Pontificia Universidad Católica de Chile

Facultad de Letras

Curso de traducción automática de lenguas naturales

César Antonio Aguilar
Facultad de Lenguas y Letras
14/08/2015

Cesar.Aguilar72@gmail.com

Introducción (1)

La traducción automática es una de las áreas de investigación más relevantes del procesamiento del lenguaje natural (PLN). De hecho, históricamente puede ser vista como una de sus líneas fundadoras: los primeros trabajos en traducción hecha por máquinas datan por lo menos desde hace 50 años.

Hoy en día existen un gran número de recursos electrónicos que permiten realizar traducciones con un grado de calidad aceptable.

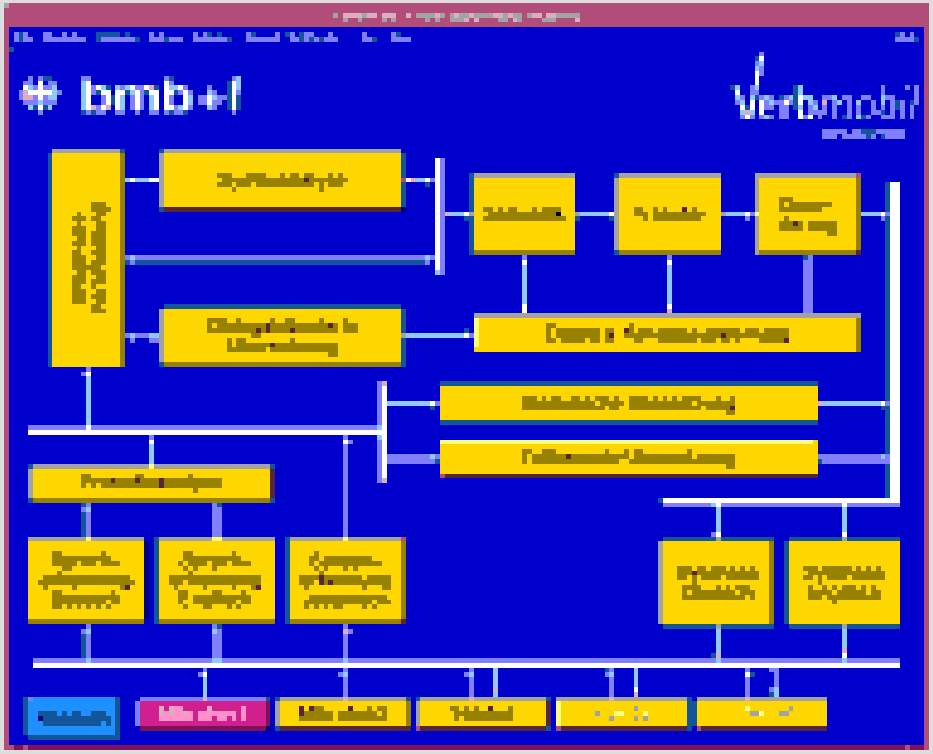
(Users view translations on screen and/or listen to voice readouts)

Introducción (2)

Un buen ejemplo de lo que es un sistema de traducción automático en tiempo real es el proyecto **Verbmobil**, desarrollado bajo el auspicio del Ministerio Federal Alemán para la Educación, la Ciencia, la Investigación y la Tecnología en 1996.

Liga al video: www.youtube.com/watch?v=noZBab-Lmss

Mayor información: http://verbmobil.dfki.de/



Introducción (3)

A pesar de los avances logrados por medio de la implementación de sistemas como **Verbmobil**, existen todavía muchos problemas para lograr una buena calidad en las traducciones. A grandes rasgos, algunas de estas dificultades a considerar son:

1. El gran incremento de información en la WEB, lo que obliga a generar algoritmos y métodos más eficaces y rápidos de traducción.
2. La necesidad de inferir dentro de un texto diferentes niveles de información, ya sea general o especializada.
3. La falta de recursos computacionales para lenguas minoritarias, tales como lenguas indígenas.

Un poco de historia (1)

Una de las ideas más interesantes que subyace en la traducción automática (y en general, en toda propuesta relacionada con la traducción) es el supuesto de que todas las lenguas humanas se derivan de un idioma universal.

Bajo este supuesto, se considera que todas las lenguas son de alguna forma una *derivación* de un lenguaje subyacente, el cual puede ser inferido y representado a través de un sistema simbólico.

Un poco de historia (2)

A lo largo de la historia, este sistema simbólico se ha representado de distintas maneras: usando notaciones matemáticas, formalismos lógicos, e incluso inventando lenguas como el **esperanto**, creado por Ludwik L.Zamenhof (1859-1917), el cual es el ejemplo más conocido de lo que es una lengua artificial.

English	Esperanto	IPA
Hello	Saluton	[sa.ˈlu.ton]
Yes	Jes	[ˈjes]
No	Ne	[ˈne]
Good morning	Bonan matenon	[ˈbo.nan ma.ˈte.non]
Good evening	Bonan vesperon	[ˈbo.nan ves.ˈpe.ron]
Good night	Bonan nokton	[ˈbo.nan ˈnok.ton]
Goodbye	Ĝis (la) revido	[ˈdʒis (la) re.ˈvi.do]
What is your name?	Kio estas via nomo?	[ˈki.o ˌes.tas ˌvi.a ˈno.mo]
My name is Marc.	Mia nomo estas Marko.	[ˌmi.a ˈno.mo ˌes.tas ˈmar.ko]
How are you?	Kiel vi fartas?	[ˈki.el vi ˈfar.tas]
I am well.	Mi fartas bone.	[mi ˈfar.tas ˈbo.ne]
Do you speak Esperanto?	Ĉu vi parolas Esperante?	[ˈtʃu vi pa.ˈro.las ˌes.pe.ˈran.te]

Un poco de historia (3)

Tras estos supuestos y primeros intentos por crear una lengua universal, será hasta las décadas de1920 y 1930 que se desarrollarán proyectos de traducción automática de secuencias lingüísticas a códigos matemáticos, dando origen a un proceso computacional que se conoce como **encriptación**.

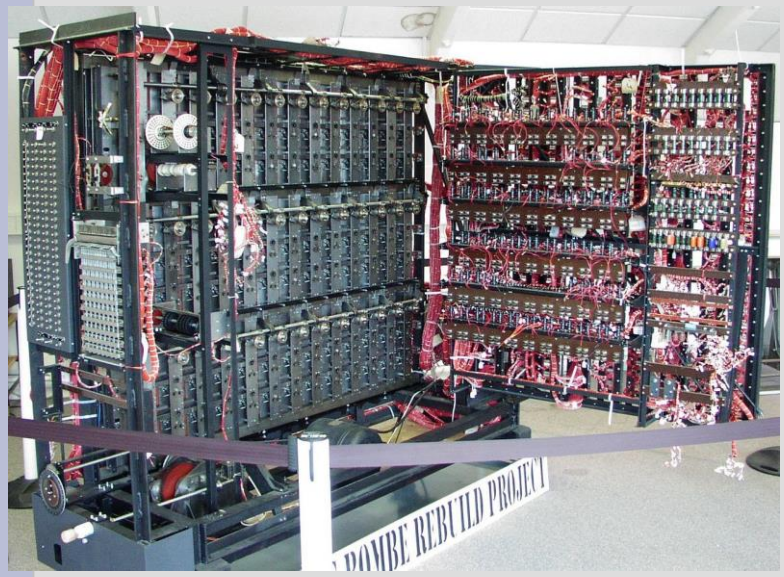
El ejemplo más conocido sobre procesos de encriptación se da en plena Segunda Guerra Mundial, cuando las fuerzas nazis comienzan a producir máquinas electrónicas identificadas con el nombre de **Enigma**.

El codificado que lograban estas máquinas llegaba a un nivel de complejidad relevante, y pasaron varios años antes de que los ejércitos aliados pudieran comprender su funcionamiento.

Un poco de historia (4)

Alan Turing
(1912-1954)

Después de varios esfuerzos llevados en colaboración entre matemáticos polacos y británicos, será hasta 1940 que se logrará avanzar el tarea de *romper el código Enigma*, esto es, en concretar un sistema de interpretación y decodificación de los mensajes cifrados por estas máquinas.



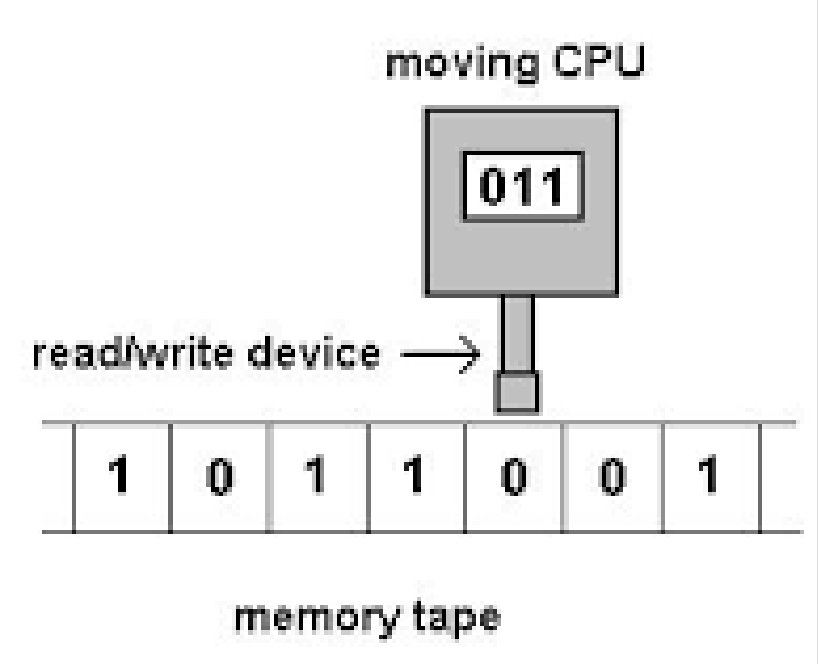
El mayor éxito logrado en esta tarea será el fruto de un equipo británico coordinado por **Alan Turing**. Para quebrar este código, Turing tuvo que concebir, en un plano netamente teórico, una máquina capaz de computar y resolver cálculos matemáticos de manera autónoma, esto es, la **máquina de Turing**.

Un poco de historia (5)

¿Qué es exactamente una *máquina de Turing*? En términos generales, se puede decir que es un dispositivo que manipula símbolos sobre una tira de cinta de acuerdo a una tabla de reglas.

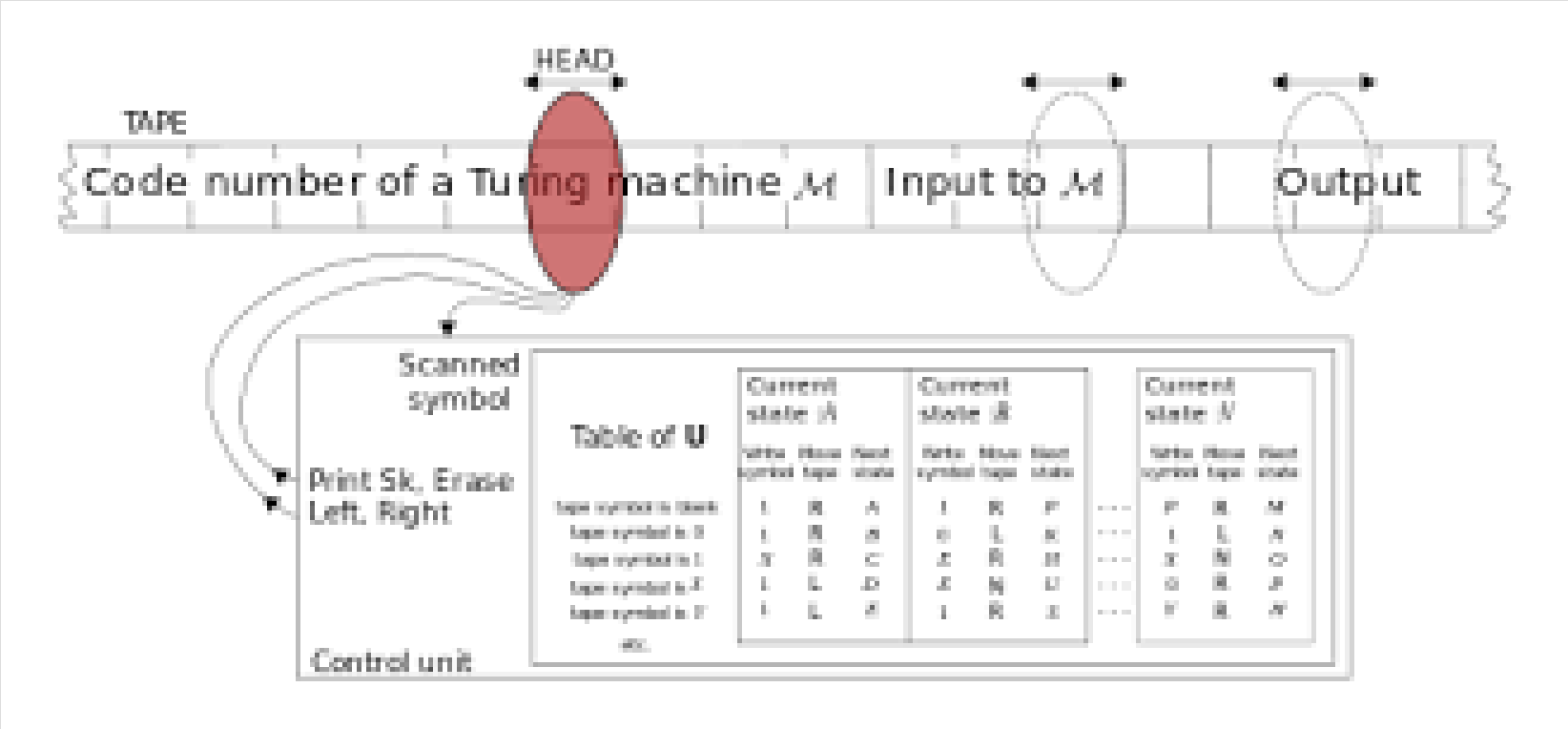
De acuerdo con la figura, lo que tenemos es una larga cinta (o secuencia) con signos específicos (en este caso, números), la cual es recorrida por una **cabeza lectora** que se desplaza conforme a una regla previamente determinada.

Cada uno de los números que aparecen grabados en la cinta representa cualquier cosa, p. e., una serie de procesos que deben ser completados para realizar una acción, digamos algo tan trivial como: *ponga a hervir la pasta antes de servirla con la salsa bolognesa*.



La propuesta de Weaver (1)

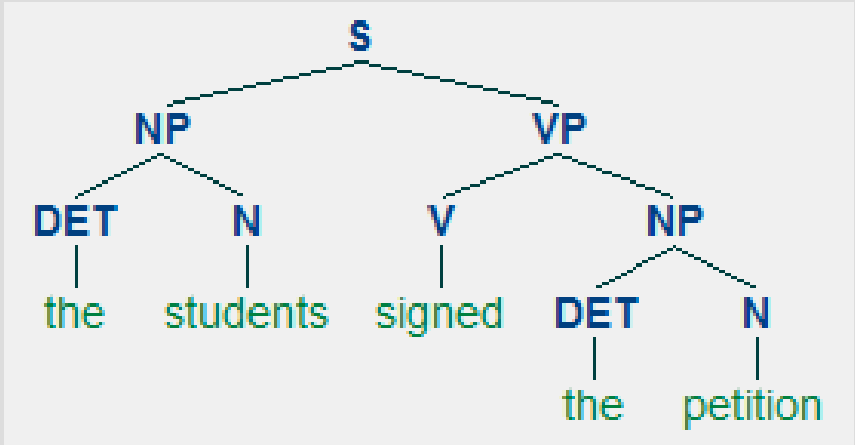
Ahora, pensemos en algo más lingüístico: imaginen que tenemos dos cintas y un cabeza lectora doble: en una cinta hay una secuencia de palabras en inglés, y en el otro hay una secuencia de palabras en ruso. ¿Qué tipo de solución nos daría la máquina de Turing a una tarea de traducción? ¿Pueden explicarlo?



La propuesta de Weaver (2)

Ahora, pensemos en algo más lingüístico: imaginen que tenemos dos cintas y un cabeza lectora doble: en una cinta hay una secuencia de palabras en inglés, y en el otro hay una secuencia de palabras en ruso. ¿Qué tipo de solución nos daría la máquina de Turing a una tarea de traducción? ¿Pueden explicarlo?

Más o menos ésta es la idea que subyacía en el experimento que realizaron en colaboración la Universidad de Georgetown con IBM, el cual fue coordinado por **Warren Weaver**, el cual tenía por objetivo evaluar un programa de traducción inglés-ruso. Este experimento es, para muchos, el punto de inicio de la traducción automática.



La propuesta de Weaver (3)

Weaver veía viable la traducción automática tomando como base los argumentos que desarrolló junto con **Claude Shannon** (1916-2001) para su **teoría matemática de la comunicación**.

Weaver contaba que cuando veía un artículo escrito en ruso (por poner un caso): *me digo, esto en realidad está en inglés, aunque codificado con extraños símbolos. ¡Vamos a decodificarlo ahora mismo!*

Warren Weaver
(1894-1978)

La propuesta de Weaver (4)

El experimento consistió en lo siguiente: se desarrolló un programa que contaba con una gramática compuesta por 6 reglas, así como un vocabulario de 250 palabras, tanto en ruso como en inglés.
El programa fue diseñado considerando un modelo transformacional, de tal suerte que cada palabra del vocabulario, dependiendo de la regla de construcción, tenía un equivalente en la otra lengua. Las oraciones traducidas eran las siguientes:

Russian (Romanized)
Mi pyeryedayem mislyi posryedstvom ryechyi.
Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyiusu.
Myezhdunarodnoye ponyimanyiye yavlyayetsya vazhnim faktorom v ryeshyenyiyi polyityichyeskix voprosov.

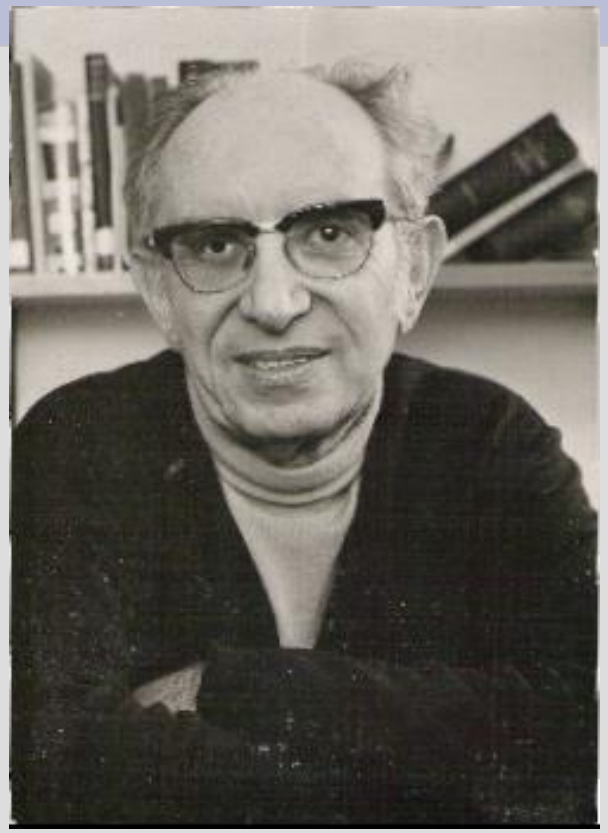
English translation
We transmit thoughts by means of speech.
Magnitude of angle is determined by the relation of length of arc to radius.
International understanding constitutes an important factor in decision of political questions.

La propuesta de Bar-Hillel (1)

La propuesta de traducción automática de Weaver y sus colegas fue mejorada gracias a los estudios desarrollados por **Yehoshua Bar-Hillel**, un filósofo del lenguaje afiliado al MIT.

Las ideas de Bar-Hillel sobre la traducción automática estaban ubicadas dentro del paradigma de la **gramática transformacional generativa**, planteada por Zellig Harris y Noam Chomsky.

Con base en este paradigma, Bar-Hillel señaló varios problemas respecto a los procesos de traducción automática, p. e., el de la ambigüedad estructural que puede existir para dar con un equivalente preciso de una lengua a otra.



Yehoshua Bar-Hillel
(1915-1975)

La propuesta de Bar-Hillel (2)

Un ejemplo respecto a este problema de ambigüedad es el siguiente. Al leer este texto:

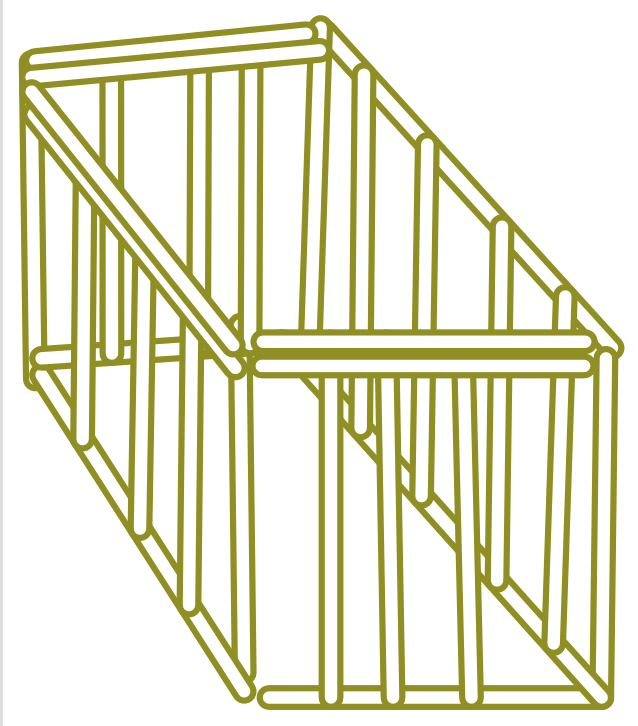
Little John was looking for his toy box. Finally he found it. *The box was in the pen*. John was very happy.

Pregunta: ¿cuál es el significado de la oración *the box was in the pen*, dentro del contexto de este texto?

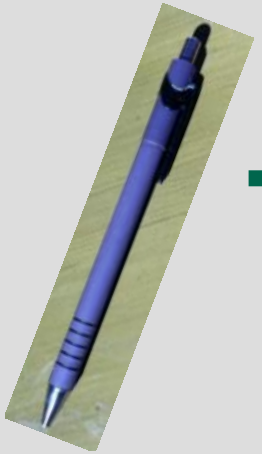
La propuesta de Bar-Hillel (3)

En teoría, tenemos dos opciones:

Opción A: la caja está en el corral =

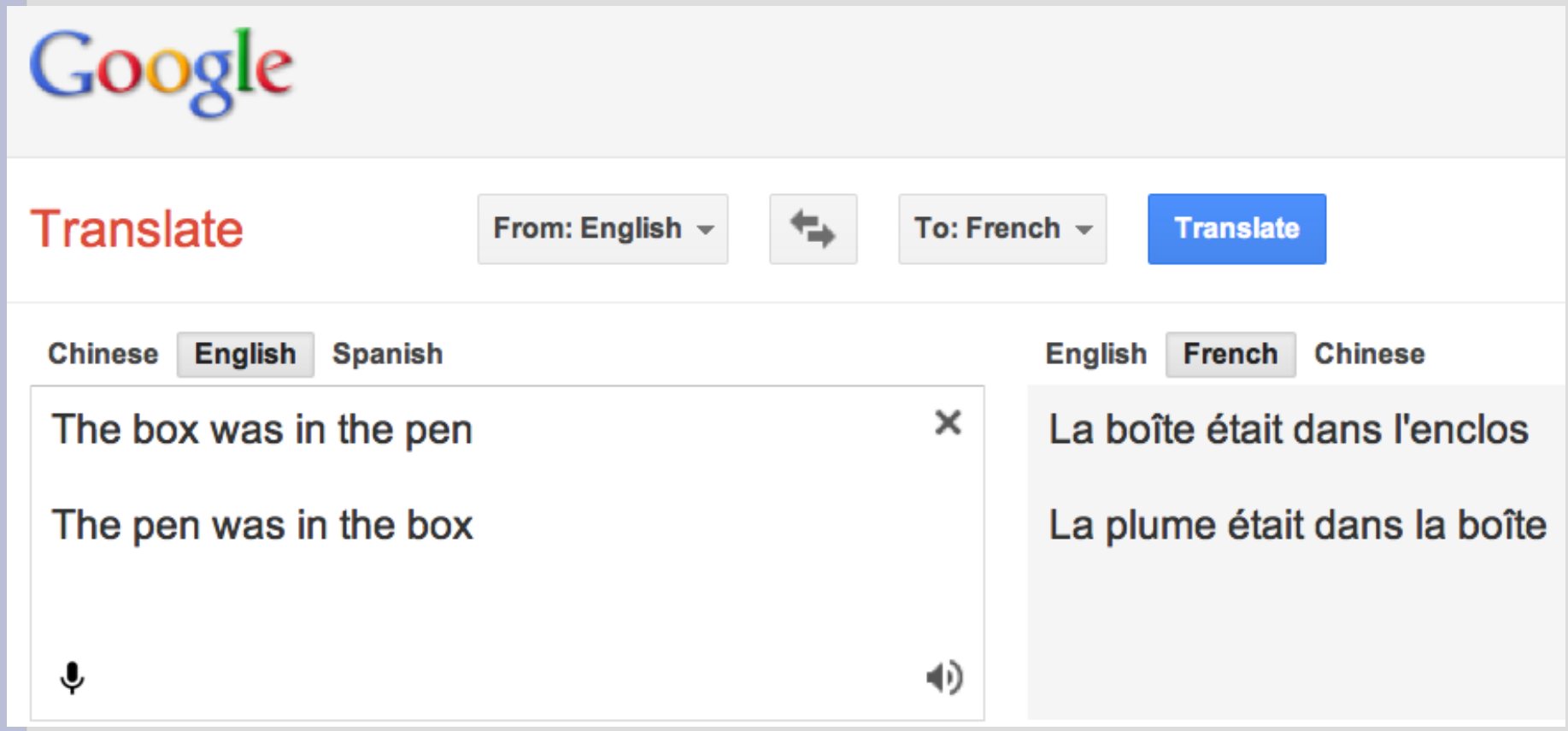


Opción B: hay un lápiz dentro de una caja, y ambos están en el corral =



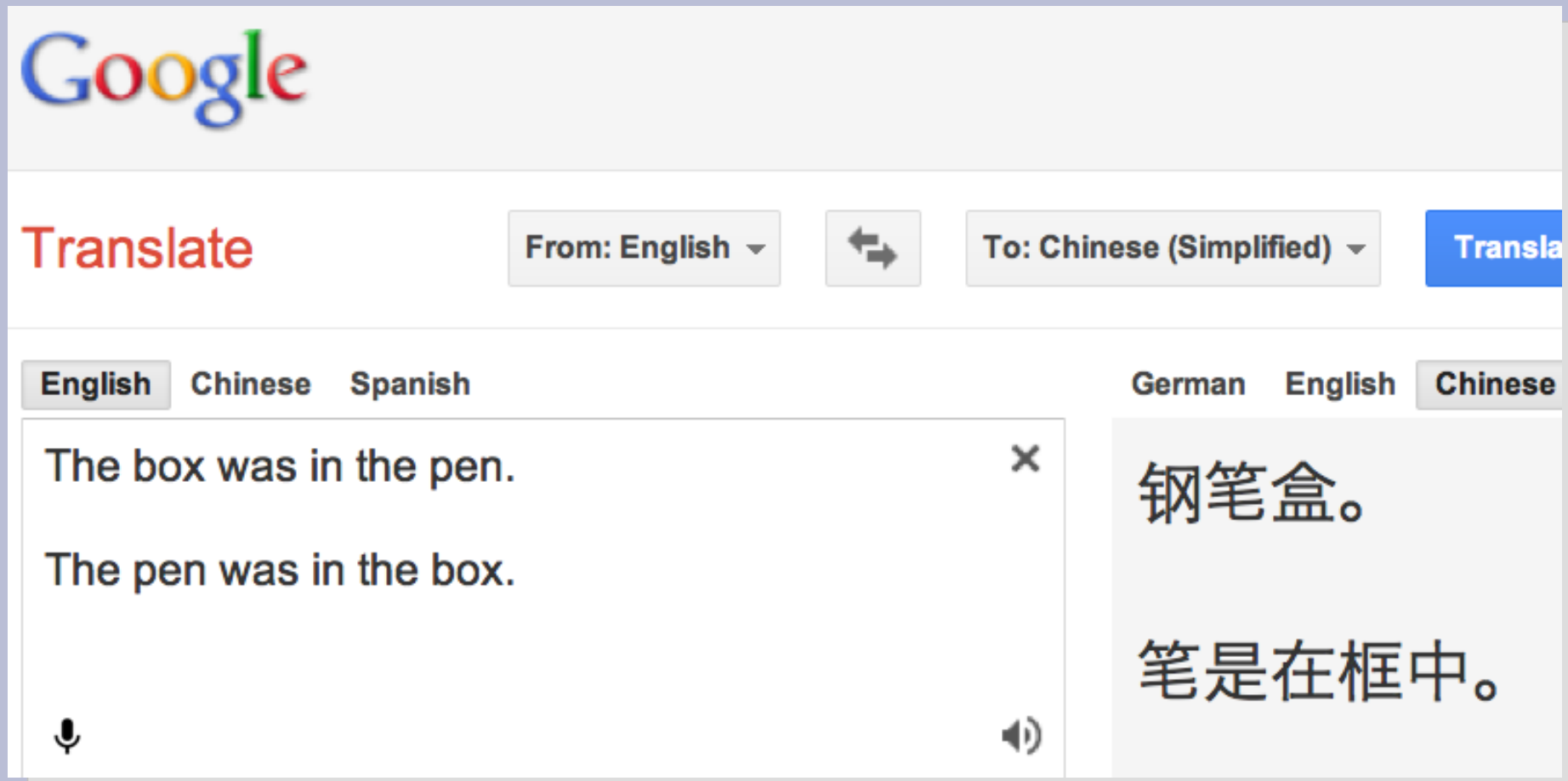
La propuesta de Bar-Hillel (4)

Veamos qué pasa si traducimos esa oración a otras lenguas, y hagamos algunas variantes. Probemos con el francés:



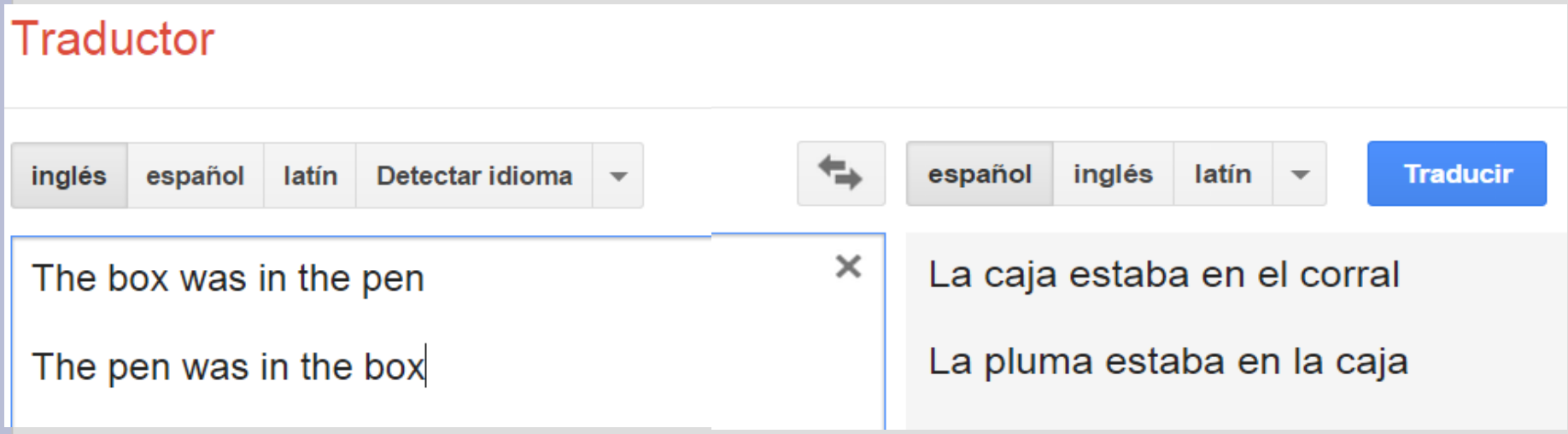
La propuesta de Bar-Hillel (5)

Y ahora, al chino:



La propuesta de Bar-Hillel (5)

Finalmente, el español:



¿Qué diferencias observan? ¿Por qué se genera esta ambigüedad al traducir esta oración?

Métodos de traducción (1)

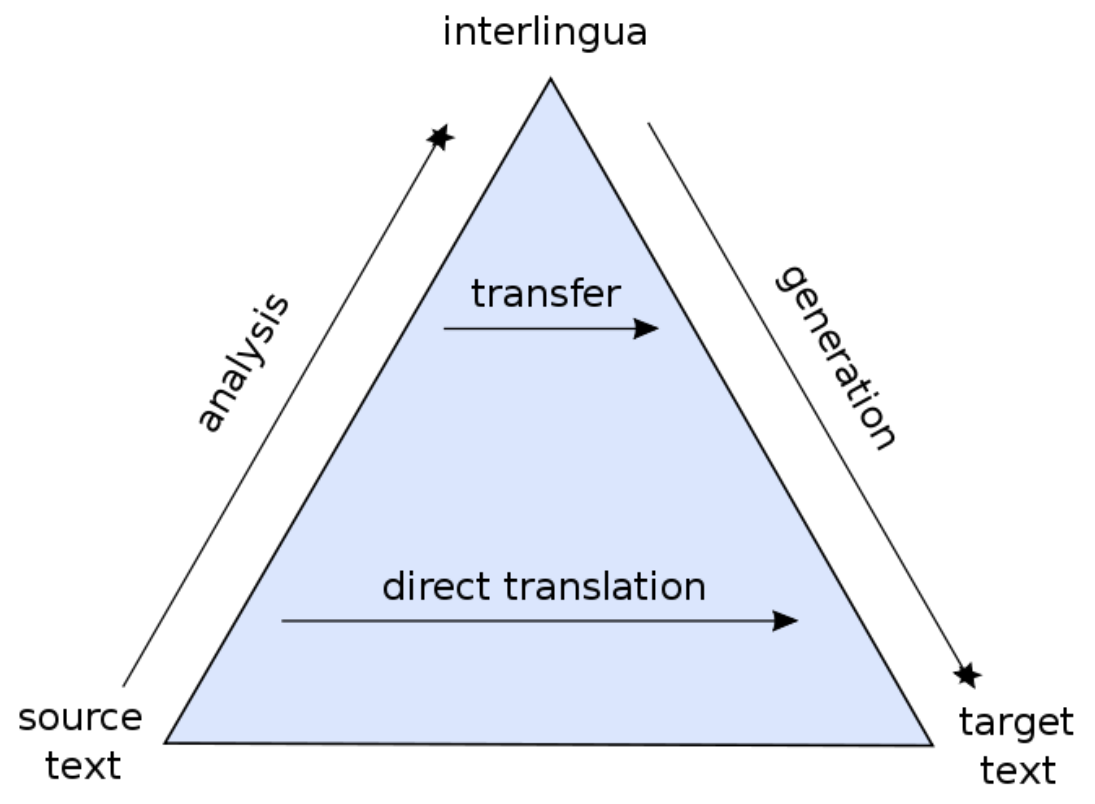
A lo largo de las experiencias acumuladas desde los años 40, la traducción automática ha delineado 4 metodologías específicas:

Traducción directa

Traducción por transferencia de significado

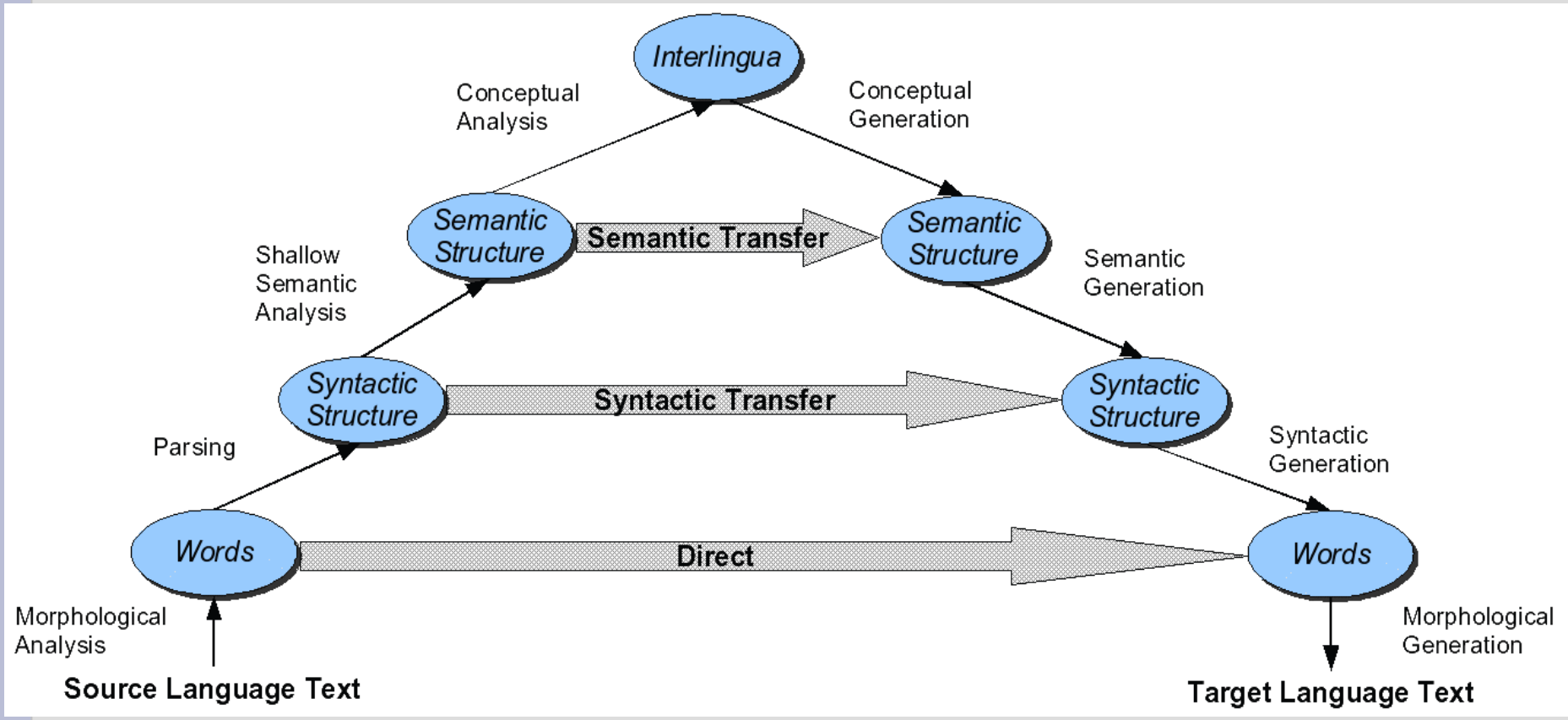
Traducción soportada por una interlengua

Traducción estadística

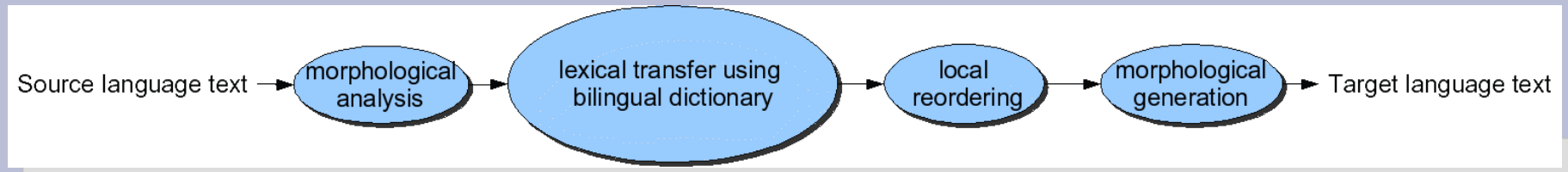


Métodos de traducción (2)

En este esquema pueden verse mejor las similitudes y diferencias entre estos métodos:



Traducción directa (1)



Proceed word-by-word through text

Translating each word

No intermediate structures except morphology

Knowledge is in the form of

- Huge bilingual dictionary
- word-to-word translation information

After word translation, can do simple reordering

- Adjective ordering English -> French/Spanish

Traducción directa (2)

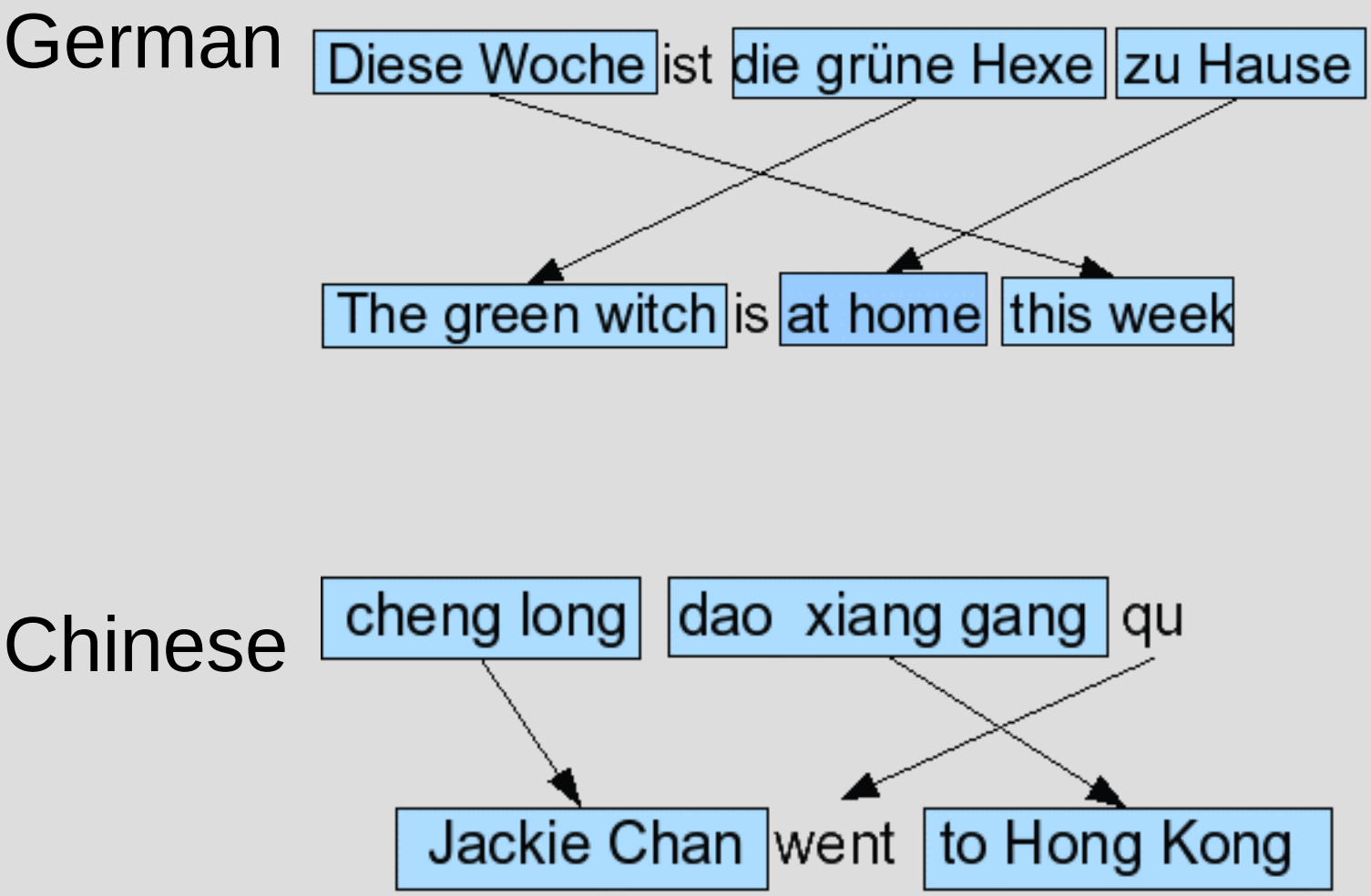
Si tuviéramos que transcribir estos criterios en un algoritmo, tendríamos algo como:

```
function DIRECT_TRANSLATE_MUCH/MANY(word) returns Russian translation

if preceding word is  how return  skol'ko
else if preceding word is  as return  stol'ko zhe
else if word is  much
    if preceding word is  very return nil
    else if following word is a noun return  mnogo
else  /* word is many */
    if preceding word is a preposition and following word is a noun return  mnogii
    else return  mnogo
```

Traducción directa (3)

Sin embargo, como muchas traducciones que se hacen palabra por palabra, este método tiene problemas importantes, p. e.:



Traducción por transferencia (1)

Idea: apply contrastive knowledge, i.e., knowledge about the difference between two languages

Steps:

Analysis: Syntactically parse source language

Transfer: Rules to turn this parse into parse for target language

Generation: Generate target sentence from parse tree

Traducción por transferencia (2)

¿Qué implica hacer una traducción considerando este método?
Veamos un ejemplo:

English: Adjective Noun

French: Noun Adjective

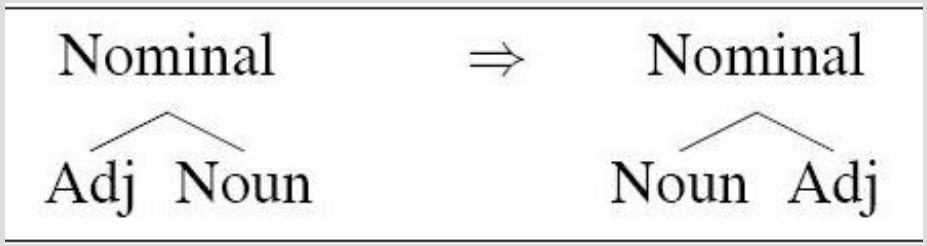
This is not always true

Route mauvaise ‘bad road, badly-paved road’

Mauvaise route ‘wrong road’

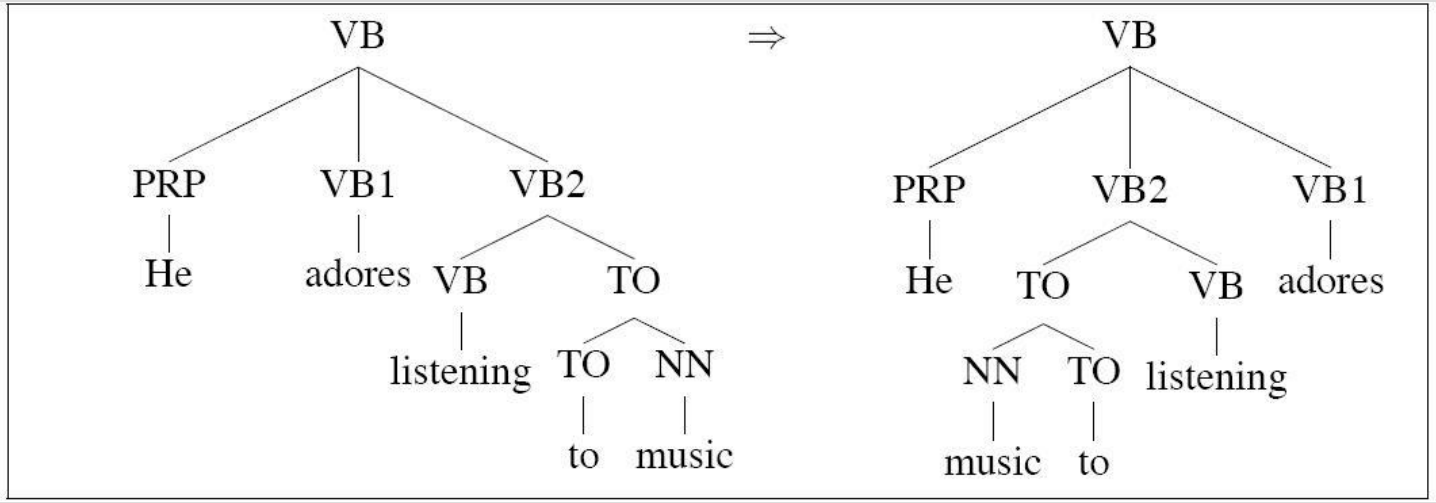
But is a reasonable first approximation

Rule:



Traducción por transferencia (3)

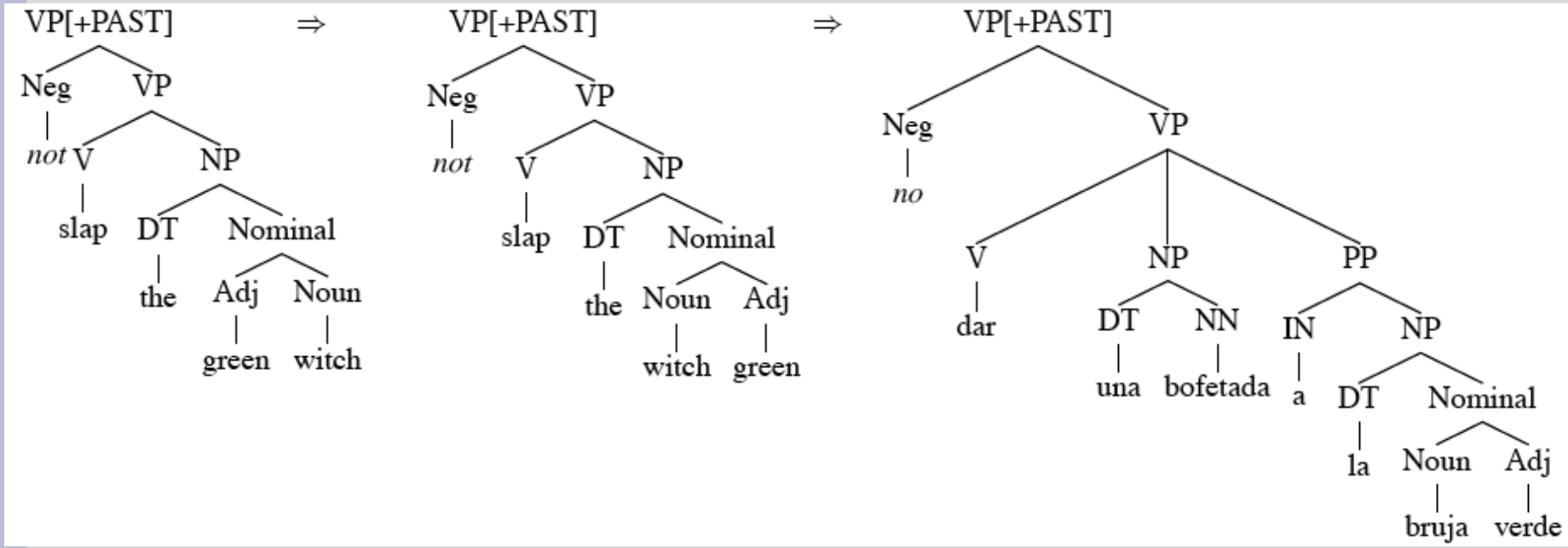
Para aplicar este método, requerimos siempre un conjunto de **reglas de transferencia**, p. e.:



	English to Spanish:		
1.	NP → Adjective$_1$ Noun$_2$	⇒	NP → Noun$_2$ Adjective$_1$
	Chinese to English:		
2.	VP → PP[+Goal] V	⇒	VP → V PP[+Goal]
	English to Japanese:		
3.	VP → V NP	⇒	VP → NP V
4.	PP → P NP	⇒	PP → NP P
5.	NP → NP$_1$ Rel. Clause$_2$	⇒	NP → Rel. Clause$_2$ NP$_1$

Traducción por transferencia (4)

La ventaja de estas reglas es que suponen la existencia de patrones universales de constitución, en este caso de estructuras sintácticas:



Traducción usando una interlengua (1)

La traducción que considera el uso de una interlengua para hacer el traspase de un idioma a otro parte de algunos supuestos como:

Instead of N^2 sets of transfer rules

Use meaning as a representation language

1. Parse source sentence into meaning representation
2. Generate target sentence from meaning.

Intuition: Use other NLP applications to do MT work

English `book` to Spanish: `libro` or `reservar`

Disambiguate book into concepts BOOKVOLUME and RESERVE

Need 2N systems (a parser and generator for each language)

Traducción usando una interlengua (2)

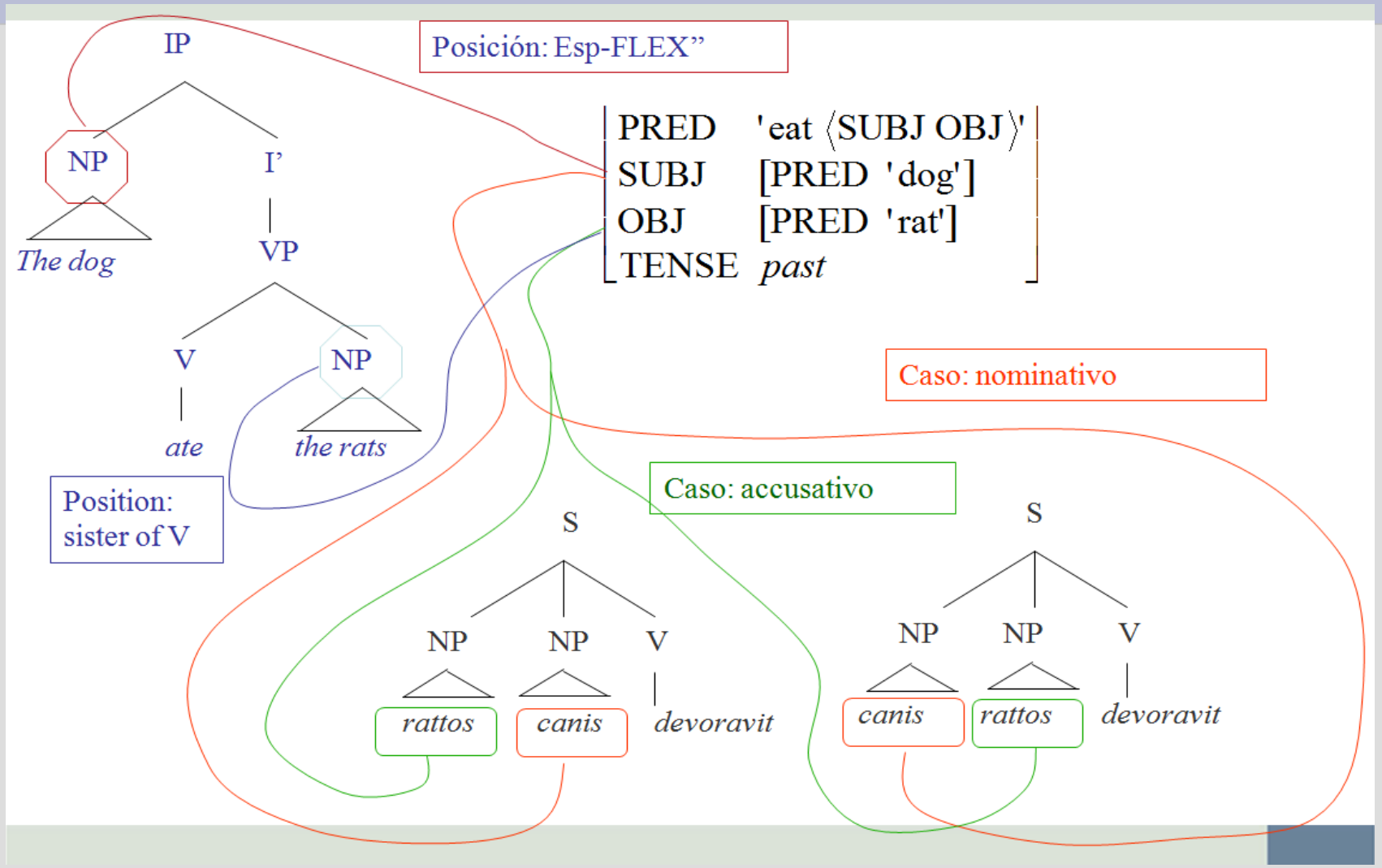
Una manera de crear una interlingua entre un idioma y otro, es ponderar el peso que tiene la información semántica de una oración. Tal información puede ser codificada en una matriz, lo que es un recurso muy usado en gramáticas léxicas como LFG o HPSG, p. e.:

Mary did not slap the green witch

$$\begin{bmatrix} \text{EVENT} & \text{SLAPPING} \\ \text{AGENT} & \text{MARY} \\ \text{TENSE} & \text{PAST} \\ \text{POLARITY} & \text{NEGATIVE} \\ \text{THEME} & \begin{bmatrix} \text{WITCH} & \\ \text{DEFINITENESS} & \text{DEF} \\ \text{ATTRIBUTES} & \begin{bmatrix} \text{HAS-COLOR} & \text{GREEN} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

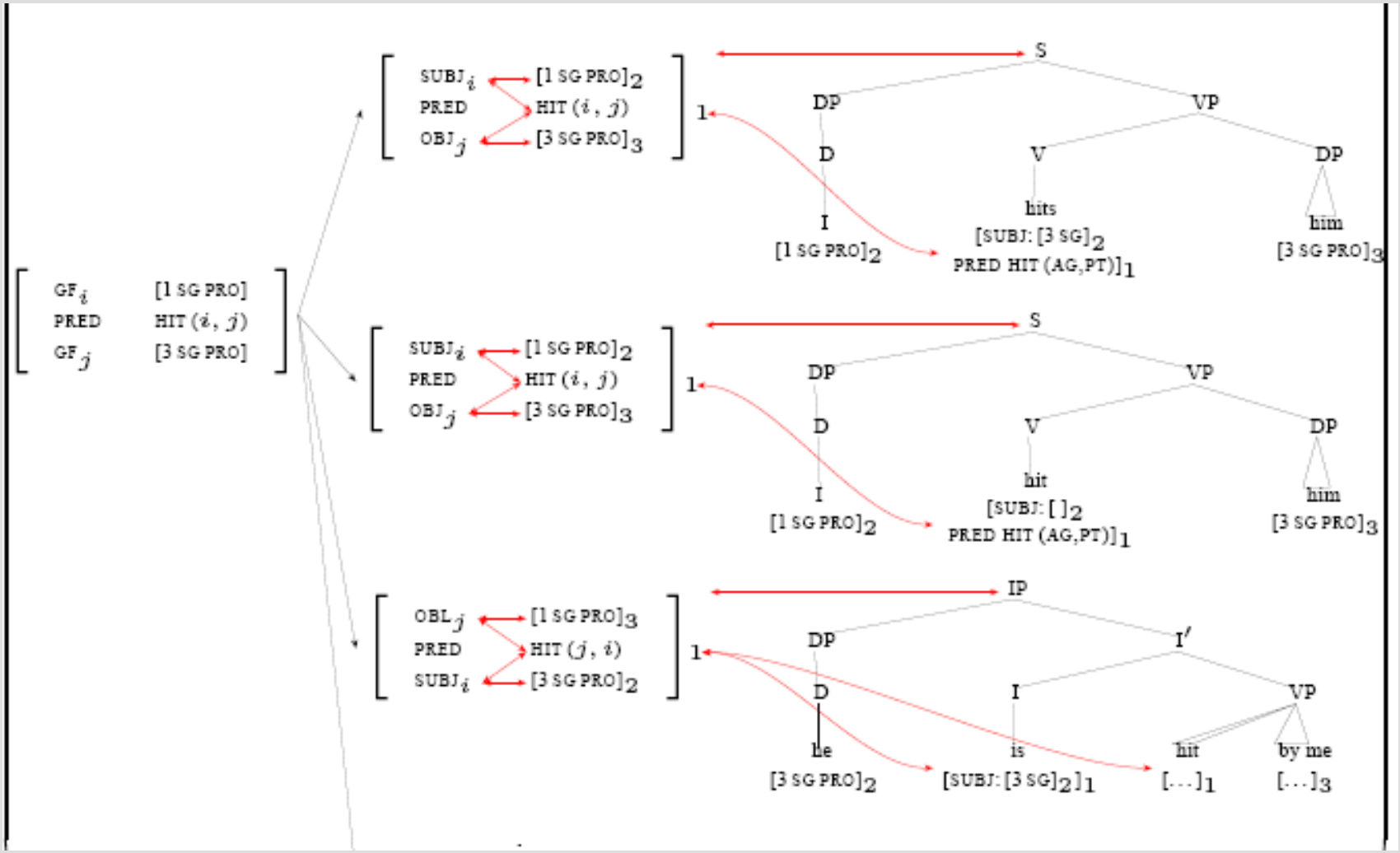
Traducción usando una interlengua (3)

Veamos un ejemplo de posible traducción de inglés al latín empleando LFG:



Traducción usando una interlengua (4)

Digamos entonces que lo que nos permite hacer una traducción es la información semántica codificada en una matriz, la cual nos permite mapear la estructura sintáctica pertinente para expresar tal información:



Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/curso-de-procesamiento-del-lenguaje-natural.html