Multilingual and Parallel Corpora

Machine Translation (ctd)

Amir Zeldes

amir.zeldes@georgetown.edu

Discussing final paper topics

- Please arrange a time in the next 3 week to come to office hours and discuss final papers:
  - Language comparison via parallel corpus (e.g. article use, tense, aspect, modality, …)
  - Translationese studies (grammatical properties of native L1 vs. TL)
  - Register studies (e.g. markers of formality in translation [subs], text normative equivalence)
  - Translation universals
  - Behavior of loanwords in translation (recent loans, integration of technical terms, development of false friends…)
  - Semantic mirror – charting translations and back-translations of words/ senses in a semantic domain (e.g. language of emotion, politics, color …)
  - Work with special corpora – non-native L2, historical data…

Recap: Direct translation model

- We match the largest possible substrings of an input
  - Not to mention
  - the problems
- Find translations
  - Ganz zu schweigen
  - die Probleme
- Assemble
  - Ganz zu schweigen + die Probleme ->

    Ganz zu schweigen **von** d**en** Problem**en**

Do we still need to understand the text?

- Unlike in rule based transfer, the meaning of sentences is no longer evaluated
- No abstract logical/interlingua representation
- Is it possible to translate a text without understanding **the source or target language**?
  - Try Arcturan : Centauri!

Homework assignment: Centauri/Arcturan [Knight 1997]

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. wat 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 .

It's actually English and Spanish! [Knight 1997]

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 .

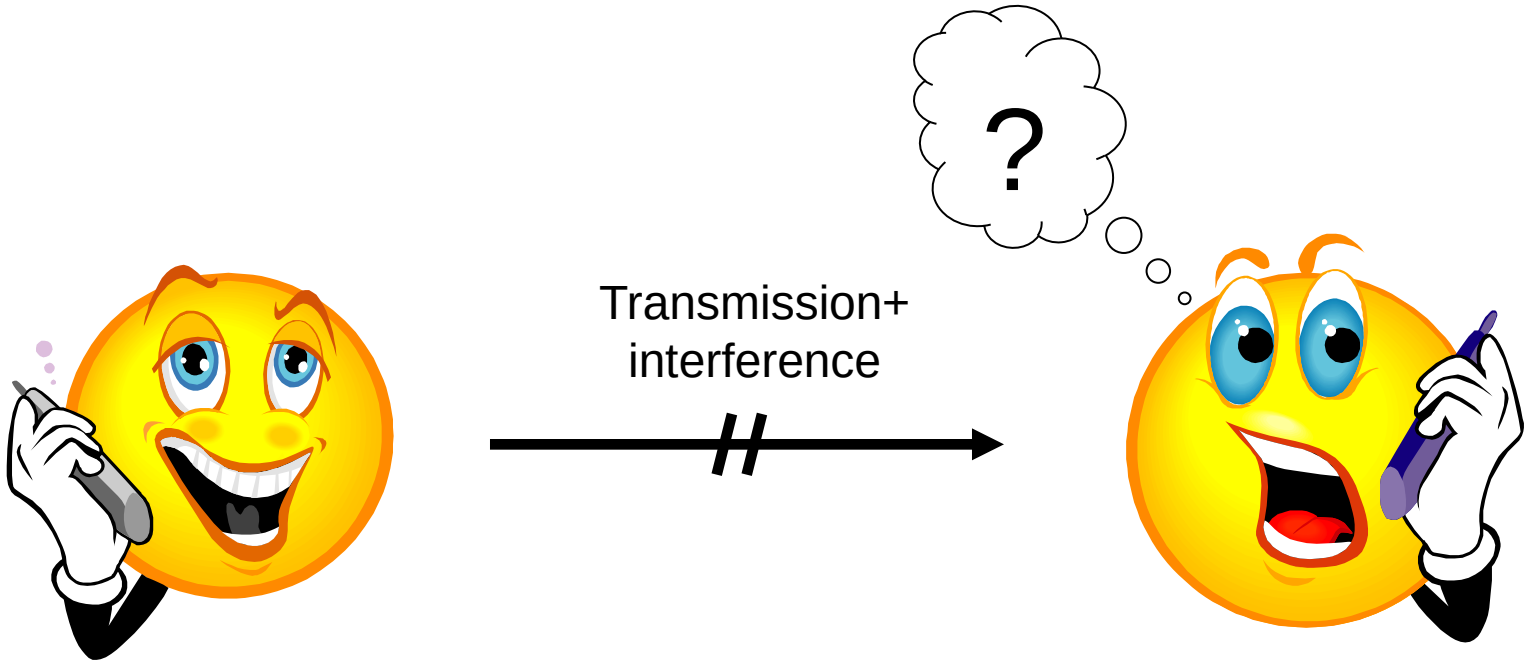
Fluency issues

- Adjustment mechanisms are not built in to direct translation (a.k.a. pure ‘transfer’ models)
- But work without **any** knowledge of SL and TL (except tokenization)
- De facto standard until the later 90s…

The Noisy Channel Model

- Originally developed in information theory for telecommunications
- Basic problem:
  - What do you do if your transmission channel (e.g. a radio or phone) has interference?

The Noisy Channel Model



- I'm running a little late
- I-… ru-… … li-…l late

What did he mean?

Input

- I-… ru-… … li-…l late

Most likely option given input

Possible outputs

- I’ve run into Lyle late
- Isle runners alive all late
- I’m running a little late
- It’s me running …

What is most likely?

- Generally in English:

  P(It's me) >

  P(I'm running) >

  …

  P(Acapulco trips germinate)

- And given an input I-… ru-… li-…l late:

  P(I'm running a little late) >

  P(I ran a little late) >

  …

Suppose German is just English with interference…

To recreate a perturbed English message we need:

- Probability of any sequence in the TL – the **Language Model**
- Probability of each translation from SL -> TL – The **Translation Model**

Language Model

- Getting probabilities for “It’s me”, “I ran late”… is easy

➢ Look in a huge corpus

➢ Caution:
If the texts we are translating **deviate** strongly from this corpus, the probabilities will be wrong!

Translation model

- How do we know how likely it is that “Ich bin’s” is translated as “it’s me”?
- Look in a huge parallel corpus?
- But:
  - Parallel corpora are smaller than monolingual corpora
  - More expensive to produce
  - We do not have examples of every sentence we need

Quick overview - probabilities

- Suppose you are a photo reporter and want to take an exclusive picture Miley Cyrus who is currently on tour (example adapted from Jonas Kuhn)
  - There are rumors that certain concerts will get cancelled
  - You want to guess what route she will take
  - Each route has a certain probability
  - Wait at a location along the route with the highest probability to take the picture

Quick overview - probabilities

- **Simple probability (Prior probability) *P*(A)**
  - You call up Miley's manager and ask whether the concert in DC will be cancelled or not
  - "60% chance the concert will take place"

P(CiDC) = 0.6
P(~CiDC) = 0.4

- **Conditional probability (Posterior probability) *P*(A|B)**
  - If the Miley has a concert in DC, how likely is it that she will visit the National Mall?
  - One out of four pop stars who gives a concert in DC also visits the Mall
  - Only 10% of stars *not giving* a concert in DC visit the Mall

P(M | CiDC) = 0.25
P(M | ~CiDC) = 0.1

So should we lurk in the Mall?

- What is *P*(Mall)?
  - We only have conditional probabilities for Miley visiting the Mall
  - We have to consider both options for the precondition

P(CiDC) = 0.6
P(~CiDC) = 0.4

P(M | CiDC) = 0.25
P(M | ~CiDC) = 0.1

- **Joint probability *P*(A,B)**
  - $P(M, CiDC) = P(CiDC) \times P(M \mid CiDC) = 0.6 \times 0.25 = 0.15$
  - $P(M, \sim CiDC) = P(\sim CiDC) \times P(M \mid \sim CiDC) = 0.4 \times 0.1 = 0.04$
  - Since CiDC and ~CiDC cover the full space of probabilities we get:
  - $P(M) = P(M, CiDC) + P(M, \sim CiDC) = 0.19$

Bayes' Law

- We just exploited the fact that joint probabilities [i.e., P(A,B)] can be calculated by multiplying prior probability for one event with the conditional probability for the other, given the first event
  - This is called the "chain rule"
  - We can go either way (because $P(A,B) = P(B,A)$):
    - $P(A,B) = P(A) \times P(B \mid A)$ or
    - $P(A,B) = P(B) \times P(A \mid B)$
- So:
  - $P(B) \times P(A \mid B) = P(A) \times P(B \mid A)$

Bayes' Law

- Another example - crime scene analogy
  - *B* is a crime scene. *A* is a person who may have committed the crime
  - Probabilities:
    - P(A|B) - look at the scene - who did it?
    - P(A) - who had a motive? Fits the crime? (profiler, etc.)
    - P(B|A) - could they have done it? (transportation, access to weapons, alibi)
  - Some people might have great motives, but no means - you need both!
- How does this apply to translation?

Bayes' Law

- Based on Bayes' Law, we can re-cast the MT problem:

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

- Separates the translation problem into two parts:
  - Finding translations
  - Assessing their *a priori* plausibility
  - Note that p(f) is a constant for any input sentence, so we can write: $p(e|f) = p(e) \times p(f|e)$

Combining parallel and monolingual

- We translate parts as best we can
- Most basic scenario: use individual words, but multiple options (p = % attestation in aligned beads)
  - P(ich|I) = 0.8 P(ich|me) = 0.2
  - P(bin|am) = 0.95
  - P(es|it) = 0.75
  - P(die|the) = 0.5
  - P(den|the) = 0.2
  - …

Combining parallel and monolingual

- We evaluate possible translations using the monolingual language model:
  - P(“it am I”) = 0
  - P(“I am it”) = 0.005
  - P(“It ’s me”) = 0.9
  - …
- Approximate P for longer chains using n-grams (Markov Model)
- Translation models propose parts to combine and initial scores, which are modulated by language model scores

Will these considerations harmonize?

- Does maximizing faithfulness (p(f|e)) always go hand in hand with maximizing fluency? (p(e))

- Example: (Kuhn 2007)
  - Japanese: *"fukaku hansei shite orimasu"*
  - Fluent translation: *"we apologize"*
  - Faithful translation: *"we are deeply reflecting (on our past behaviour, and what we did wrong, and how to avoid the problem next time)"*

IBM Model 1 (Brown et al. 1988, 1993)

- Implements the Noisy Channel Model in a **generative process** framework
  - Every word produces a word in the translation, based on translation probabilities
  - All possible **orderings** of the words are considered
  - Most likely ordering is picked based on language model

IBM Model 1

- Problems:
  - Yields same number of words in translation
  - Many orderings are implausible, waste of resources
  - Worst still, some alternative orderings are not impossible!
    - Джон ударил Билла
    - John hit Bill
    - Bill hit John – also possible

IBM Model 2

- Introduce constraint on position distance in translations
- For a word e_i prefer the translation where the corrseponding f_j has close i~j
  - Джон$_1$ ударил$_2$ Билла$_3$
  - John$_1$ hit$_2$ Bill$_3$
  - Bill$_3$ hit$_2$ John$_1$ x

Same number of words

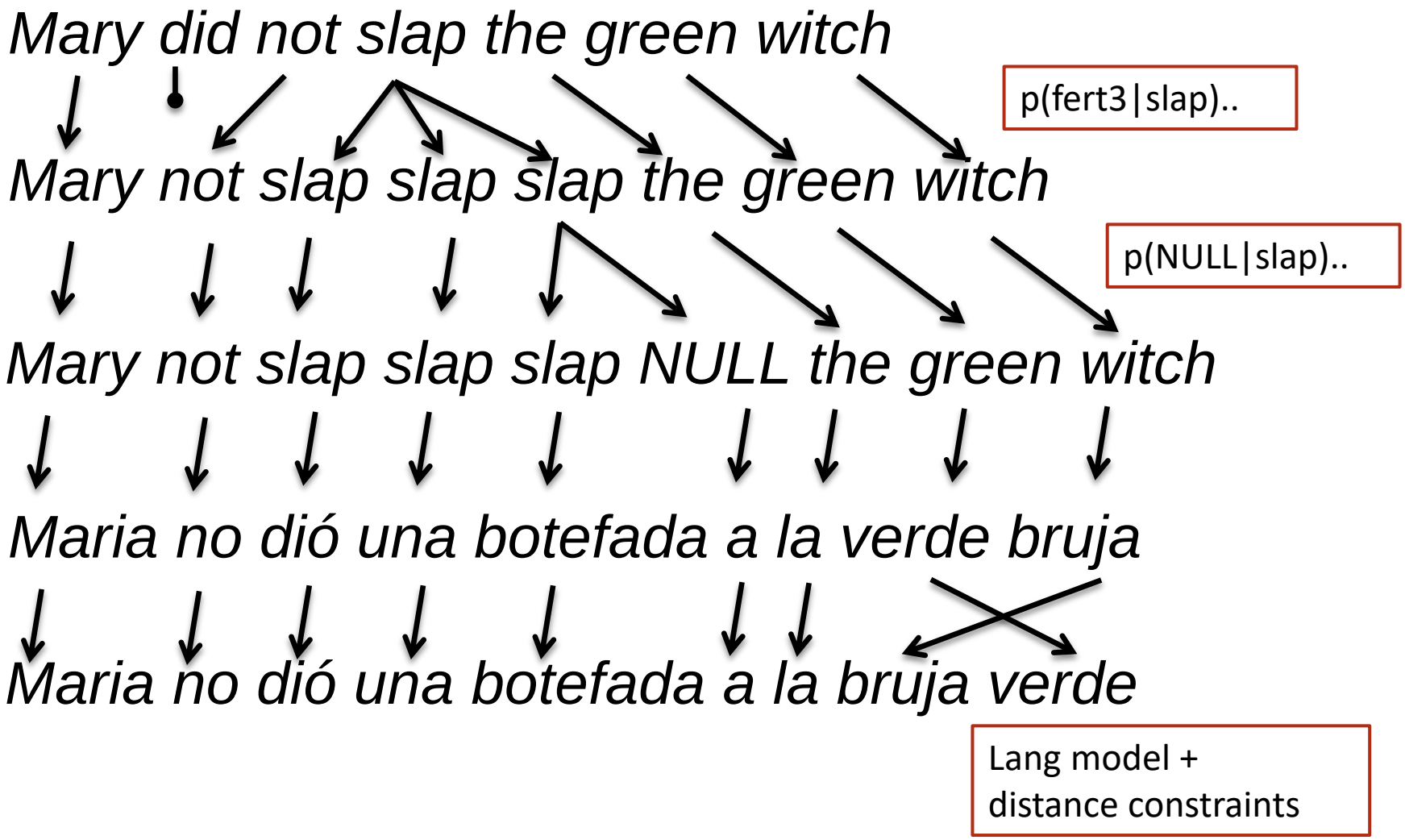
- Looking up each word's translation is problematic:
  - *John is coming*
  - *John ?? vient*

- Is it right to say that "is" corresponds to nothing?
- Inversely, is it right to say that "nothing" corresponds to "is"?

IBM Model 3

- Introduce **fertility** probabilities for each word
- **Null insertion** probabilities for words that receive no alignment
  - *Mary did not slap the green witch*
  - *Maria no dió una botefada a la bruja verde*

IBM Model 3


Mary did not slap the green witch
p(fert3|slap)..
Mary not slap slap slap the green witch
p(NULL|slap)..
Mary not slap slap slap NULL the green witch
Maria no dió una botefada a la verde bruja
Maria no dió una botefada a la bruja verde
Lang model +
distance constraints

Models 4-5

- Also add probabilities for neighboring **POS tags**
- Favor **specific positions** for target indices (able to model things like SVO – if a noun already occupies S, another noun must occupy O)

Remaining problems

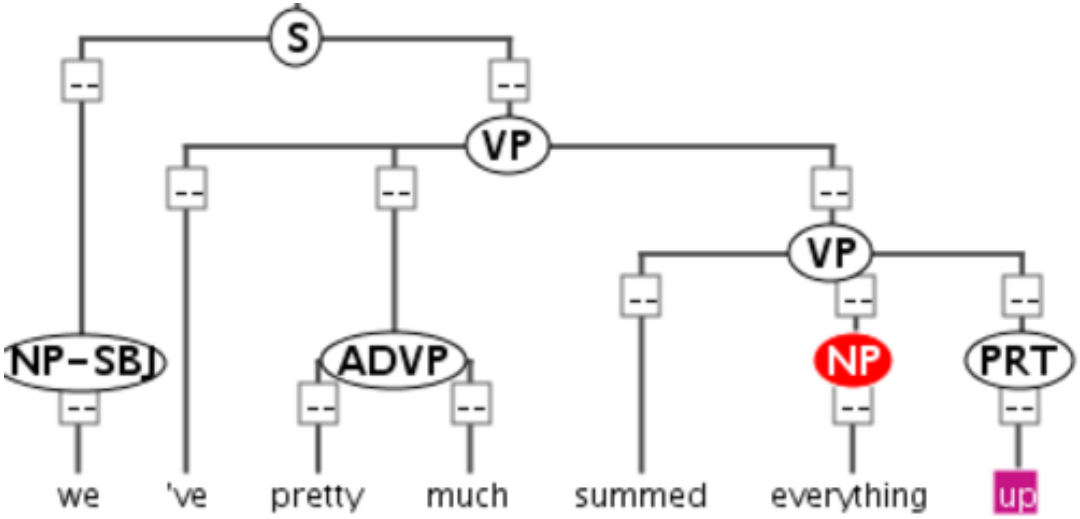
- Fertility and null insertion account for one-to-many and one-to-null alignments
- No accounting for many-to-one
- Strongly different word orders are penalized since many items must move (which belong together, e.g. VSO > SOV)
- What would we need to do?

Multiword lexicography?

- Bilingual dictionaries are also arranged around **words**
  - Always?
  - When not?

Phrase based approaches

- Hard to find ‘meaningful’ multi-word expressions based only the corpus
- Need to consider all **phrases**:



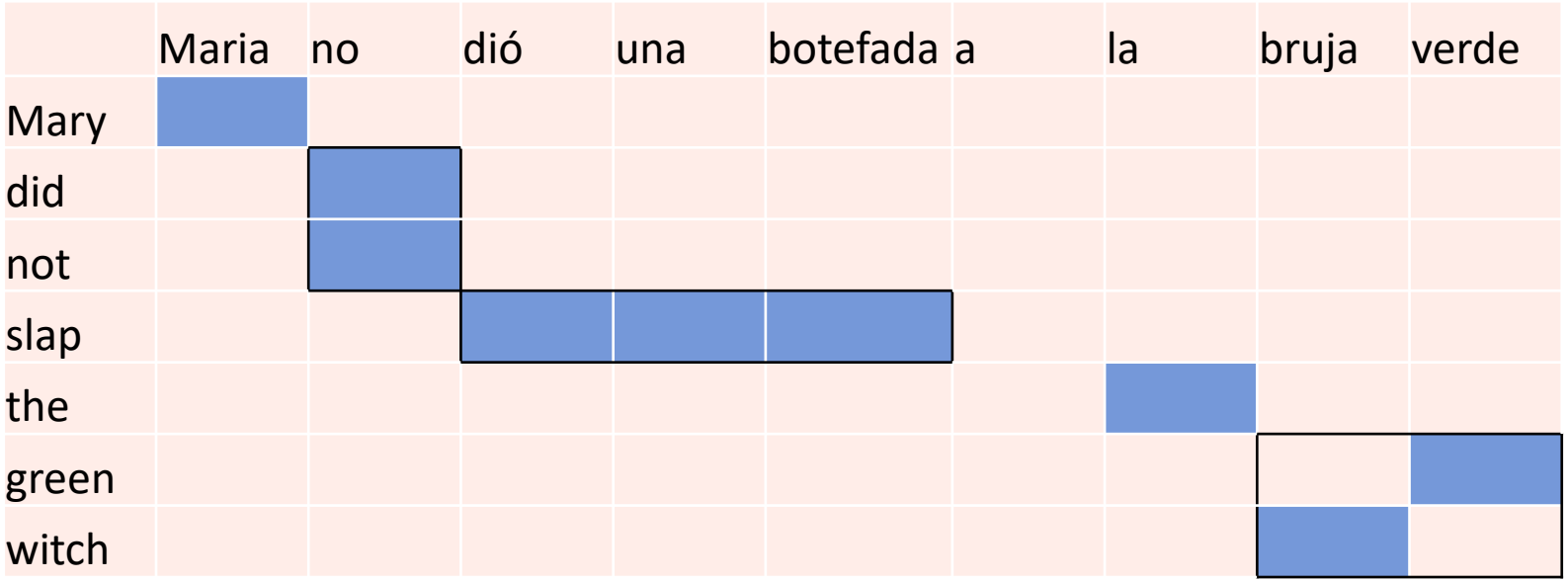
Syntactic phrases (constituents)

- Phrases are identified by three main criteria:
  - Pro-form substitution test:
    - I saw [the dog] yesterday -> I saw [it] yesterday
    - *I saw it *(to mean: the dog yesterday)*
  - Movement test:
    - The car hit [the dog] -> [The dog] was hit by the car
    - *dog was hit by the car the
  - Question test:
    - What did you see? The dog.
    - *the dog yesterday.

Phrase based approaches

- But parallel corpora do not come with hand made parse trees…
- Can we learn from running text?

Learning phrase alignment



Not aligned phrases