CS447: Natural Language Processing

*http://courses.engr.illinois.edu/cs447*

# Lecture 21: Machine Translation

Julia Hockenmaier

*juliahmr@illinois.edu*

3324 Siebel Center

# Machine Translation in 2018

## 习近平在上海考察

2018-11-07 19:41:21　来源：新华网

**习近平在上海考察时强调**

**坚定改革开放再出发信心和决心**

**加快提升城市能级和核心竞争力**

新华社上海11月7日电　中共中央总书记、国家主席、中央军委主席习近平近日在上海考察时强调，坚持以新时代中国特色社会主义思想为指导，坚决贯彻落实党中央决策部署，坚定改革开放再出发的信心和决心，坚持稳中求进工作总基调，全面贯彻新发展理念，坚持以供给侧结构性改革为主线，加快建设现代化经济体系，打好三大攻坚战，加快提升城市能级和核心竞争力，更好为全国改革发展大局服务。

## Xi Jinping inspected in Shanghai

**Xi Jinping stressed during his visit to Shanghai**

**Strengthening reform and opening up and starting to build confidence and determination**

**Accelerate the improvement of urban energy level and core competitiveness**

Xinhua News Agency, Shanghai, November 7th, Xi Jinping, general secretary of the CPC Central Committee, president of the State Council and chairman of the Central Military Commission, stressed during his recent visit to Shanghai that he should adhere to the guidance of socialism with Chinese characteristics in the new era, resolutely implement the decision-making and deployment of the Party Central Committee, and strengthen reform and opening up. Confidence and

# Machine Translation in 2012

新华新闻 地方联播 > 正文

## 家庭生活日记账：计价精确到“8分8厘3”

2012年11月09日 20:57:04
来源： 北京晚报

【字号：大 中 小】【打印】 【纠错】

“婴儿纸尿裤，384元；手机贴膜，10元；报纸，1元……”十年来的每一天，在北京这座城市，都有5000个收入不同、构成各异的家庭在细心填写着统一格式的生活账本——《城镇居民家庭生活情况日记账》。

这一行行的“针头线脑”真实记录着他们的生活轨迹，也勾勒着“城镇居民可支配收入”、“城镇住户调查收入情况”等事关国计民生的统计数据。

而长年和这些原生态数据打交道的基层调查员们，总能亲身感知到生活在这座城市中的人们赚钱、花钱的那些事……

我们每天的生活如何变成数字，数字又如何影响我们的生活？

北京有5000个记账户家庭，他们的记录将成为政府了解居民收入、生活、物价等多方面信息的渠道，为制订社会发展计划和进行科学决策提供重要依据，包括最低生活保障线、最低工资标准等等。这些家庭由统计调查单位遵循随机抽样的原则选取，记账家庭三年整体轮换一次。

他们填报的数据，经过系统的整理、汇总和分析后，每个月都会形成《城镇居民人均可支配收入》等多份数据报告，公众可以进入北京统计信息网（www.bjstats.gov.cn）查询。

**Google Translate**
**translate.google.com**

## Journal of family life pricing accurate to 8 points 8% 3

2012 at 17:28 on November 9th Beijing Evening News Reviews

"Baby diaper, 384 yuan; cell phone foil, 10 yuan; newspaper, 1 yuan ..." Every day in the past decade, in the city of Beijing, there are 5,000 different income, constitute a different family carefully fill in a unified format Living Book - "Journal of Urban Households Living Journal."

This line of "needle and thread" really records their life trajectory, but also outlines the "urban residents disposable income", "urban household survey income" and other statistics on people's livelihood.

For many years, grass-roots investigators who deal with these original ecological data can always personally know those people who live in the city make money and spend money ...

How do our daily lives become numbers, and how do numbers affect our lives?

Beijing has 5,000 households with book-keeping households. Their records will be the channels through which the government can understand various aspects of residents' incomes, living standards, prices and other information, and provide an important basis for formulating social development plans and making scientific decisions, including the minimum living standard and the minimum wage and many more. These families are selected by the statistical investigation unit following the principle of random sampling, and the bookkeeping households are rotated in their entirety in three years.

The data they fill in, after systematically arranging, aggregating and analyzing, each month will form a number of data reports such as "per capita disposable income of urban residents" and

# Why is MT difficult?

# Some examples

John loves Mary.
*Jean aime Marie.*

John told Mary a story.
*Jean a raconté une histoire à Marie.*

John is a computer scientist.
*Jean est informaticien.*

John swam across the lake.
*Jean a traversé le lac à la nage*.

# Correspondences

John loves Mary.

*Jean aime Marie.*

John told Mary a story.

*Jean [a raconté ] une histoire [à Marie].*

John is a [computer scientist].

*Jean est informaticien.*

John [swam across] the lake.

*Jean [a traversé] le lac [à la nage].*

# Correspondences

One-to-one:

John = *Jean*, aime = *loves*, Mary=*Marie*

One-to-many/many-to-one:

Mary = [*à Marie*]

[a computer scientist] = *informaticien*

Many-to-many:

[*swam across*__ ] = [*a traversé* ___ *à la nage*]

Reordering required:

told Mary$_1$ [a story]$_2$ = *a raconté* [*une histoire*]$_2$ [*à Marie*]$_1$

# Lexical divergences

The different senses of **homonymous words** generally have different translations:

English-German: (river) bank - Ufer
(financial) bank - Bank

The different senses of **polysemous words** may also have different translations:

I **know that** he bought the book: Je **sais qu**'il a acheté le livre.
I **know** Peter: Je **connais** Peter.
I **know** math: Je **m'y connais en** maths**.**

# Lexical divergences

## Lexical specificity

German ***Kürbis*** = English ***pumpkin*** or ***(winter) squash***

English ***brother*** = Chinese ***gege*** (older) or ***didi*** (younger)

## Morphological divergences

English: ***new*** *book(s),* ***new*** *story/stories*

French: un **nouveau** *livre (sg.m), une* **nouvelle** *histoire (sg.f),*
*des* **nouveaux** *livres (pl.m), des* **nouvelles** *histoires (pl.f)*

- How much **inflection** does a language have?
  (cf. Chinese vs.Finnish)
- How many **morphemes** does each word have?
- How easily can the morphemes be **separated**?

# Syntactic divergences

Word order: fixed or free?

If fixed, which one? [SVO (Sbj-Verb-Obj), SOV, VSO,… ]

Head-marking vs. dependent-marking

**Dependent-marking** (English) *the man'**s** house*

**Head-marking** (Hungarian) *the man house-**his***

Pro-drop languages can omit pronouns:

Italian (with inflection): *I eat = mangi**o**; he eats = mangi**a***

Chinese (without inflection): *I/he eat: chīfàn*

# Syntactic divergences: negation

| | Normal | Negated | |
|---|---|---|---|
| **English** | *I drank coffee.* | *I **didn't drink (any)** coffee.* | *do*-support, *any* |
| **French** | *J'ai bu du café* | *Je **n'**ai **pas** bu **de** café.* | ne..pas<br>*du* → *de* |
| **German** | *Ich habe Kaffee getrunken* | *Ich habe **keinen** Kaffee getrunken* | *keinen Kaffee*<br>=<br>'no coffee' |

# Semantic differences

## Aspect:

- English has a **progressive aspect**:
  *'Peter swims' vs. 'Peter is swimming'*
- German can only express this with **an adverb**:
  *'Peter schwimmt' vs. 'Peter schwimmt gerade'* ('swims currently')

## **Motion events** have two properties:

- **manner** of motion (*swimming*)
- **direction** of motion (*across* the lake)

Languages express either the manner with a verb
and the direction with a 'satellite' or vice versa (L. Talmy):

English (satellite-framed): *He* [*swam*]$_{\text{MANNER}}$ **[*across*]$_{\text{DIR}}$** *the lake*

French (verb-framed): *Il a* **[*traversé* ]$_{\text{DIR}}$** *le lac* **[*à la nage* ]**$_{\text{MANNER}}$

# An exercise

# Knight's Centauri and Arctuan

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.

# The original corpus

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 están enfadados.

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

7a. the clients and the associates are enemies.
7b. los clientes 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 están en Europa.

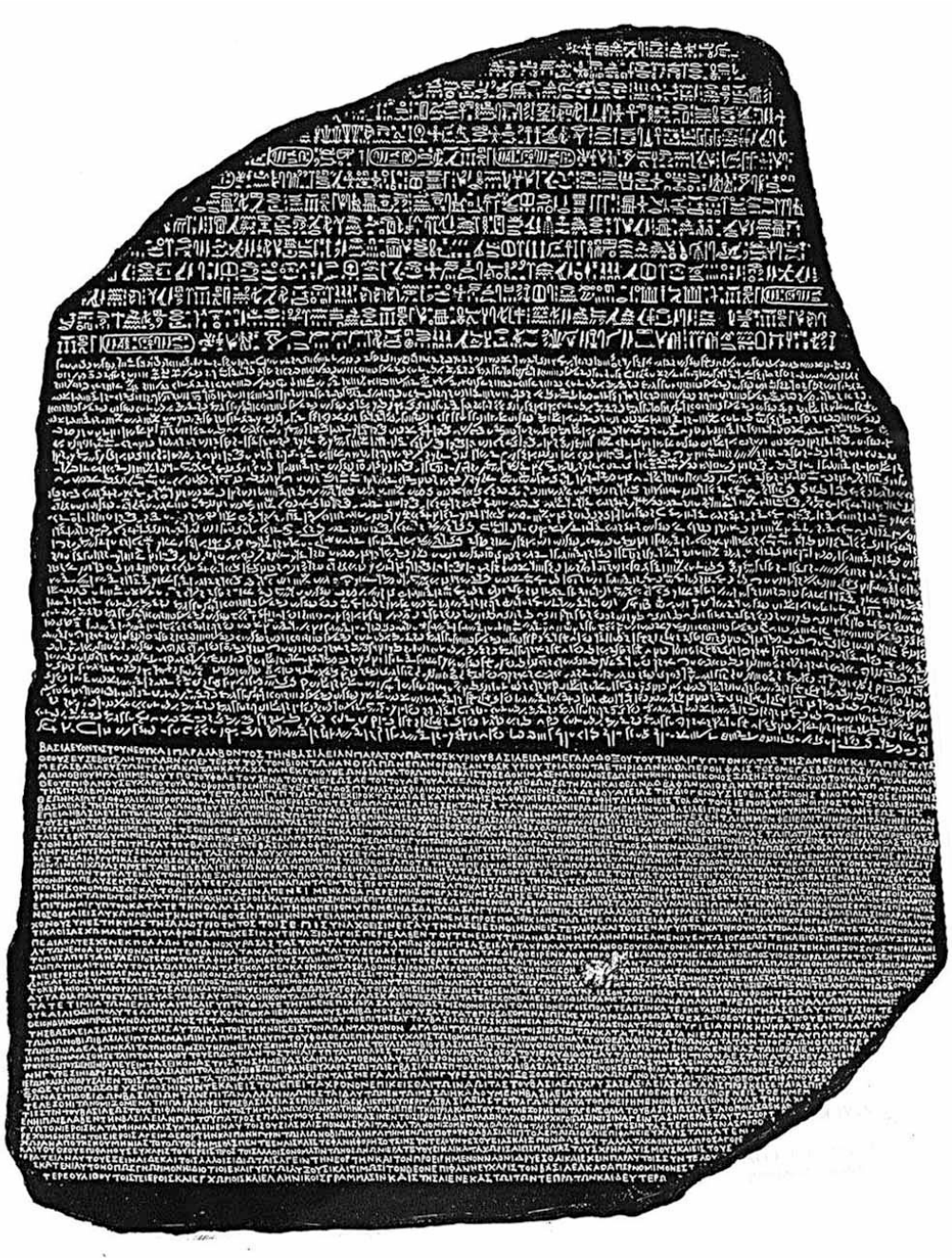
10a. the modern groups sell strong pharmaceuticals.
10b. los grupos modernos venden medicinas fuertes.

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

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

| | |
|---|---|
| 1a. **Garcia** and **associates**. | 1a. ok-**voon** ororok **sprok**. |
| 1b. **Garcia** y **asociados**. | 1b. at-**voon** bichat **dat**. |
| 2a. Carlos **Garcia** has three **associates**. | 2a. ok-drubel ok-**voon** anok plok **sprok**. |
| 2b. Carlos **Garcia** tiene tres **asociados**. | 2b. at-drubel at-**voon** pippat rrat **dat**. |
| 3a. **his associates are not** strong. | 3a. **erok sprok izok hihok** ghirok. |
| 3b. **sus asociados no son** fuertes. | 3b. **totat dat arrat vat** hilat. |
| 4a. **Garcia** has a company **also**. | 4a. ok-**voon** anok drok brok **jok**. |
| 4b. **Garcia tambien** tiene una empresa. | 4b. at-**voon krat** pippat sat lat. |
| 5a. **its** clients **are** angry. | 5a. **wiwok** farok **izok** stok. |
| 5b. **sus** clientes **están** enfadados. | 5b. **totat** jjat **quat** cat. |
| 6a. **the associates are also** angry. | 6a. **lalok sprok izok jok** stok. |
| 6b. **los asociados tambien están** enfadados. | 6b. **wat dat krat quat** cat. |
| 7a. **the** clients and **the associates are** enemies. | 7a. **lalok** farok ororok **lalok sprok izok** enemok |
| 7b. **los** clientes y **los asociados son** enemigos. | 7b. **wat** jjat bichat **wat dat vat** eneat. |
| 8a. **the** company has three **groups**. | 8a. **lalok** brok anok plok **nok**. |
| 8b. **la** empresa tiene tres **grupos**. | 8b. **iat** lat pippat rrat **nnat**. |
| 9a. **its groups are** in Europe. | 9a. **wiwok nok izok** kantok ok-yurp. |
| 9b. **sus grupos están** en Europa. | 9b. **totat nnat quat** oloat at-yurp. |
| 10a. **the** modern **groups** sell strong pharmaceuticals | 10a. **lalok** mok **nok** yorok ghirok clok. |
| 10b. **los grupos** modernos venden medicinas fuertes | 10b. **wat nnat** gat mat bat hilat. |
| 11a. **the groups do not** sell zanzanine. | 11a. **lalok nok crrrok hihok** yorok zanzanok. |
| 11b. **los grupos no** venden zanzanina. | 11b. **wat nnat arrat** mat zanzanat. |
| 12a. **the** small **groups are not** modern. | 12a. **lalok** rarok **nok izok hihok** mok. |
| 12b. **los grupos** pequeños **no son** modernos. | 12b. **wat nnat** forat **arrat vat** gat. |

# Machine translation approaches



# The Rosetta Stone

Three different translations of the same text:

- Hieroglyphic Egyptian (used by priests)
- Demotic Egyptian (used for daily purposes)
- Classical Greek (used by the administration)

Instrumental in our understanding of ancient Egyptian

This is an instance of parallel text:

The Greek inscription allowed scholars
to decipher the hieroglyphs

# MT History

WW II: Code-breaking efforts at Bletchley Park, England (Alan Turing)

1948: Shannon/Weaver: Information theory

1949: Weaver's memorandum defines the task

1954: IBM/Georgetown demo: 60 sentences Russian-English

1960: Bar-Hillel: MT to difficult

1966: ALPAC report: human translation is far cheaper and better:
kills MT for a long time

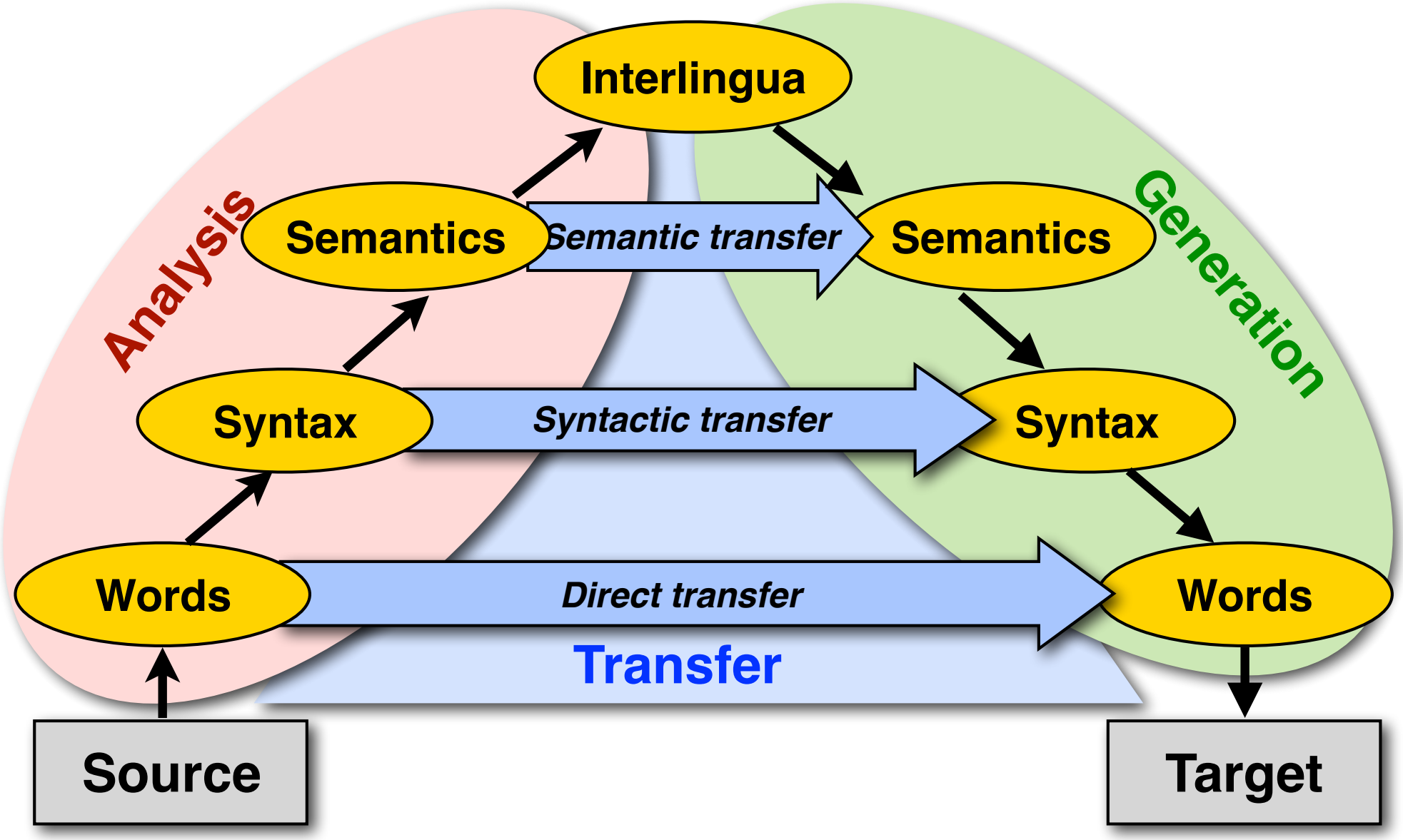
1980s/90s: Transfer and interlingua-based approaches

1990: IBM's CANDIDE system (first modern statistical MT system)

2000s: Huge interest and progress in wide-coverage statistical MT:
phrase-based MT, syntax-based MT, open-source tools

Now: Neural machine translation

# The Vauquois triangle



# Direct translation

Maria non dió una bofetada a la bruja verde.

**1. Morphological analysis of source string**

Maria $\underline{\text{non}_{\text{Neg}}}$ $\underline{\text{dar}_{\text{3sgF-Past}}}$ una bofetada a la bruja verde

(usually, a complete morphological analysis)

**2. Lexical transfer (using a translation dictionary):**

Mary not $\text{slap}_{\text{3sgF-Past}}$ to the witch green.

**3. Local reordering:**

Mary not $\text{slap}_{\text{3sgF-Past}}$ the $\underline{\text{green}}$ $\underline{\text{witch}}$.

**4. Morphology:**

Mary $\underline{\text{did not slap}}$ the green witch.

# Limits of direct translation: Phrasal reordering

**Adverb placement in German:**

The green witch is at home this week.

Diese Woche ist die grüne Hexe zuhause.

**Japanese SOV order:**

He adores listening to music

Kare ha ongaku wo kiku no ga daisuki desu

**PPs in Chinese:**

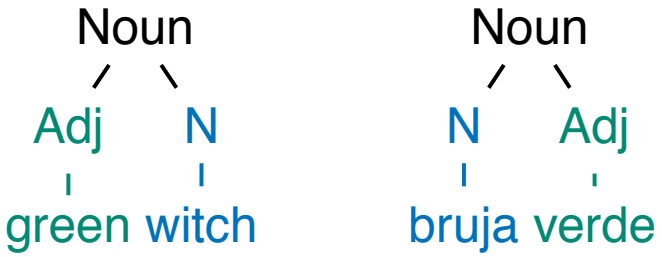
Jackie Cheng went to Hong Kong

Cheng Long dao Xianggang qu

# Syntactic transfer

Requires a syntactic parse of the source language, followed by reordering of the tree

**Local reordering:**




**Nonlocal reordering:**

# Semantic transfer

Done at the level of **predicate-argument structure** (some people call this syntactic transfer too…):

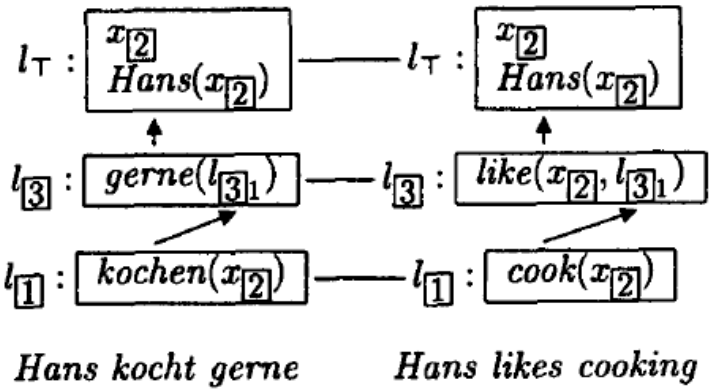
$$\begin{bmatrix} \text{SUBJ} & [\text{PRED} \quad \text{HANS}]\boxed{2} \\ \text{PRED} & \text{KOCHEN}\langle \uparrow \text{SUBJ} \rangle \\ \text{ADJN} & \{[\text{PRED} \quad \text{GERNE}]\boxed{3}\} \end{bmatrix}\boxed{1}$$

*Hans kocht gerne*

$$\begin{bmatrix} \text{SUBJ} & [\text{PRED} \quad \text{HANS}]\boxed{2} \\ \text{PRED} & \text{LIKE}\langle \uparrow \text{SUBJ}, \uparrow \text{XCOMP} \rangle \\ \text{XCOMP} & \begin{bmatrix} \text{SUBJ} & [\text{PRED} \quad \text{HANS}]\boxed{2} \\ \text{PRED} & \text{COOK}\langle \uparrow \text{SUBJ} \rangle \end{bmatrix}\boxed{1} \end{bmatrix}\boxed{3}$$

*Hans likes cooking*

or at the level of **semantic representations** (e.g. DRSs):

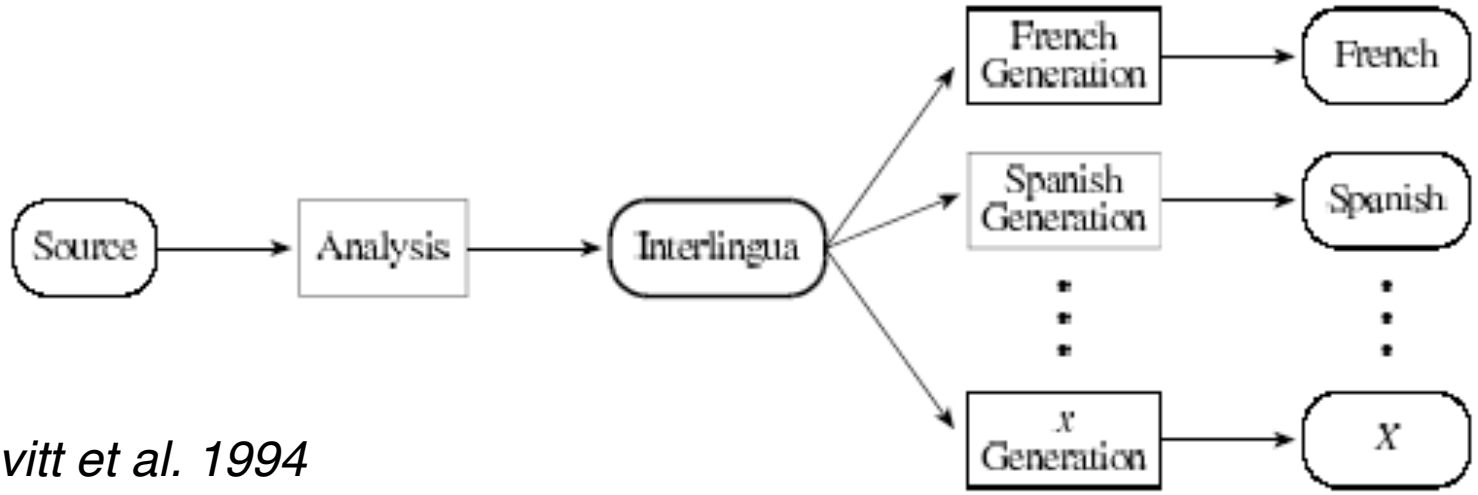



*Dorna et al. 1998*

# Interlingua approaches

Based on the assumption that there is one **common meaning representation** (e.g. predicate logic) that abstracts away from *any* difference in surface realization.

- Semantic transfer: each language produces its own meaning representation
- Was thought useful for multilingual translation




*Leavitt et al. 1994*

# Statistical Machine Translation

# Statistical Machine Translation

We want the best (most likely) [English] translation for the [Chinese] input:

$$\text{argmax}_{\text{English}}\ P(\ \text{English} \mid \text{Chinese}\ )$$

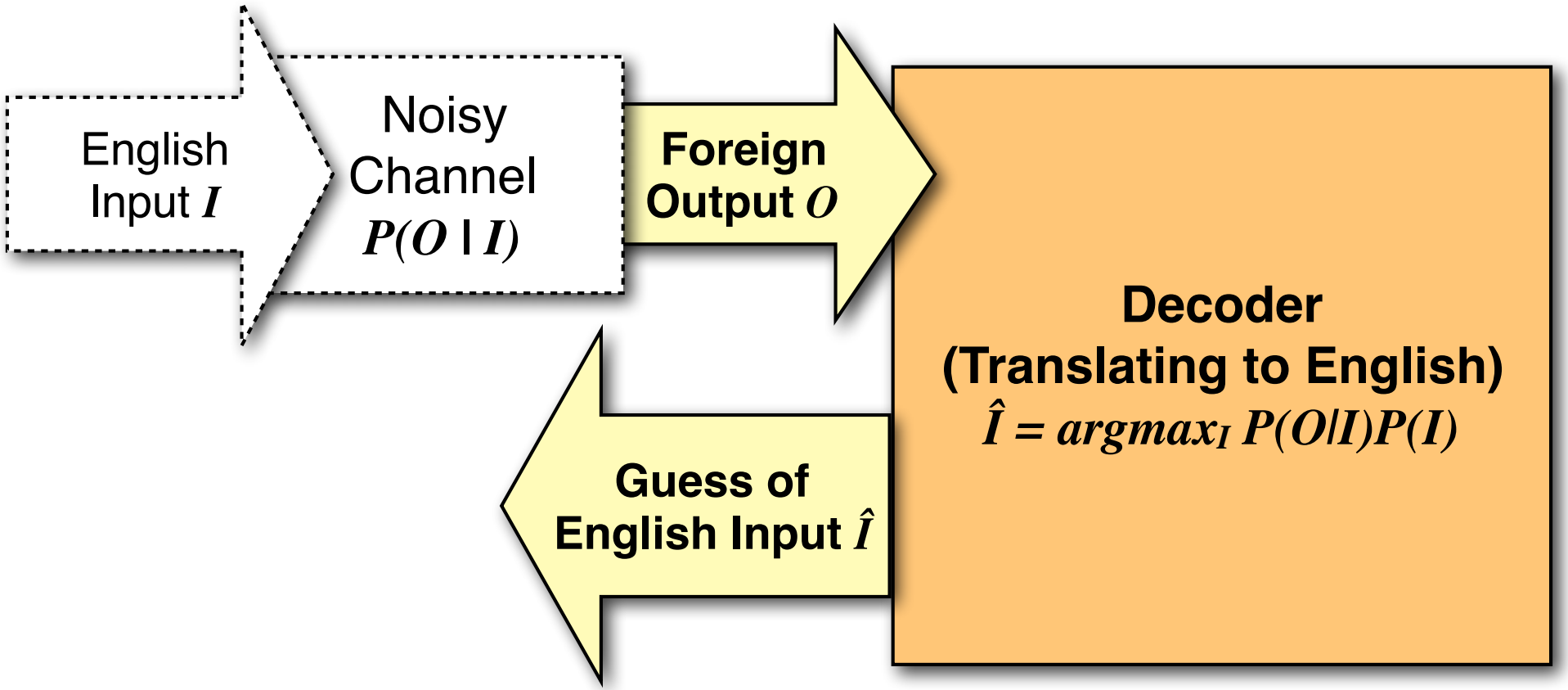
We can either model this probability directly,
or we can apply Bayes Rule.
Using Bayes Rule leads to the “noisy channel” model.

As with sequence labeling, Bayes Rule simplifies the modeling task, so this was the first approach for statistical MT.

# The noisy channel model

**Translating from Chinese to English:**

$$\text{argmax}_{Eng} P(Eng|Chin) = \text{argmax}_{Eng} \underbrace{P(Chin|Eng)}_{\text{Translation Model}} \times \underbrace{P(Eng)}_{\text{LanguageModel}}$$



# The noisy channel model

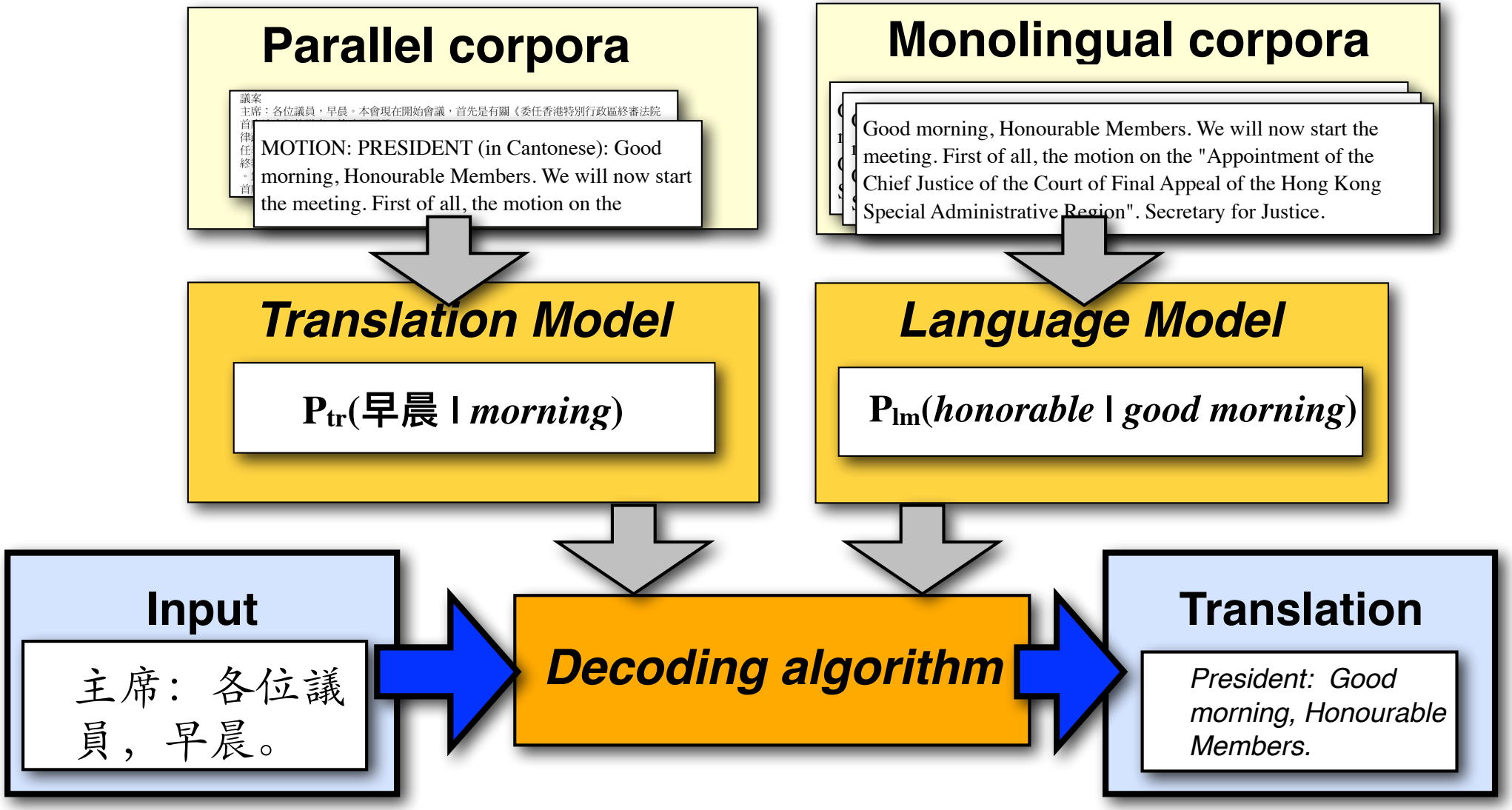
This is really just an application of **Bayes' rule**:

$$\begin{aligned}
\hat{E} &= \arg\max_E P(E|F) \\
&= \arg\max_E \frac{P(F|E) \times P(E)}{P(F)} \\
&= \arg\max_E \underbrace{P(F|E)}_{\text{Translation Model}} \times \underbrace{P(E)}_{\text{Language Model}}
\end{aligned}$$

The **translation model** $P(F \mid E)$ is intended to capture the **faithfulness of the translation**.
It needs to be trained on a **parallel corpus**

The **language model** $P(E)$ is intended to capture the **fluency of the translation**.
It can be trained on a (very large) **monolingual corpus**

# Statistical MT



# *n*-gram language models for MT

With training on data from the web and clever parallel processing (MapReduce/Bloom filters), *n* can be quite large

- Google (2007) uses 5-grams to 7-grams,
- This results in huge models, but the effect on translation quality levels off quickly:

**Size of models**

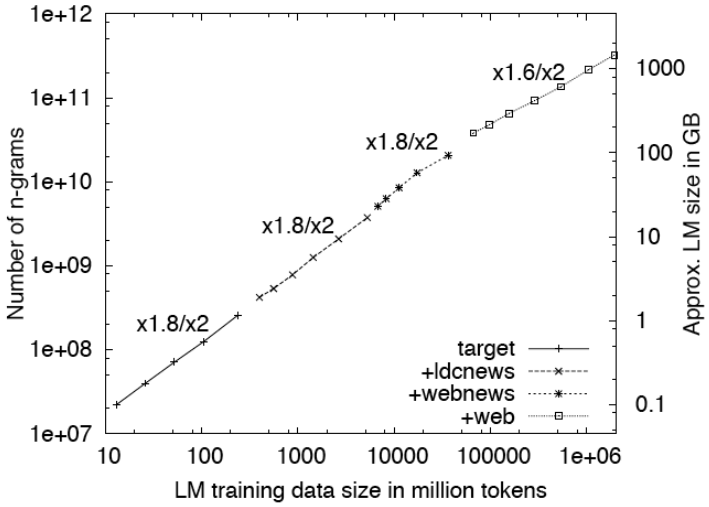



Figure 3: Number of $n$-grams (sum of unigrams to 5-grams) for varying amounts of training data.

**Effect on translation quality**

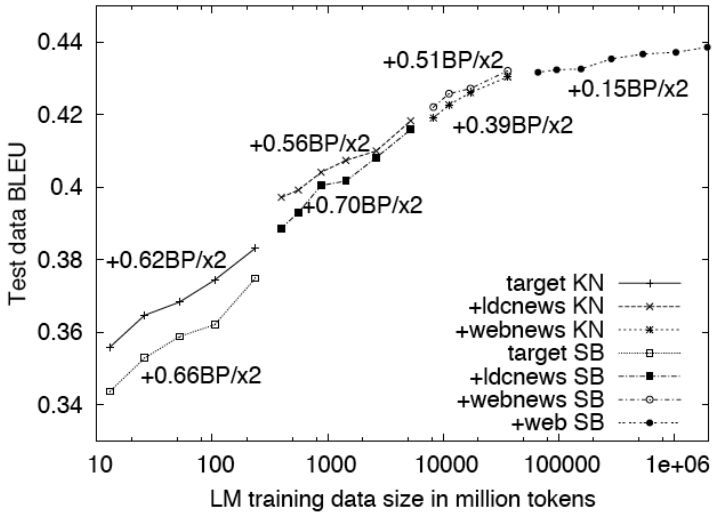



Figure 5: BLEU scores for varying amounts of data using Kneser-Ney (KN) and Stupid Backoff (SB).

# Translation probability $P(fp_i \mid ep_i)$

Phrase translation probabilities can be obtained from a **phrase table:**

| EP | FP | count |
|---|---|---|
| green witch | grüne Hexe | … |
| at home | zuhause | 10534 |
| at home | daheim | 9890 |
| is | ist | 598012 |
| this week | diese Woche | …. |

This requires **phrase alignment** on a parallel corpus.

# Creating parallel corpora

A **parallel corpus** consists of the same text in two (or more) languages.

Examples: Parliamentary debates: Canadian Hansards; Hong Kong Hansards, Europarl; Movie subtitles (OpenSubtitles)

In order to train translation models, we need to **align the sentences** (Church & Gale '93)

| English | French |
|---|---|
| According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees. | Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte e grande partie de l'accroissement du volum des ventes. L'emploi et les investissemen ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté. |

| English | French |
|---|---|
| According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates. | Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment. |
| The higher turnover was largely due to an increase in the sales volume. | La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. |
| Employment and investment levels also climbed. | L'emploi et les investissements ont également augmenté. |
| Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees. | La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté. |

# Today's key concepts

Why is machine translation hard?

- Linguistic divergences: morphology, syntax, semantics

Different approaches to machine translation:

- Vauquois triangle
- Statistical MT (more on this next time)