# Lecture reviews — Week 04

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

# Purpose of these lecture reviews

- Improve/deepend your learning
- Answer your questions
- Save you practice/revision time

Why are these sessions not recorded?

1. the intention is to have *appropriate*/*adapted*/*personalized* face-to-face interaction
2. recording them would lead to an extra 2 hours/week video lecture
   (which is too much *passive* content)

# Content

1. Big picture:
   What did you retain? What keypoints do you remember?
2. Questions?
3. More examples


J.-C. Chappelier & M. Rajman

# Week 4 keypoints

- Words / tokens (def, pay attention, ...)
- n-grams
  - learn/train : parameters : $P(X_1 \dots X_n)$
  - usage : $\hat{\prod}_{i=1} P(X_i \mid X_{i-k} \dots X_{i-1})$, $k = n-1$
  - ↳ char
  - ↳ tokens : language models
- learning: MLE vs smoothing (add $\alpha$)
- OoV forms

# Week 4 keypoints

- Words vs. tokens
- *n*-gram models
- MLE and add-one smoothing are bad (in NLP)
- Language Identification
- Out-of-Vocabulary froms:
  - OoV forms do matter
  - 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors


J.-C. Chappelier & M. Rajman

usage vs training

$$f_\theta(x)$$

$$\prod_{i=n}^{|w|} P(x_i \mid \underbrace{x_{i-(n-1)} \cdots x_{i-1}}_{n-1})$$

($n$ spans $x_{i-(n-1)} \cdots x_i$)

$$P(x_{i-(n-1)} \cdots x_i)$$

$$\frac{P(x_{i-(n-1)} \cdots x_{i-1}\, x_i)}{P(x_{i-(n-1)} \cdots x_{i-1})}$$

$$P(x_{i-(n-1)} \cdots x_{i-1}) = \sum_x P(x_{i-(n-1)} \cdots x_{i-1}\, x)$$

# Week 4 keypoints

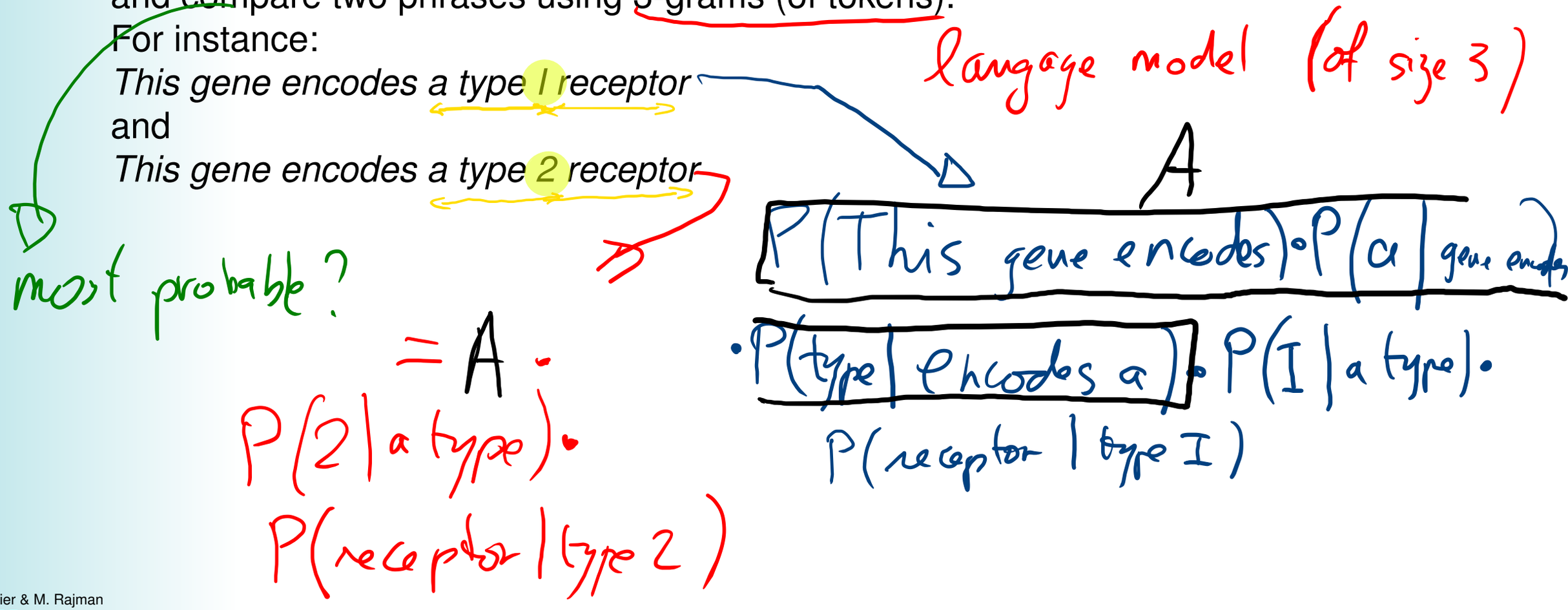
- Words vs. tokens
- *n*-gram models
- MLE and add-one smoothing are bad (in NLP)
- Language Identification
- Out-of-Vocabulary froms:
  - OoV forms do matter
  - 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors

Questions?


J.-C. Chappelier & M. Rajman

# Week 4 review example

Take a random Wikipedia page (e.g. https://en.wikipedia.org/wiki/ACVRL1) and compare two phrases using 3-grams (of tokens).
For instance:
*This gene encodes a type I receptor*
and
*This gene encodes a type 2 receptor*



# Week 4 review example

Take a random Wikipedia page (e.g. https://en.wikipedia.org/wiki/ACVRL1) and compare two phrases using 3-grams (of tokens).
For instance:
*This gene encodes a type I receptor*
and
*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?


J.-C. Chappelier & M. Rajman

$$P(xyz)$$

$$P(xy)P(z|xy).$$

1. where do we start from

2. tokenize

3. learn $P(token_i \mid token_{i+1}\ token_{i+2})$

↳ This vs this

↓

this | BegOfSent

This

≡ <BOS>|this

learning <BOS> Serine / theronie

1st param: $P(\text{<BOS>}, \text{serine}, /)$

$$= \frac{1 + \alpha}{N + \alpha \cdot M}$$

⚠ MLE? ✗

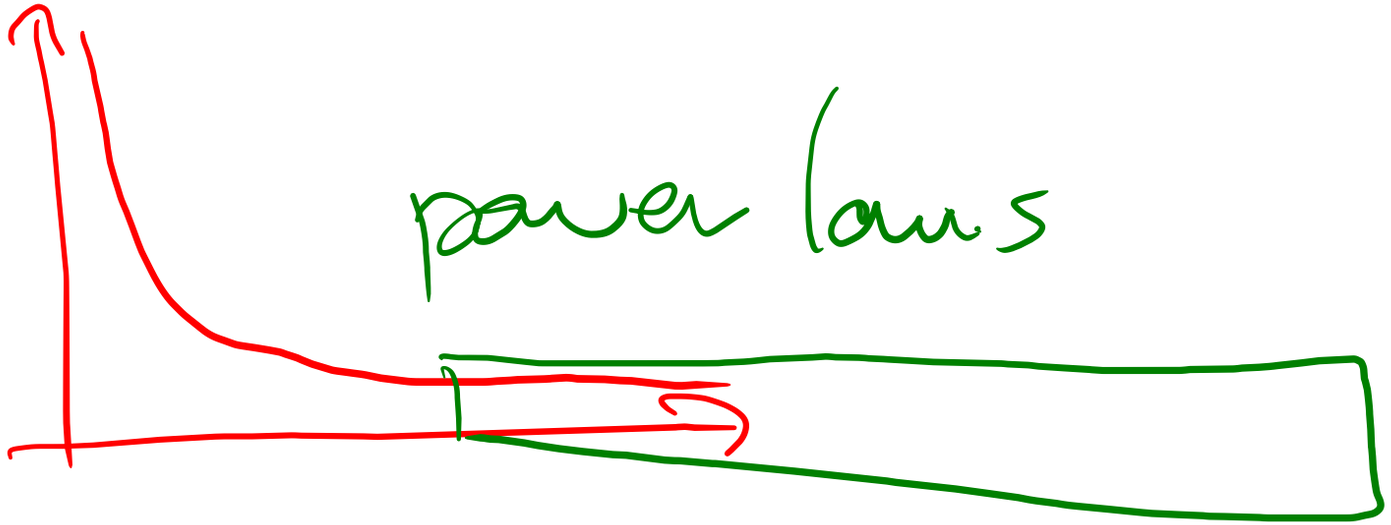
$N$: # 3-gram in corpus

$M$: # of possible 3-grams (huge!)

MLE

are bad

with



# Week 4 review example

Take a random Wikipedia page (e.g. https://en.wikipedia.org/wiki/ACVRL1) and compare two phrases using 3-grams (of tokens).
For instance:
*This gene encodes a type I receptor*
and
*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. "*Serine/threonine-protein kinase recept*")


J.-C. Chappelier & M. Rajman

# Week 4 review example

Take a random Wikipedia page (e.g. https://en.wikipedia.org/wiki/ACVRL1) and compare two phrases using 3-grams (of tokens).
For instance:
*This gene encodes a type I receptor*
and
*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. "*Serine/threonine-protein kinase recept*")
3. How to deal with upper-/lowercase? (e.g. "*This*")


J.-C. Chappelier & M. Rajman

# Week 4 review example

Take a random Wikipedia page (e.g. https://en.wikipedia.org/wiki/ACVRL1) and compare two phrases using 3-grams (of tokens).
For instance:
*This gene encodes a type I receptor*
and
*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. "*Serine/threonine-protein kinase recept*")
3. How to deal with upper-/lowercase? (e.g. "*This*")
4. What estimates? (MLE? Smoothing?)


J.-C. Chappelier & M. Rajman