Statistics to the Rescue!

- Rests on primary data
- No linguistic/nonlinguistic distinction
- Treats all phenomena impartially
- Deterministic
- Local
- Rapid development cycle
- People annotate rather than analyze
- Good enough results for government work

Doing it by numbers

- What words are most likely to occur in a translation of this sentence, given the source words that it contains and the translations we have seen?
- What order should they be in, given what we know about other sentences in the target language?

The Statistical Approach: Training

The translation model

Find pairs of words ("phrases") that have a high probability of occurring opposite one another in sentences that are translations of one another.

The Language Model

Find short sequences of words (N-grams) that have a high probability of occurring together.

Other stuff

Fertility

Distortion ...

Model Evaluation

Compare translations to human gold standard(s) using a similarity measure.

"Bleu" score—number of trigrams shared by candidate and gold standard(s)

N.B. The better the system gets, the less reliable the measure becomes.

Unfortunately we have …

Zipf's law

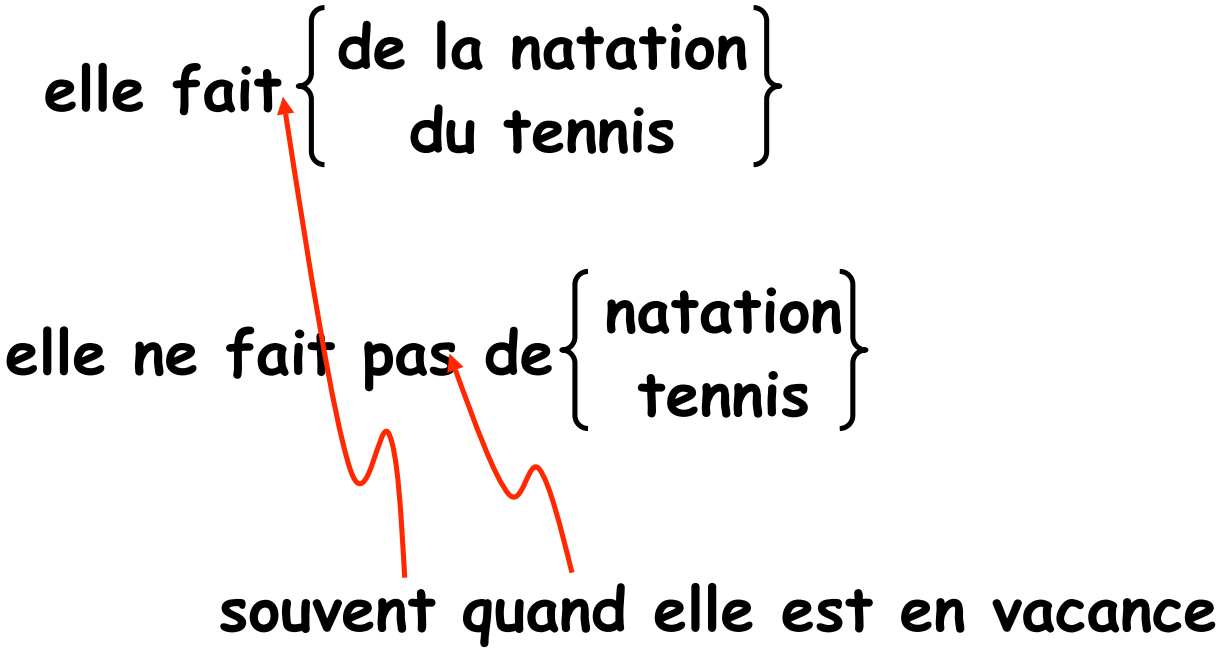
Locality

Emergent Properties

AI

Bleu score

Linguistic Facts—Locality



Facts about translation

… are not all reflected in emergent properties of translations

Does this train go to Endville?

Est-ce que c'est ta cousine?

I just got back from Texas/Utah. I had forgotten how good beer tastes.

Ich hatte vergeßen, wie gut[es] Bier schmekt.

It may be necessary to reduce condenser steam side pressure

pression latérale de la vapeur

pression côté vapeur

Pick up the red token off the table

Puts it in the box

Proposals

- Hybrids
- Monolingual human consultants
  - Reflective Editing
- Triangulation

Reflective Editing

Produce many translations

Display one of them—the best one.

The editor changes it into …

A version that the system had already foreseen, but not chosen as the preferred version.

∴ We know what choices the system would have had to make to reach that version.

∴ We will make those choices when translating into the next language.

Triangulation

There are three windows in the room

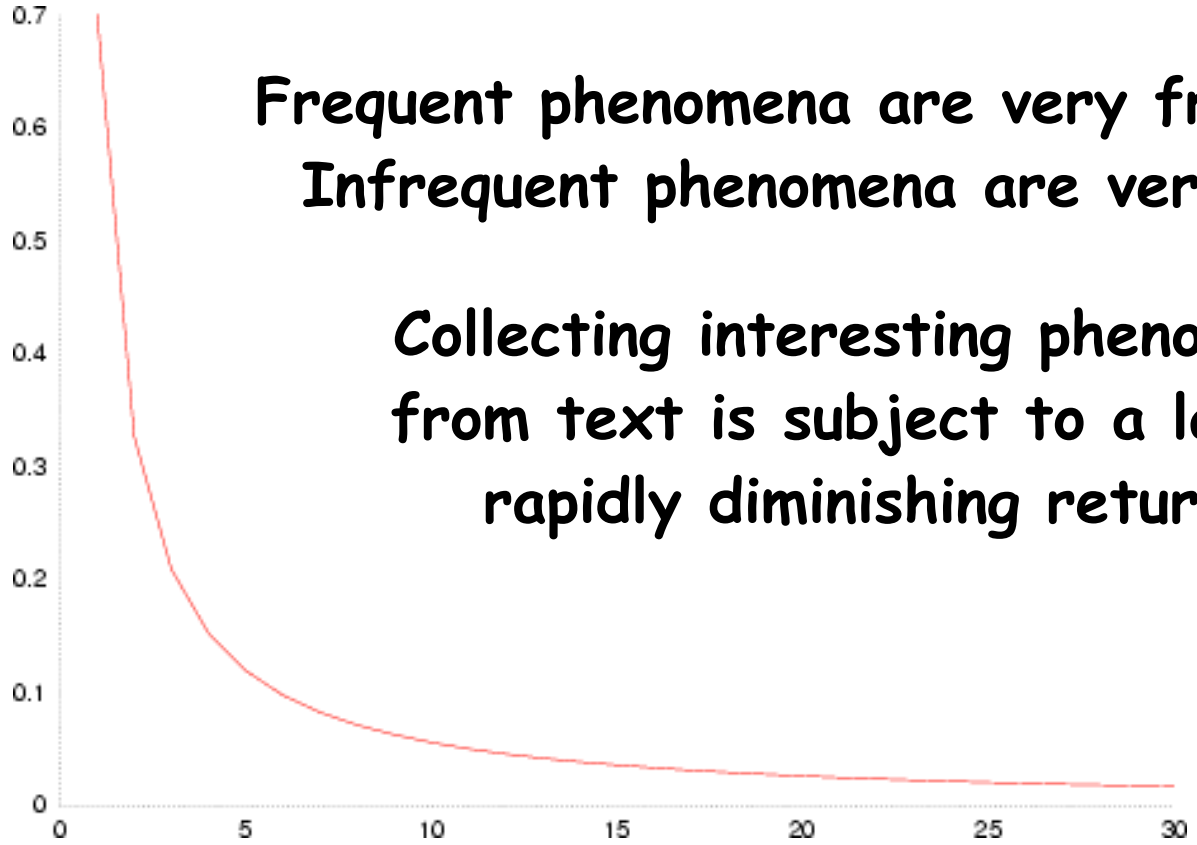
Il y a trois fenêtres dans la salle.

Il y a trois guichets dans la salle.

Es gibt drei Fenster in dem Zimmer.
Es gibt drei Schalter in dem Zimmer.

fenêtre ~ Fenster
guichet ~ Schalter

Zipf's Law



Frequent phenomena are very frequent;
Infrequent phenomena are very rare

Collecting interesting phenomena from text is subject to a law of rapidly diminishing returns

Emergent Properties

The important facts about language may not be emergent properties of text.

L'arbitraire du signe

The important facts about translation may not all be emergent properties of translations.

The End

Fin **Ende**