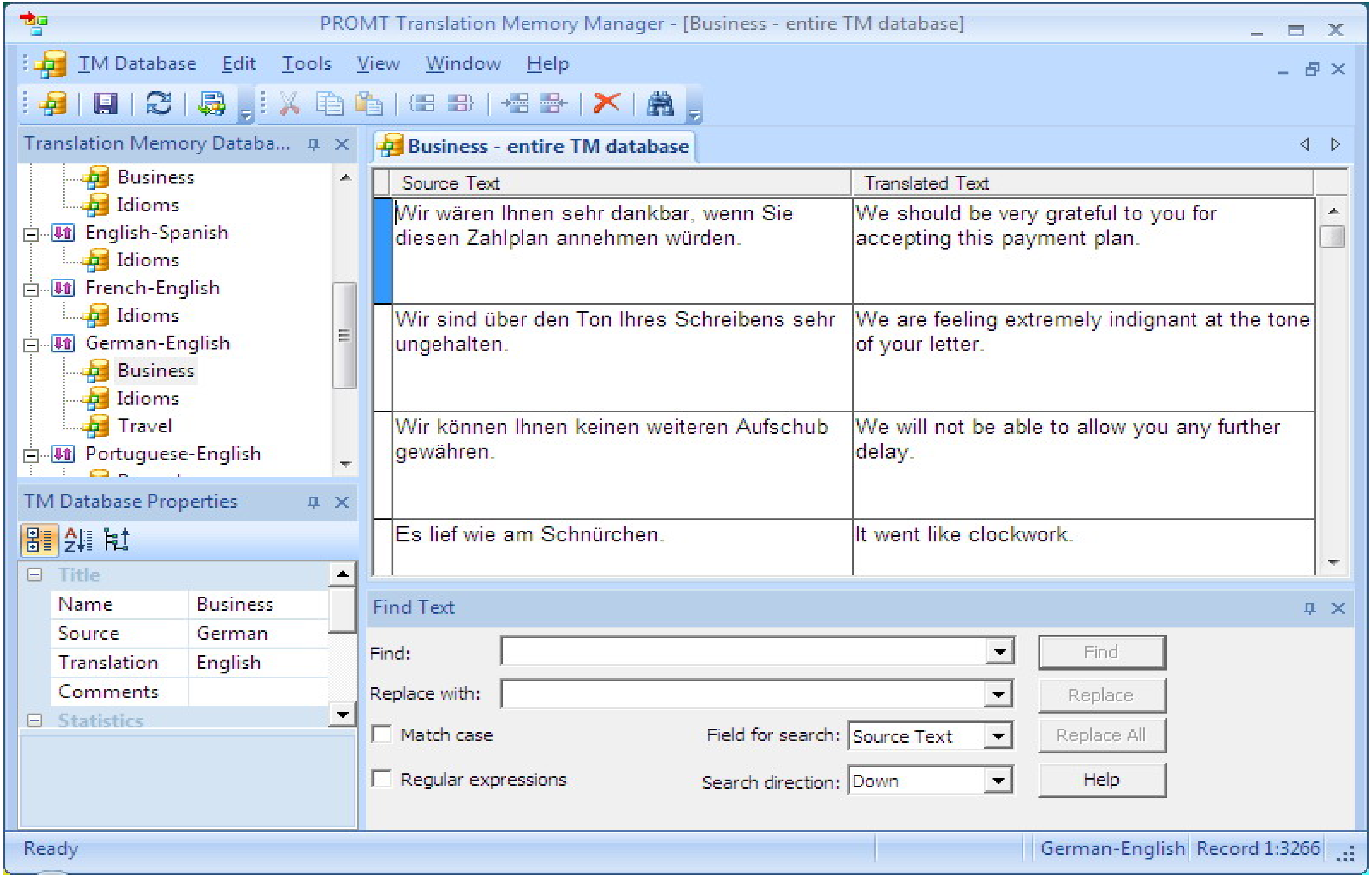
# Prerequisite

A prerequisite for Data-Driven MT (and also TM, which is *not* MT, but rather CAT):

- Example-Based MT (EBMT)
- Statistical MT (SMT) & Neural MT (NMT)
- Hybrid Models which use some probabilistic processing

is a *parallel corpus* (or *bitext*) of aligned sentences.

# Parallel data prerequisite for corpus-based MT



| Source Text | Translated Text |
|---|---|
| Wir wären Ihnen sehr dankbar, wenn Sie diesen Zahlplan annehmen würden. | We should be very grateful to you for accepting this payment plan. |
| Wir sind über den Ton Ihres Schreibens sehr ungehalten. | We are feeling extremely indignant at the tone of your letter. |
| Wir können Ihnen keinen weiteren Aufschub gewähren. | We will not be able to allow you any further delay. |
| Es lief wie am Schnürchen. | It went like clockwork. |

Title

| | |
|---|---|
| Name | Business |
| Source | German |
| Translation | English |
| Comments | |

Statistics

Find:

Replace with:

□ Match case

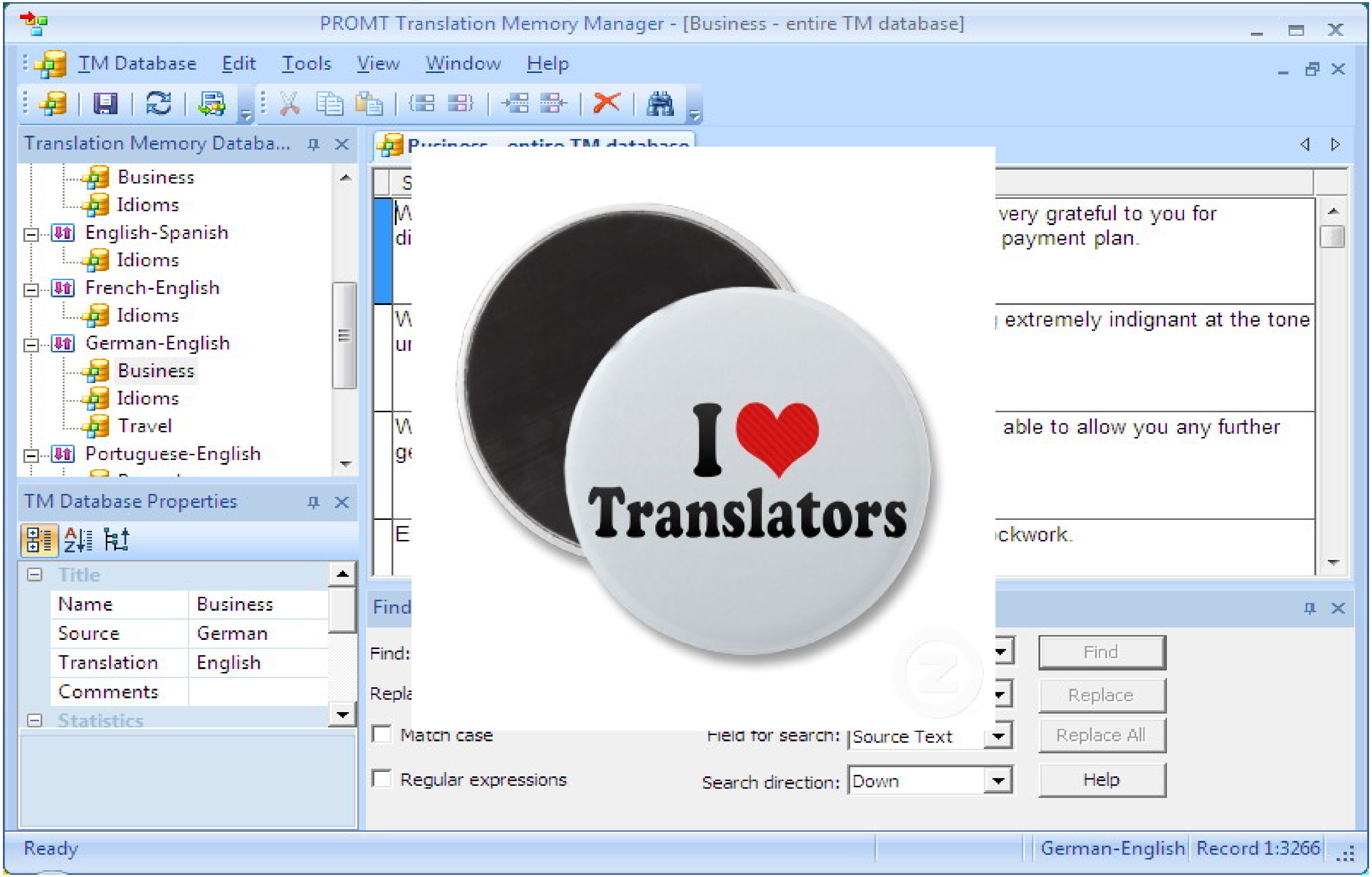
□ Regular expressions

Field for search: Source Text

Search direction: Down

German-English Record 1:3266

# Parallel data prerequisite for corpus-based MT


PROMT Translation Memory Manager - [Business - entire TM database]
TM Database
Edit
Tools
View
Window
Help
Translation Memory Databa...
Business - entire TM database
TM Database Properties
Find
Replace
Replace All
Help
I ♥ Translators

- Business
- Idioms
- English-Spanish
  - Idioms
- French-English
  - Idioms
- German-English
  - Business
  - Idioms
  - Travel
- Portuguese-English

| | |
|---|---|
| Title | |
| Name | Business |
| Source | German |
| Translation | English |
| Comments | |
| Statistics | |

very grateful to you for payment plan.

extremely indignant at the tone

able to allow you any further

ckwork.

Find:

Repla

□ Match case

□ Regular expressions

Field for search: Source Text

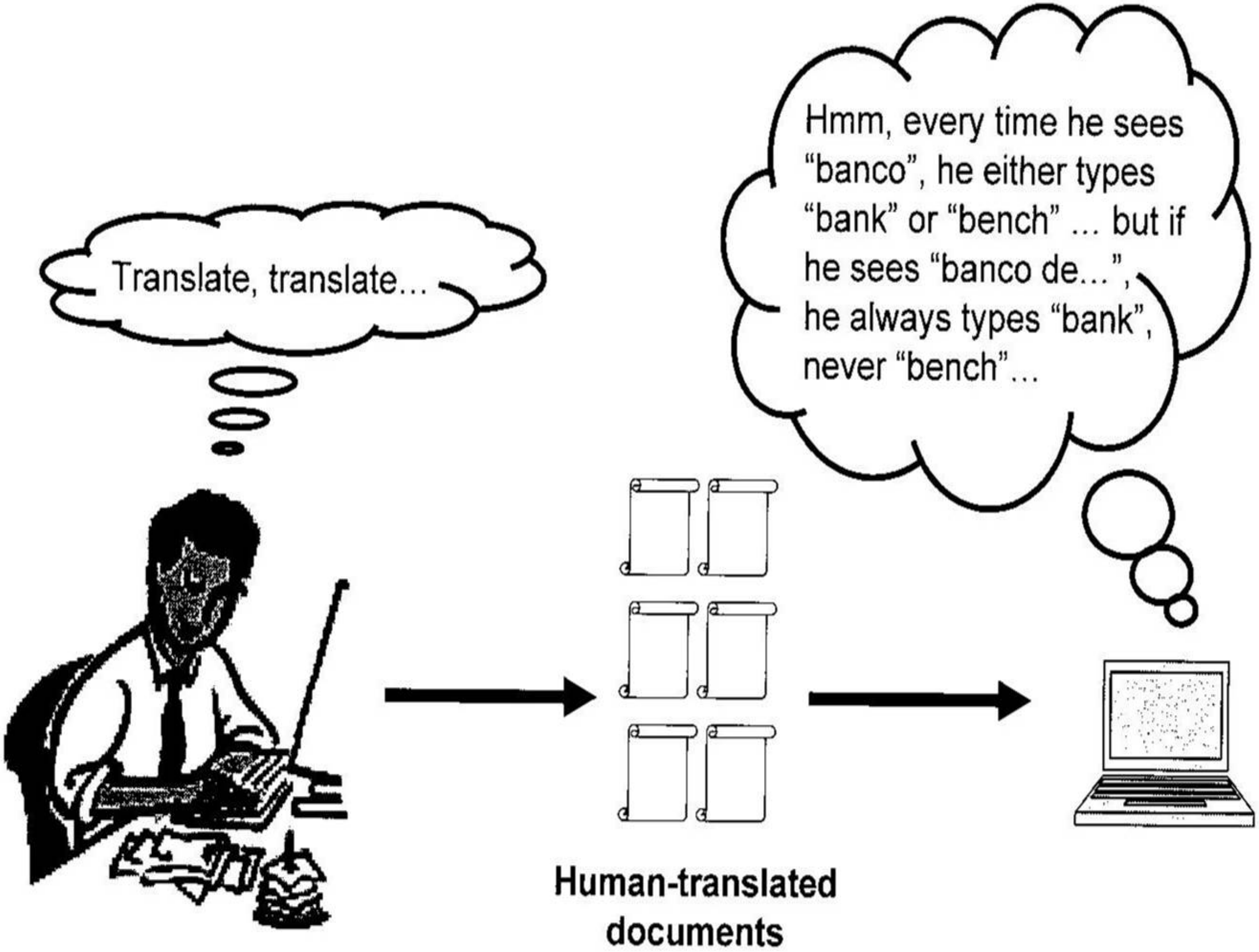
Search direction: Down

Ready | German-English | Record 1:3266

# So how does SMT work?

How might *you* go about translating between two languages you know nothing about?!

# Statistical Machine Translation


Translate, translate…
Hmm, every time he sees "banco", he either types "bank" or "bench" … but if he sees "banco de…", he always types "bank", never "bench"…
Human-translated documents

*Thanks to Kevin Knight …*

# Centauri/Arcturan [Knight, 1997]

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

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

# Centauri/Arcturan [Knight, 1997]

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

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

# Centauri/Arcturan [Knight, 1997]

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

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

# Centauri/Arcturan [Knight, 1997]

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

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

# Centauri/Arcturan [Knight, 1997]

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

12a. lalok rarok nok izok **hihok** mok .

12b. wat nnat forat arrat vat gat .

# Centauri/Arcturan [Knight, 1997]

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

12a. lalok rarok nok izok **hihok** mok .

12b. wat nnat forat arrat vat gat .

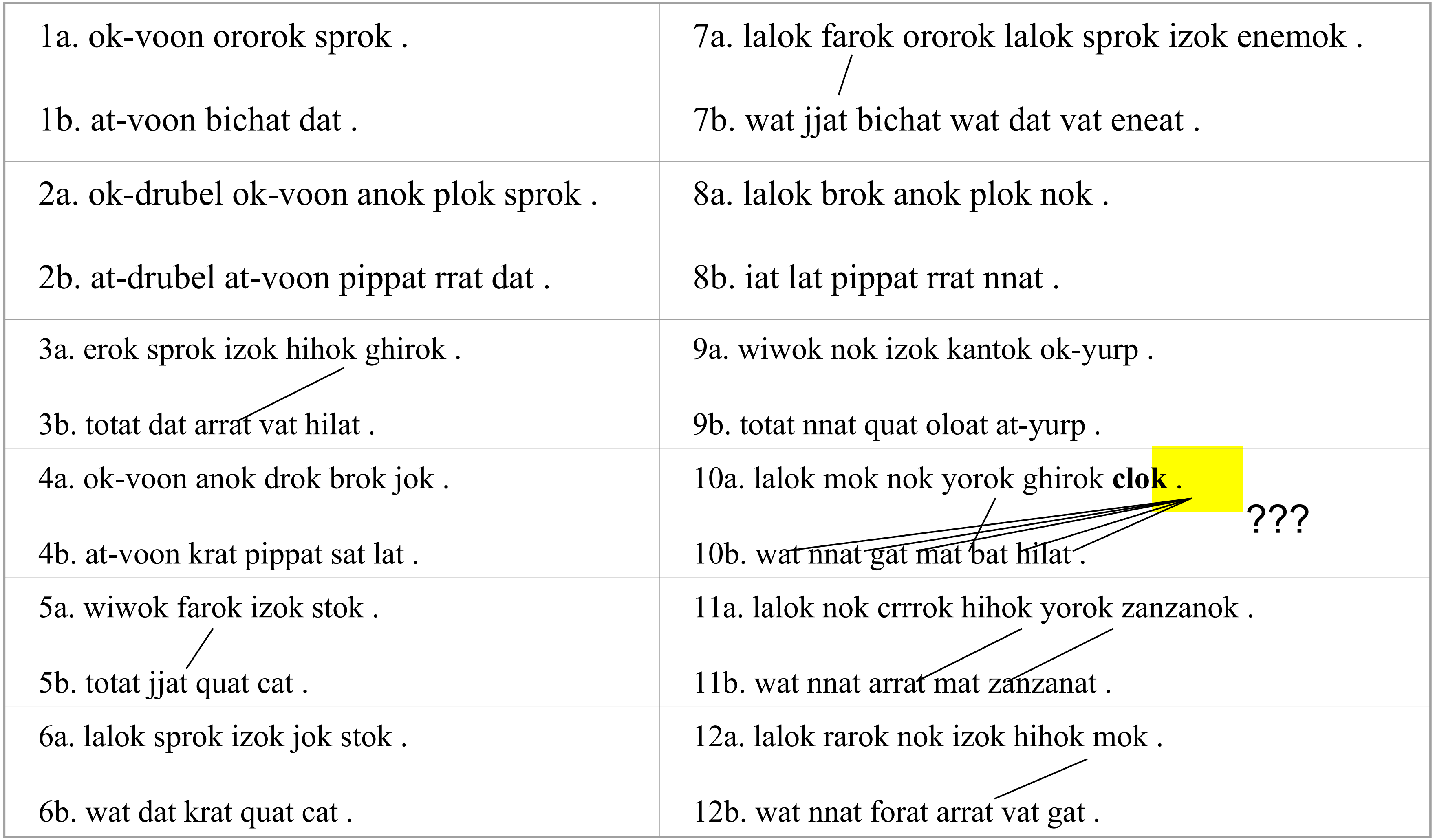
# Centauri/Arcturan [Knight, 1997]

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

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

# Centauri/Arcturan [Knight, 1997]

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

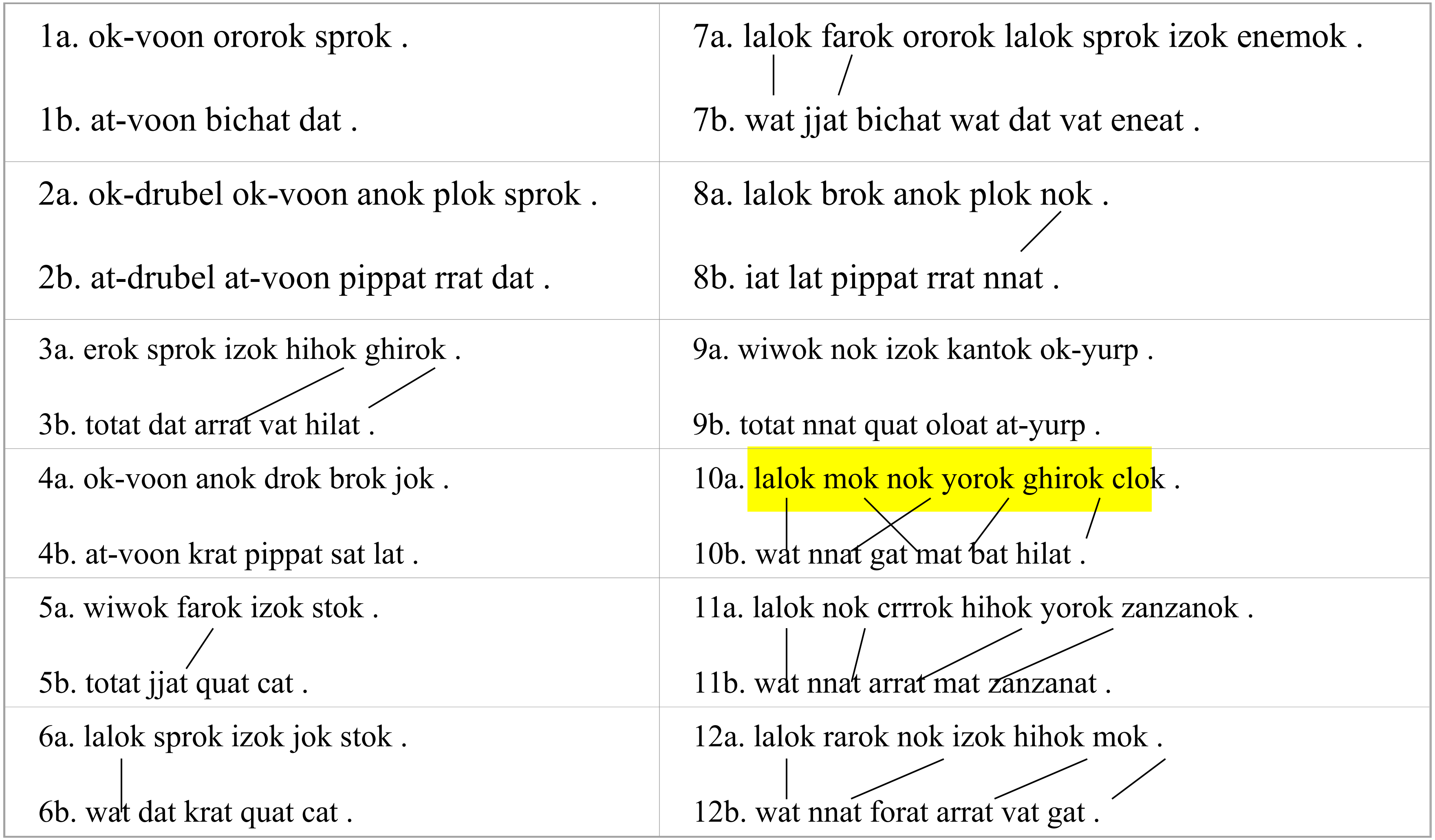
11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

# Centauri/Arcturan [Knight, 1997]

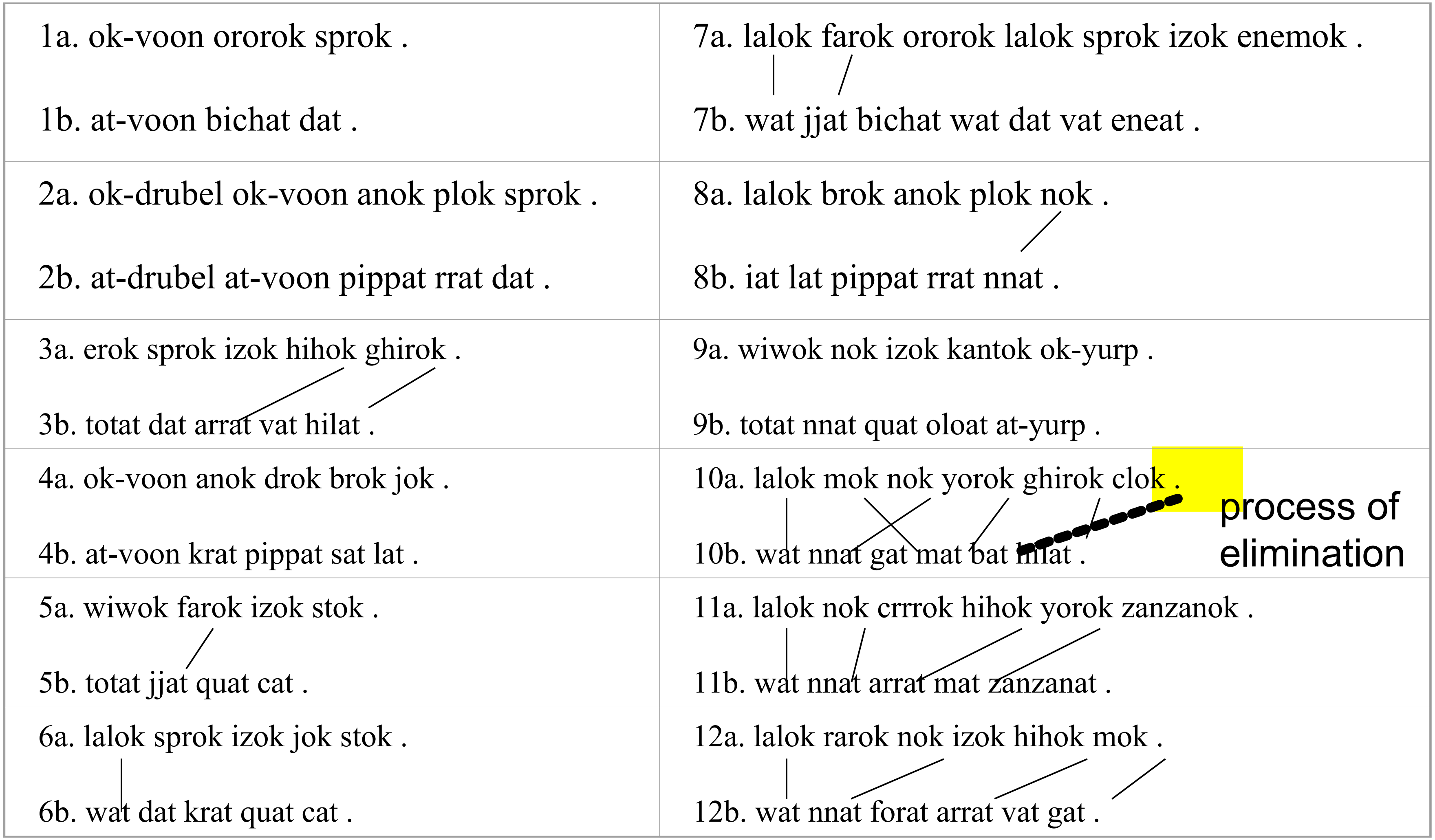
Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp



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

# Centauri/Arcturan [Knight, 1997]

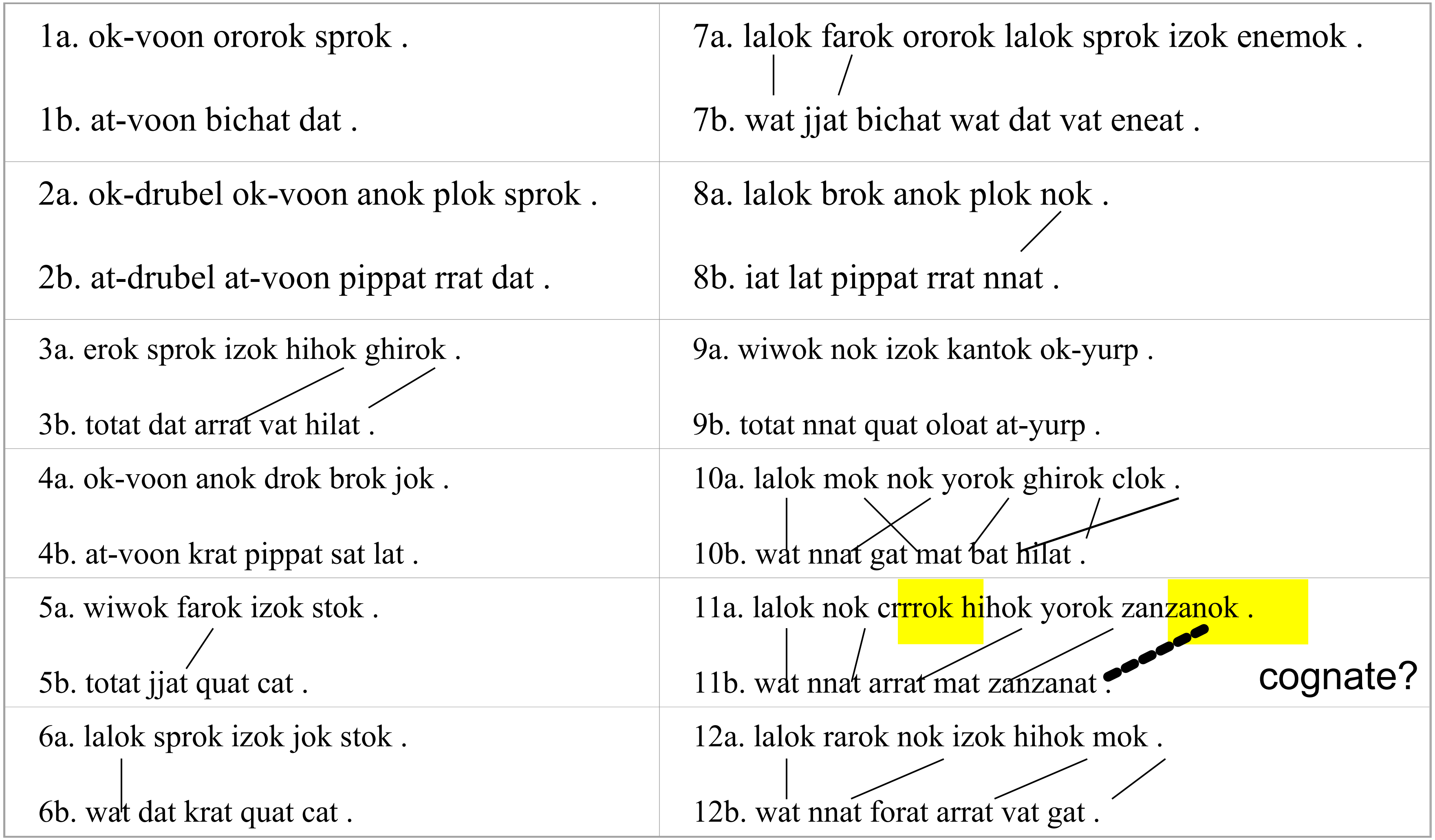
Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp



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

# Centauri/Arcturan [Knight, 1997]

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



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

# Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

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

# Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order:
{ **jjat, arrat, mat, bat, oloat, at-yurp** }

- There are 6! different orders possible, so 720 different translations.
- Best order (according to placement in TL side of the corpus is as given above):
  - Not just unigrams, but *n*-grams also …

# It’s Really Spanish—English!

**Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa**

1a. Garcia and associates .
1b. Garcia y asociados .

2a. Carlos Garcia has three associates .
2b. Carlos Garcia tiene tres asociados .

3a. his associates are not strong .
3b. sus asociados no son fuertes .

4a. Garcia has a company also .
4b. Garcia tambien tiene una empresa .

5a. its clients are angry .
5b. sus clientes estan enfadados .

6a. the associates are also angry .
6b. los asociados tambien estan enfadados .

7a. the clients and the associates are enemies .
7b. los clients y los asociados son enemigos .

8a. the company has three groups .
8b. la empresa tiene tres grupos .

9a. its groups are in Europe .
9b. sus grupos estan en Europa .

10a. the modern groups sell strong pharmaceuticals .
10b. los grupos modernos venden medicinas fuertes .

11a. the groups do not sell zenzanine .
11b. los grupos no venden zanzanina .

12a. the small groups are not modern .
12b. los grupos pequenos no son modernos .

# Some more to try …

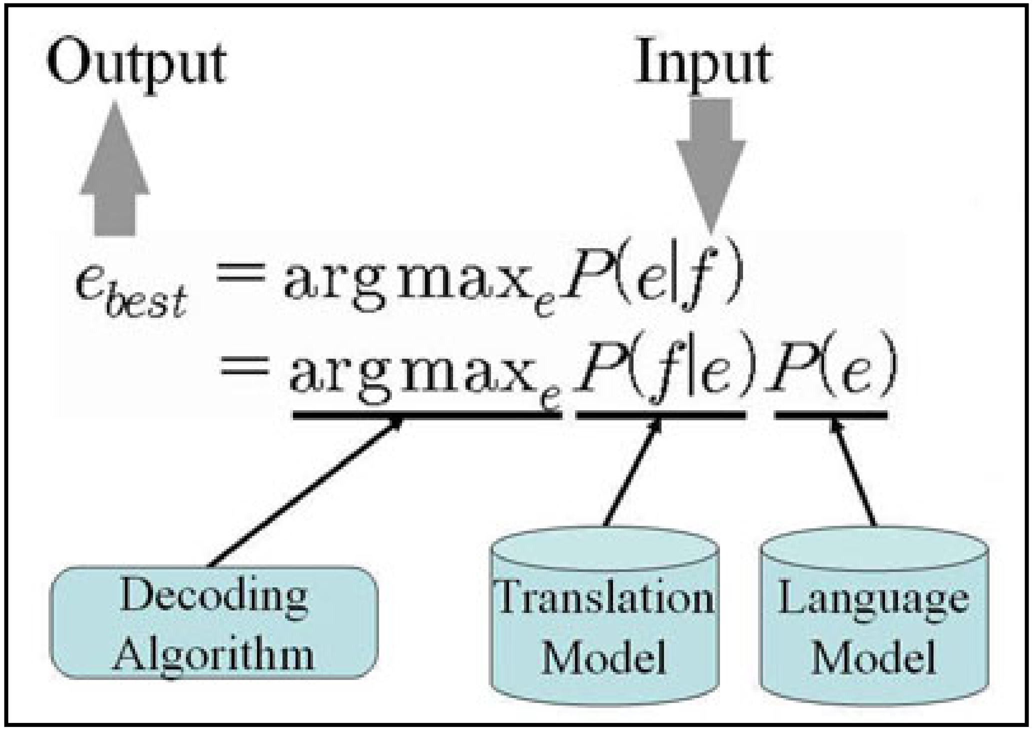
- iat lat pippat eneat hilat oloat at-yurp.
- totat nnat forat arrat mat bat.
- wat dat quat cat uskrat at-drubel.

# Some more to try …

- iat lat pippat eneat hilat oloat at-yurp.
- totat nnat forat arrat mat bat.
- wat dat quat cat uskrat at-drubel.

… if you have trouble sleeping at night!

# How does SMT Work?



$$e_{best} = \arg\max_e P(e|f)$$

$$= \underbrace{\arg\max_e}_{\text{Decoding Algorithm}} \underbrace{P(f|e)}_{\text{Translation Model}} \underbrace{P(e)}_{\text{Language Model}}$$

- No(t much) maths today …
- Instead:
  - The story of SMT in pictures …
  - It’s (mostly) all about the **Data** …

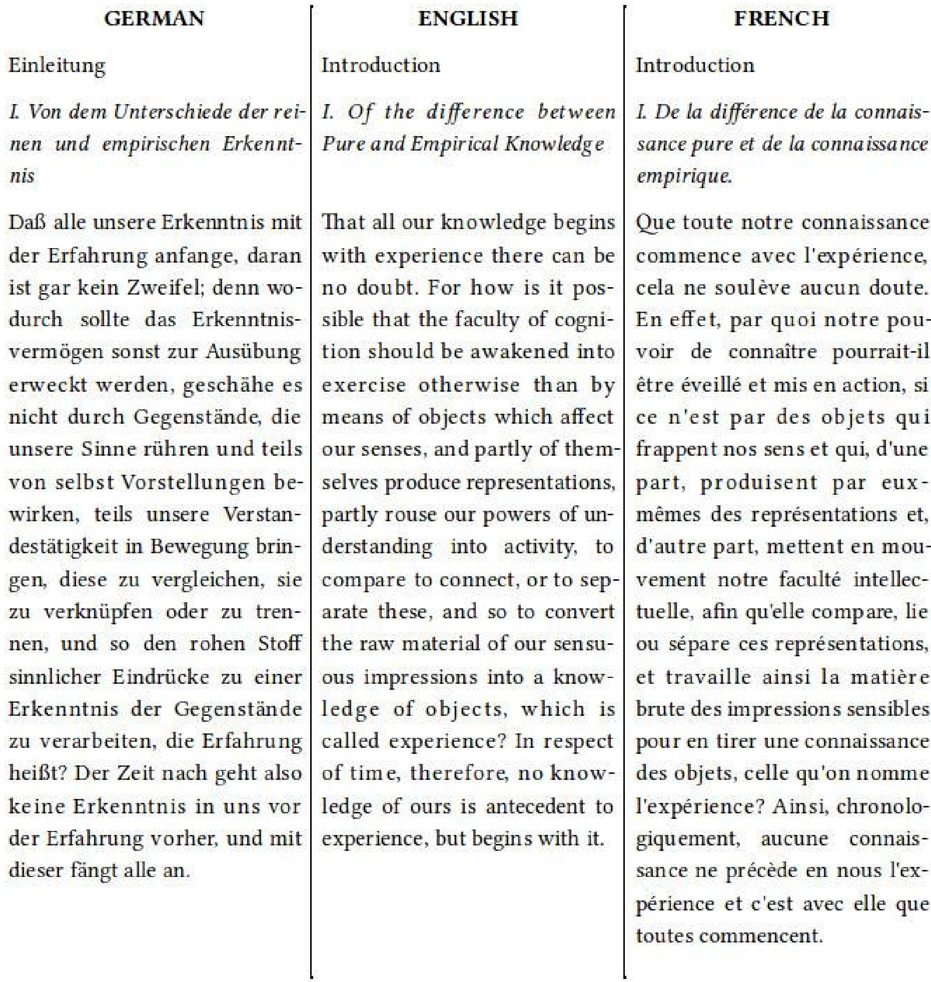
# How does SMT Work?

Statistical MT learns from data

Two kinds of data:

- Source documents and their human translations
- Target language collections

- 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 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.

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 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.

FRENCH

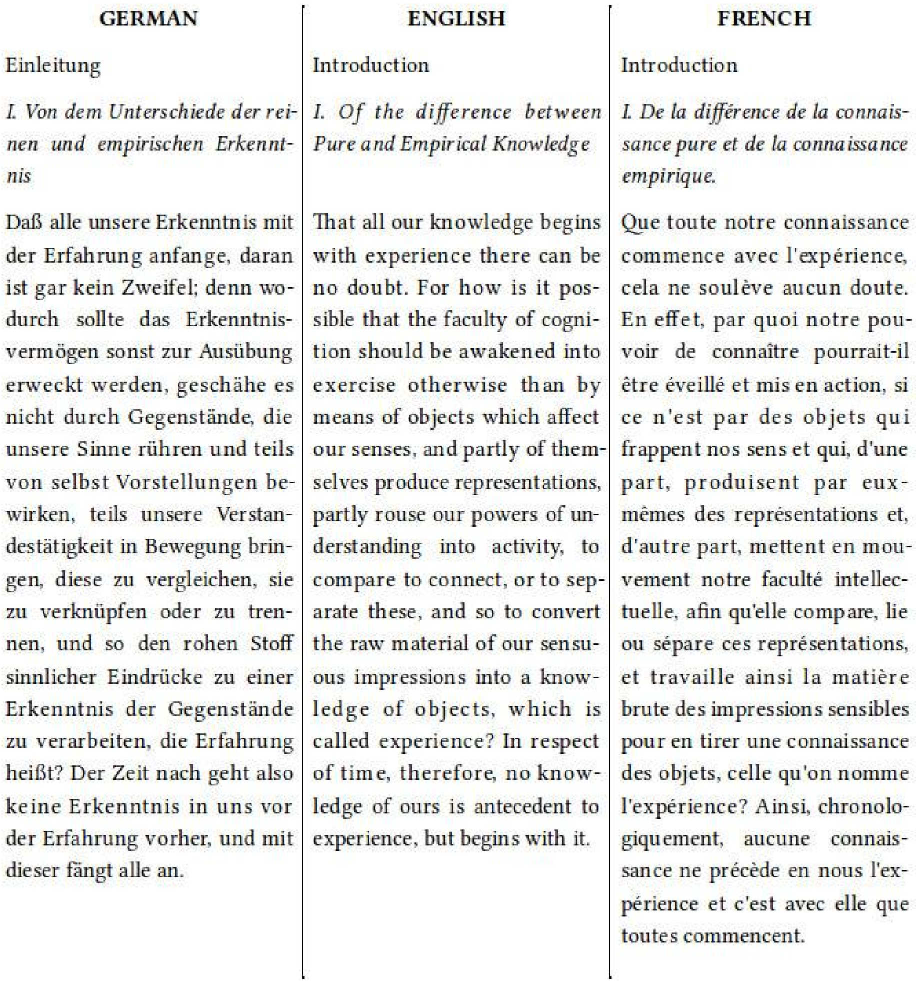
Introduction

*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.

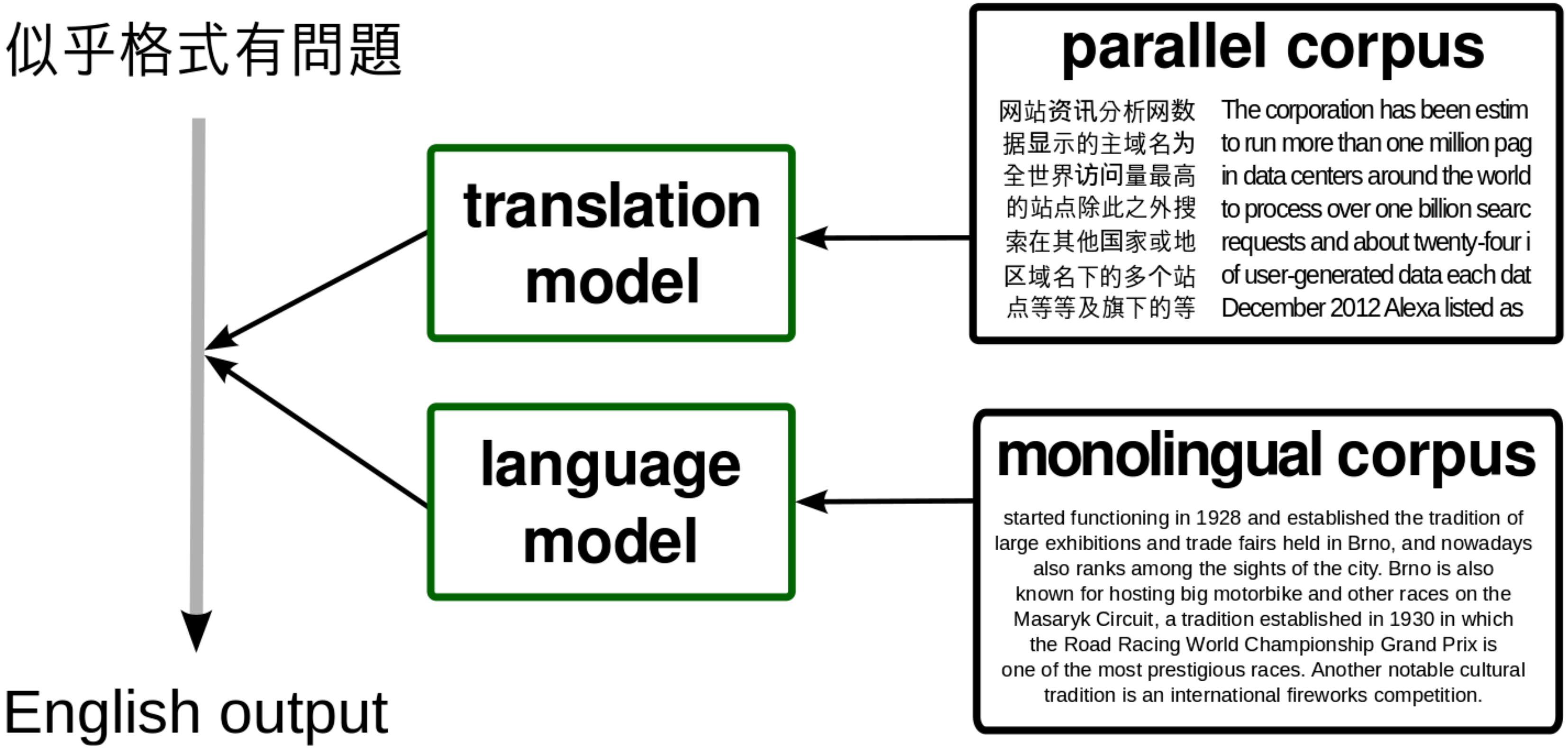
# What can/do we learn from Data?

- Which sentences translate as which: sentence alignment
- Which words translate as which: word alignment + translation probabilities => translation model
- What do good target sentences look 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. |

# How does SMT Work?



似乎格式有問題

translation model

language model

English output

**parallel corpus**

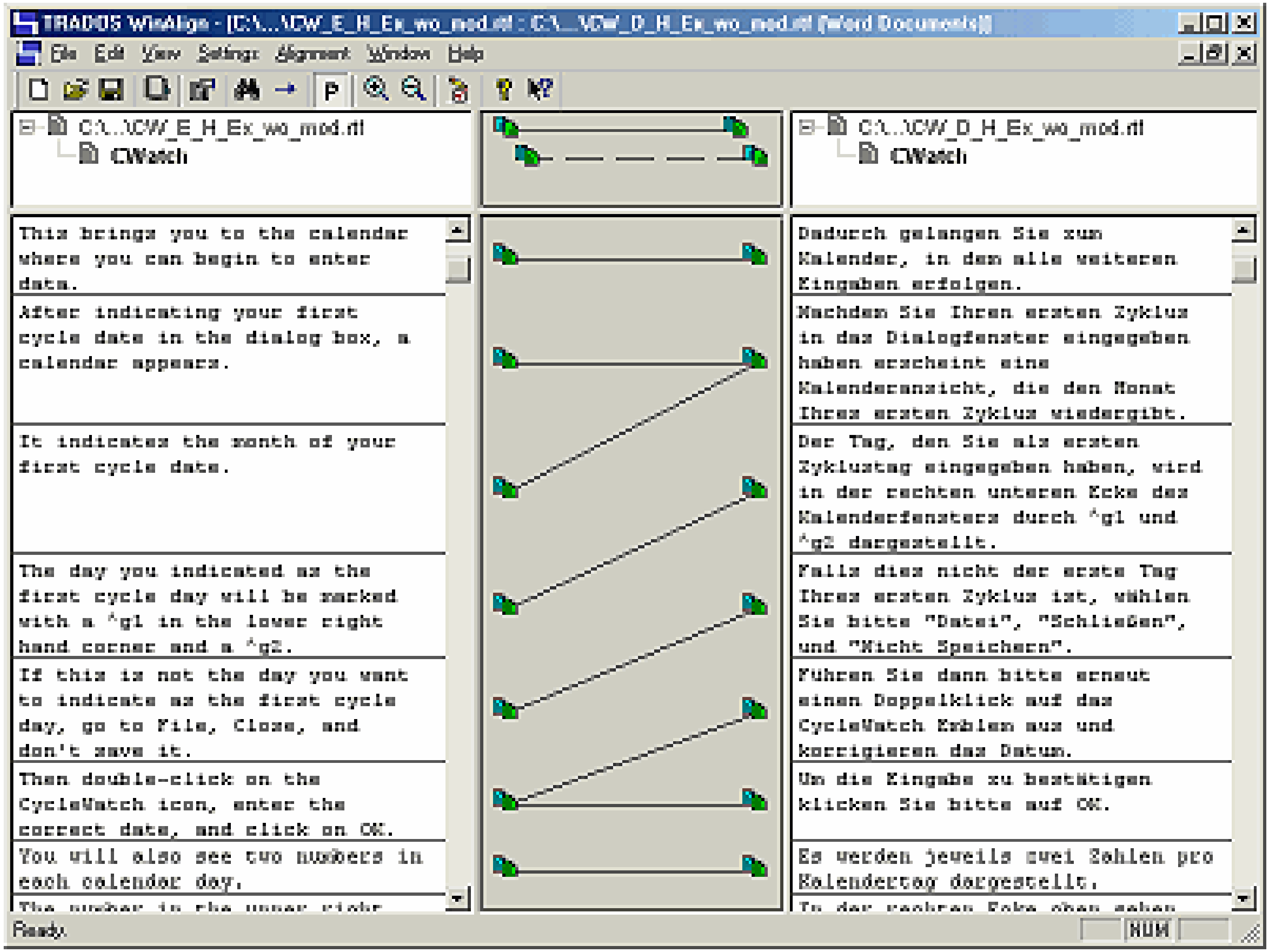
网站资讯分析网数
据显示的主域名为
全世界访问量最高
的站点除此之外搜
索在其他国家或地
区域名下的多个站
点等等及旗下的等

The corporation has been estim
to run more than one million pag
in data centers around the world
to process over one billion searc
requests and about twenty-four i
of user-generated data each dat
December 2012 Alexa listed as

**monolingual corpus**

started functioning in 1928 and established the tradition of large exhibitions and trade fairs held in Brno, and nowadays also ranks among the sights of the city. Brno is also known for hosting big motorbike and other races on the Masaryk Circuit, a tradition established in 1930 in which the Road Racing World Championship Grand Prix is one of the most prestigious races. Another notable cultural tradition is an international fireworks competition.

# Sentence Alignment

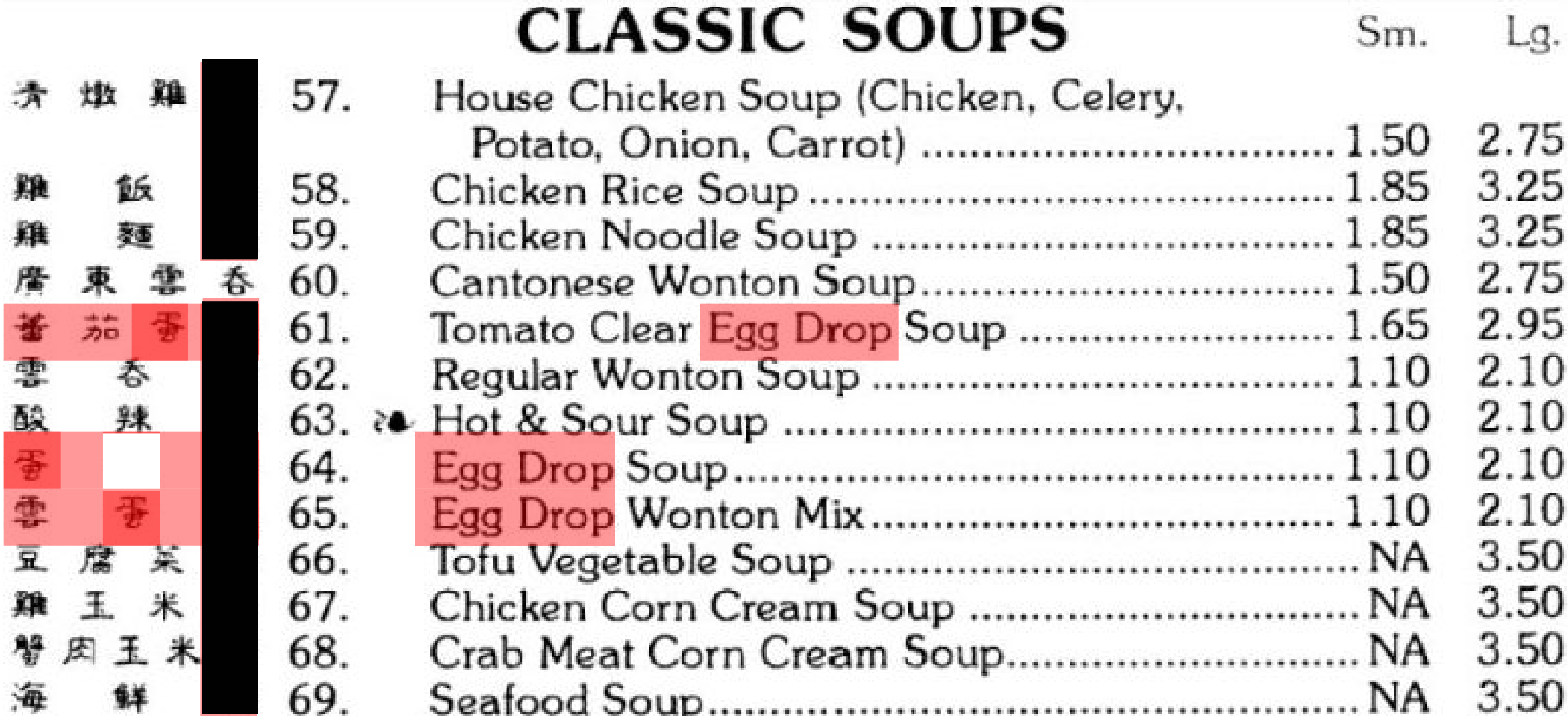


TRADOS WinAlign - [C:\...\CW_E_H_Ex_wo_mod.rtf : C:\...\CW_D_H_Ex_wo_mod.rtf (Word Documents)]
File Edit View Settings Alignment Window Help
P
C:\...\CW_E_H_Ex_wo_mod.rtf
CWatch
C:\...\CW_D_H_Ex_wo_mod.rtf
CWatch
Ready.
NUM

| | |
|---|---|
| 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 |

Let’s try it for a language pair that *someone* in the class might know …

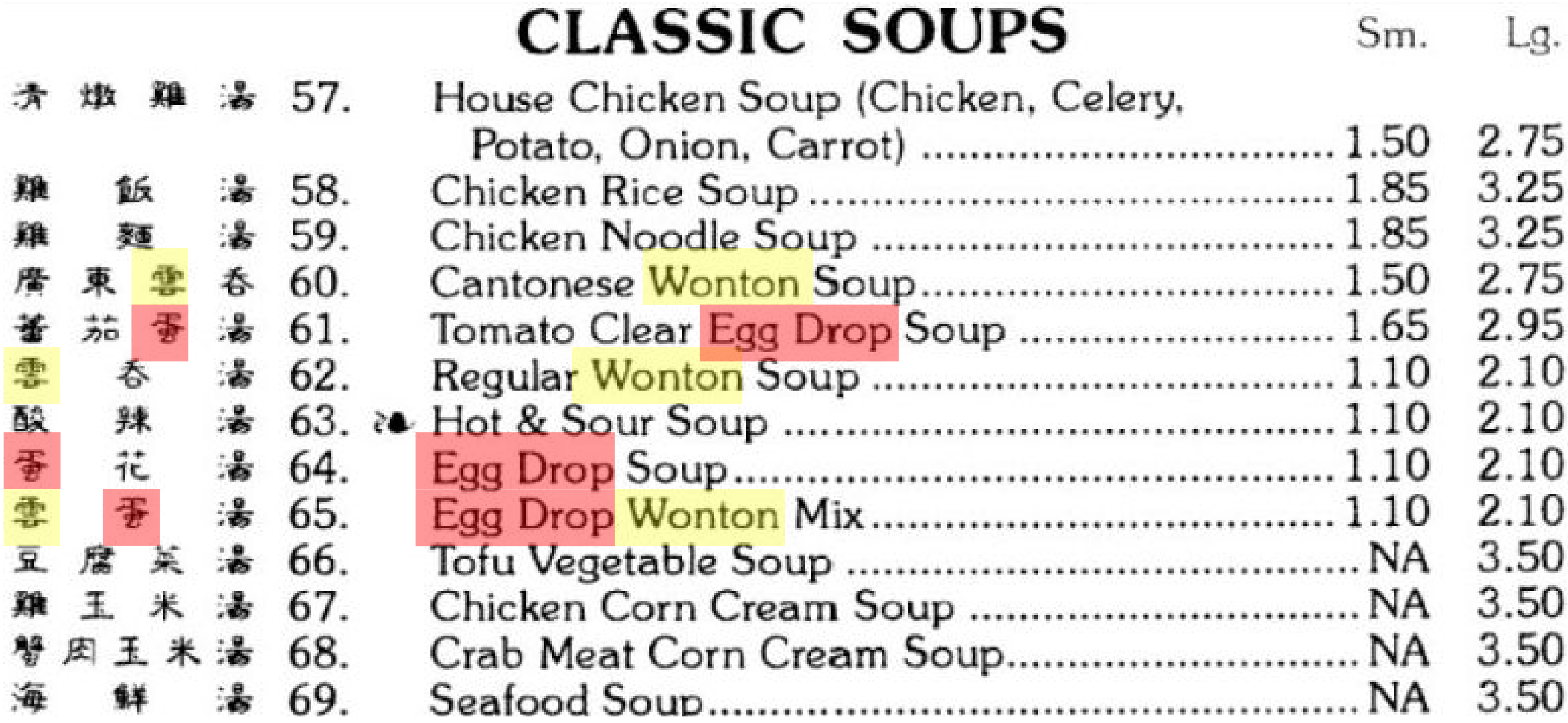
# Word Alignment



## CLASSIC SOUPS

| | | | | Sm. | Lg. |
|---|---|---|---|---|---|
| 清燉雞 | [redacted] | 57. | House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot) | 1.50 | 2.75 |
| 雞飯 | [redacted] | 58. | Chicken Rice Soup | 1.85 | 3.25 |
| 雞麵 | [redacted] | 59. | Chicken Noodle Soup | 1.85 | 3.25 |
| 廣東雲 | 吞 | 60. | Cantonese Wonton Soup | 1.50 | 2.75 |
| 蕃茄蛋 | [redacted] | 61. | Tomato Clear Egg Drop Soup | 1.65 | 2.95 |
| 雲吞 | [redacted] | 62. | Regular Wonton Soup | 1.10 | 2.10 |
| 酸辣 | [redacted] | 63. | Hot & Sour Soup | 1.10 | 2.10 |
| 蛋 | [redacted] | 64. | Egg Drop Soup | 1.10 | 2.10 |
| 雲蛋 | [redacted] | 65. | Egg Drop Wonton Mix | 1.10 | 2.10 |
| 豆腐菜 | [redacted] | 66. | Tofu Vegetable Soup | NA | 3.50 |
| 雞玉米 | [redacted] | 67. | Chicken Corn Cream Soup | NA | 3.50 |
| 蟹肉玉米 | [redacted] | 68. | Crab Meat Corn Cream Soup | NA | 3.50 |
| 海鮮 | [redacted] | 69. | Seafood Soup | NA | 3.50 |

# Word Alignment

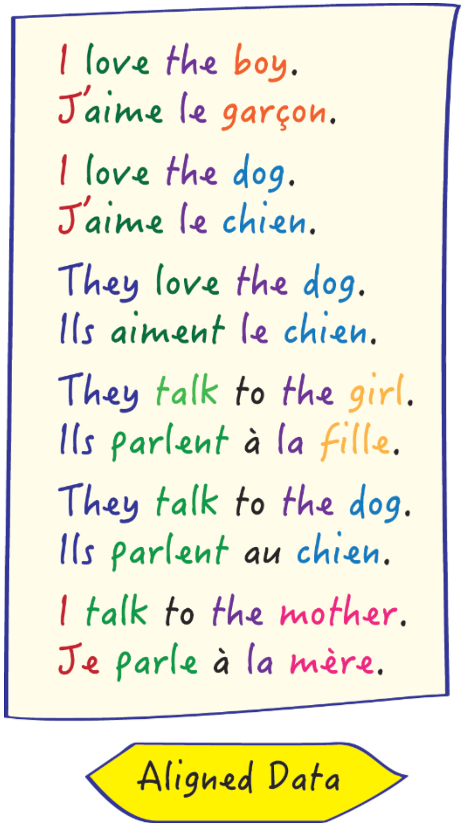


**CLASSIC SOUPS**

| | | | 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 |

Let’s try it for a language pair that *more of you* in the class might know …

# Statistical MT



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

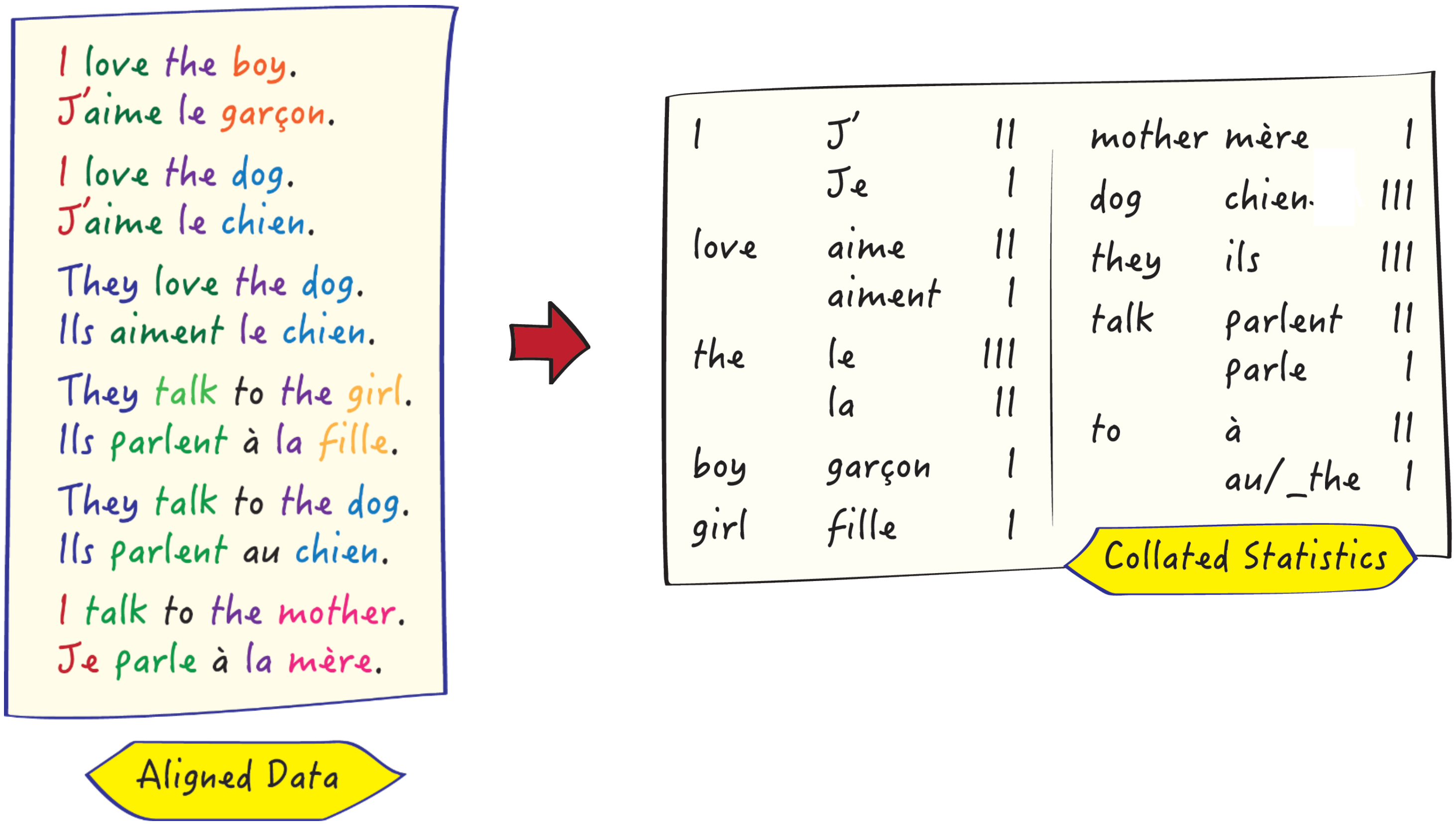
They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

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

Aligned Data

# Statistical MT



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.

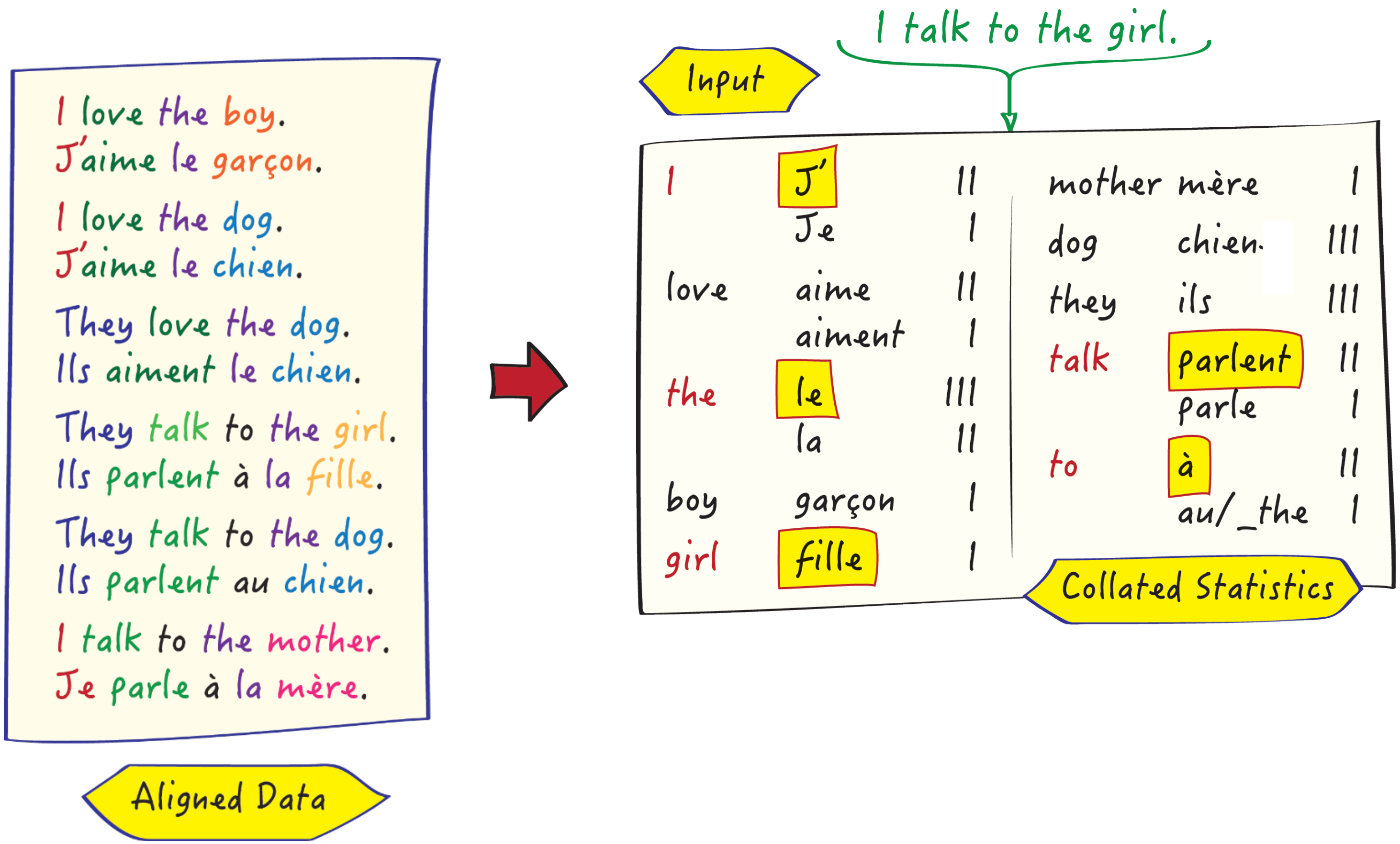
I talk to the mother.
Je parle à la mère.

Aligned Data

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

Collated Statistics

# Statistical MT



I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

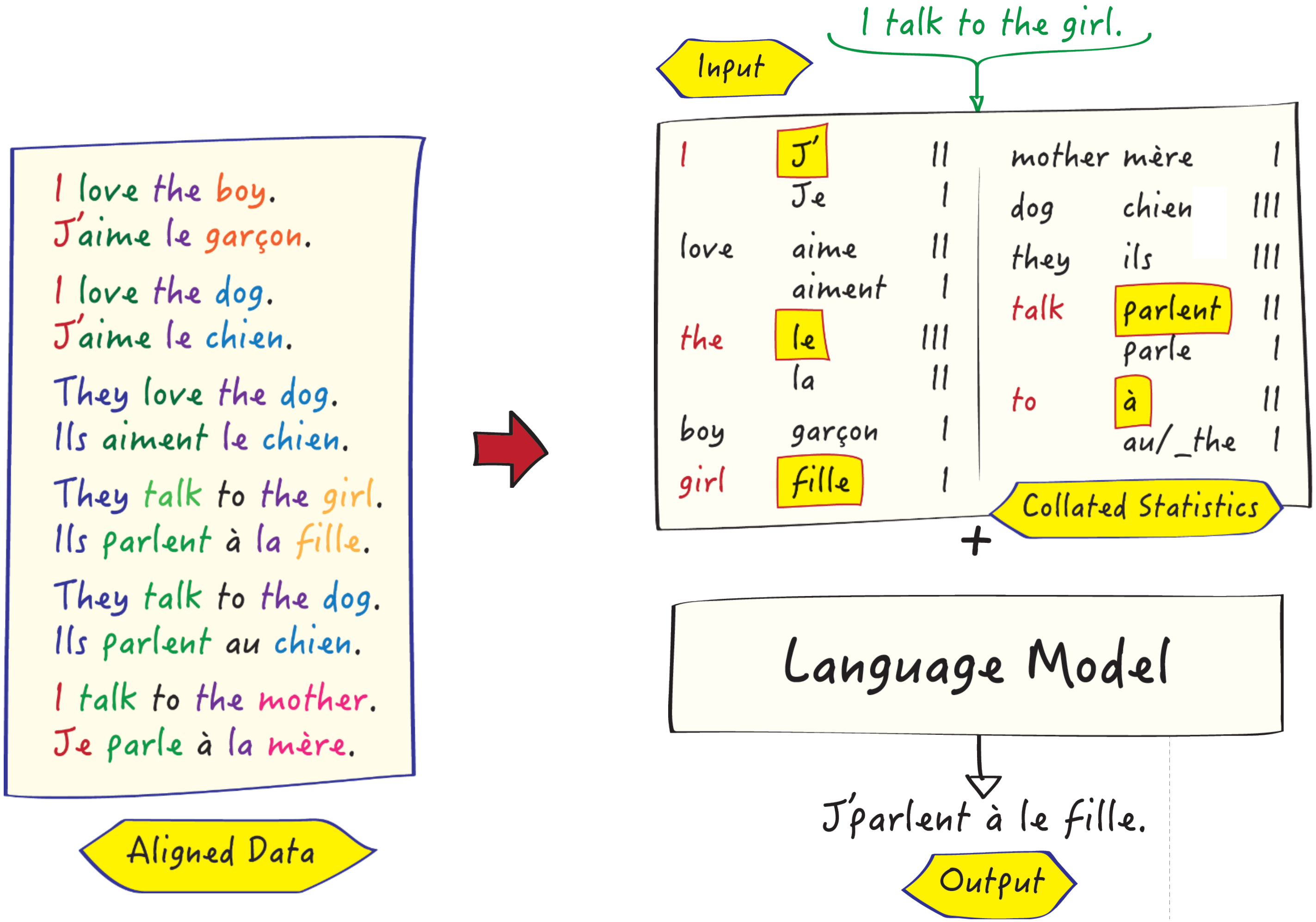
They talk to the girl.
Ils parlent à la fille.

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 | chien | III |
| they | ils | III |
| talk | parlent | II |
| | parle | I |
| to | à | II |
| | au/_the | I |

# Statistical MT



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.

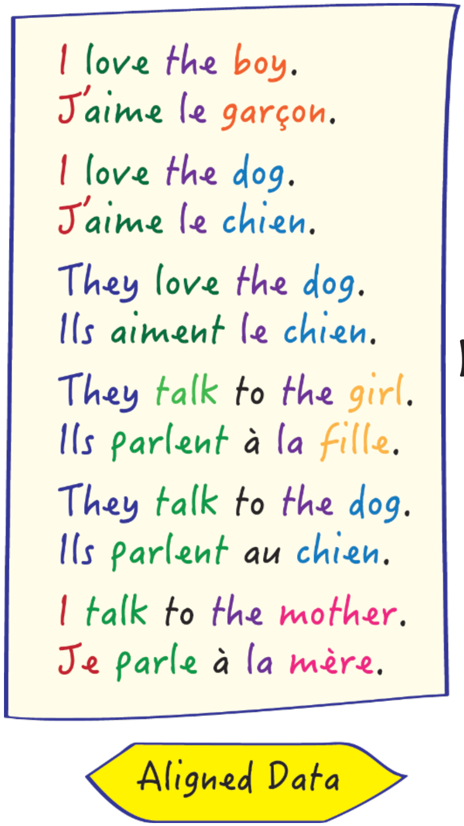
They talk to the dog.
Ils parlent au chien.

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

| | | | | | |
|---|---|---|---|---|---|
| I | J' | II | mother | mère | I |
| | Je | I | dog | chien | 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 MT



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 | |
| J’ | parlentau | | 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.

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

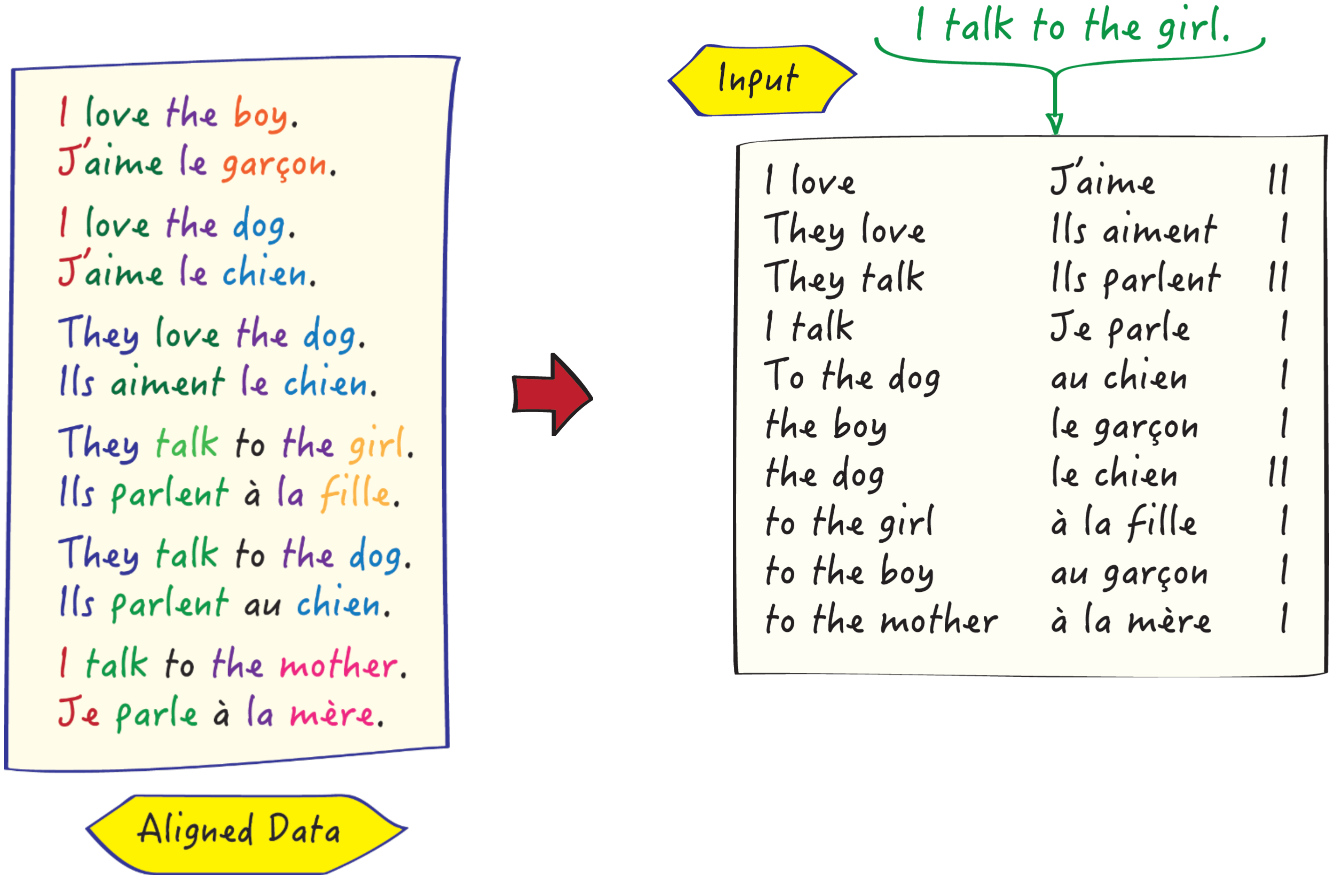
# Phrase-Based SMT

- 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 : à la fille
  - I talk : Je parle

# Phrase-Based SMT



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.

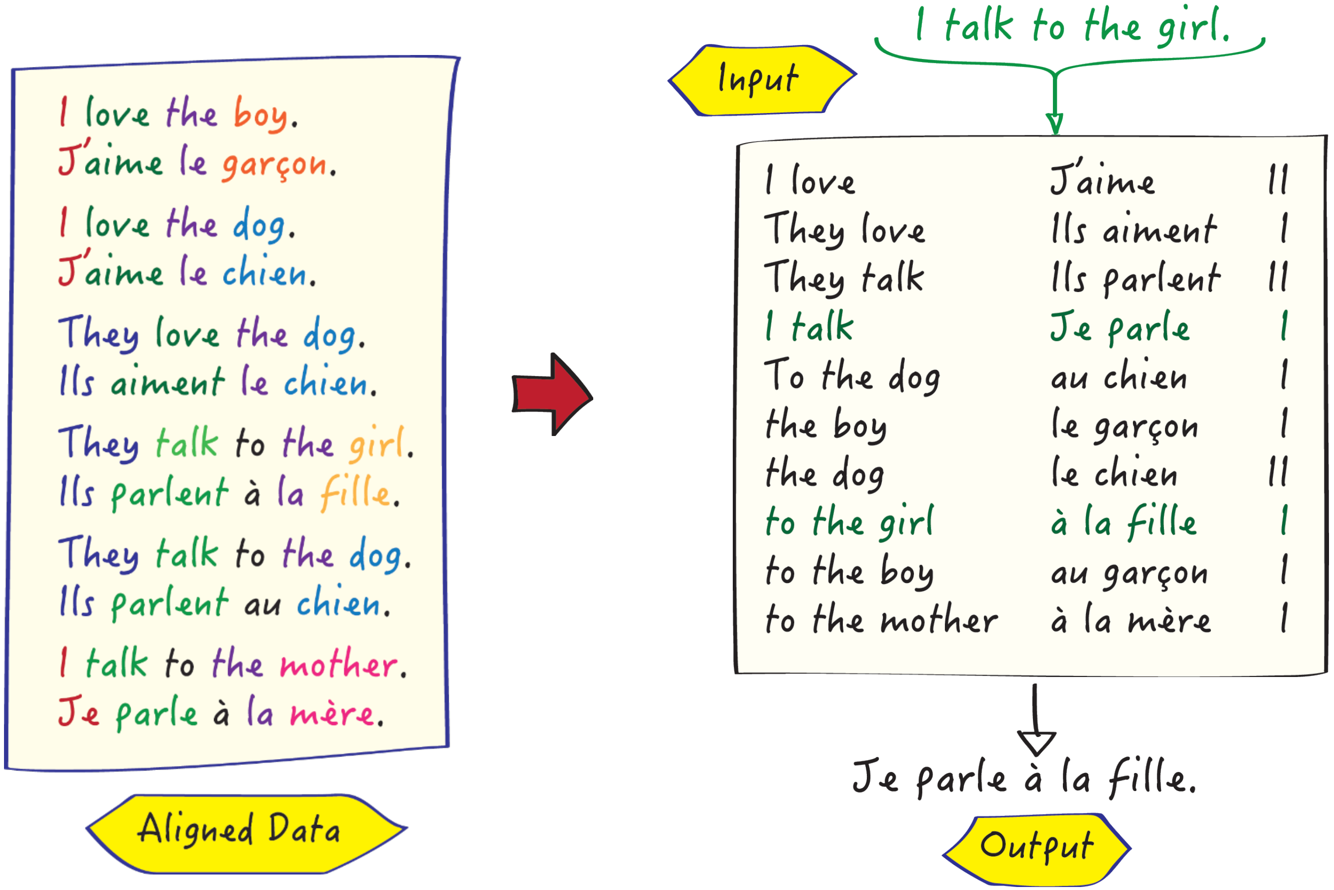
They talk to the dog.
Ils parlent au chien.

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

I talk to the girl.

| | | |
|---|---|---|
| I 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 SMT



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 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 |

Je parle à la fille.

# Phrase-Based SMT

- *Much* better than word-based SMT!
- (Was) standard technology: Google, Microsoft, Baidu, global Localisation & Translation industry

- Moses open-source PB-SMT toolkit
- Most widely used SMT platform
- Research funded by EC
- (Was) used by EC DGT's MT@EC

MOSES CORE

# What did we learn?

- what parallel corpora look like (more on this soon);
- viewing parallel corpora through the ‘eyes’ of a computer;
- how relevant  parallel corpora are for MT;
- how to build bilingual dictionaries from parallel corpora;
- how cognate information may be useful in MT;
- how to do word alignment;
- about the ‘chicken-and-egg’ nature of dictionaries (which enable word alignments) and word alignments (which enable dictionary writing) …

# What else do we need to know?

- about word alignment and dictionary writing on a larger scale;
- about phrasal alignment, the norm in real translation data;
- about unalignable words;
- the importance of knowing the target language (vs. source) in making fluent translations;
- the importance of short sentence pairs (where alignment possibilities are restricted) in helping disambiguate/align longer sentence pairs;
- about locality in word order shifts;
- how to guess the meanings/translations of unknown words;
- about how much uncertainty the machine faces in working with limited data;
- …

# Can such methods be scaled to ‘real’ MT?

- Availability of monolingual and bilingual corpora?
- Possibility of sentence-aligning bilingual corpora?
- Can we write an algorithm to extract the translation dictionary?
- Can we write an algorithm to extract the monolingual word pair counts?
- Can we write an algorithm to generate translations using our translation dictionary and word pair counts?

- WILL THE TRANSLATIONS PRODUCED BE ANY GOOD?

# Parallel Corpora

- Hugely important … but not available in a wide range of language pairs:
  - Chinese—English: Hong Kong data
  - French—English: Canadian Hansards
  - Older EU pairs: Europarl [Koehn 04]
  - Newer EU pairs: JRC-Acquis Communautaire
  - Arabic—English: LDC Data
  - NIST, IWSLT, TC-STAR Evaluations

# Good Quality Language & Translation Models

- Any statistical approach to MT requires the availability of aligned bilingual corpora which are:
    - large;
    - good-quality;
    - representative.

# Corpus 1

Mary and John have two children.
The children that Mary and John have are aged 3 and 4.
John has blue eyes.

Question 1: what’s P(have) vs. P(has) in a corpus?

Question 2: what’s P(have | John) vs. P(has | John) in a corpus?

Question 3: what’s P(have) vs. P(has) in *this* corpus? What's their *relative* probability?

Question 4: what’s P(have | John) vs. P(has | John) in *this* corpus?

# Corpus 2

Am I right, or am I wrong?
Peter and I are seldom wrong.
I am sometimes right.
Sam and I are often mistaken.

Question 5: What two generalisations would a probabilistic language model (based on *bigrams*, say) infer from this data, which are not true of English as a whole? Are there any other generalisations that could be inferred?

Question 6: Try to think of some trigrams (and 4-grams, if you can) that cannot be ‘discovered’ by a bigram model? What you’re looking for here is a phrase where the third (or subsequent) word depends on the first word, which in a bigram model is ‘too far away’ ...

# Some Observations

- Note that all the sentences in these corpora are well-formed.
- If, on the other hand, the corpus contains ill-formed input, then that too will skew our probability models …

… and our translations will be affected!

# Corpus 1 Revisited

- Using Google, I got:
  - # 'have' = 380,000,000
  - # 'has' = 244,000,000
  - # 'John has' = 227,000
  - # 'John have' = 25,700
- Revisit the Questions and calculate the *actual* probabilities! How accurate/inaccurate were the original models that we derived?

# Corpus 2 Revisited

- Using Google, I got:
    - # ‘am I’ = 3,690,000
    - # ‘I am’ = 8,060,000
    - # ‘I are’ = 1,230,000
- Revisit the Questions and calculate the *actual* probabilities! How accurate/inaccurate were the original models that we derived?

# Bilingual Corpora

All this applies to bitexts too!

Q: of what English word are these possible French translations (from the *Canadian Hansards*, note)?

Q: what's ???

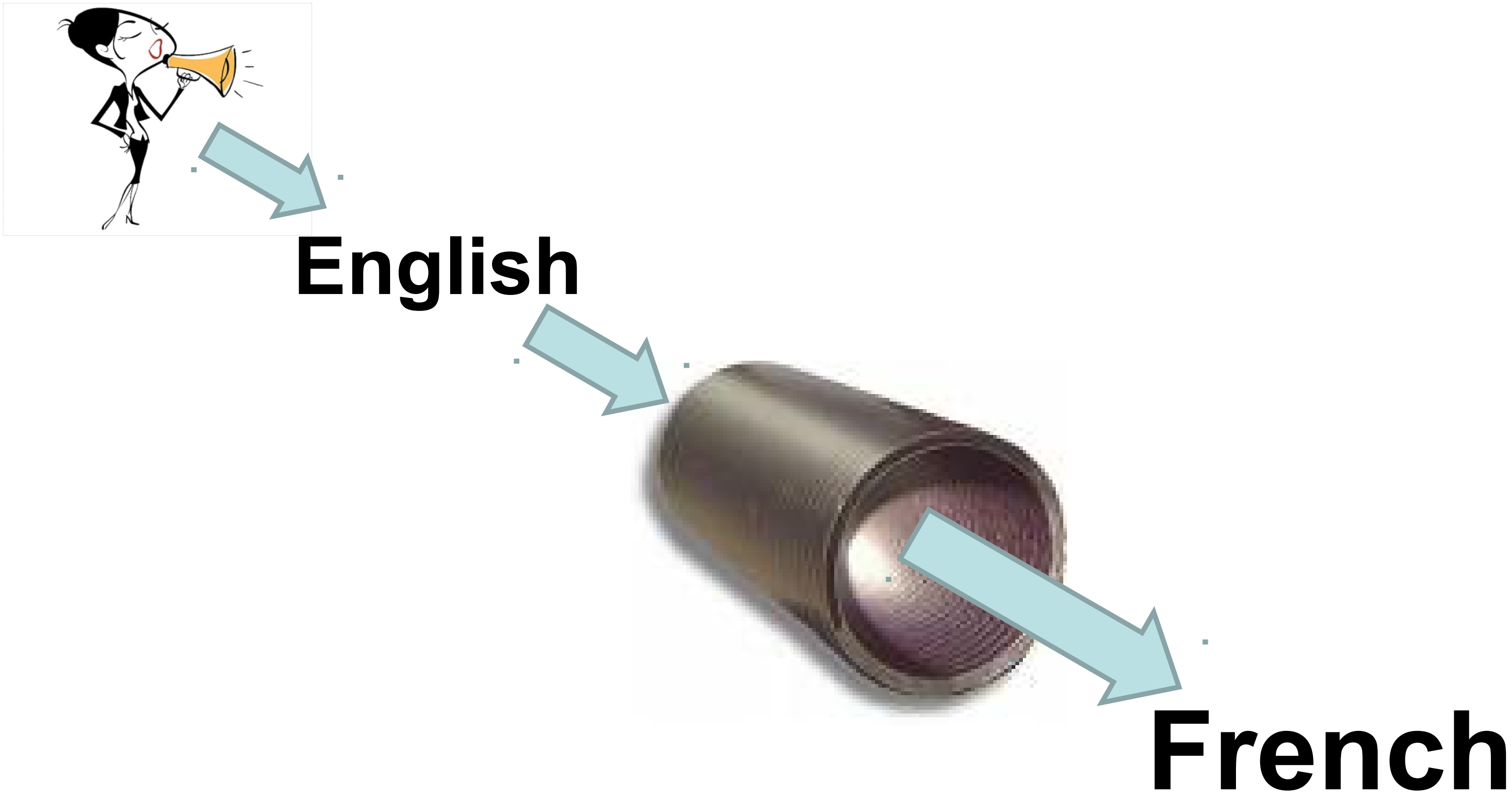
| **French** | **Probability** |
|---|---|
| ??? | .808 |
| entendre | .079 |
| entendu | .026 |
| entends | .024 |
| entendons | .013 |

# *Caveat interpres*!

- Beware of sparse data!
- Beware of unrepresentative corpora!
- Beware of poor quality language!

If the corpora are small, or of poor quality, or are unrepresentative, then our statistical language models will be poor, so any results we achieve will be poor.

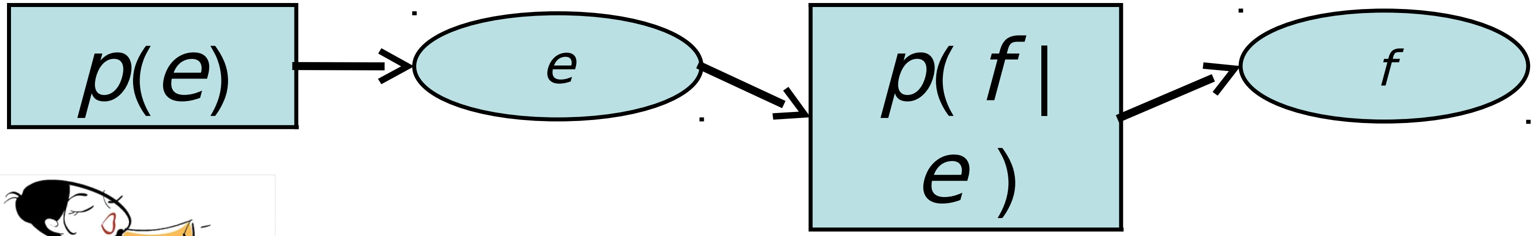
# Noisy Channel Framework



English

French

# Noisy Channel Framework

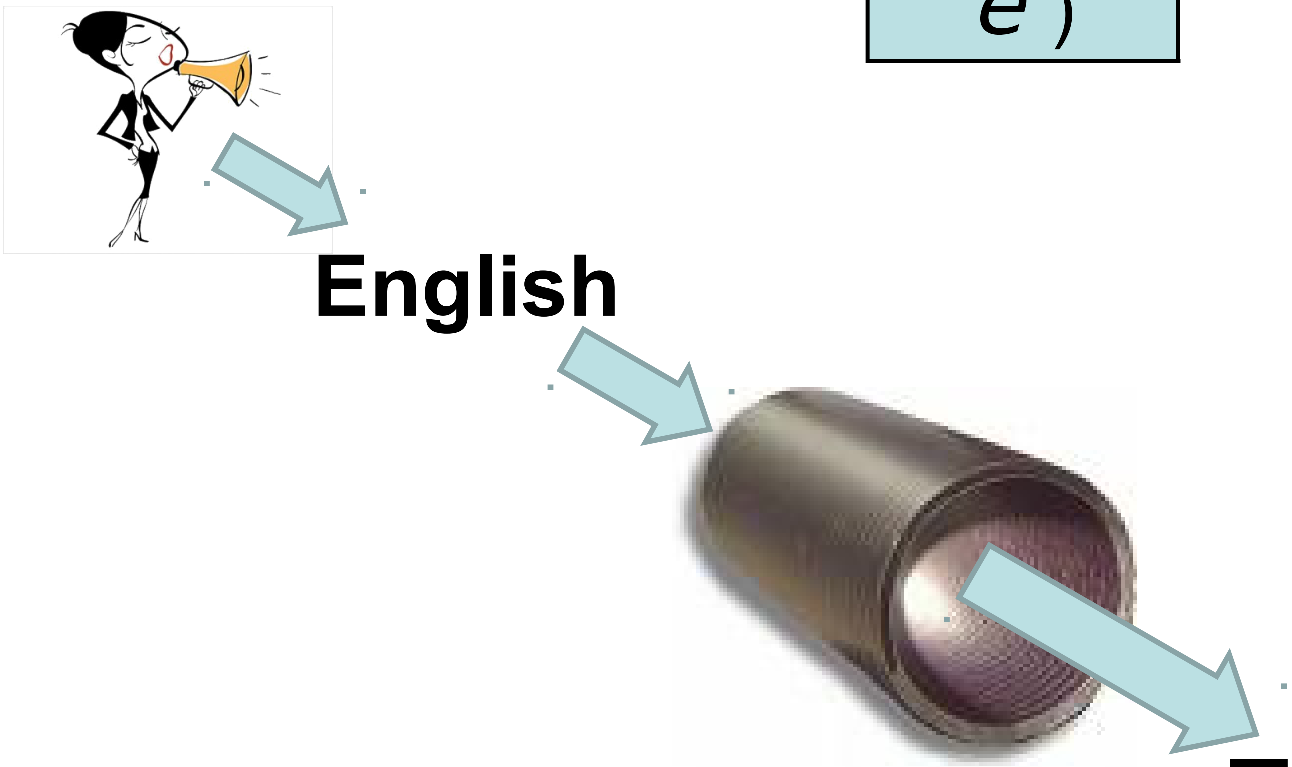


$p(e)$

$e$

$p(f \mid e)$

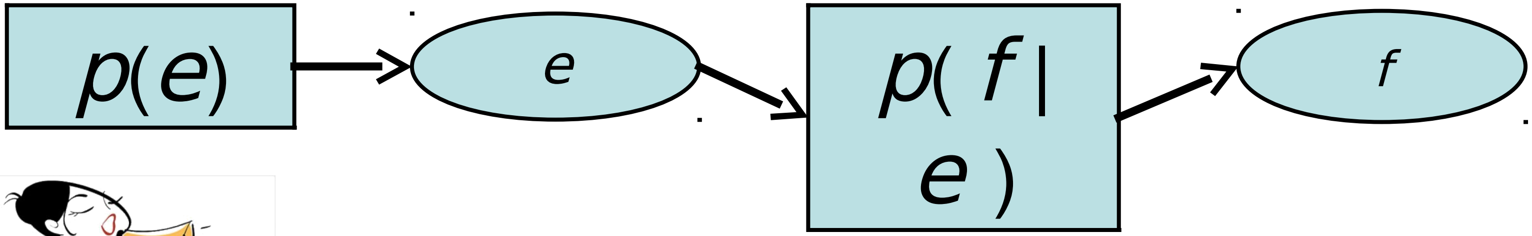
$f$



**English**

**French**

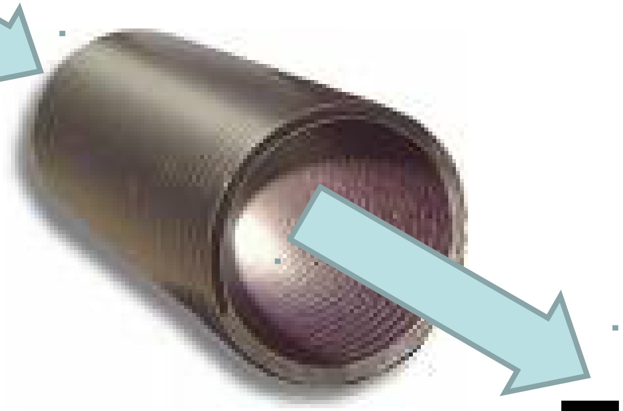
# Noisy Channel Framework



$p(e)$ → $e$ → $p(f \mid e)$ → $f$

$$p(f) = p(e)p(f \mid e)$$

**English**



**French**

# Noisy Channel Framework

Applying Bayes' Rule, we have:

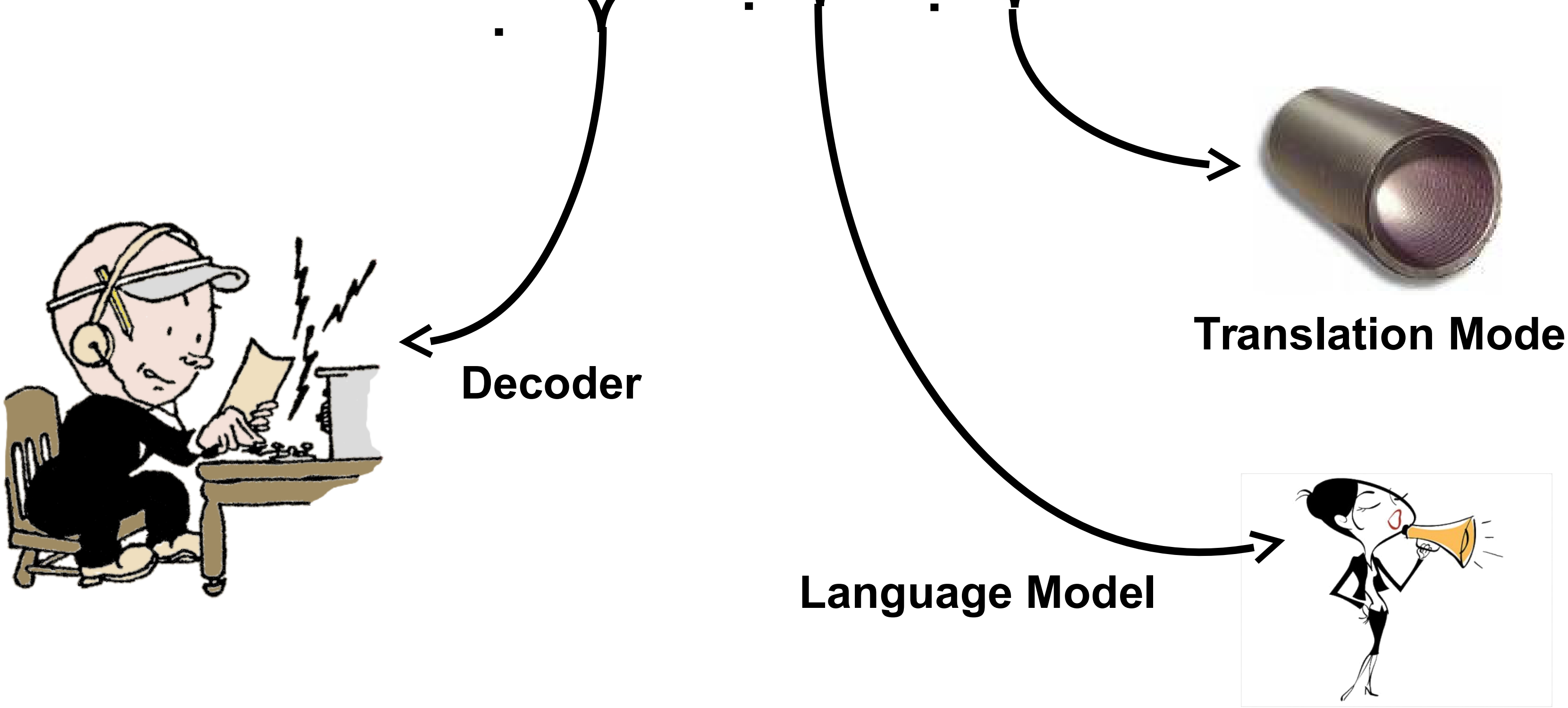
$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

Thus:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f) = \underset{e}{\operatorname{argmax}}\, p(e)p(f|e)$$

# SMT Components

$$\hat{e} = \underbrace{\operatorname*{argmax}_{e}}\,\underbrace{p(e)}\,\underbrace{p(f|e)}$$



Decoder

Translation Model

Language Model

# Noisy Channel Framework

- The *translation model* models how likely it is that $f$ is a translation of $e$ – adequacy.
- The *language model* models how likely it is that $e$ is an acceptable sentence – fluency.
- The *decoder* searches for the most likely $e$.

# Fluency versus Adequacy

(towards MT Evaluation)

Source Sentence:

Le chat entre dans la chambre.

- Adequate fluent translation:

  The cat enters the bedroom.

- Adequate disfluent translation:

  The cat enters in the bedroom.

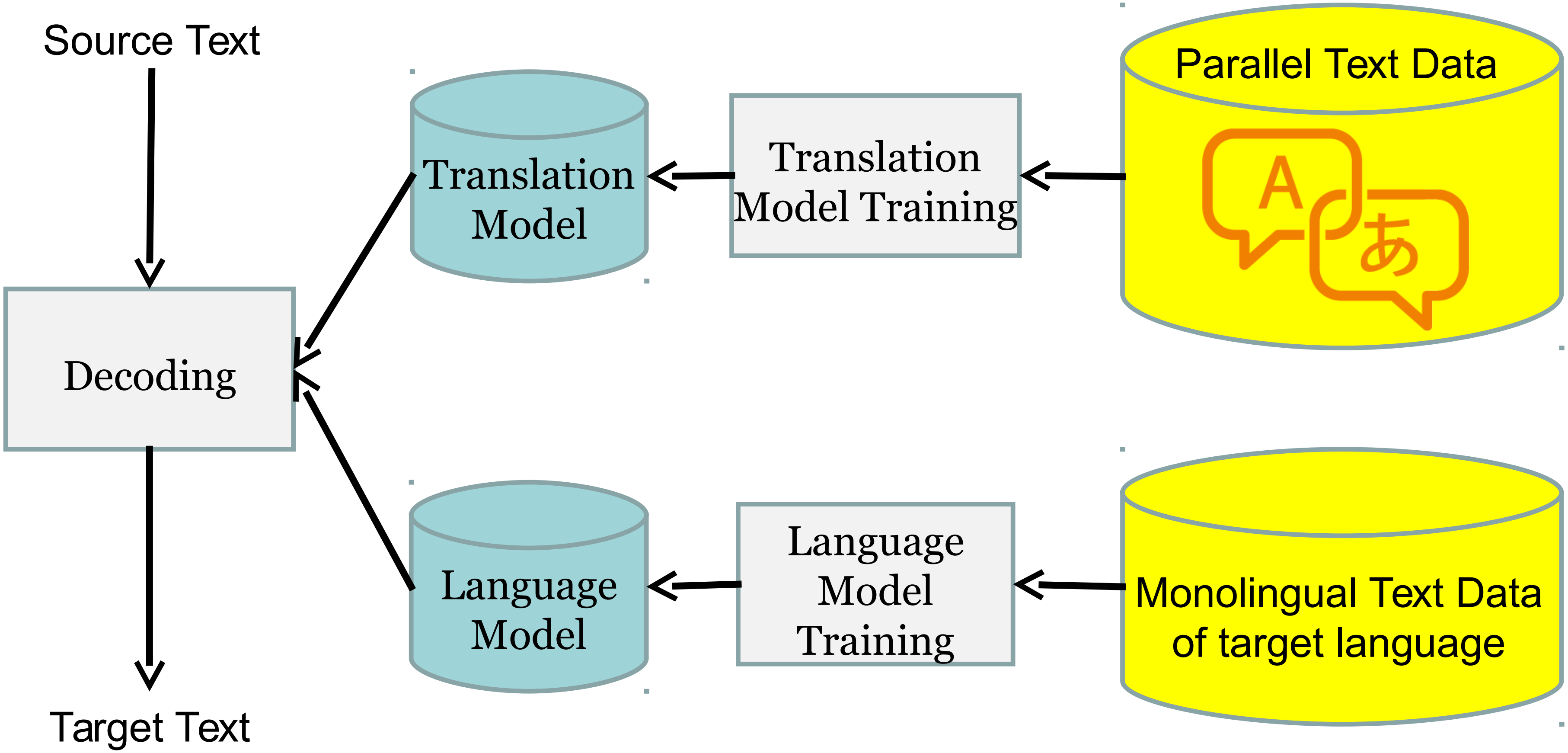
- Fluent inadequate translation:

  My Granny plays the piano.

- Disfluent inadequate translation:

  piano Granny the plays My

# SMT Flow



Source Text

Decoding

Target Text

Translation Model

Translation Model Training

Parallel Text Data

Language Model

Language Model Training

Monolingual Text Data of target language

# OK, so that’s set the scene

I hope that’s enough to
get you
started/interested in
SMT (and soon NMT) …

THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
BÏYAN
SHUKRIA
BOLZÏN
MERCI
GRAZIE
MEHRBANI
JUSPAXAR
GOZAIMASHITA
EFCHARISTO
PALDIES