Introducción al Procesamiento de Lenguaje Natural

Grupo de PLN – InCo

Traducción Automática

Introducción

Uno de los primeros problemas de PLN

¿Por qué es difícil?

- Tipologías lingüísticas: SVO vs SOV
- Divergencia léxica: pata vs pierna vs leg
- Diferencias morfológicas: aglutinante vs fusional
- Densidad referencial: sujetos omitidos

Traducción Automática Basada en Reglas

Traducción Directa

Enfoque más simple

Traducir palabra a palabra

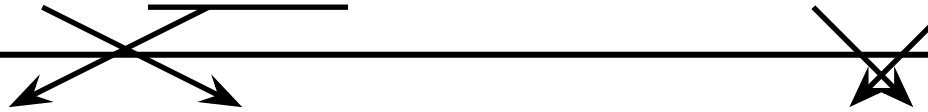
- Diccionario bilingüe

Podemos agregar algunas reglas simples de posprocesamiento

- Nom-Adj vs. Adj-Nom
- Generación morfológica

Traducción Directa

Entrada	Mary didn't slap the green witch
Análisis morfológico	Mary <DO-PAST> not slap the green witch
Transferencia léxica	María <PAST> no dar una bofetada a la verde bruja
Reordenamiento local	María no dar <PAST> una bofetada a la bruja verde
Generación morfológica	María no dio una bofetada a la bruja verde

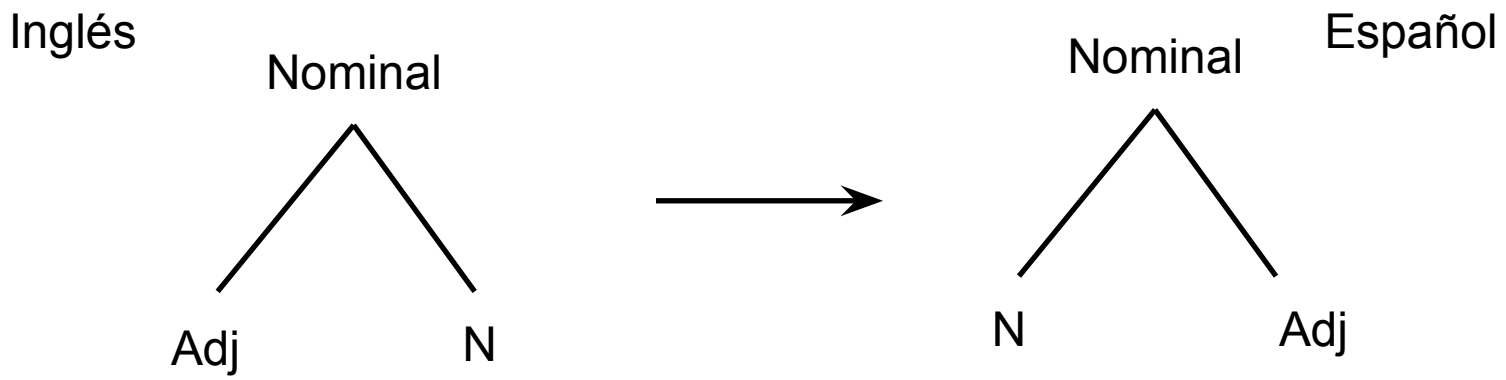


Transferencia Sintáctica

Parsing del lenguaje origen

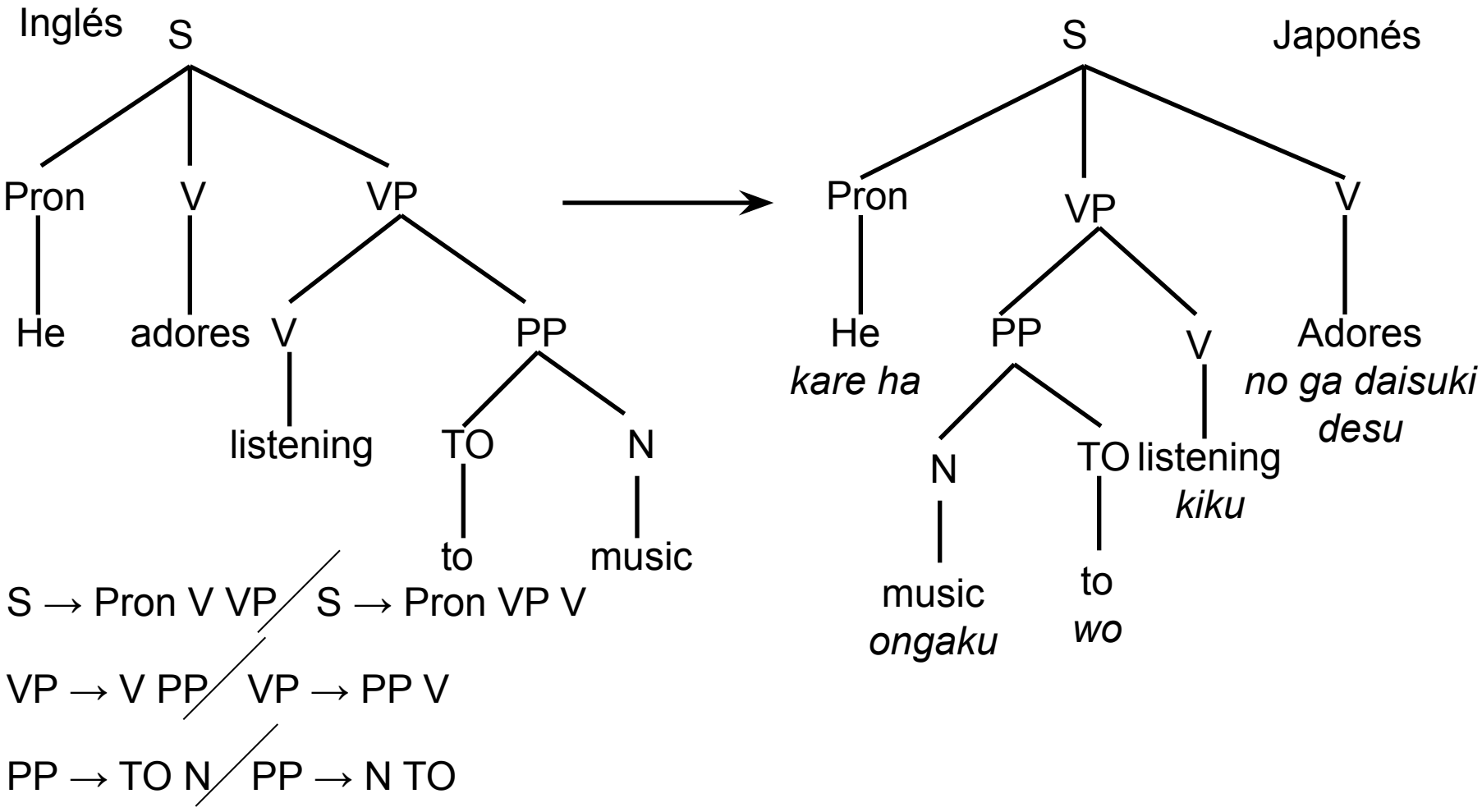
Generación en en lenguaje destino

Reglas de transferencia entre árboles y subárboles



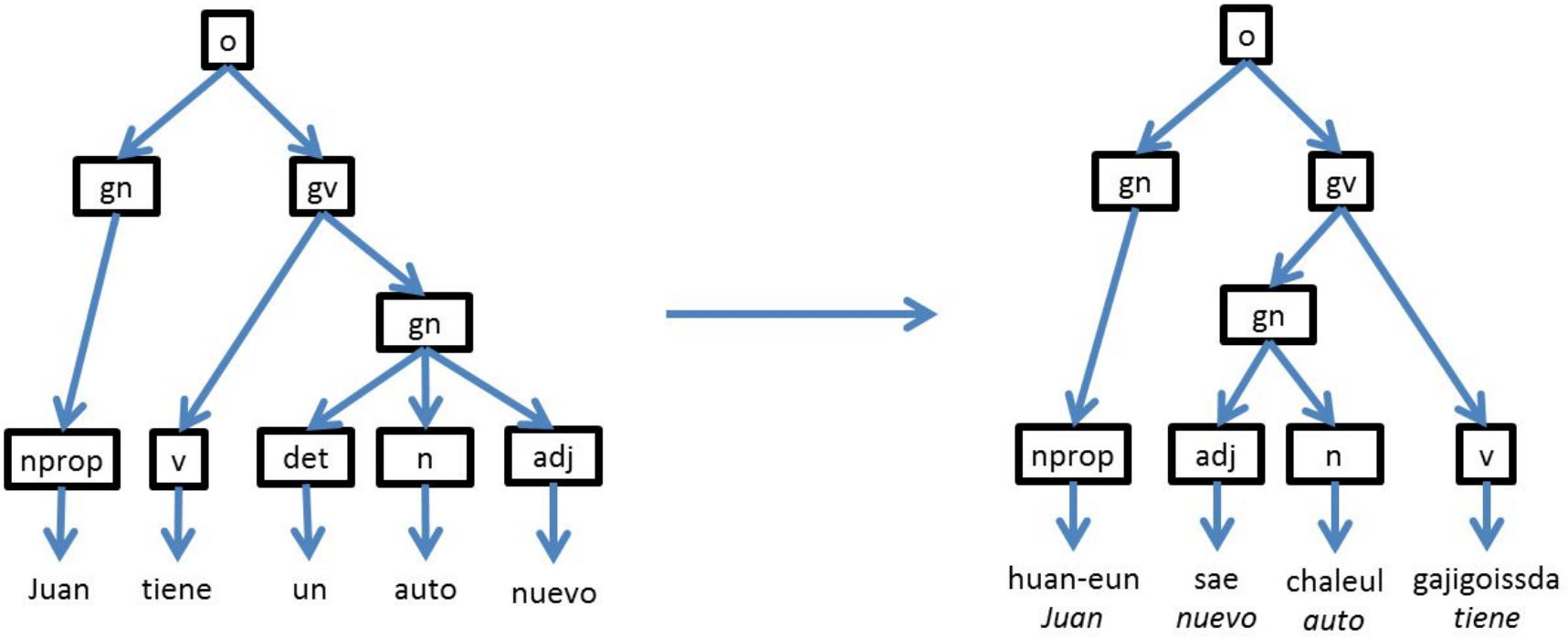
$$\text{Nom} \rightarrow \text{Adj N} \;/\; \text{Nom} \rightarrow \text{N Adj}$$

Transferencia Sintáctica



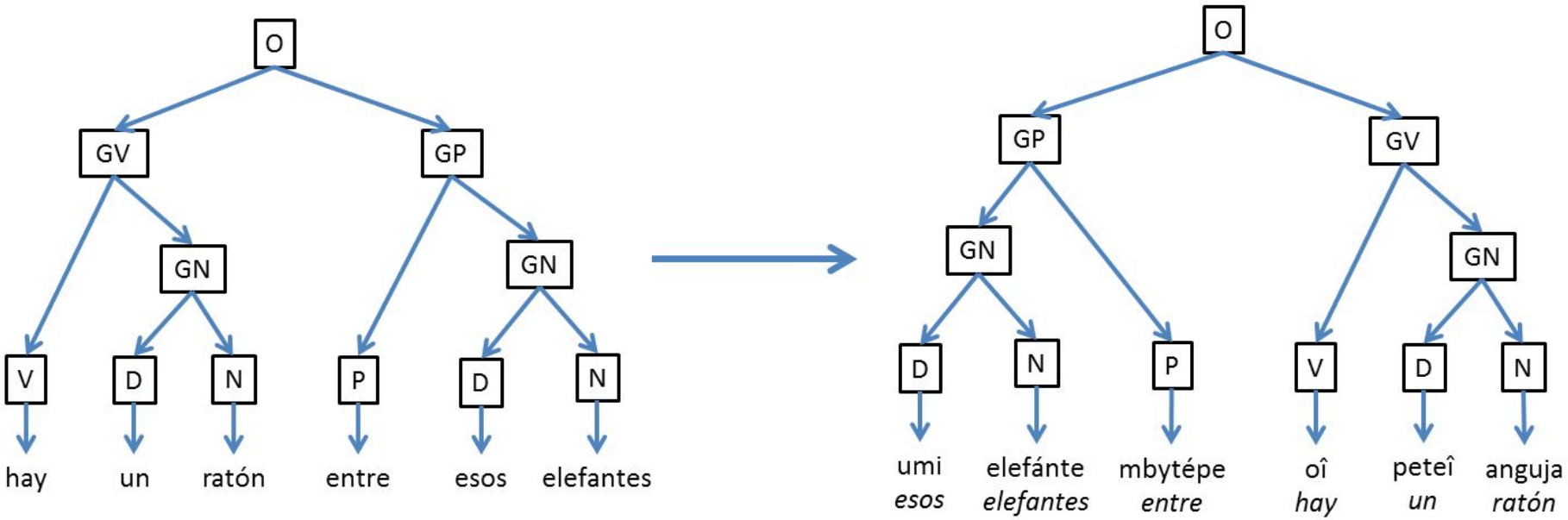
Ejercicio

(Febrero 2015) Suponga que se quiere construir un sistema de traducción automática del español al coreano basado en transferencia sintáctica. Infiera las reglas de transferencia que utilizaría dicho sistema, basándose en la siguiente traducción de ejemplo:

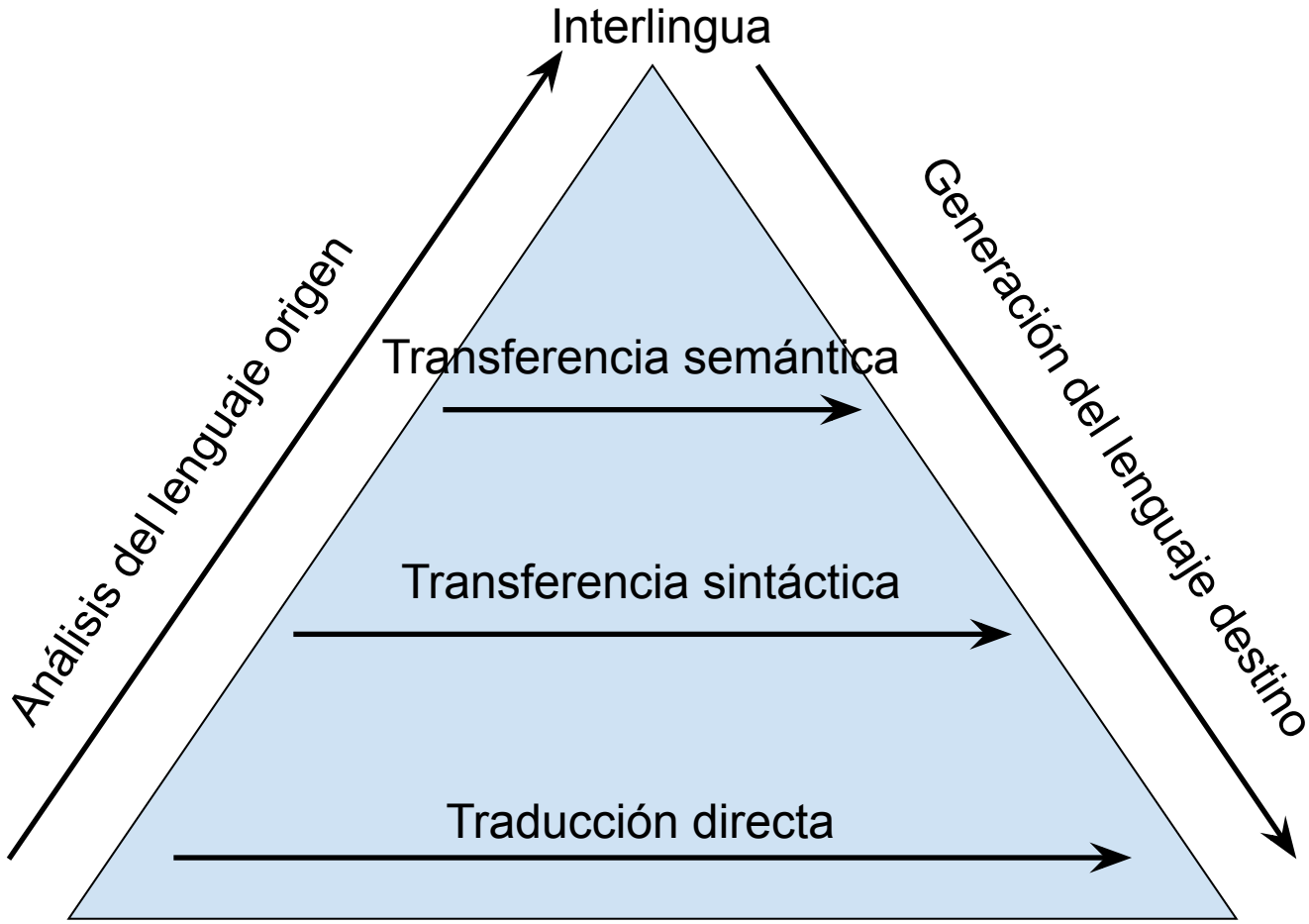


Ejercicio

(Febrero 2018) Suponga que se quiere construir un sistema de traducción automática del español al guaraní basado en transferencia sintáctica. Infiera las reglas de transferencia que utilizaría dicho sistema, basándose en la siguiente traducción de ejemplo:



Triángulo de Vauquois



Interlingua

IDEA: buscar una representación semántica común para todos los lenguajes

Capaz de captar el significado de todos → complicado

¡No necesitamos transferencia!

Interlingua

¿Qué representación utilizar como interlingua?

- Lógica de primer orden
- Minimum recursion semantics
- Frames de eventos

$$\begin{bmatrix} \text{EVENT} & \text{Slapping} \\ \text{AGENT} & \text{Mary} \\ \text{TENSE} & \text{Past} \\ \text{POLARITY} & \text{Negative} \\ \text{THEME} & \begin{bmatrix} \text{Witch} & & \\ \text{DEFINITENESS} & \text{Def} & \\ \text{ATTRIBUTES} & [\text{HAS}-\text{COLOR} & \text{Green}] \end{bmatrix} \end{bmatrix}$$

Interlingua

La representación conjunta debe modelar las características de todos los idiomas a la vez

- en chino existen palabras diferentes para "*hermano mayor*" y "*hermano menor*"
- esa información no es útil cuando traducimos de inglés a español

En la práctica se utilizan para dominios acotados

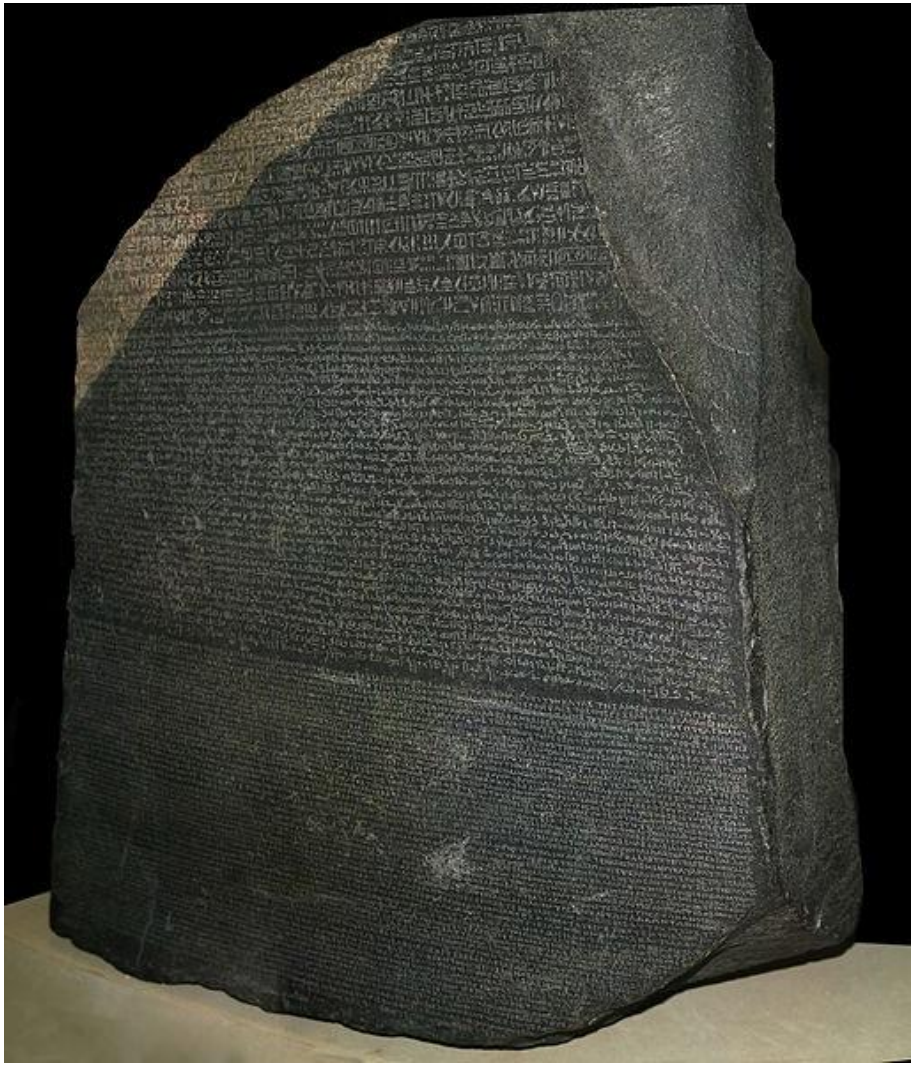
- Meteorología
- Manuales técnicos

Corpus Paralelos

Corpus Paralelos

Un ejemplo famoso de corpus paralelo:

La Piedra de Rosetta.



Corpus Paralelos

Conjuntos de pares de textos

- Un texto en el idioma origen y otro en el idioma destino
- A diferencia de los corpus monolingües, no es tan común que existan "naturalmente"

Dónde aparecen? Países multilingües y entidades multinacionales

- Francés-Inglés, Chino-Inglés
- Europarl: 11 idiomas más usados de la UE, 44M de palabras por idioma
- United Nations Parallel Corpus: 10M palabras en árabe, chino, español, francés, inglés, ruso

Hay alrededor de 7000 idiomas en el mundo

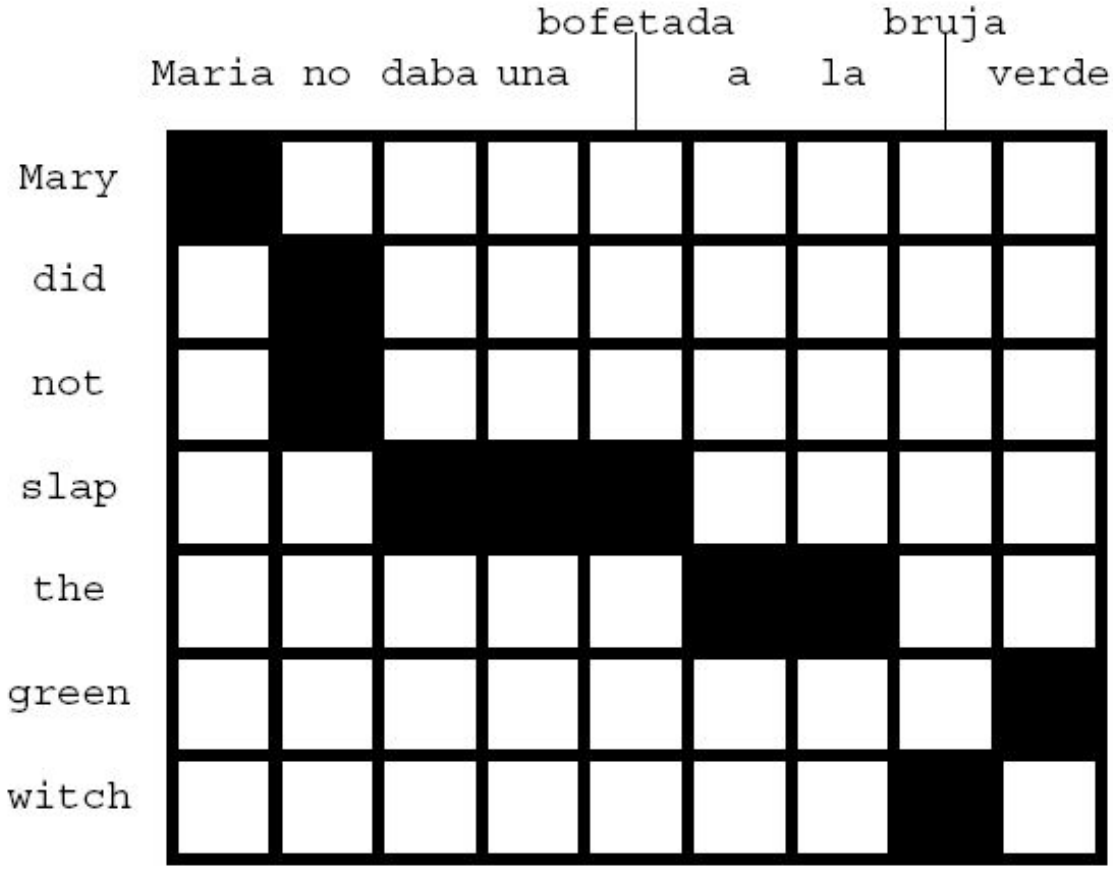
- La mayoría no tienen ningún tipo de corpus, mucho menos paralelo!

Corpus Paralelos

Diferentes tipos de *alineación*:

- Alineados a nivel de documento
- Alineados a nivel de oración
  - Algoritmo de Gale y Church
  - Distancia coseno con embeddings multilingües
- Alineados a nivel de palabra
  - Este es el ideal, pero en general no existen

Alineación


bofetada
bruja
Maria no daba una a la verde
Mary
did
not
slap
the
green
witch

Traducción Automática Estadística

MT Estadística

Tomamos una frase en español uruguayo:

Ese cuento es más viejo que el agujero del mate

¿Cómo la traducimos al inglés?

→Traducimos el significado de la metáfora:
"*That's a very old tale*"

→Traducimos literalmente, como se pueda:
"That tale is as old as the hole made to a type of calabash gourd that is used to drink a stimulant tea with a metal straw"

MT Estadística

La traducción que queremos debe:

Ser fiel a la oración original → **Adecuación** (Adequacy, Fidelity)

Sonar natural en el lenguaje destino → **Fluidez** (Fluency)

Este es el problema general al que se enfrentan los traductores humanos

- Es imposible obtener ambas en la práctica
- Hay que transar en un punto intermedio cómodo

Canal ruidoso

Lenguaje origen *F*, lenguaje destino *E*.

Oración origen $f=f_1,f_2,..,f_m$

Queremos encontrar la mejor oración en el lenguaje destino $\hat{e}=e_1,e_2,..,e_n$ que maximice:

$$\begin{aligned}\hat{e} &= \arg\max_e P(e \mid f) \\ &= \arg\max_e \frac{P(f \mid e)P(e)}{P(f)} \\ &= \arg\max_e P(f \mid e)P(e)\end{aligned}$$

MT Estadística

Quiero traducir una oración del idioma F al idioma E

$$\hat{e} = \arg\max_e P(f \mid e)P(e)$$

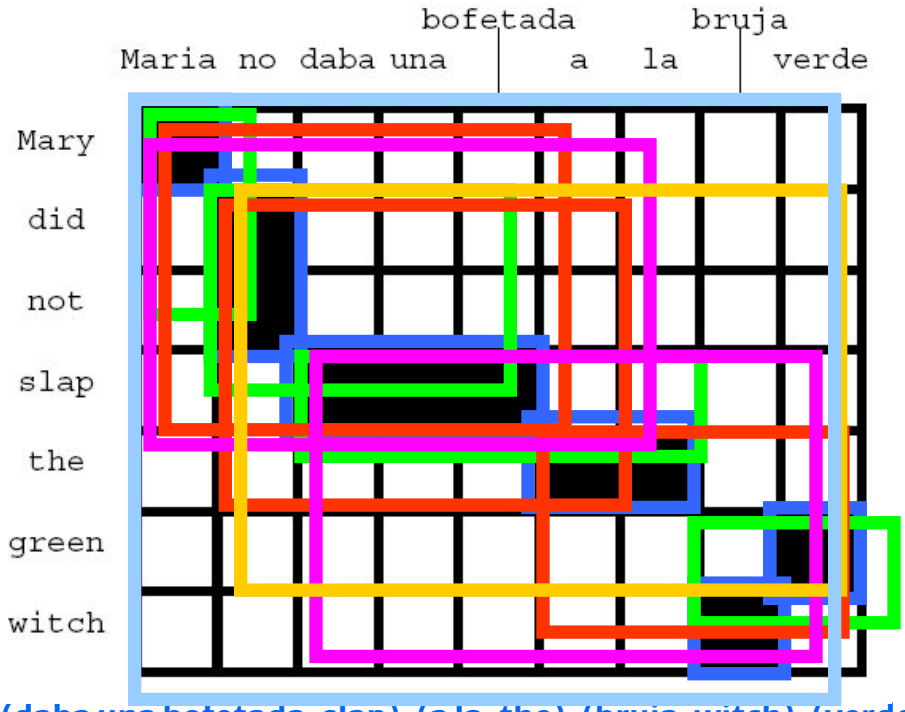
Para resolverla necesitamos:

- Un **modelo de lenguaje** P(*e*)
- Un **modelo de traducción** P(*f*|*e*)
- Un **decodificador**: proceso para buscar, de todas las oraciones *e* del idioma destino, cuáles son las más probables dado *f*

MT Estadística

- Modelos de lenguaje P(e)
  - n-gramas
- Modelos de traducción P($f|e$)
  - Uso de corpus paralelos (bilingües)
  - Basados en palabras
  - Basados en frases
- Decodificador:
  - Beam search

Modelo basado en frases



(María, Mary), (no, did not), (daba una bofetada, slap), (a la, the), (bruja, witch), (verde, green)

(María no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

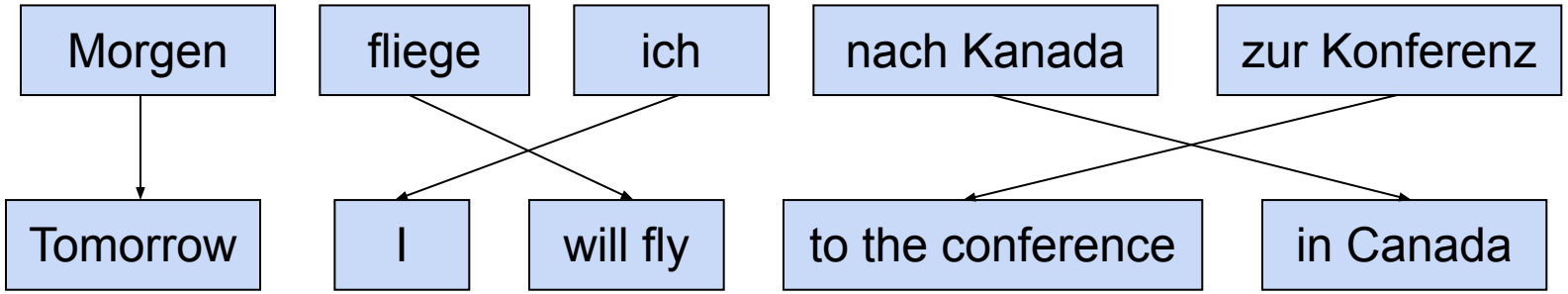
(María no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

(María no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch)

(no daba una bofetada a la bruja verde, did not slap the green witch),

(María no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Modelo basado en frases



La entrada en idioma origen se segmenta en frases

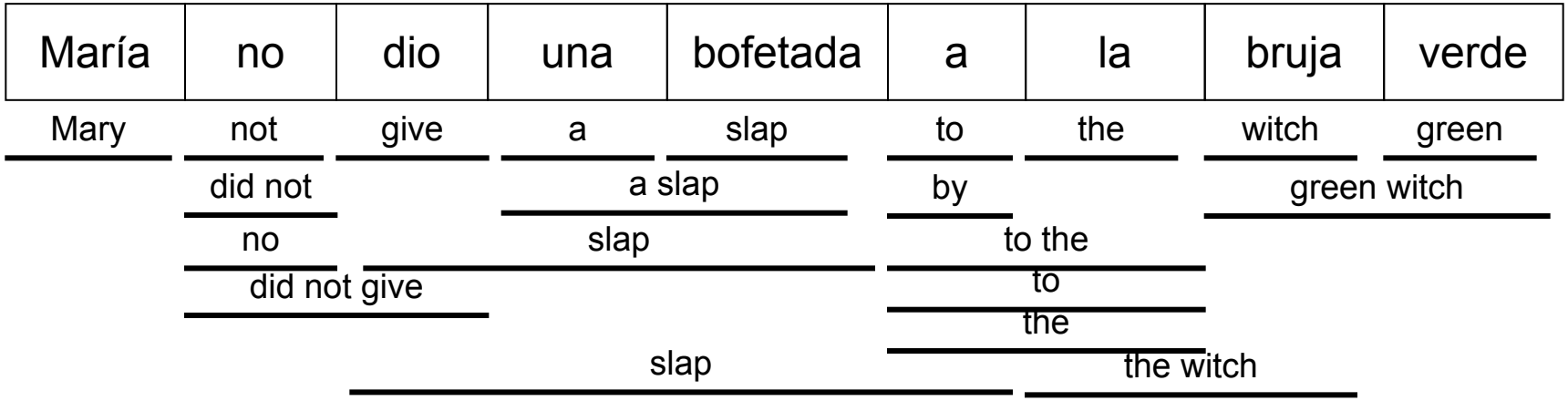
Cualquier secuencia de palabras, no es necesario que tenga significado lingüístico

Cada frase se traduce al idioma destino $\Phi(f,e)$

Se reordenan las frases $\Omega(f,e)$

$P(f|e) = \Phi(f,e)\,\Omega(f,e)$

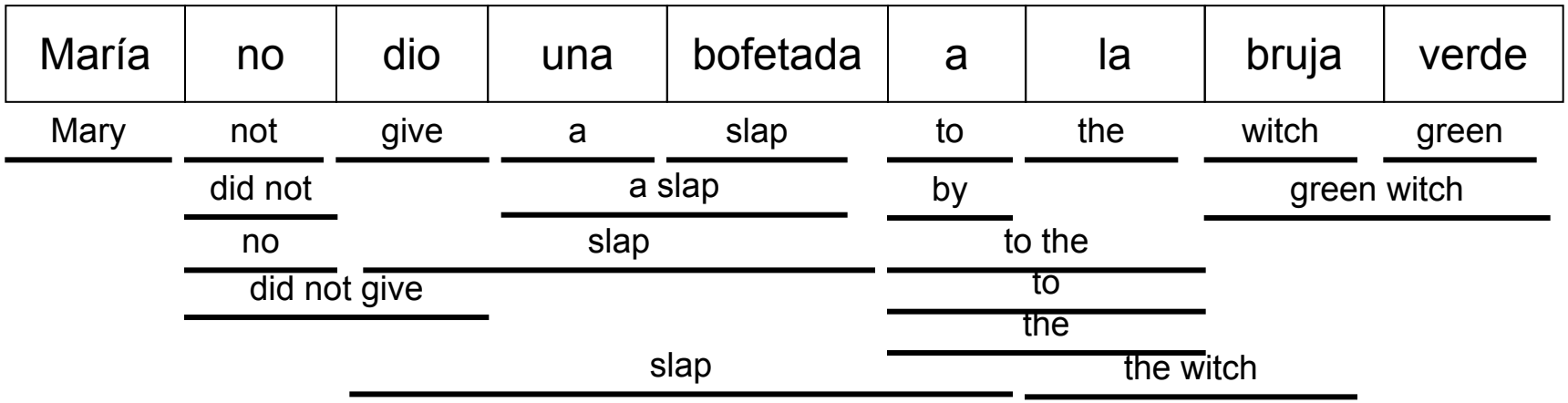
Expansión de hipótesis



Buscar *posibles traducciones de frase*

- varias formas de *segmentar* las palabras en frases
- varias formas de *traducir* cada frase

Expansión de hipótesis

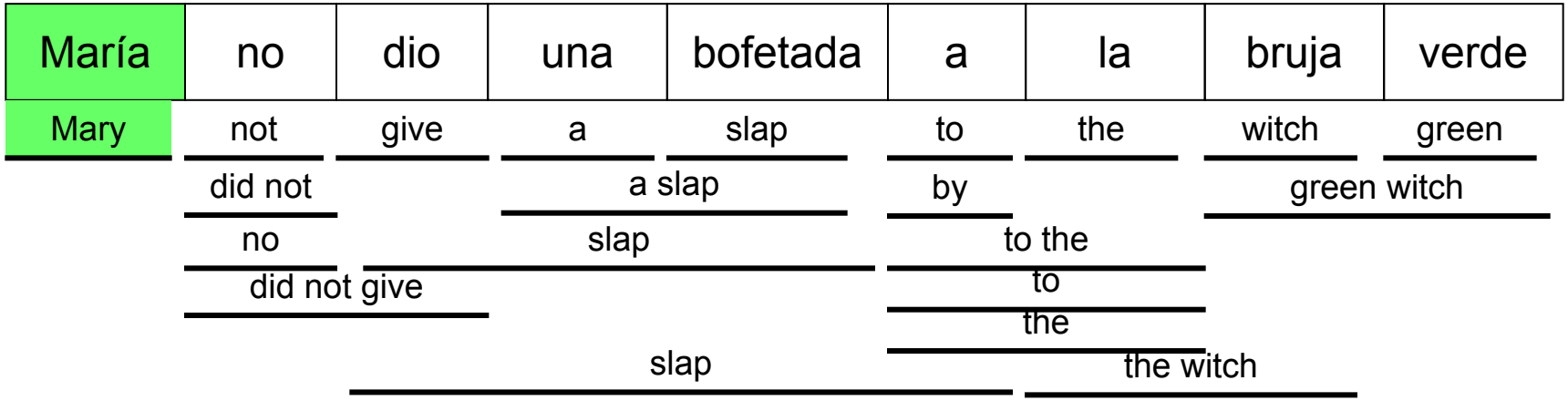


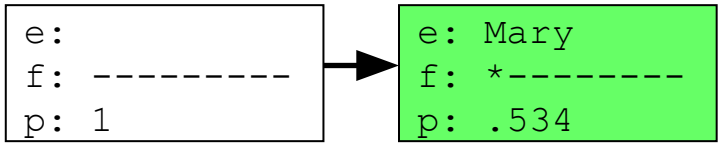

```
e:
f: ---------
p: 1
```

Comenzar con la hipótesis vacía

- e: no hay palabras en inglés
- f: no se cubre ninguna palabra foránea
- p: probabilidad 1

Expansión de hipótesis

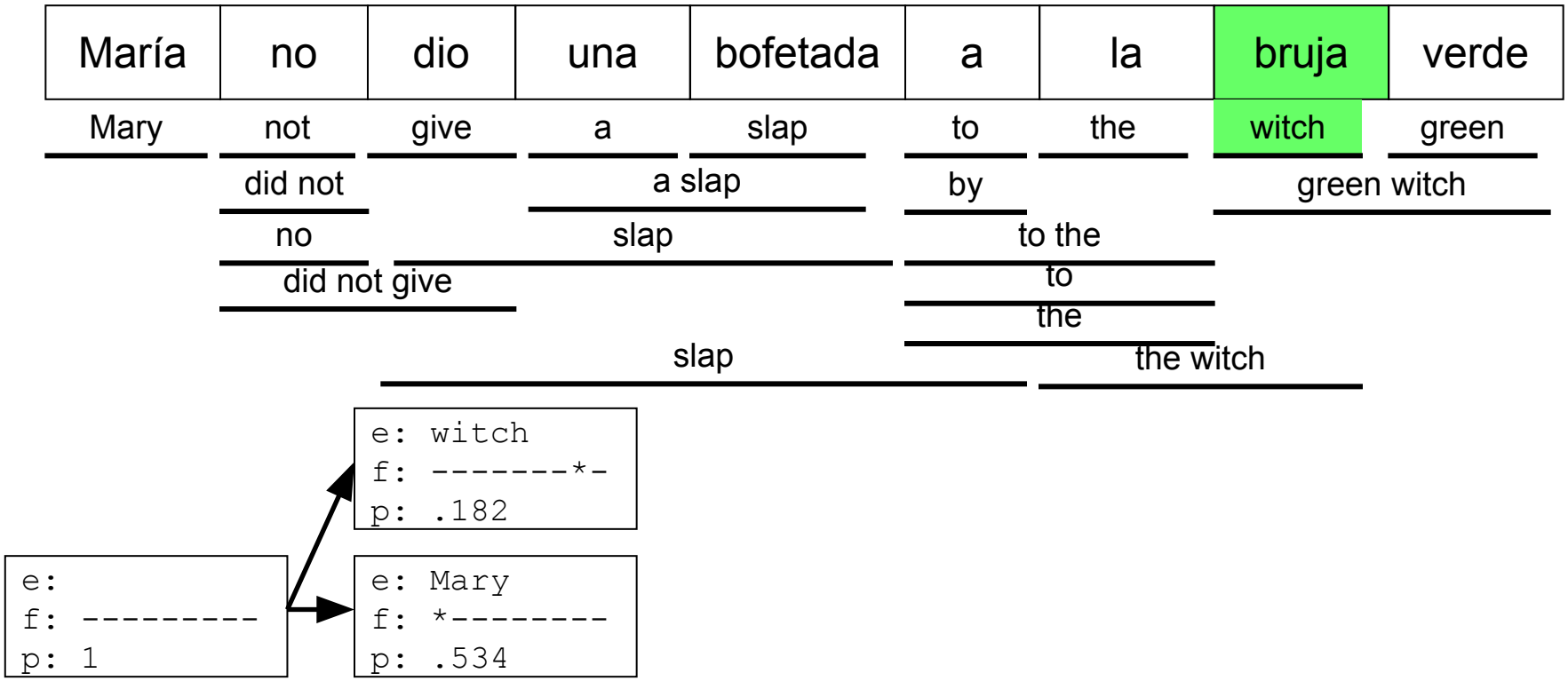




Elegir una opción de traducción:

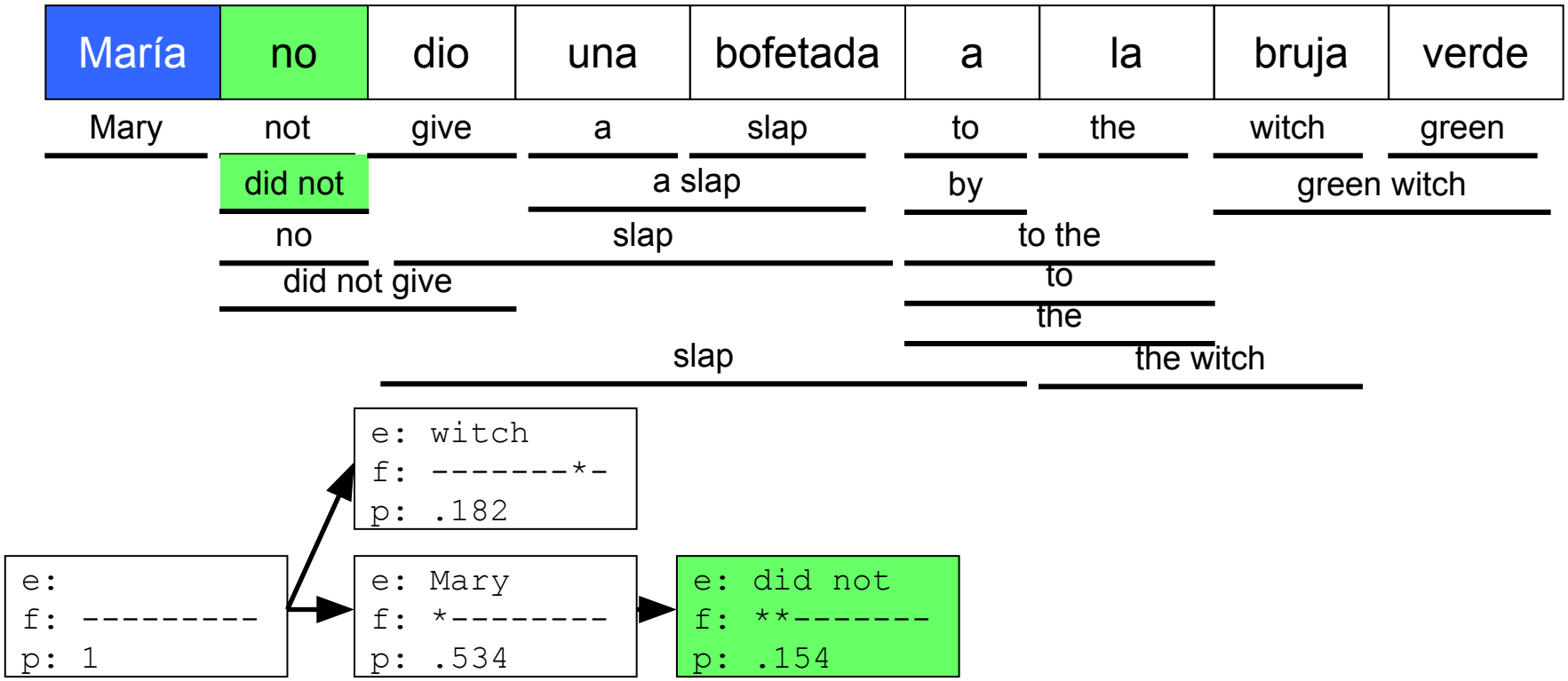
- **e: insertamos la frase en inglés Mary**
- **f: se cubre la primera palabra foránea**
- **p: probabilidad 0.534**

Expansión de hipótesis



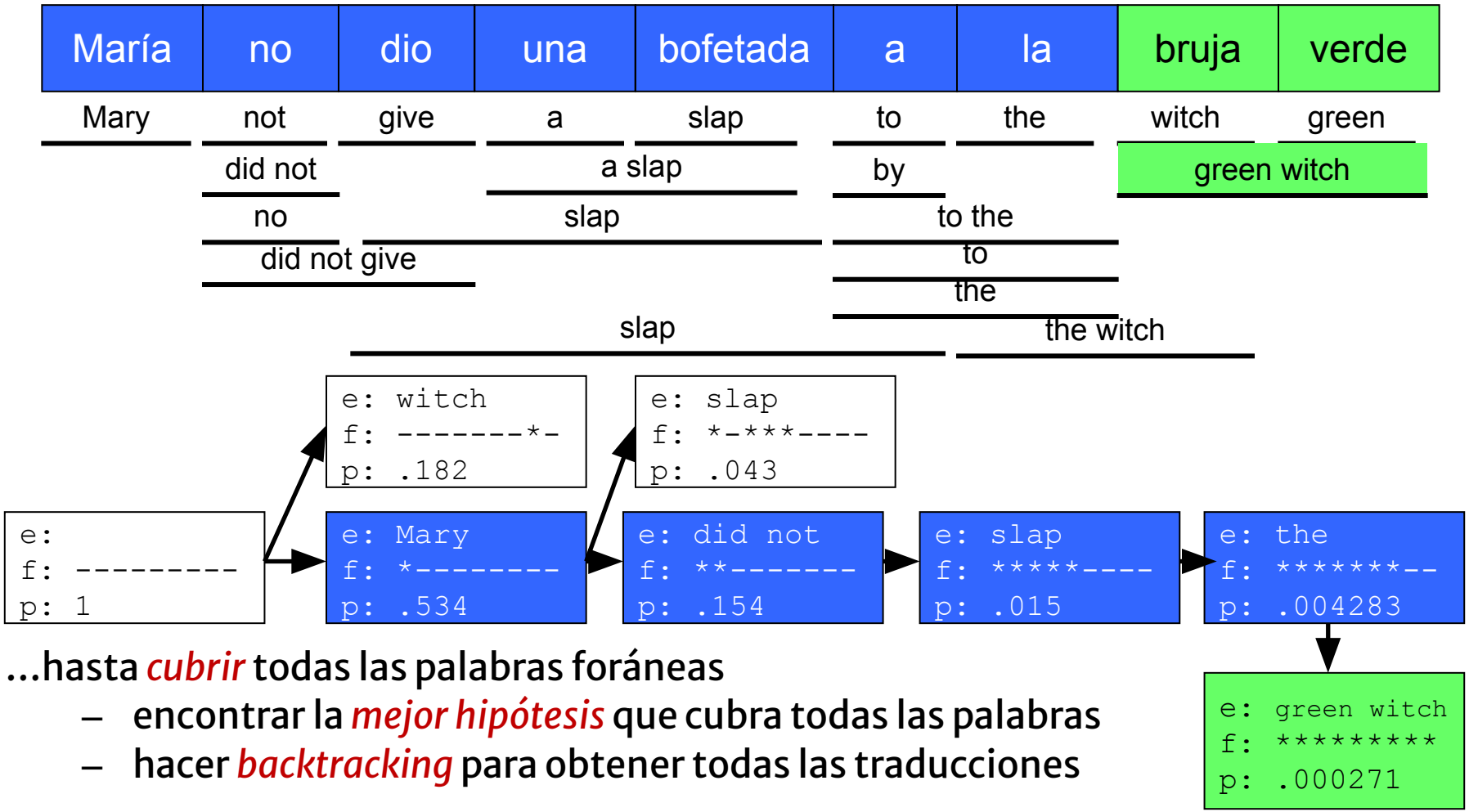
Insertar otra *hipótesis*

Expansión de hipótesis



Se va insertando una nueva hipótesis en cada paso del algoritmo

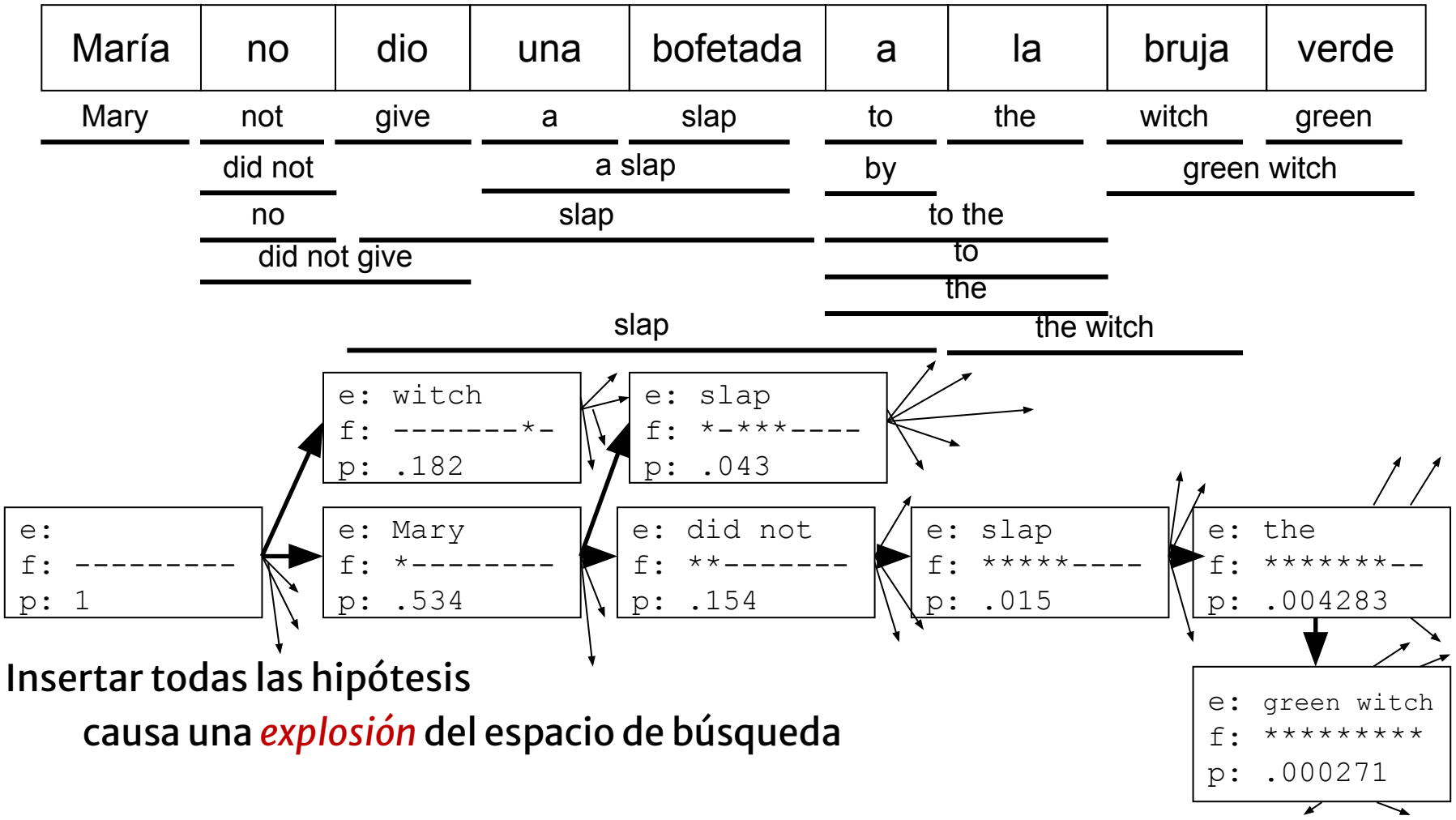
Expansión de hipótesis



...hasta *cubrir* todas las palabras foráneas

- encontrar la *mejor hipótesis* que cubra todas las palabras
- hacer *backtracking* para obtener todas las traducciones

Expansión de hipótesis



Insertar todas las hipótesis

causa una *explosión* del espacio de búsqueda

Explosión del espacio de búsqueda

Número de hipótesis *exponencial* respecto al largo de la oración

Reducción del espacio de búsqueda:

- sin riesgo: *recombinación* de hipótesis
- con riesgo: *podado* de hipótesis
  - las *n* mejores a cada paso
  - las de diferencia máxima α respecto a la mejor

Traducción Automática Neuronal

Traducción Automática Neuronal

La traducción automática es un problema de correspondencia entre pares de secuencias

- Respuestas a preguntas
- Resúmenes automáticos
- Otros...

Redes tipo seq2seq (secuencia a secuencia)

Traducción Automática Neuronal

Modelo codificador-decodificador (encoder-decoder)

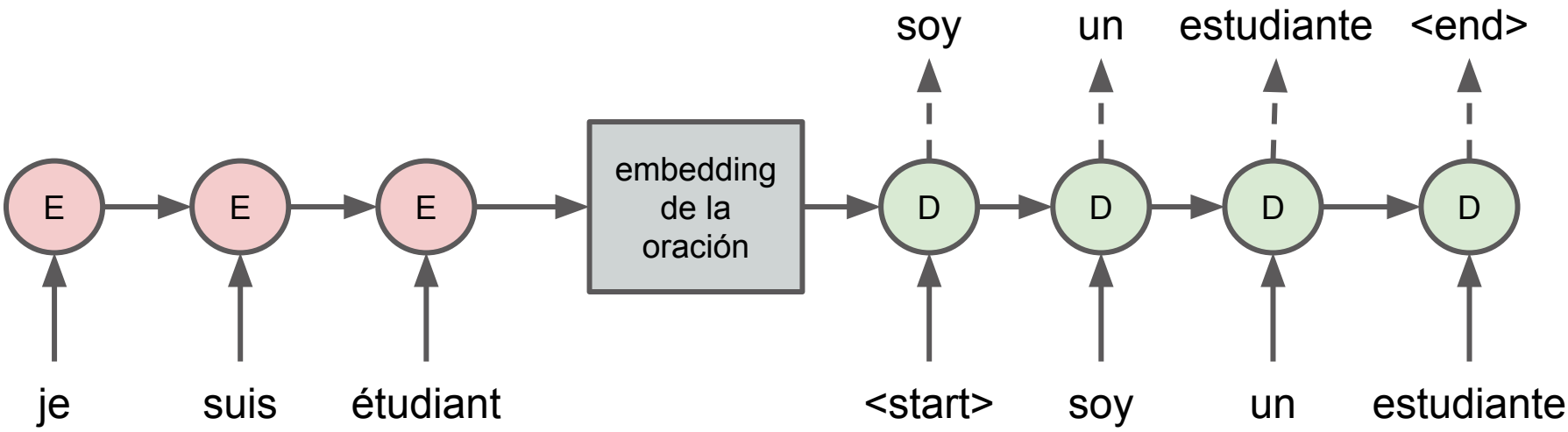
Una red codifica la oración en el idioma origen

Una red genera la decodificación en el idioma destino

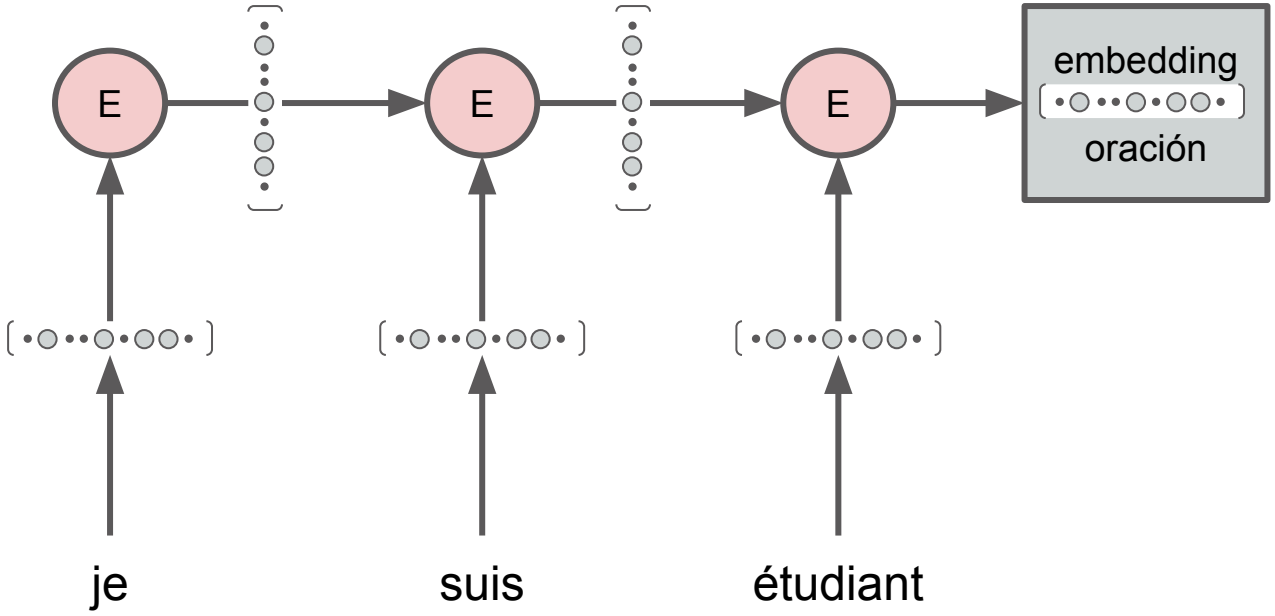
Se utilizan arquitecturas secuenciales:

- RNN/LSTM
- Transformer

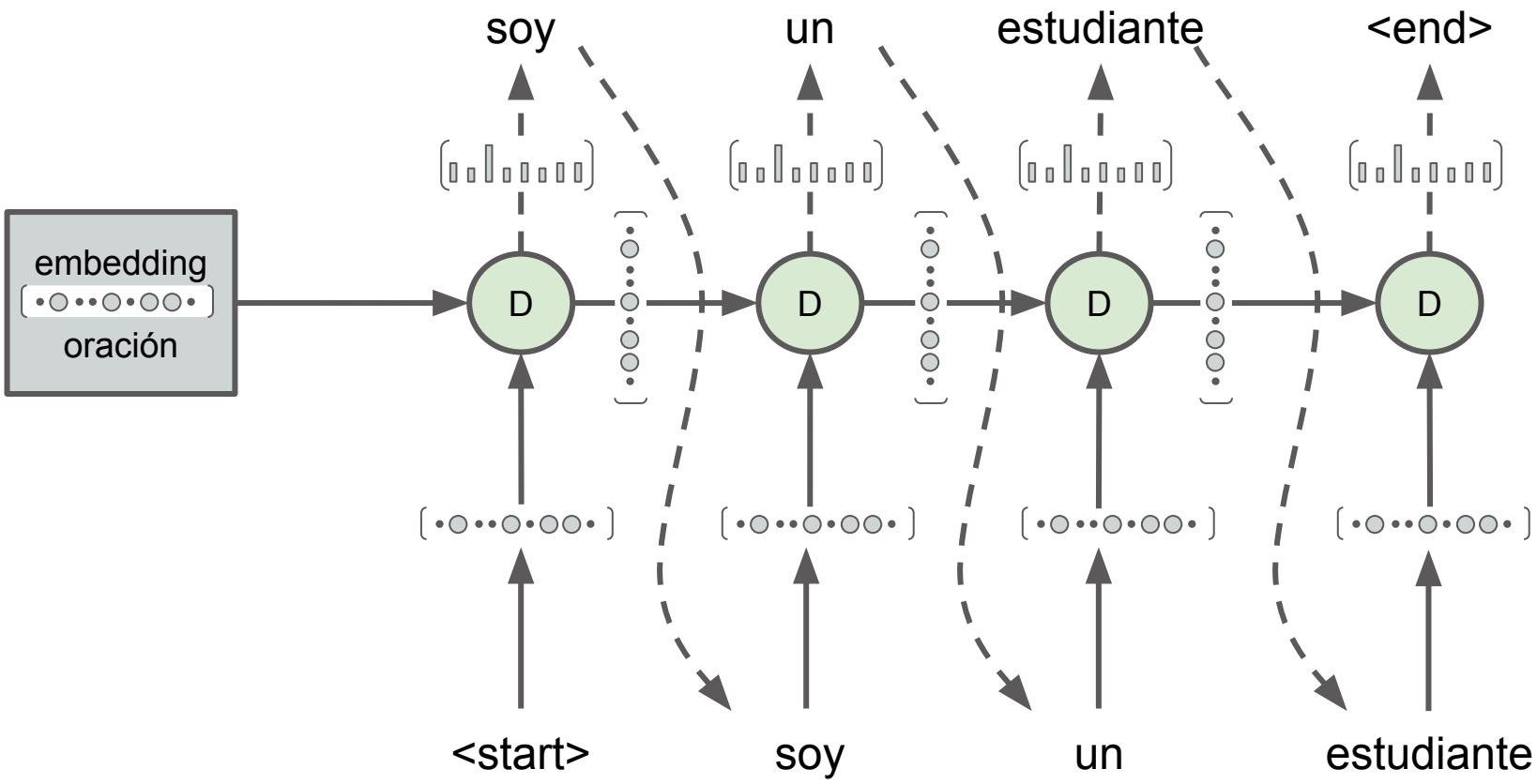
Modelo Encoder-Decoder



Encoder



Decoder


soy
un
estudiante
<end>
embedding
oración
D
D
D
D
<start>
soy
un
estudiante

Entrenamiento

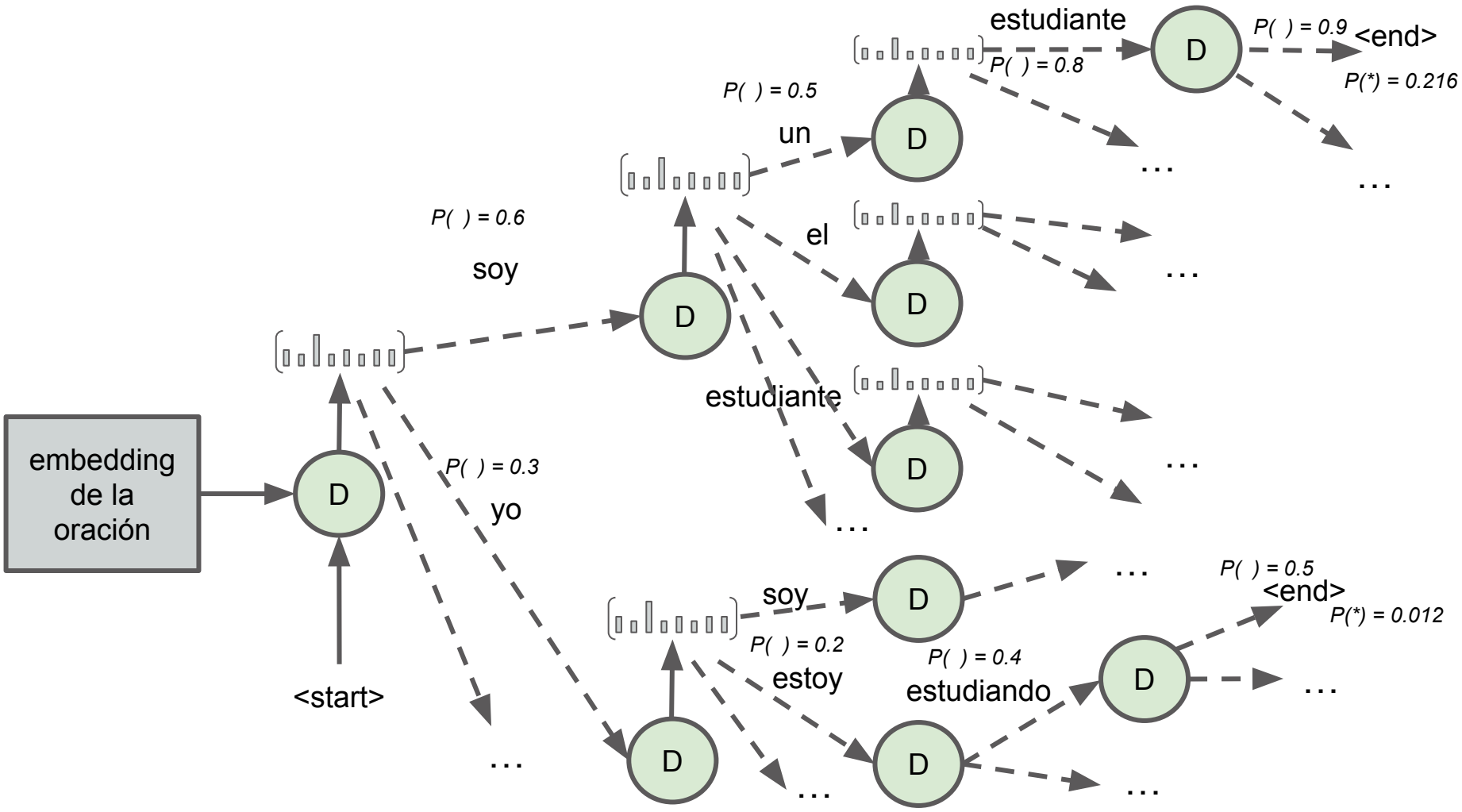
End-to-end: usamos pares de oraciones alineados como ejemplos de entrada y salida esperadas

Cada entrada del decoder usa la palabra correcta esperada, y no la salida del paso anterior

Tokenización:

- Se define un vocabulario fijo para cada idioma
- O se puede usar un vocabulario compartido entre idiomas
- Esto permite usar embeddings compartidos para inicialización

Beam Search

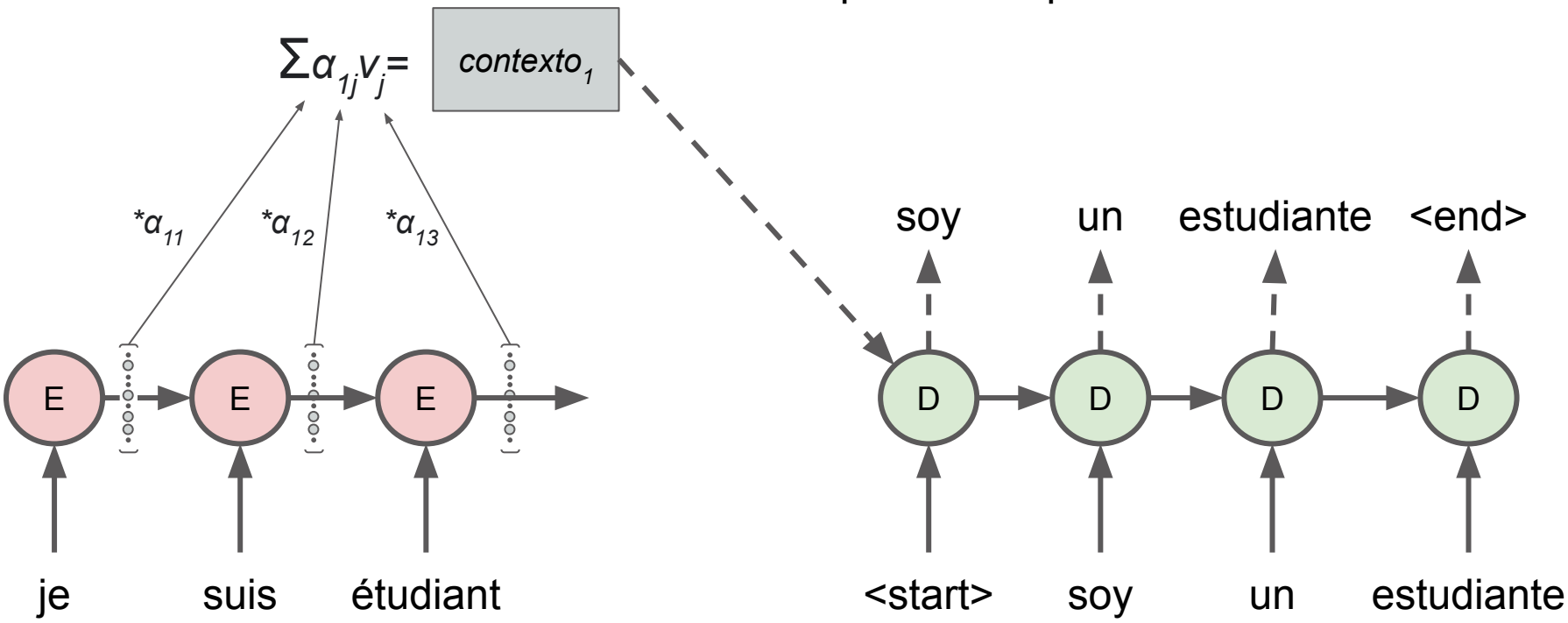

embedding de la oración
<start>
D
P( ) = 0.3
yo
P( ) = 0.6
soy
P( ) = 0.5
un
el
estudiante
P( ) = 0.8
estudiante
P( ) = 0.9
<end>
P(*) = 0.216
soy
P( ) = 0.2
estoy
P( ) = 0.4
estudiando
P( ) = 0.5
<end>
P(*) = 0.012
…

Modelo Atencional

Se crea un vector de contexto nuevo en cada paso del decoder

Promedio ponderado de los embeddings del encoder

La forma de ponderar se aprende durante el entrenamiento

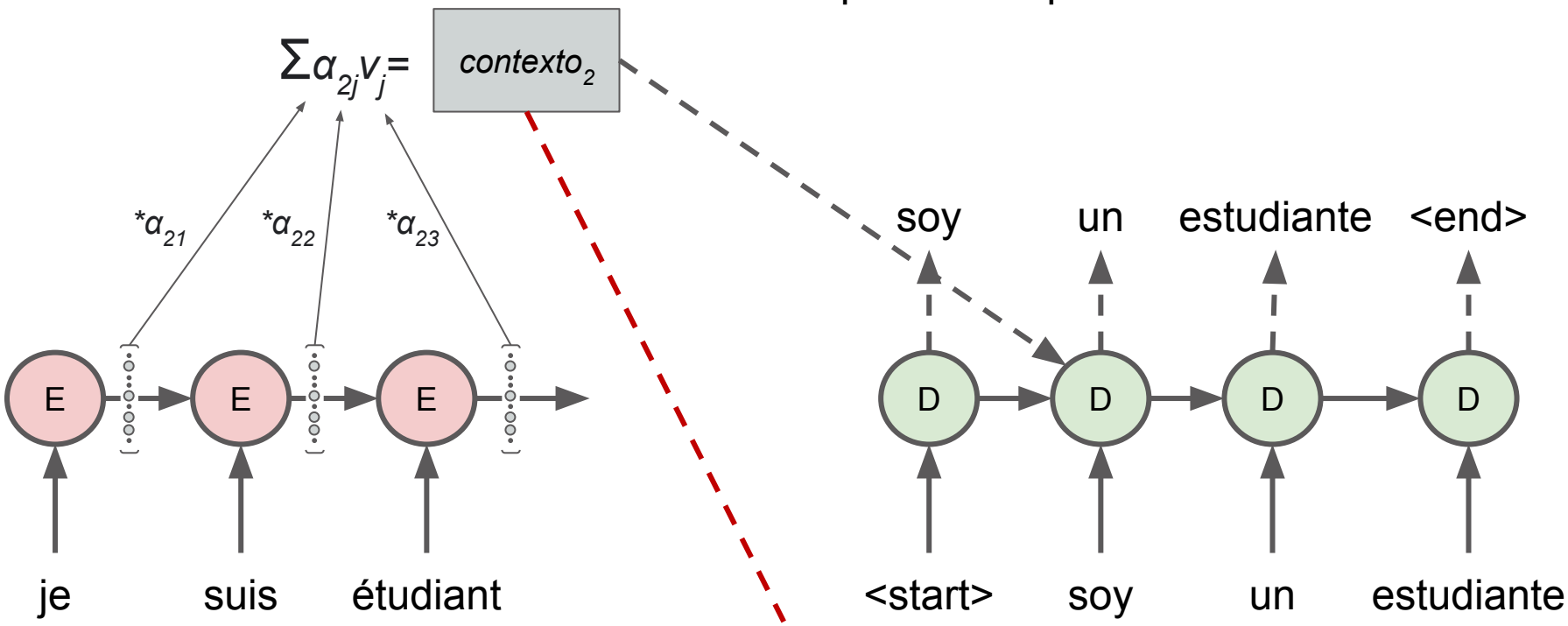


Modelo Atencional

Se crea un vector de contexto nuevo en cada paso del decoder

Promedio ponderado de los embeddings del encoder

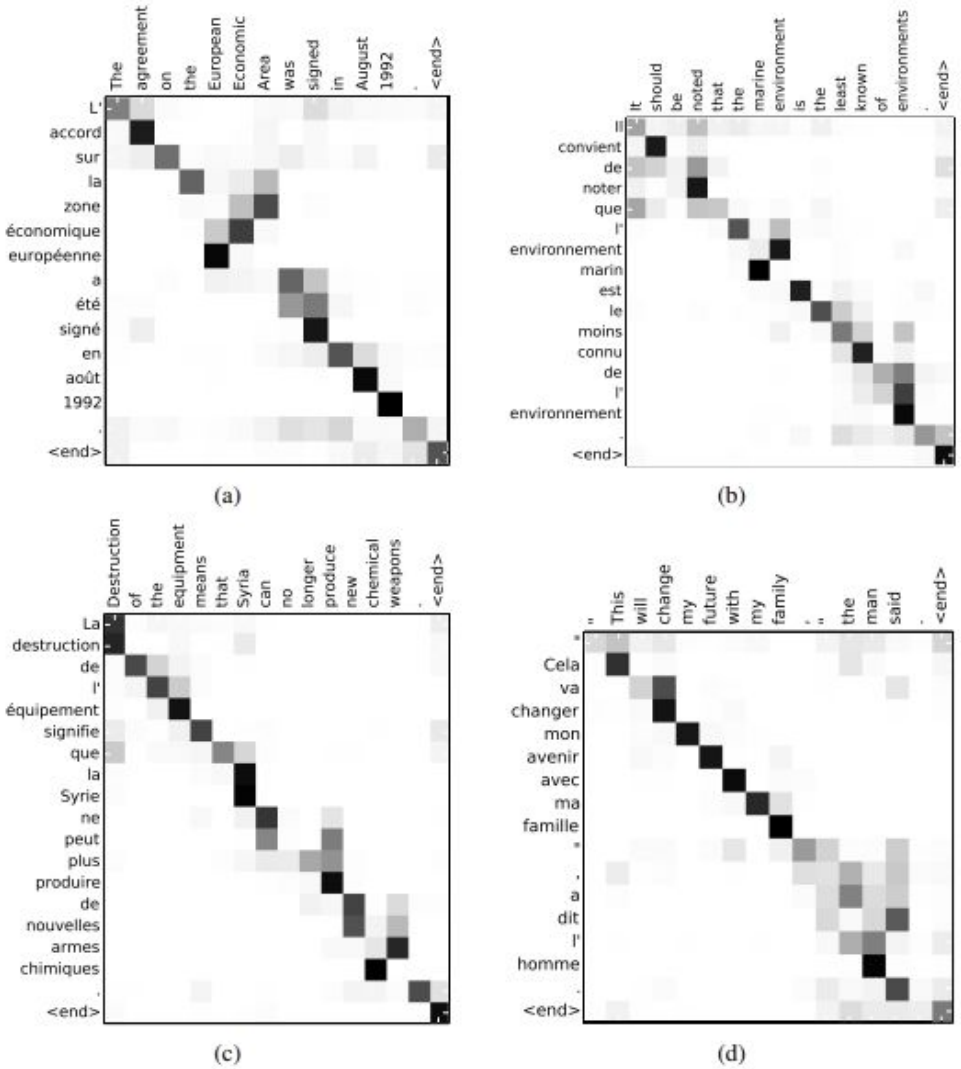
La forma de ponderar se aprende durante el entrenamiento



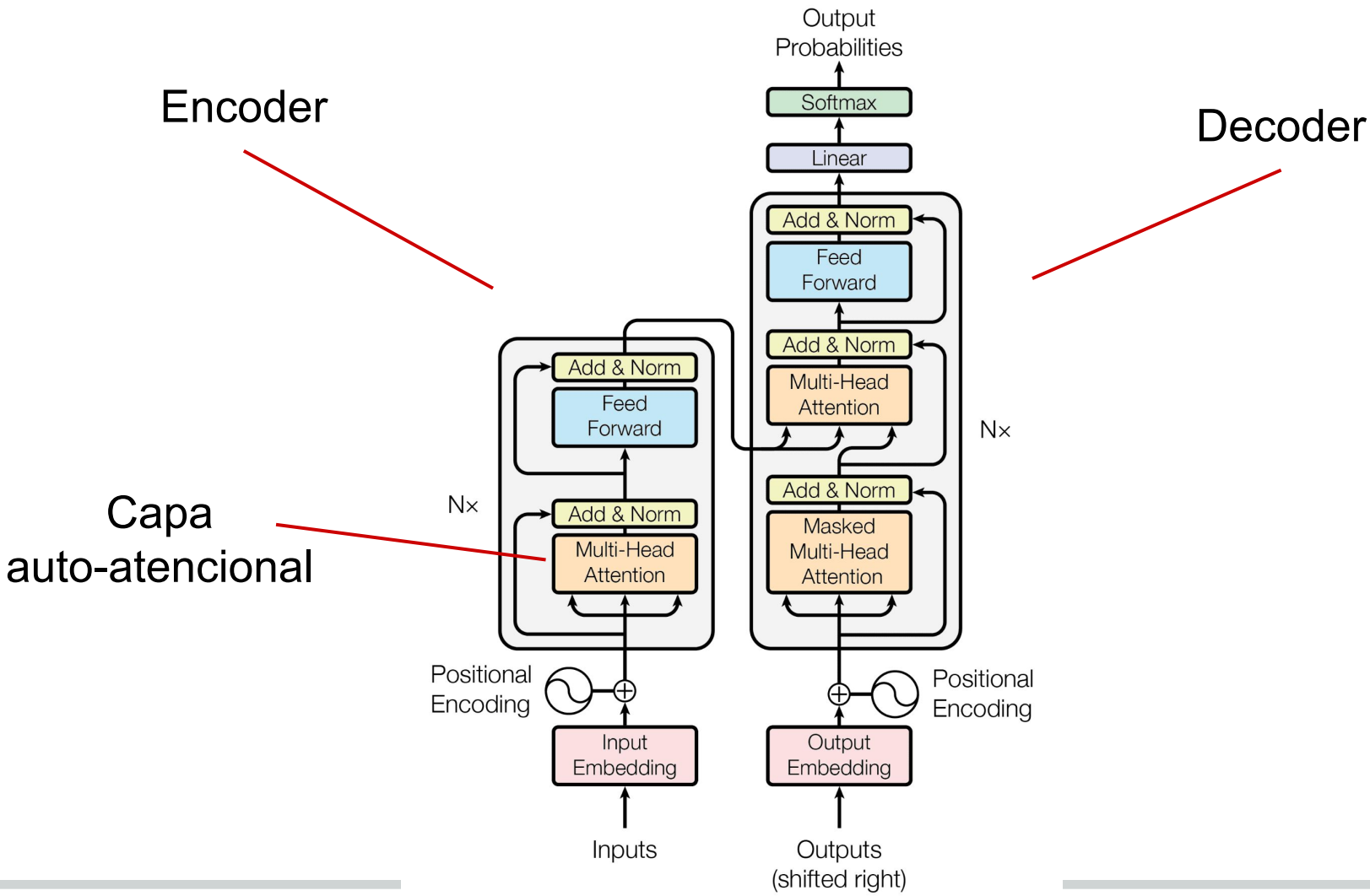
¿Cuántas ponderaciones α_{ij} hay?

*largo oración origen * largo oración destino*

Modelo Atencional



Transformer



Evaluación

Evaluación

Medida BLEU

$$BLEU = BP \exp\left(\sum_{1}^{N} w_n \log p_n \right)$$

- Compara un conjunto de traducciones candidatas con un conjunto de traducciones de referencia
- Cuenta n-gramas presentes en los candidatos que también estén en las referencias (n = 1,2,3,4)
- Incluye una penalización por brevedad (BP) para que las traducciones demasiado cortas tengan menos puntos

Evaluación

Definición de BP:

$$BP = \begin{cases} 1 & (C' \geq R') \\ \exp\left(1 - \frac{R'}{C'}\right) & (C' < R') \end{cases}$$

Donde R' es el largo total (en palabras) de todas las referencias (documento referencia) y C' es la el largo total de las traducciones candidatas (documento candidato)

Ejercicio

Calcular BLEU para los siguientes candidatos a traducción inglés-español:

Original	Referencia	Candidato 1	Candidato 2
Yesterday it was raining very heavily	Ayer estaba lloviendo muy fuerte	Llovía muy fuerte ayer	Ayer estaba lloviendo pesadamente
I wasn't carrying an umbrella	No llevaba paraguas	Yo no llevaba un paraguas	Él llevaba paraguas
Then I took shelter under a roof	Entonces me refugié bajo un techo	Entonces me refugié debajo de un techo	Entonces refugié bajo techo

Considere n-gramas hasta n=3 y pesos w_i equiprobables

Evaluación

Medida chrF

A diferencia de BLEU, que cuenta n-gramas de tokens, chrF utiliza n-gramas de caracteres

$$\mathrm{chrF}\beta = (1+\beta)\frac{\mathrm{chrP}\cdot\mathrm{chrR}}{\beta^2\cdot\mathrm{chrP}+\mathrm{chrR}}$$

- chrP: cantidad de n-gramas de caracteres de la hipótesis que están en la referencia
- chrR: cantidad de n-gramas de caracteres de la referencia que están en la hipótesis
- En general $n \leq 4$ o 6, y $\beta = 3$

Evaluación

Tanto BLEU como chrF toman valores entre 0 y 1 como BLEU

Las dos penalizan traducciones que no estén entre las referencias

En general se correlacionan con la evaluación subjetiva humana, pero el número resultante en sí es difícil de interpretar

BLEU es más estricta que chrF porque cuenta palabras exactas, mientras que chrF igual da algunos puntos si se tradujo una sub-palabra

chrF es mejor para hacer comparaciones con idiomas morfológicamente ricos, ya que puede haber muchas flexiones de una misma palabra

Herramientas

- Sistemas basados en reglas
  - Apertium: http://www.apertium.org/
- Sistemas estadísticos
  - Moses: http://www.statmt.org/moses/
  - IRSTML Toolkit http://hlt.fbk.eu/en/irstlm
  - GIZA++ https://code.google.com/p/giza-pp/
- Sistemas neuronales
  - OpenNMT http://opennmt.net/
  - MarianNMT https://marian-nmt.github.io/

Ejemplo:
Traducción Guaraní–Español

Idioma Guaraní

- Lengua indígena de América del Sur
- Hablada por entre 6 y 10 millones de personas

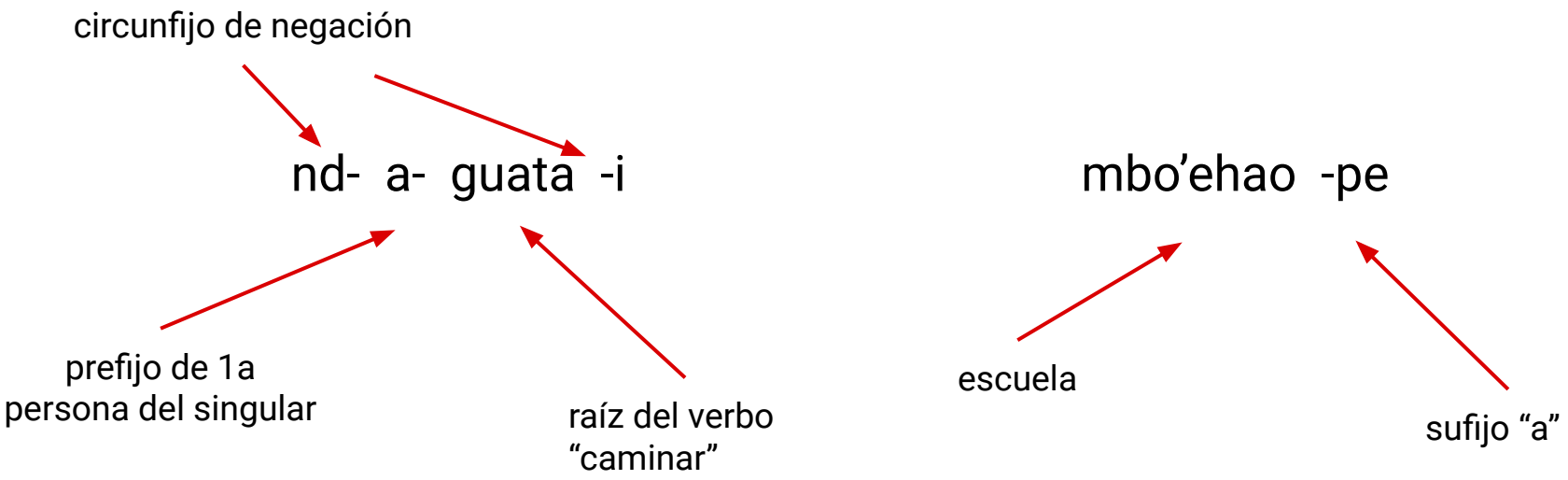
  Paraguay, Argentina, Bolivia, Brasil

- Contacto con español y otras lenguas europeas por alrededor de 500 años
- Hablada por toda la sociedad, no solo población indígena

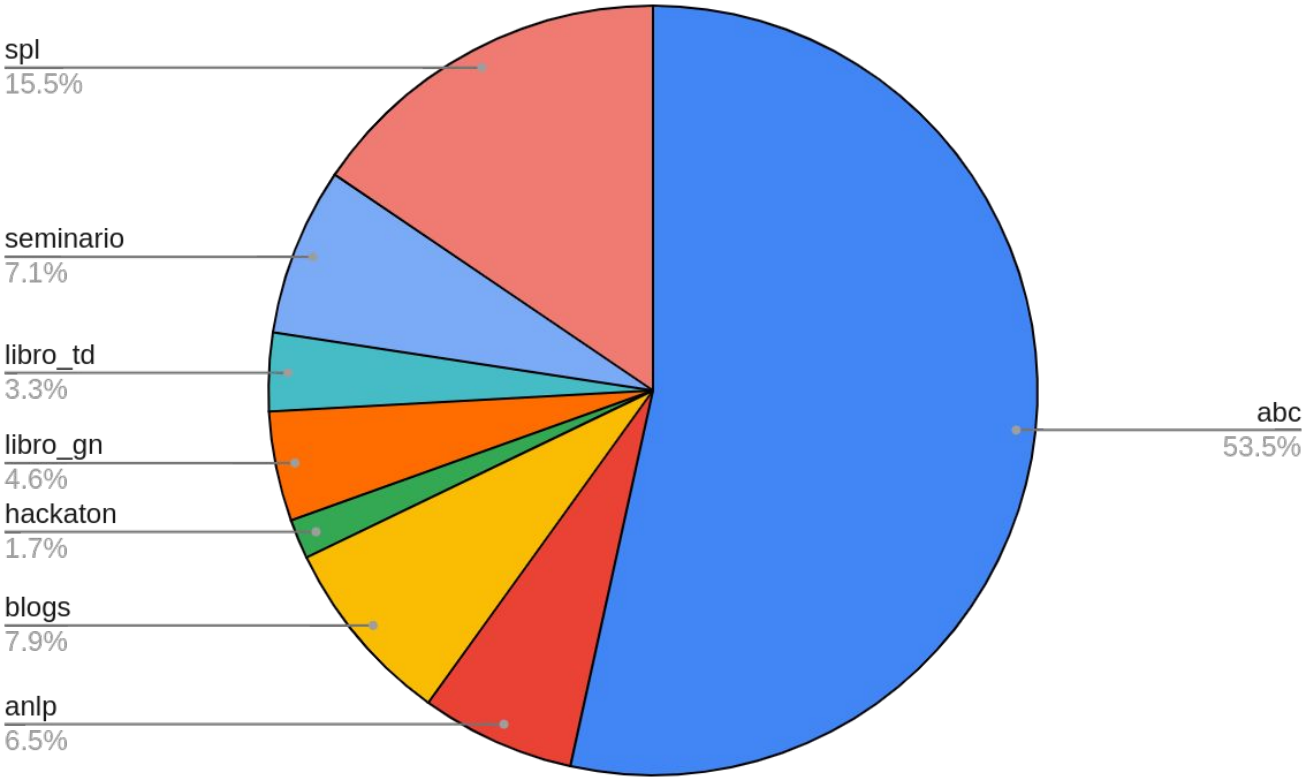
Sintaxis del Guaraní

ndaguatái mbo'ehaópe

No camino a la escuela



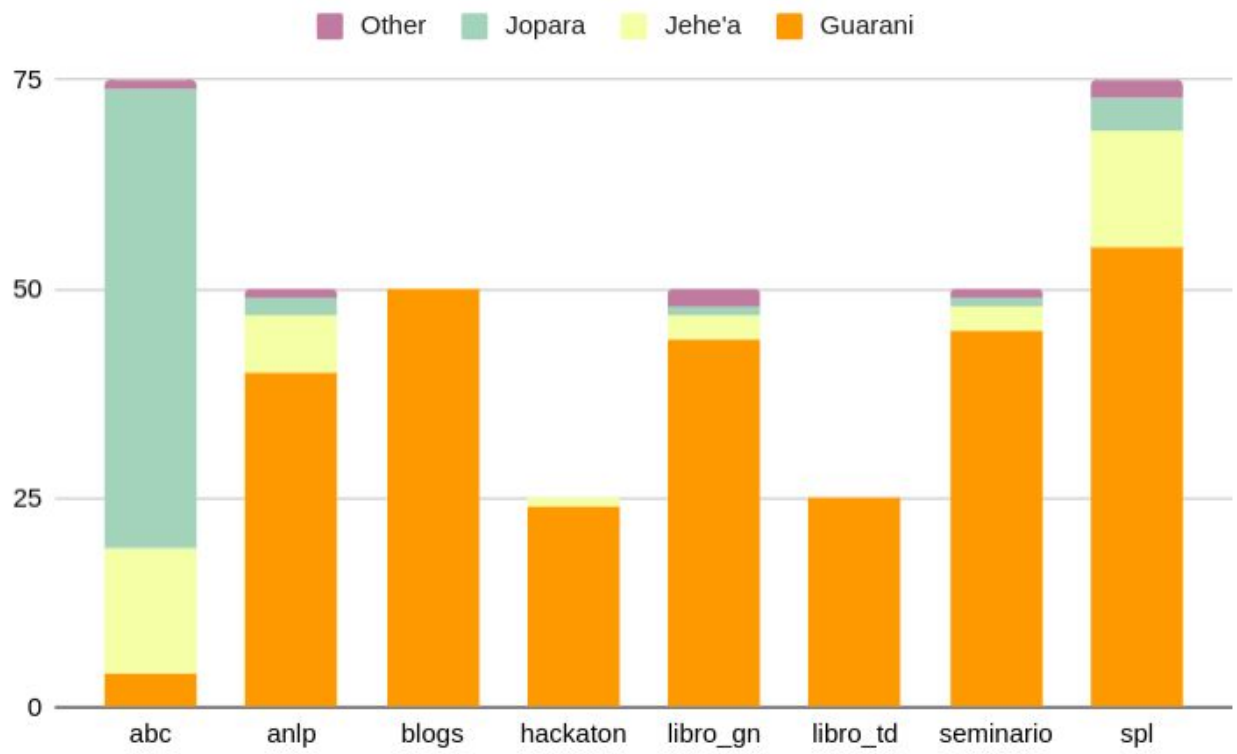
Corpus Jojajovai



Alrededor de 30 mil pares en total

https://github.com/pln-fing-udelar/jojajovai

Variedades



El code-switching entre guaraní y español es muy frecuente

Variedades

- *Embohasamína ko marandu umi rehayhuvévape...*

  Por favor, pasa este mensaje a las personas que estimas...

- *Afara orenunsiáta ko'êrõ*

  Afara renuncia mañana

- *Ojuhúma 52 allanamiento Argentina gotyo ha 21 detenido, 200.000 munición ha 2.500 fusil ojokóva.*

  En Argentina ya han realizado unos 52 allanamientos, 21 detenidos, 200.000 municiones con 2.500 fusiles secuestrados.

Alineación

Itaugua omokyre'ÿ "omopotî" Congreso

En Itauguá promueven "limpiar" el Congreso

Omopotîvo hikuái tetãme vicio política, ko'ã itaugüeño he'íva ombotovévo pokarême umi elemento omopotîva.

Con el propósito de limpiar al país de los vicios de la política, los itaugüeños expresan su repudio a los corruptos con elementos de limpieza.

Ko'ã 50 tapicha oñembyaty parroquía Virgen del Rosario plazoleta-pe ko distrito onemanifestávo político pokarême.

Unas 50 personas se encuentran en la plazoleta de la parroquía Virgen del Rosario de este distrito manifestando su repudio hacia los políticos corruptos.

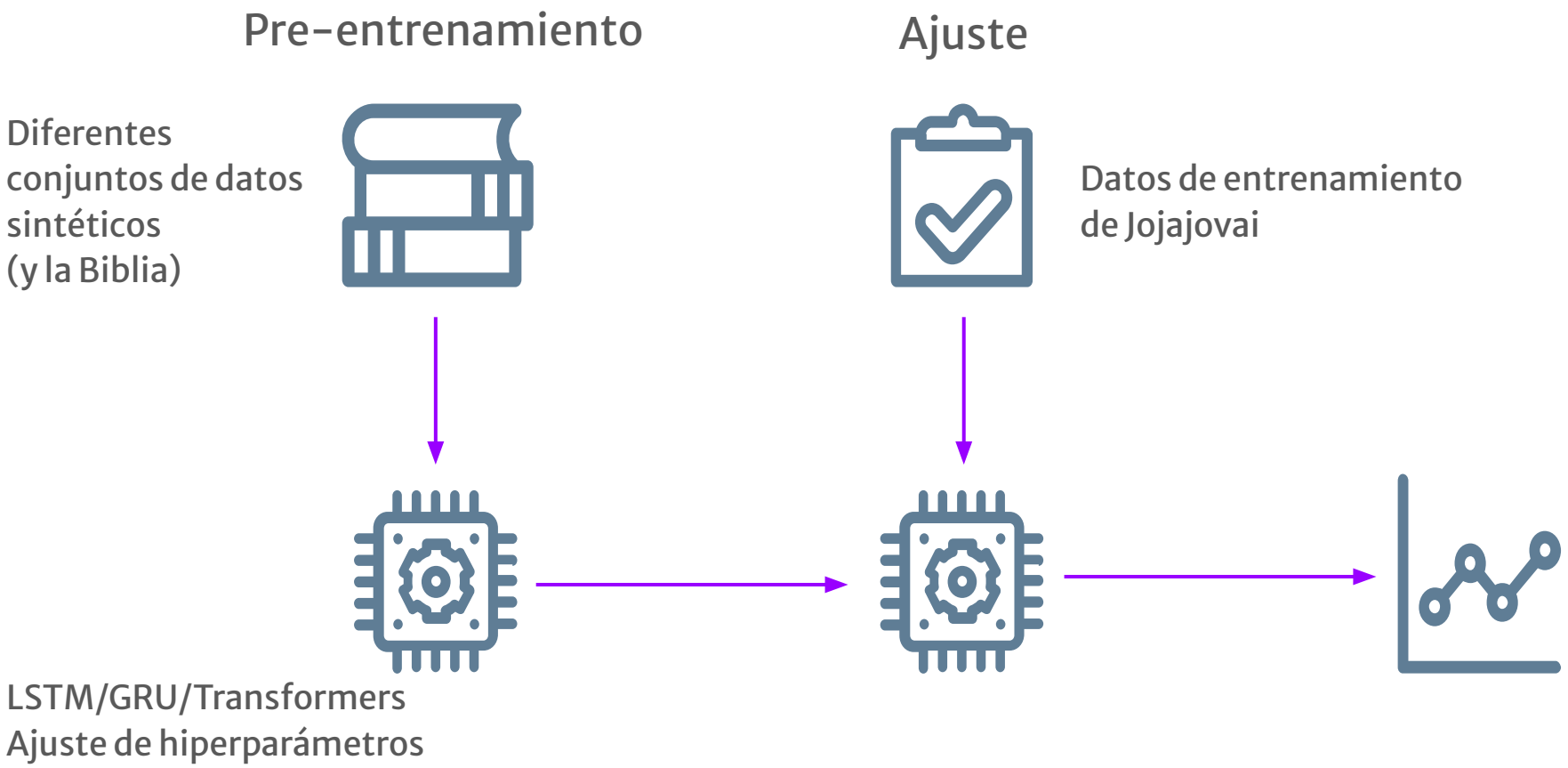
Durante el encuentro "limpiaron" un muñeco de un diputado acusado de corrupción.

Itaugüeño oipotáva ohechauka ipotîha itáva ha ikatúha paraguayo ikatu omopotî parlamento ha upévare hi'agui, orekóva yvyra orepasa haguã, trapo de piso, typycha ha lavandina oguahëva plaza parroquial rovái.

Los itaugüeños quieren demostrar que los paraguayos pueden limpiar el parlamento y por eso se acercaron, con palos de repasar, trapos de piso, escobas y lavandina hasta la plaza parroquial.

La manifestación fue acompañada por aplausos y vítores por parte de los vecinos.

Entrenamiento



Grammar-based Data Augmentation for Low-Resource Languages: The Case of Guarani-Spanish Neural Machine Translation
Lucas, Baladón, Pardiñas, Agüero-Torales, Góngora, Chiruzzo. 2024

Gramáticas de Rasgos

Spanish	Guarani
Part-of-speech: V (verb)	
Type: M (main), A (auxiliary), S (semi-auxiliary)	
Mood: I (indicative), S (subjunctive), M (imperative), P (participle), G (gerund), N (infinitive)	
Tense: P (present), I (past imperfective), F (future), S (past perfect), C (conditional)	
Person: 1 (first), 2 (second), 3 (third)	
Number: S (singular), P (plural)	
Gender: M (male), F (female), C (common)	Inclusiveness (only for first-person plural): I (inclusive), E (exclusive)
	Pronoun position (only for third-person plural): B (before the verb), A (after the verb), 0 (not relevant)
Transitivity: I (intransitive), T (transitive), D (ditransitive)	

Verbo

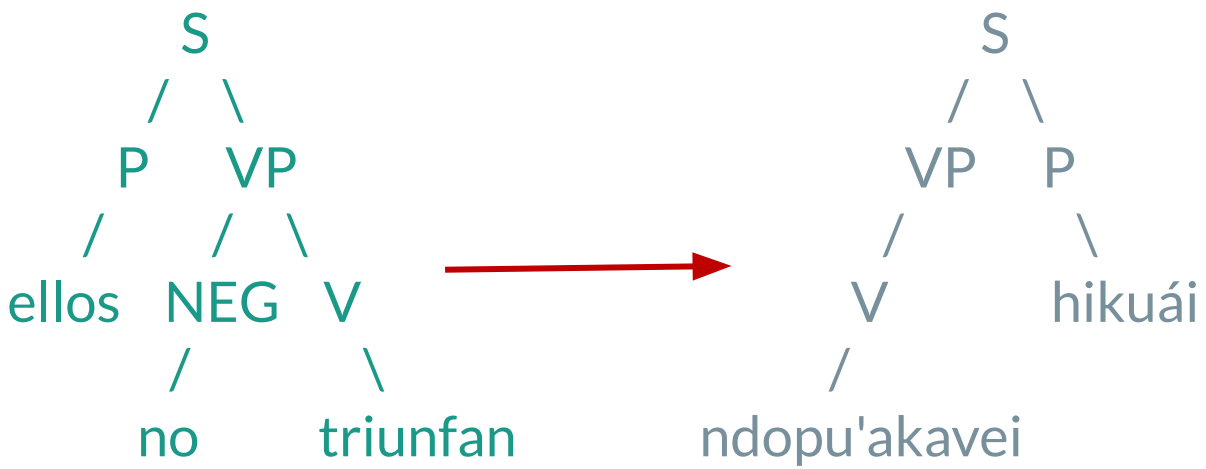
miro / amaña
mira / omaña
miré / amañakuri
miró / omañakuri
...

Spanish	Guarani
Part-of-speech: N (noun)	
Type: C (common), P (proper)	
Gender: M (male), F (female), C (common)	Gender: 0 (there is no gender for nouns)
Number: S (singular), P (plural), N (invariable)	
	Nasalization: N (nasal), O (oral)

Sustantivo

perro / jagua
perros / jaguakuéra
río / ysyry
amistad / ñoirũ
piedra / ita
piedras / itakuéra
...

Transferencia Sintáctica



Reglas de transferencia entre idiomas

```
"VP -> NEG V": [
        "VP[AGR='?a', POS='?p'] -> V[AGR='?a', NEG='1', POS='?p']"
    ]
```

Corpus Sintético

Generar oraciones aleatorias en español

Transferir a guaraní

Creamos tantos datos como queramos!

Set paralelo con más de 1M de palabras en guaraní

Pero…

Una piedra mintió / *Peteĩ ita oñe'ẽreikuri*

Él no apretará nuestras sopas / *Ha'e nomombeita ore jukysykuéra*

Corpus AnCora

Conjunto de noticias en español con 14,000 oraciones, 500,000 palabras

Detectar y transferir porciones traducibles con nuestra gramática

Josep y Angel Ortiz desbordan la ilusión en su mirada

Josep ha Angel Ortiz ochovi ilusión ima'eme

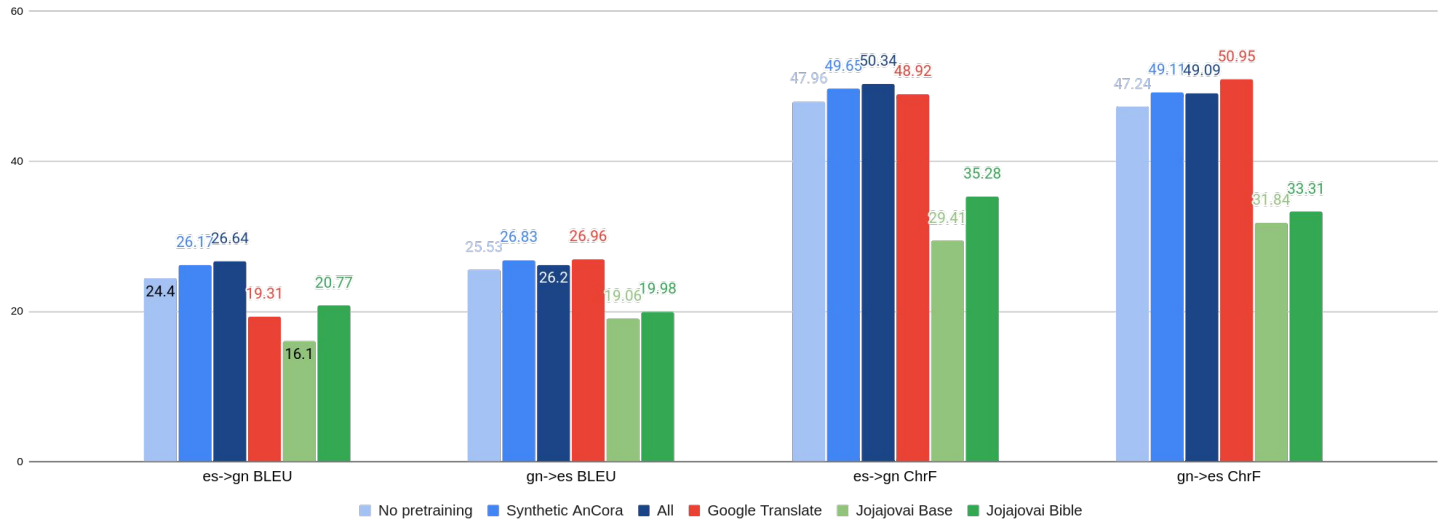
Oraciones más realistas

Genera un code-switching artificial!

Experimentos de Traducción Automática

Dir	Model	ChrF	BLEU
es→gn	jojajovai base	29.41	16.10
	jojajovai + bible	35.28	20.77
	Google	**48.92**	**19.31**
	no-pre	47.96	24.40
	pre-ancora	49.65	26.17
	pre-all	**50.34**	**26.64**

Dir	Model	ChrF	BLEU
gn→es	jojajovai base	31.84	19.06
	jojajovai + bible	33.31	19.98
	Google	**50.95**	**26.96**
	no-pre	47.24	25.53
	pre-ancora	**49.11**	**26.83**
	pre-all	49.09	26.20



Experimentos de Traducción Automática

Dir	Metric	Model	abc	anlp	blogs	hackathon	libro_gn	libro_td	seminario	spl
es→gn	ChrF	s2s - All	**58.76**	24.58	32.30	34.69	**30.16**	**39.38**	28.88	48.50
		s2s - AnCora	58.34	23.59	31.55	31.65	28.93	37.00	29.71	46.99
		Google Translate	56.61	**37.05**	**39.38**	**41.71**	28.82	28.15	**35.94**	**49.49**
		Jojajovai Base	37.44	14.10	21.35	20.02	16.98	24.10	19.83	37.49
		Jojajovai Bible	46.14	18.67	25.45	23.39	19.15	28.25	22.32	39.63
	BLEU	s2s - All	**31.45**	3.01	16.10	5.47	7.72	**10.49**	7.78	29.58
		s2s - AnCora	31.16	2.66	15.34	3.67	**10.86**	8.63	8.76	28.38
		Google Translate	23.56	**6.01**	**16.27**	**5.75**	8.30	3.09	**9.00**	**30.01**
		Jojajovai Base	18.24	0.75	7.73	3.09	3.44	5.15	3.02	20.73
		Jojajovai Bible	24.48	1.76	11.26	3.06	7.46	3.38	5.15	23.51
gn→es	ChrF	s2s - All	56.17	21.48	34.54	31.09	28.56	36.02	30.15	**48.61**
		s2s - AnCora	56.31	21.17	33.37	30.34	30.36	**37.69**	30.64	48.58
		Google Translate	**56.73**	**42.04**	**45.25**	**46.32**	**31.88**	29.62	**36.73**	44.49
		Jojajovai Base	40.25	14.77	24.71	19.35	17.15	24.02	23.15	41.68
		Jojajovai Bible	42.03	17.19	25.40	23.58	19.08	26.45	23.05	41.24
	BLEU	s2s - All	30.06	4.33	18.44	14.69	9.70	15.69	9.95	30.59
		s2s - AnCora	**30.83**	4.04	18.13	10.86	10.50	**18.41**	10.10	**31.21**
		Google Translate	30.81	**19.80**	**24.45**	**18.44**	**11.29**	9.02	**13.16**	23.58
		Jojajovai Base	20.84	1.55	11.89	6.45	5.40	10.25	6.37	25.93
		Jojajovai Bible	22.14	2.52	12.50	6.48	7.80	8.56	6.80	25.83