Tema 2: Traducción automática: Nociones básicas

Margarita Alonso Ramos

Master LUP

2012

Organización de la exposición

- Métodos de TA
  - basada en reglas
  - basada en corpus / basada en analogías
- Método directo
- Métodos indirectos
  - Interlingua
  - Transferencia
- Métodos estadísticos
- Comparación
- Traductores on line

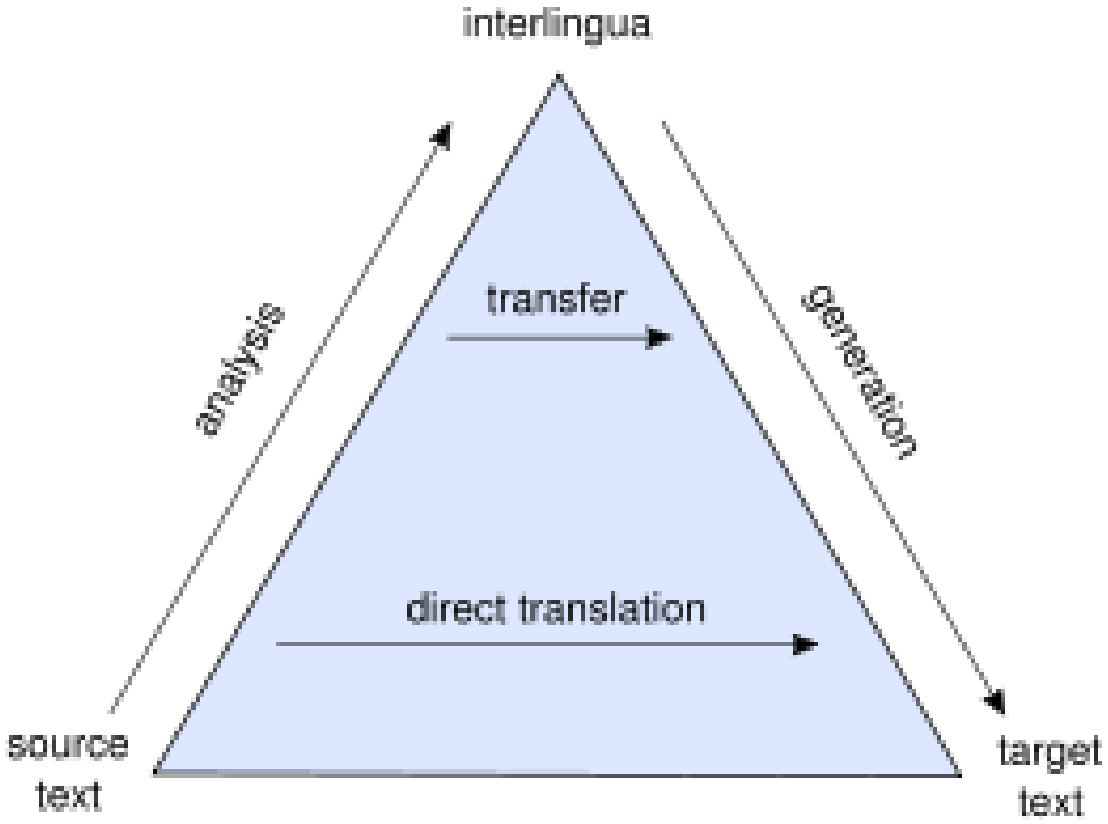
1. Características básicas de un sistema de TA

- unidireccional vs. bidireccional
- bilingüe vs. multilingüe
- interactivo vs. completamente automático
- métodos o estrategias
  - basadas en reglas
    - método directo
    - método de interlingua
    - método de transferencia
  - basadas en analogías o corpus
- teoría lingüística en la que se basa el sistema de TA

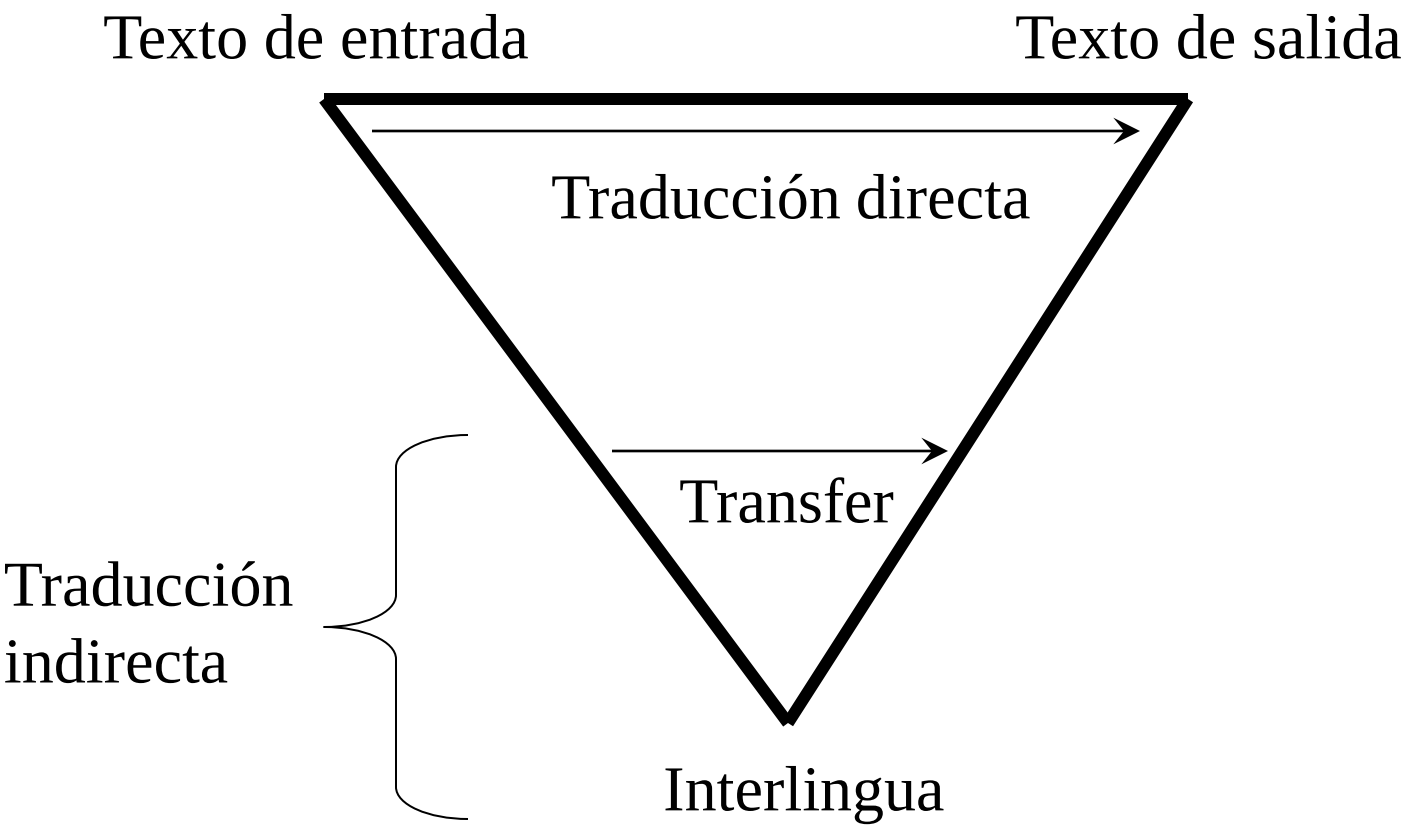
Sistemas de traducción interactivos

- **TA asistida por el humano**
  - **preedición** (identificación y resolución de problemas potenciales)
    - marcado de nombres propios
    - etiquetado POS en casos de ambigüedad (*Flying planes can be dangerous)*
    - introducción de palabras desconocidas en los diccionarios
  - **postedición**
    - corrección del resultado del proceso automático
- **TH asistida por el ordenador**
  - entornos para **traductores profesionales en una compañía**
    - bases de datos terminológicas multilingües
    - memorias de traducción
    - un servidor que gestiona las bases de datos comunes, distribución de tareas, gestión de la versión traducida, etc.
  - entornos para traductores profesionales **independientes** y traductores ocasionales
    - equipamiento como el precedente pero sin las herramientas de gestión
    - diccionarios varios: 1) con términos temporales y personales; 2) terminológicos, 3) vocabulario general
    - tesauros, correctores ortográficos, conjugadores, lematizadores, etc.

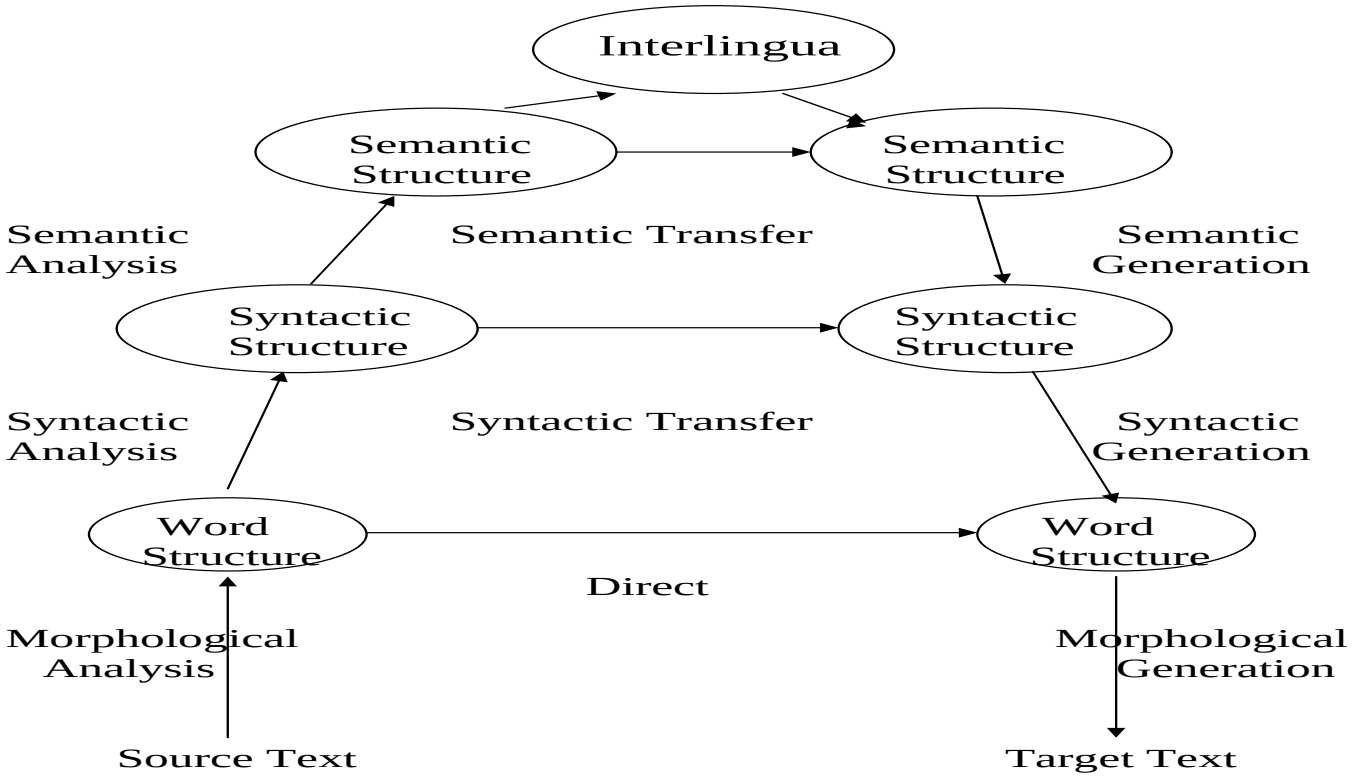
Pirámide de Vauquois: transferencia e interlingua



Pirámide invertida



La pirámide en detalle (Dorr et al., 1999)

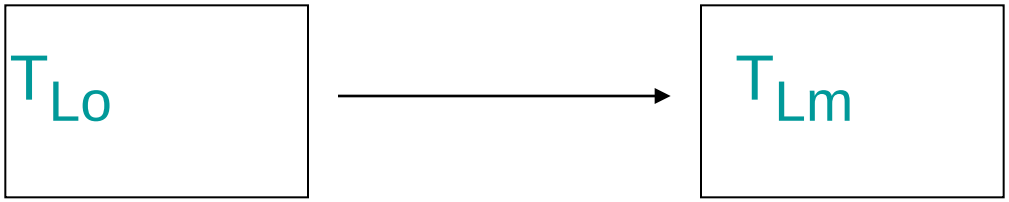


Criterios para escoger

- Diferencias tipológicas entre la L_O y la L_M
- Disponibilidad de recursos (léxicos y gramáticas de análisis y de generación)
- ¿Cuántas lenguas están implicadas?
- Disponibilidad de corpus paralelos
- Calidad de traducción esperada
- Restricciones de tiempo, etc.

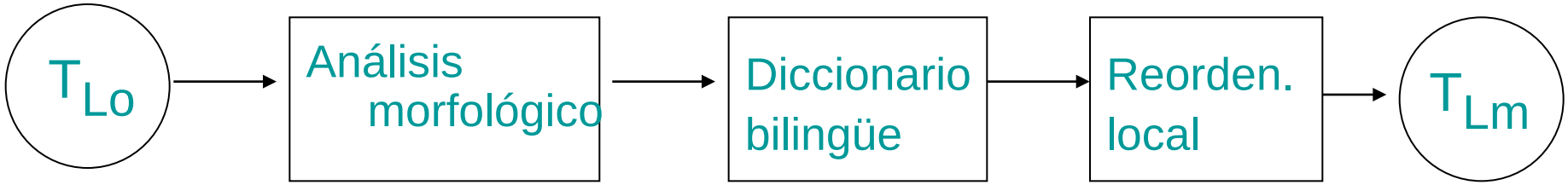
2. Método directo

- análisis mínimo del T_{LO}
- traducción palabra a palabra
- carece de cualquier fase intermedia
- procesamiento del T_{LO} conduce directamente al T_{LM}
- frecuentes errores léxicos y estructuras propias de la L_M



2. Método directo

- los programas explotan las similitudes tanto de la estructura como del léxico de las dos lenguas
- La principal tarea consiste en la consulta del diccionario bilingüe
- TAUM-Météo



Problemas con laTraducción directa

What´s the time? ⟶ **Qué es el tiempo?*

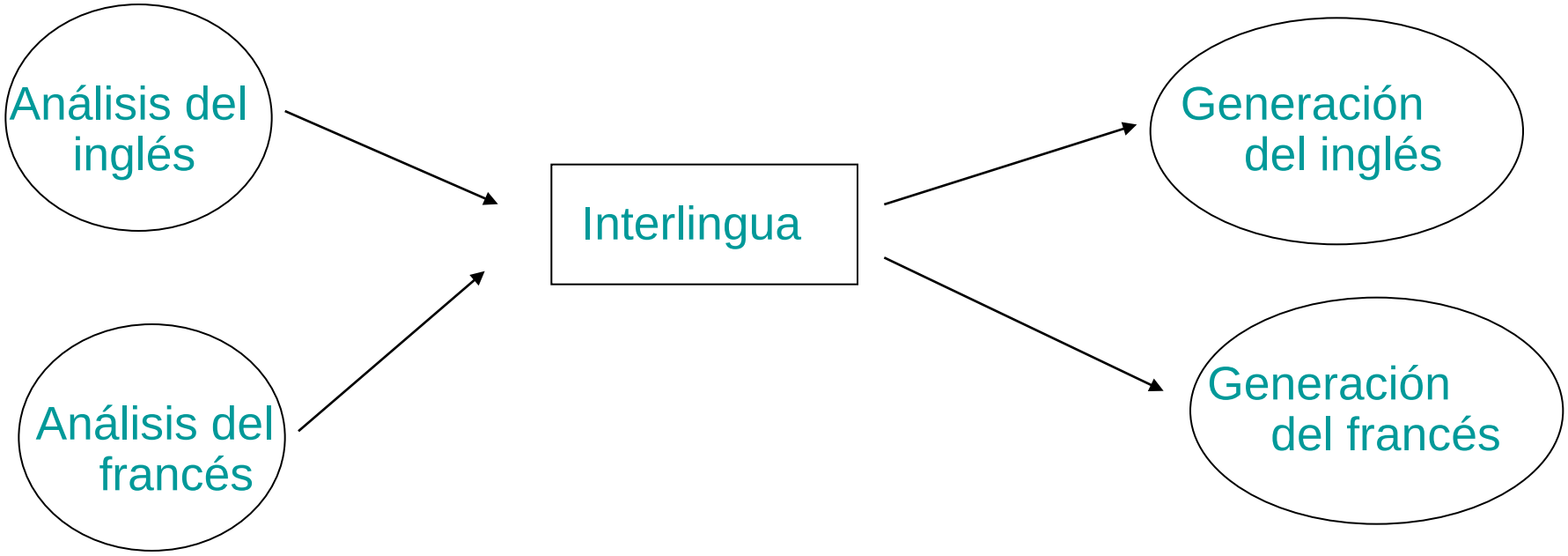
Qué hora es? ⟶ **What hour is?*

3. Métodos indirectos

- los programas de segunda generación
- Sistemas de interlingua
- Sistemas de transferencia

3. Métodos indirectos: sistemas de interlingua

- Dos fases:
  - Fase de análisis: del T_{LO} a la interlingua
  - Fase de generación: de la interlingua al T_{LM}
  - Interlingua: representación semántica de carácter “universal” común a las lenguas que se van a traducir



3. Métodos indirectos: sistemas de interlingua

- Una nueva lengua: dos módulos más (uno de análisis y otro de generación)
- Módulo de Análisis de español: 4 sentidos de traducción
  - inglés-francés, francés-inglés
  - español-francés, español-inglés
- Módulo de Generación del español: 6 sentidos de traducción
  - Además de los anteriores
  - francés-español, inglés-español
- Teóricamente, se permite la traducción desde y a la misma lengua
- Mayor dificultad: definir la interlingua

Problemas con la interlingua (1)

- Japonés:

  ‘hermano mayor’ vs. ‘hermano menor’

  ‘mi madre’ vs. ‘tu madre’ vs. ‘madre en general’
- Ruso:

  zhenit´sja ‘casarse (un hombre)’ vs. *vyxodit´zamuzh* ‘casarse (una mujer)’

Problemas con la interlingua (2)

- Alemán:

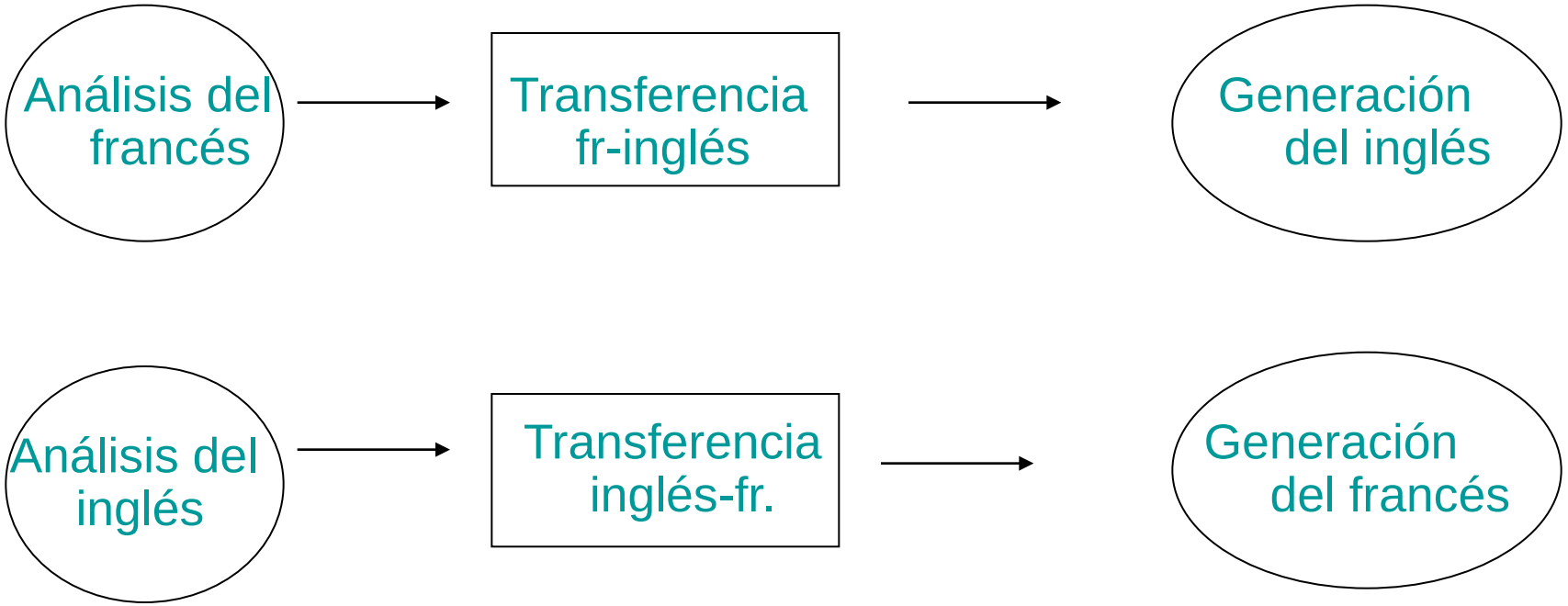
  essen ‘comer (las personas)’ vs. fressen ‘comer (los animales)’

- Francés:

  rivière ‘río (en general)’ vs. *fleuve* ‘río grande desembocando en el mar’

3. Métodos indirectos: sistemas de transferencia

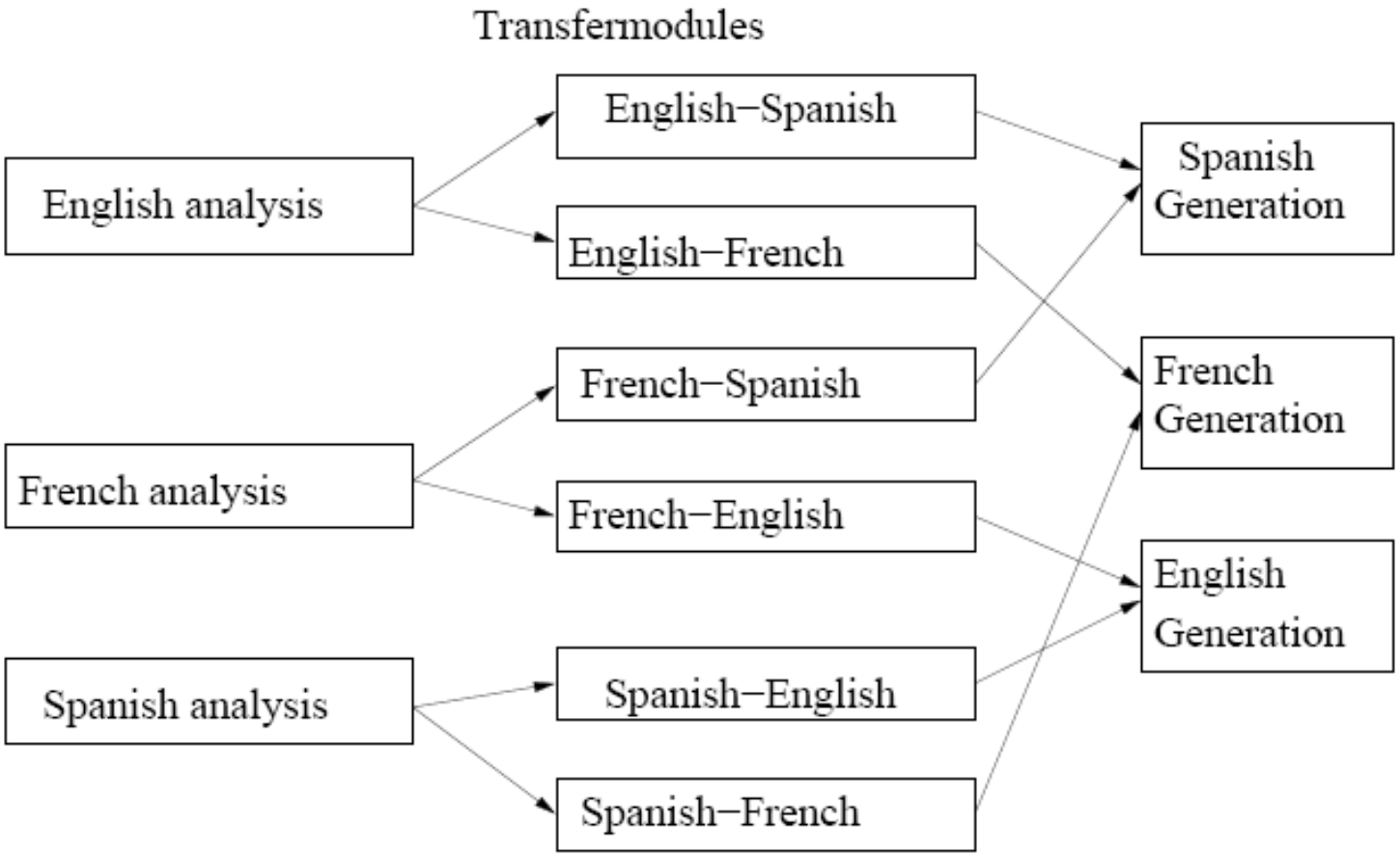
- Dos representaciones intermedias, una por cada lengua del par de traducción
- las representaciones son dependientes de la lengua que caracterizan



3. Métodos indirectos: sistemas de transferencia

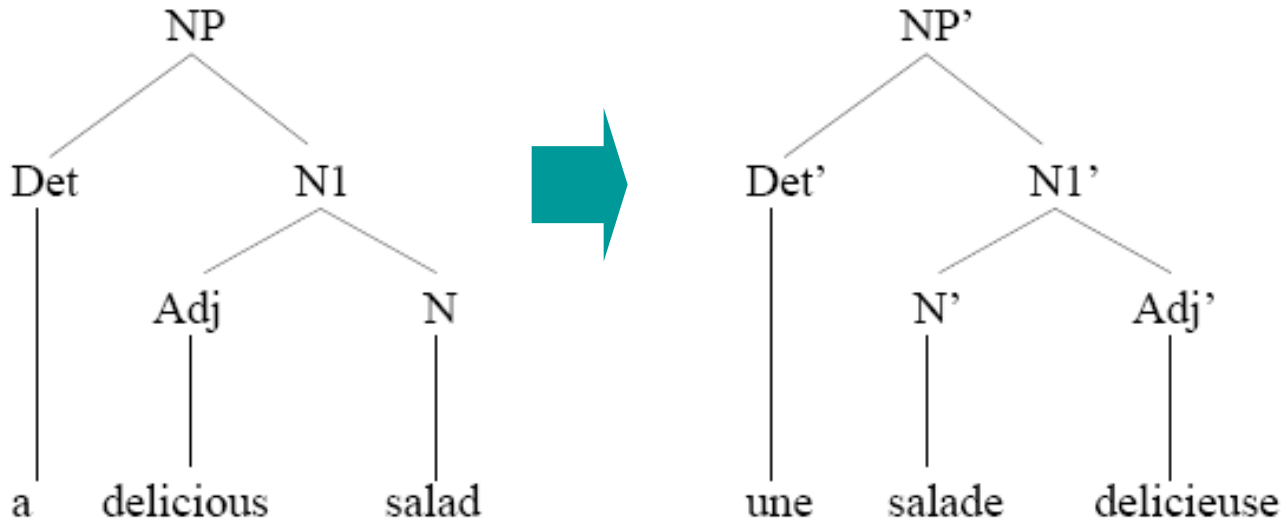
- el resultado del análisis es una representación abstracta del T_O
- el punto de partida para la generación es una rep. del T_M
- la función del módulo de transferencia es convertir las Rep. intermedias de L_O en Rep. intermedias de L_M
- la transferencia se puede realizar a nivel léxico, sintáctico o semántico
- Problema: una tercera lengua supone 4 nuevos módulos de transferencia
  - francés-español, español-francés
  - inglés-español, español-inglés
- Fórmula de módulos de transferencia: n(n-1)
- Fórmula de módulos totales: n(n+1)

Sistema de transferencia



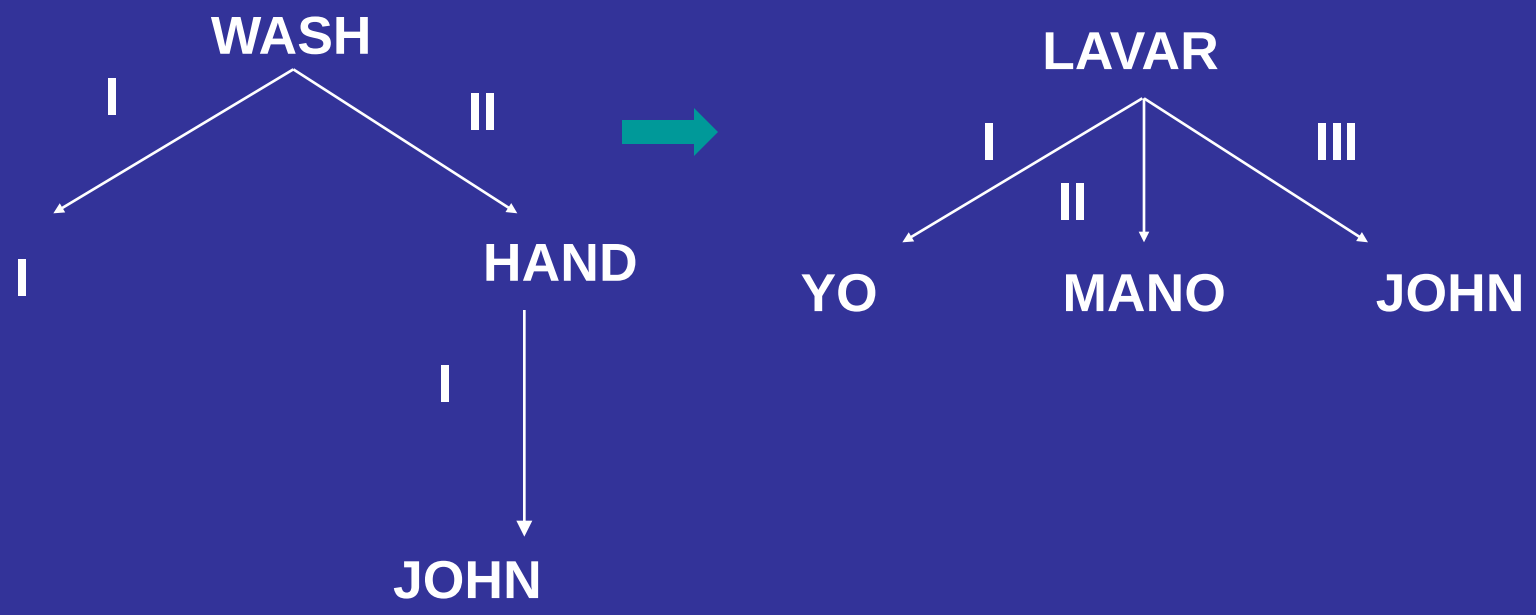
4. Ejemplos de Transferencia estructural

$ESintL_O$ y $ESintL_M$ en un Sistema de transferencia (1)



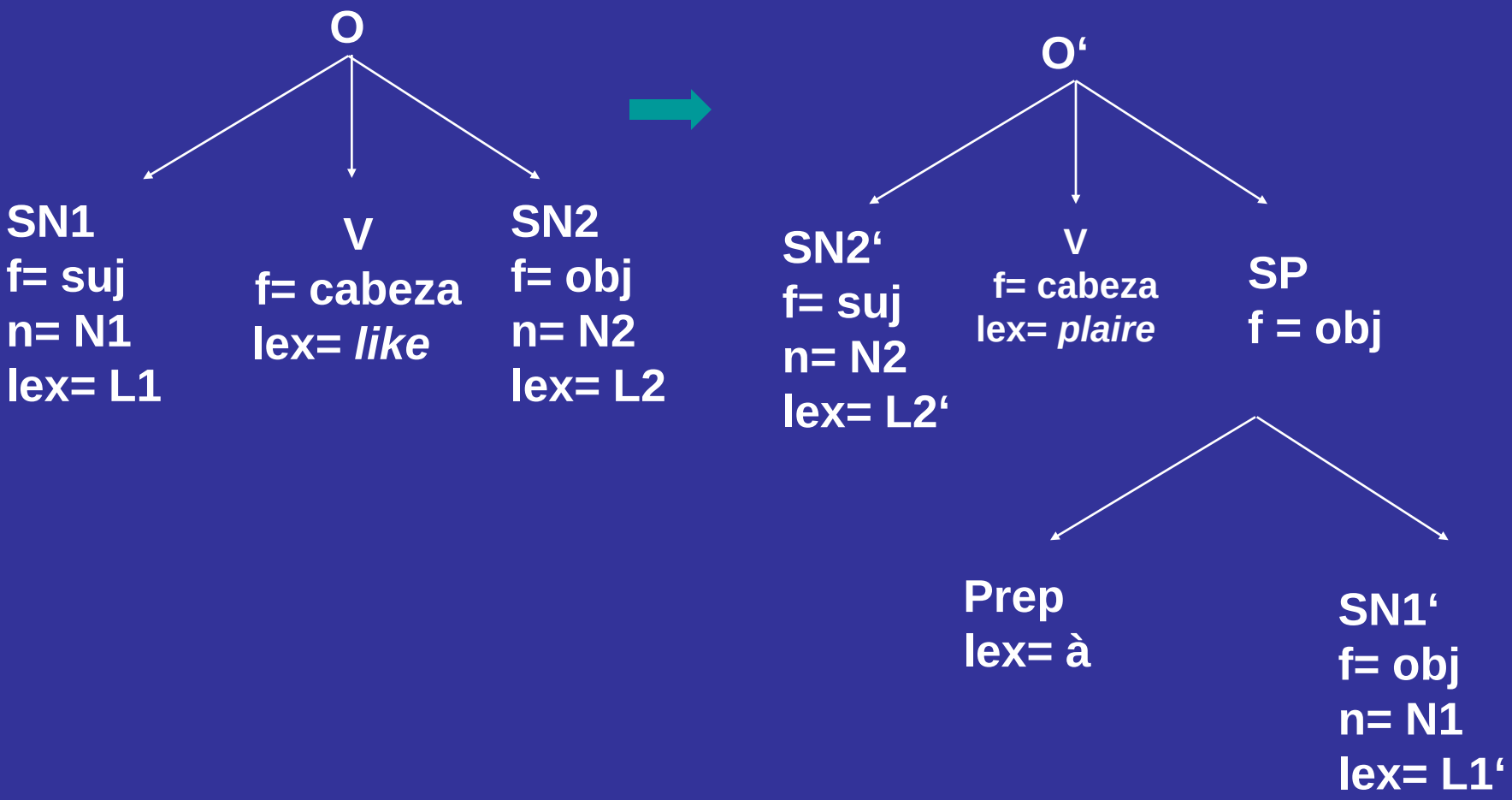
$ESintL_O$ y $ESintL_M$ en un sistema de transferencia estructural (2)

- I washed John's hands ←→ Le lavé las manos a John



Regla de transferencia estructural (1)

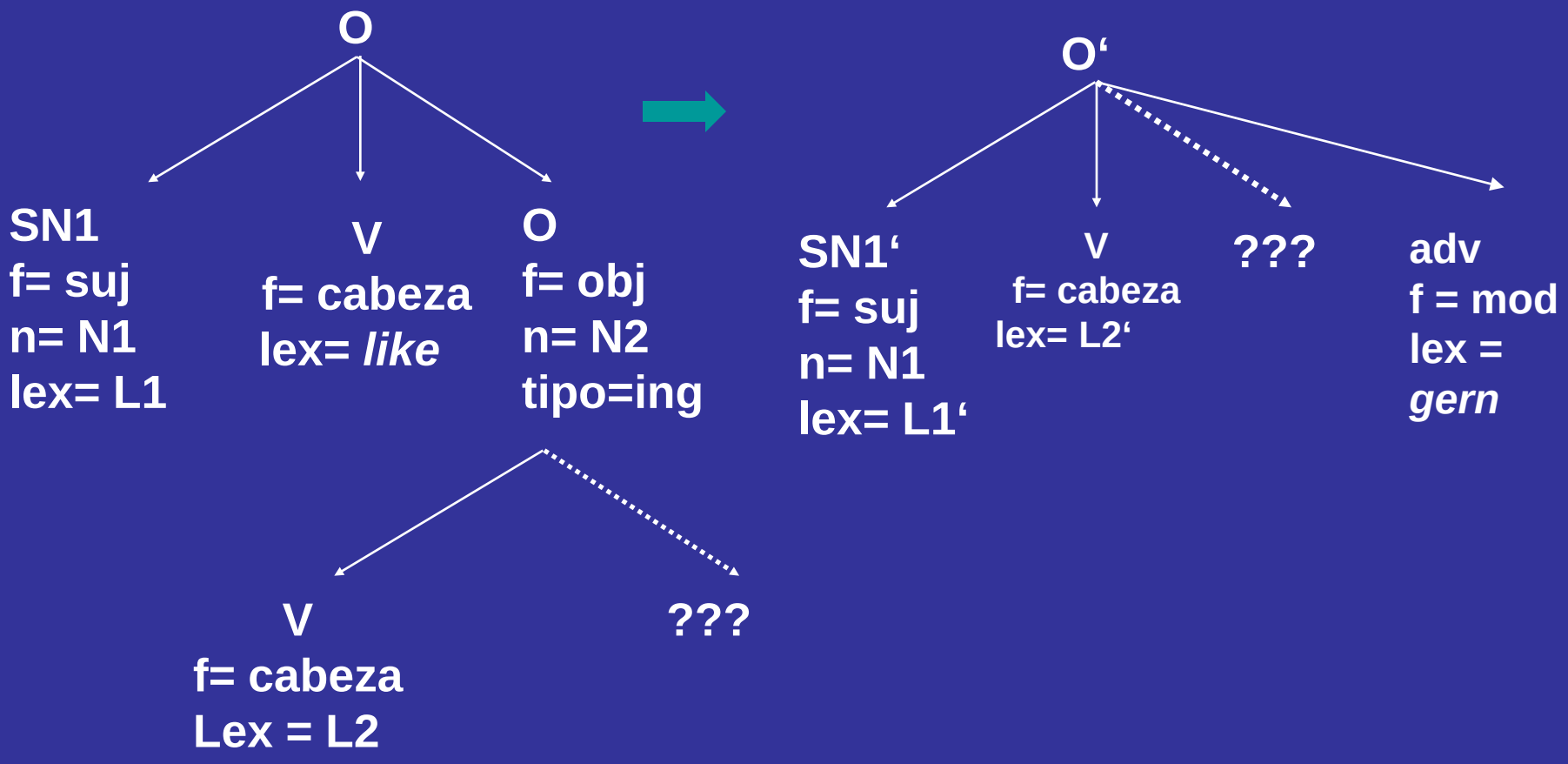
- John likes cheese ←→ Le fromage plaît à John
- like ←→plaire



Regla de transferencia estructural (2)

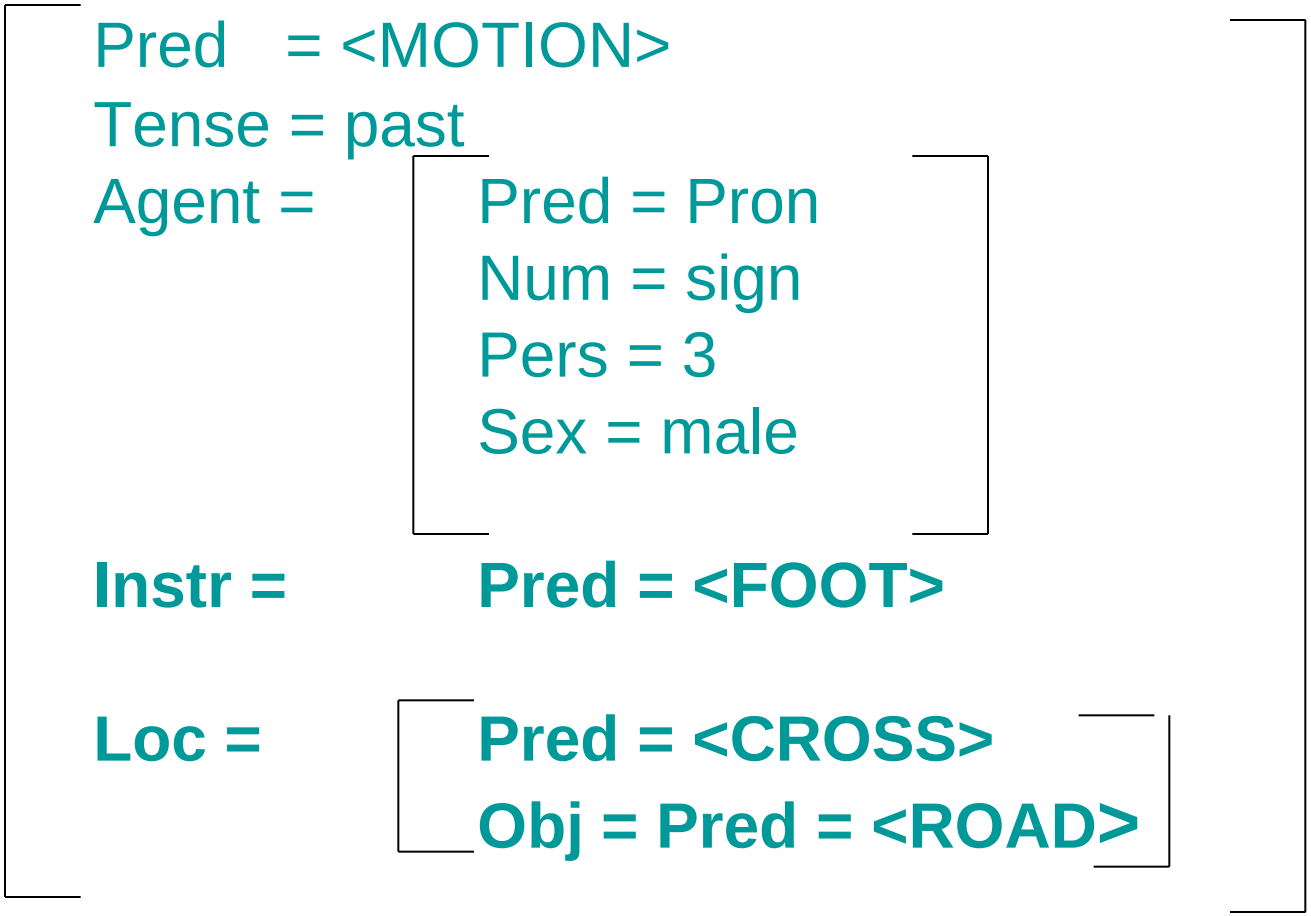
John likes swimming in the summer ←→ Johm schwimmt in Sommer gern

- like ←→gern



Ejemplos de Representación interlingüe (1)

He walked across the road ⇐> Cruzó la calle a pie



Ejemplo de Representación interlingue (2)

Periodically, clean the ventilation slots with your vacuum cleaner.

```
(*E-CLEAN
   (MOOD IMP)
   (EVENT-FREQUENCY &PERIODICALLY)
   (THEME (*O-VENTILATION-SLOT
               (NUMBER PL)
               (REFERENCE DEFINITE)))
   (INSTRUMENT (*O-VACUUM-CLEANER
               (PERSON SECOND)
               (POSSESSIVE +))))
```

Léxico en un sistema de interlingua

Correspondencias léxicas a interlingua

```
(:ROOT “clean”
    :CAT V
    :CLASS verbs-of-cleaning
    :HEAD *E-CLEAN)

(:ROOT “periodically”
    :CAT ADV
    :HEAD NIL
    SEM ((EVENT-FREQUENCY &PERIODICALLY)))

(:ROOT “vacuum cleaner”
    :CAT N
    :HEAD *O-VACUUM CLEANER)
```

5. Métodos estadísticos

- la traducción basada en analogías ha sido posible gracias a la recopilación de grandes corpus
- Corpus bilingües alineados
  - Alinear es hacer explícitas las relaciones de correspondencia entre segmentos del corpus bilingüe
  - Enfoque estadístico: aprovecha la similitud de rasgos cuantitativos como la longitud de las oraciones, el número de palabras o de caracteres
  - Enfoque lingüístico: emparejamiento previo de unidades sintagmáticas
- Investigaciones de IBM con el corpus Hansard
  - Cálculo de probabilidades de que una palabra cualquiera corresponda a dos, una o ninguna palabra en la oración traducida en la otra lengua
  - “the” corresponde a “le” con un índice del 61%, a “la” con 17,8%, a “l’” con 8,3%, a “les” con 2,3%

5. Métodos estadísticos

- Memorias de traducción
  - se almacenan traducciones realizadas manualmente y validadas para reutilizarlas en la traducción de textos similares
  - paquetes de software que incluyen los módulos de gestión, además de programas para crear y mantener bases de datos terminológicas, alineadores automáticos, etc.
  - **Programas más conocidos**: **Déjà vu** [http://www.atril.com] **Trados**
  - **Artículo http://www.asetrad.org/index.asp?op=24&detalle=12&pag=1**

5. Métodos estadísticos

- Algunas características de las memorias de traducción
  - ¿cómo habíamos traducido antes *real ale* en un texto sobre gastronomía
  - ¿y *benchmark* en un texto financiero? ¿y en uno de ingeniería?
  - Permiten “propagar” cambios a los archivos traducidos y recuerdan estos cambios para traducciones futuras
  - Uso y generación de glosarios

6. TA basada en reglas

Se basa en la descripción completa de ambas lenguas en todos los niveles:

- Léxico
- Morfológico
- Sintáctico
- Semántico

Hasta mediados de los 90, el enfoque más popular

TA basada en reglas

Pros

- Buena para empezar de cero
- Permite hacer generalizaciones
- Si el sistema dedujo que “is” es precedido, a menudo, por un nombre, basándose en casos como *the man/dog/horse is* …

entonces, no necesitaría encontrar todas las combinaciones de “nombre is” para saber que ‘previously-unseen-noun is’ es más probable que ’the is’.

CONTRAs

- Lleva tiempo producir textos de calidad
- Considerada cara
- Los recursos no son fácilmente usados para otros pares de lenguas

TA basada en corpus

Todo el conocimiento lingüístico es recuperado del **corpus paralelo**

El esfuerzo que suponen las representaciones intermedias es minimizado o eliminado

Las probabilidades que describen correspondencias entre la L_O y la L_M se adquieren a partir de un corpus bilingüe paralelo y las probabilidades de un modelo lingüístico se adquieren a partir de un texto monolingüe en la L_M

La alineación es de GRAN importancia

Encarna la idea de traducción por analogía (y no por análisis lingüístico (profundo): la gente traduce en base a traducciones previas

TA estadística basada en ejemplos

PROS

- lleva menos tiempo producir resultados de calidad
- No requiere representaciones ling. intermedias caras para producir resultados razonables
- considerada relativamente barata
- muchas herramientas reusables (menos fácilmente entre lenguas distantes)

CONTRAS

- se basa en **corpus bilingües**
- pocos métodos para controlar los errores
- recursos (dic. bilingües y gramáticas) son necesarios para mejorar la calidad

Parece que los enfoques “ortodoxos”

están limitados de uno u otro modo

pero pueden complementarse mutuamente.

Estudios de caso que pertenecen al enfoque HÍBRIDO

Experimentos de comparación

- Pares de lenguas:
  - Se esperaba que sueco-> holandés funcionara mejor que sueco-portugués, pero no fue así
- Corpus más grandes producen mejores resultados
- Resultados mejores con el enfoque de reglas donde se comprende el comportamiento del sistema y se pueden corregir los errores mejorando los léxicos y las gramáticas
- Los sistemas basados en la estadística hacen errores impredecibles
- Algunos aprenden a corregir errores basándose en las corrección humana de textos traducidos automáticamente
- Conclusión: optar por los basados en reglas pero apoyados estadísticamente

- **nada impedirá que Internet se convierta en un inmenso depósito abierto de traducciones.**
- **Es fácil imaginar motores de búsqueda similares en cobertura y potencia a Google, pero con un campo de acción especializado en la búsqueda en corpora multilingües.**
- **Traducir sería entonces tan sencillo como encontrar una equivalencia en la lengua deseada al texto de búsqueda.**
- **Linguee**

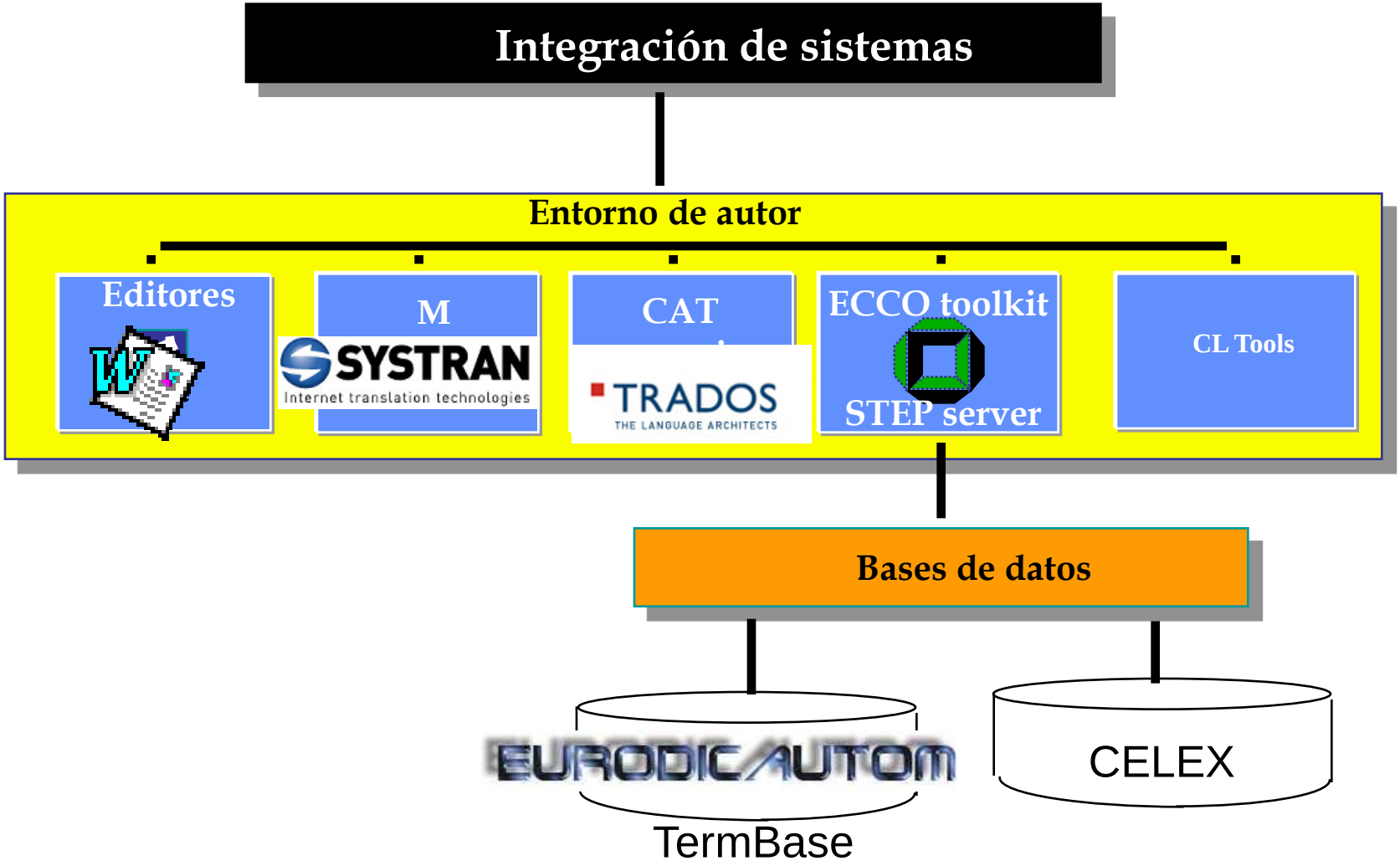
Corpus multilingües en TMX

- **Formato TMX (translation memory exchange format)**
- **Basado en XML**
- **TuMatXa con demos sobre Memorias de traducción:**
- **http://www.tumatxa.com/es/**
- **http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/ehu_uda01.htm**

Segmentación

<TU>
<TUV lang="EN" creationdate="1600" creationid="William Shakespeare" changedate="1951" changeid="Peter Alexander/Collins">
<SEG>**Exeunt marching. A peal of ordance shot off.**</SEG></TUV>
<TUV lang="ES" creationdate="1929" creationid="Luis Astrana Marín/Aguilar" >
<SEG>**Marcha fúnebre. Salen, llevándose los cadáveres. Después se oye una descarga de artillería.**</SEG></TUV>
<TUV lang="ES" creationdate="1994" creationid="José María Valverde/Planeta">
<SEG>**Se van marchando; después, se disparan salvas de artillería.**</SEG></TUV>
<TUV lang="ES" creationdate="1994" creationid="Ángel-Luis Pujante/Espasa">
<SEG>**Salen en marcha solemne, seguida de una salva de cañón.**</SEG></TUV>
</TU>

Herramientas: *The ideal workstation*



Direcciones con información sobre TA o TAO

- **Artículo de Guinovart de 1995, aunque antiguo, hay cosas todavía vigentes: http://www.elprofesionaldelainformacion.com/contenidos/1995/mayo/informtica_y_traduccin_estado_del_arte.html**

- **http://ddd.uab.cat/pub/quaderns/11385790n2p143.pdf**

- **Empresas**

  http://www.star-spain.com/es/inicio/index.php

  http://www.hermestrans.com/old/es/servicios.html

  http://www.sdl.com/es/

Imaxin, Proxecto Carvalho

http://www.youtube.com/watch?v=QW-9TrJ8qFQ

Direcciones web

http://es.wikipedia.org/wiki/Traducci%C3%B3n_autom%C3%A1tica

http://es.wikipedia.org/wiki/Memoria_de_traducci%C3%B3n

Corpus paralelos UVigo
http://sli.uvigo.es/CLUVI/index.html#lega
Traductor de Google:
http://www.google.com/intl/es/help/faq_translation.html
http://translate.google.com/about/intl/es_ALL/
Traductor de Microsoft:
http://www.bing.com/translator/
Traductor de El País:
www.elpais.es/traductor/index.html
El mundo:
www.elmundo.es/traductor
Instituto Cervantes:
http://traductor.cervantes.es/traduccion.htm

Opentrad:
- OpenTrad apertium: http:://sli.uvigo.es/tradutor

Sishitra
- http://sishitra.iti.es/

Ensayos

For Jean Rommes, the crisis came five years ago, on a Monday morning when she had planned to go to work but wound up in the hospital, barely able to breathe. She was 59, the president of a small company in Iowa. Although she had quit smoking a decade earlier, 30 years of cigarettes had taken their toll.

Con Google en 2007:

Para Jean Rommes, la crisis llegó hace cinco años, en un lunes por la mañana cuando ella tenía previsto ir a trabajar, pero acabó en el hospital, apenas pueden respirar. Ella era 59, el presidente de una pequeña empresa en Iowa. Aunque había dejado de fumar antes de una década, 30 años de cigarrillos habían cobrado su tributo

Con Google en 2008:

Para Jean Rommes, la crisis llegó hace cinco años, en un lunes por la mañana cuando tenía previsto ir a trabajar, pero la herida en el hospital, apenas capaz de respirar. Ella fue de 59, el presidente de una pequeña empresa de Iowa. Aunque había dejado de fumar de una década antes, de 30 años de cigarrillos habían tomado su peaje

Con Google en 2009:

Para Jean Rommes, la crisis se produjo hace cinco años, en una mañana de lunes, cuando ella había planeado para ir a trabajar, pero terminó en el hospital, apenas podía respirar. Ella fue de 59, el presidente de una pequeña empresa de Iowa. A pesar de que había dejado de fumar hace una década, de 30 años de cigarrillos habían tomado su peaje.

Ensayos

For Jean Rommes, the crisis came five years ago, on a Monday morning when she had planned to go to work but wound up in the hospital, barely able to breathe. She was 59, the president of a small company in Iowa. Although she had quit smoking a decade earlier, 30 years of cigarettes had taken their toll.

Con Google en 2010:

Para Rommes Jean, la crisis llegó hace cinco años, en una mañana de lunes, cuando ella tenía planeado ir a trabajar, pero terminó en el hospital, casi sin poder respirar. Ella fue de 59, el presidente de una pequeña empresa de Illinois. A pesar de que había dejado de fumar hace una década, de 30 años de cigarrillos había tomado su peaje

Con Google en 2011:

Para Jean Rommes, la crisis llegó hace cinco años, en una mañana de lunes, cuando ella había planeado para ir a trabajar, pero terminó en el hospital, casi sin poder respirar. Ella fue de 59, el presidente de una pequeña empresa de Iowa. A pesar de que había dejado de fumar hace una década, de 30 años de cigarrillos habían dejado su huella.

Con Google en octubre 2012

Para Jean Rommes, la crisis llegó hace cinco años, en una mañana de lunes, cuando ella tenía planeado ir a trabajar, pero terminó en el hospital, casi sin poder respirar. Ella era de 59 años, presidente de una pequeña empresa de Iowa. A pesar de que había dejado de fumar hace una década, 30 años de cigarrillos había dejado su huella.