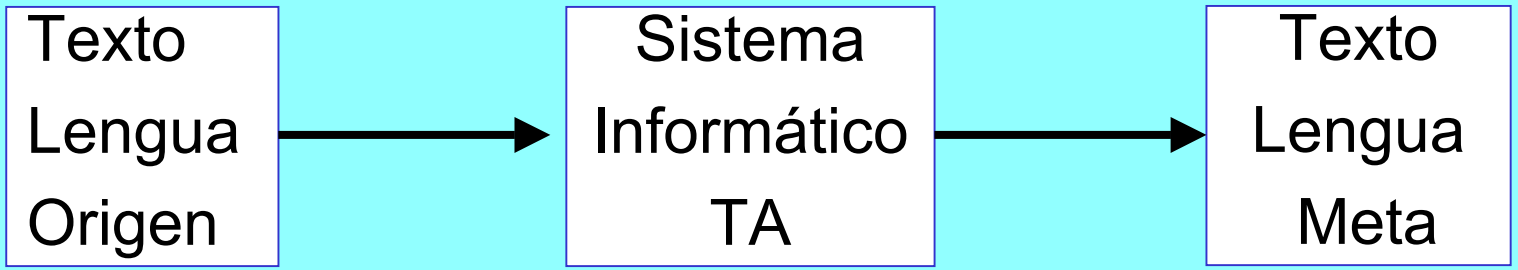
Traducción Automática

1.- Traducción automática (TA)

- 1.1.- ¿Qué es la TA?
- 1.2.- Breve historia
- 1.3.- Expectativas de la TA
- 1.4.- Métodos básicos en TA
- 1.5.- Principales problemas de la TA (análisis)
- 1.6.- Problemas de transferencia e interlingüa
- 1.7.- La generación
- 1.8 .- Algunos programas comerciales de TA

1.1.- ¿Qué es la Traducción Automática?

Traducción Automática proviene de ***Machine Translation***



Traducción Automática **(TA): sistemas informáticos que llevan a cabo traducciones de una lengua a otra con o sin intervención humana**

Núcleo de la TA: **la automatización del proceso de traducción en su totalidad**

1.1.- ¿Qué es la Traducción Automática?

El proceso automático de traducción suele requerir intervención humana:

- Sistemas con **postedición**
- Sistemas con **preedición**
- Sistemas **interactivos**

1.1.- ¿Qué es la Traducción Automática?

- Sistemas con **postedición**:

  - Revisión de la traducción obtenida automáticamente

    - el carácter de la revisión depende del destinatario

(primer borrador -pretraducción-, traducción de cierta calidad, versión para un especialista en el tema)

- Sistemas con **preedición**:

  -Escritura del texto origen en un lenguaje controlado para:

    - reducir las ambigüedades potenciales

    - restringir la complejidad sintáctica de las oraciones

1.1.- ¿Qué es la Traducción Automática?

- Sistemas **interactivos**:

  - El programa indica los problemas de ambigüedad y de selección y el usuario los resuelve durante la traducción

1.2 Breve historia

Primeras ideas en el siglo XVII

- Descartes, Leibniz: formulan teorías sobre la elaboración de diccionarios basados en códigos numéricos universales
- Cave Beck, Athanasius Kircher, Johann Becher: trabajan en una “lengua universal” sin ambigüedades que se base en principios lógicos y símbolos icónicos
- John Wilkins: elabora una interlingüa en “Essay towards a Real Character and a Philosophical Language” (1668)

1.2 Breve historia

Siglos XVIII, XIX

- Otras muchas propuestas de lenguas internacionales

  (*esperanto* es el ejemplo más conocido)

Primera mitad del XX

1933 George Artsouni: dispositivo de almacenamiento en banda de papel de una especie de diccionario multilingue

Petr Smirnov-Troyanskii: se anticipa a su época con sus teorías sobre la TA

1.2 Breve historia

Petr Smirnov-Troyanskii: concibe 3 fases en la TA

1.- Análisis "lógico" de las palabras, reduciéndolas a:

- formas básicas y
- funciones sintácticas

2.- Transformación de secuencias de formas básicas y funciones en secuencias equivalentes en la lengua meta

3.- Conversión del producto de las 2 fases anteriores en las formas normales de la lengua meta

Patentó una máquina para la fase 2.

1.2 Breve historia

1949 Weaver plantea:

- la posibilidad de utilizar ordenadores para llevar a cabo traducciones
- métodos
  - técnicas criptográficas (usadas en la guerra)
  - análisis estadísticos
  - teoría de la información de Shannon
  - exploración de la lógica subyacente
  - exploración de las características universales del lenguaje

1.2 Breve historia

1952 Primer simposio sobre TA:

- necesidad de pre y postedición
- construcción de sistemas para sublenguajes
- demostrar la viabilidad técnica de la TA

1954 Primera demostración pública de un sistema de TA:

- traducción ruso-inglés
- vocabulario restringido (250 palabras) y 6 reglas gramaticales
- experiencia de escaso valor científico pero que estimuló la financiación de proyectos de TA en USA y URSS

1.2 Breve historia

Década de los 60

Surgen 2 tendencias a la hora de abordar la TA

- Métodos empíricos con base estadística (Universidad de Washington, IBM, Universidad Georgetown)
- Planteamientos teóricos basados en el estudio de lingüística fundamental (MIT, Harvard, Berkeley, Univ. de Leningrado)

1.2 Breve historia

- La investigación en la década de los 60 influyó notablemente no solo en la TA sino en:
  - lingüística computacional
  - inteligencia artificial

- La comunidad científica fue tomado conciencia de la complejidad de los problemas lingüísticos

1.2 Breve historia

1966 Informe ALPAC (Automatic Language Processing Advisory Commitee) concluye:

- "*no existe expectativa inmediata o previsible de que la TA resulte útil*"
- desaconseja realizar inversiones en investigación en TA
- recomienda:
  - desarrollo de herramientas para los traductores
  - apoyo al estudio en lingüística computacional

Consecuencia: abandono casi total de la investigación en TA en USA

1.2 Breve historia

Década de los 70

- La investigación en TA se traslada a Canadá y Europa Occidental

1976 sistema de traducción de partes meteorológicos (inglés-francés, lenguaje controlado) METEO

Finales de los 70 La Comunidad Europea plantea el proyecto EUROTRA

Objetivo: creación de un sistema de TA de diseño avanzado capaz de trabajar con todas las lenguas de la CE

1.2 Breve historia

Década de los 80

- Además del modelo de transferencia toman vigencia los sistemas basados en **interlingüa** y los **sistemas basados en conocimiento** (fundados en la investigación en IA sobre comprensión del lenguaje natural)

Década de los 90

- A los modelos anteriores se suma el modelo estadístico pudiendose hablar de la TA estadística

Desde los 80

- incorporación en los canales comerciales de programas de TA (en general baja calidad lingüística)

1.3 Expectativas de la TA

Primeros objetivos (década de los 50):

☺ Traducir de forma automática cualquier texto

🕐 Tras los primeros resultados ...

☹ Toma de conciencia de:

- problemas lingüísticos
- problemas de formalización y modelización del LN

1.3 Expectativas de la TA

Objetivos (década de los 70 hasta los 90):

- Mismos objetivos
- Proyectos internacionales generosamente financiados

☹ Resultado = No consecución de los objetivos

Objetivos (década de los 90):

- TA en dominios específicos
- TA con postedición en traducciones de calidad

☺ Algunos investigadores tienen esperanzas en la consecución del objetivo inicial pero no a corto plazo

1.3 Expectativas de la TA

Cuello de botella de la sociedad de la información

Sobrecarga informativa

- "En los albores de una *nueva era* que seguirá siendo plurilingüe, la traducción es el principal cuello de botella para la pretendida *globalización* de la información..."
- "Comparado con las rotativas más modernas capaces de producir unos **20 millones de páginas por hora**, un traductor manual puede llegar a rendir, en los casos más favorables, a un ritmo de **20 páginas por día**,..."
- La incorporación de nuevos estados miembros a la Unión Europea plantea graves problemas de traducción. Fuente:

  The Journal of Record for Human Language Technology (Sept. 1999)

1.3 Expectativas de la TA

- “La productividad de los traductores puede ser muy variable, de las menos de 100 páginas por traductor/año según la Secretaría de Estado Canadiense, a las más de 700 páginas que la Comisión Europea otorga a sus traductores ...

- “... si se tiene en cuenta que sólo un 3% del volumen total de páginas traducidas corresponde a obras literarias, existen motivos más que suficientes para ensayar la mecanización a gran escala de la producción plurilingüe del grueso de publicaciones diplomáticas, administrativas, comerciales y técnicas, cuyas traducciones son, por su propia naturaleza, mecánicas y rutinarias, ...”

1.3 Expectativas de la TA

Para muchos investigadores los objetivos "realistas" de la TA:

- Traducciones en borrador en áreas bien delimitadas
- Textos sin valor literario

1.3 Expectativas de la TA

Lenguajes de especialidad

- Los lenguajes formales comparten dos características esenciales con los lenguajes utilizados en las áreas de especialidad:

  - son precisos y

  - están sujetos a la normalización

- Las evaluaciones realizadas sobre los sistemas de TA:

  los mejores resultados cualitativos se obtienen precisamente aprovechado las propiedades de los lenguajes de especialidad (***sublenguajes***)

1.4 Métodos básicos en la TA

Se pueden distinguir dos enfoques:

- basados en la lingüística computacional y en IA
- basados en el estudio y procesamiento de corpus

Métodos basados en la lingüística computacional y en IA

- Método directo
- Métodos indirectos
  - interlingüa
  - transferencia

1.4 Métodos básicos en la TA

Método directo

Características:

- Utilizado en los años 50
- Carece de fase intermedia
- Producto final ≈ traducción por palabra
- No hay un análisis sintáctico ni semántico
- Pueden darse errores léxicos

1.4 Métodos básicos en la TA

Método directo

Características:

- Dan lugar a estructuras sintácticas inapropiadas
- Carencia de una base lingüística apropiada

Uso actual:

- En sistemas de traducción bilingüe unidireccional en las estructuras similares entre la lengua origen y meta

1.4 Métodos básicos en la TA

Métodos indirectos:

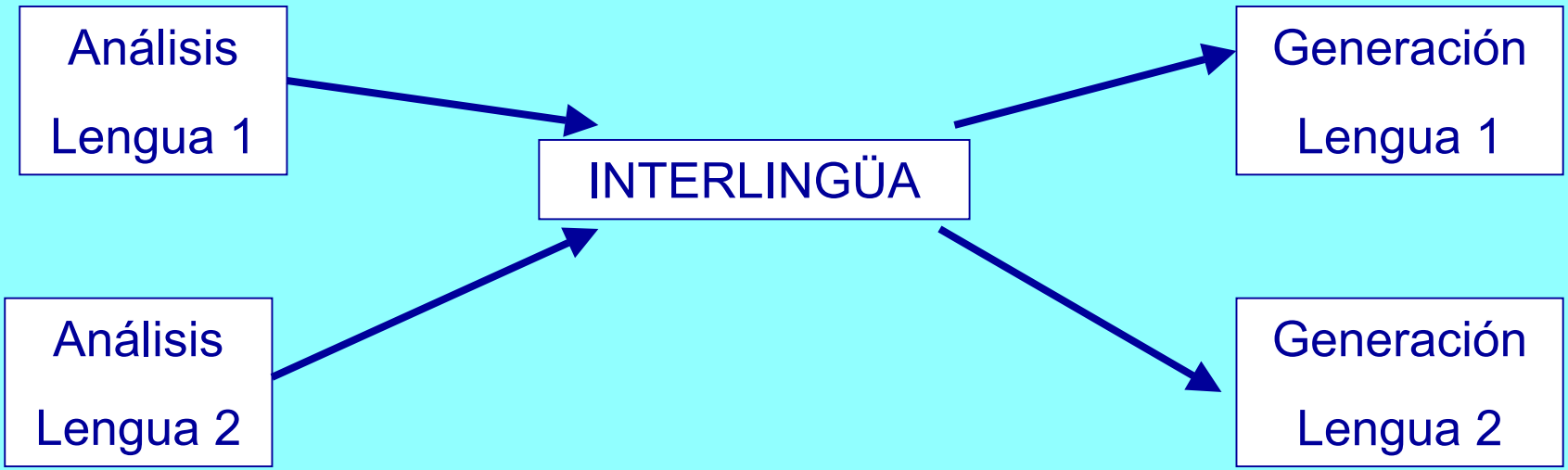
utilizan una representación intermedia a partir de la que se genera el texto en la lengua meta

Dos métodos indirectos:

- interlingüa
- transferencia

1.4 Métodos básicos en la TA

Método indirecto: interlingüa



Representación interlingüa:

- abstracta
- independiente de las lenguas origen y meta

1.4 Métodos básicos en la TA

Método indirecto: interlingüa

Dificultad:

- definición de una representación interlingüa “universal” que pueda ser una representación intermedia entre cualesquiera lenguas

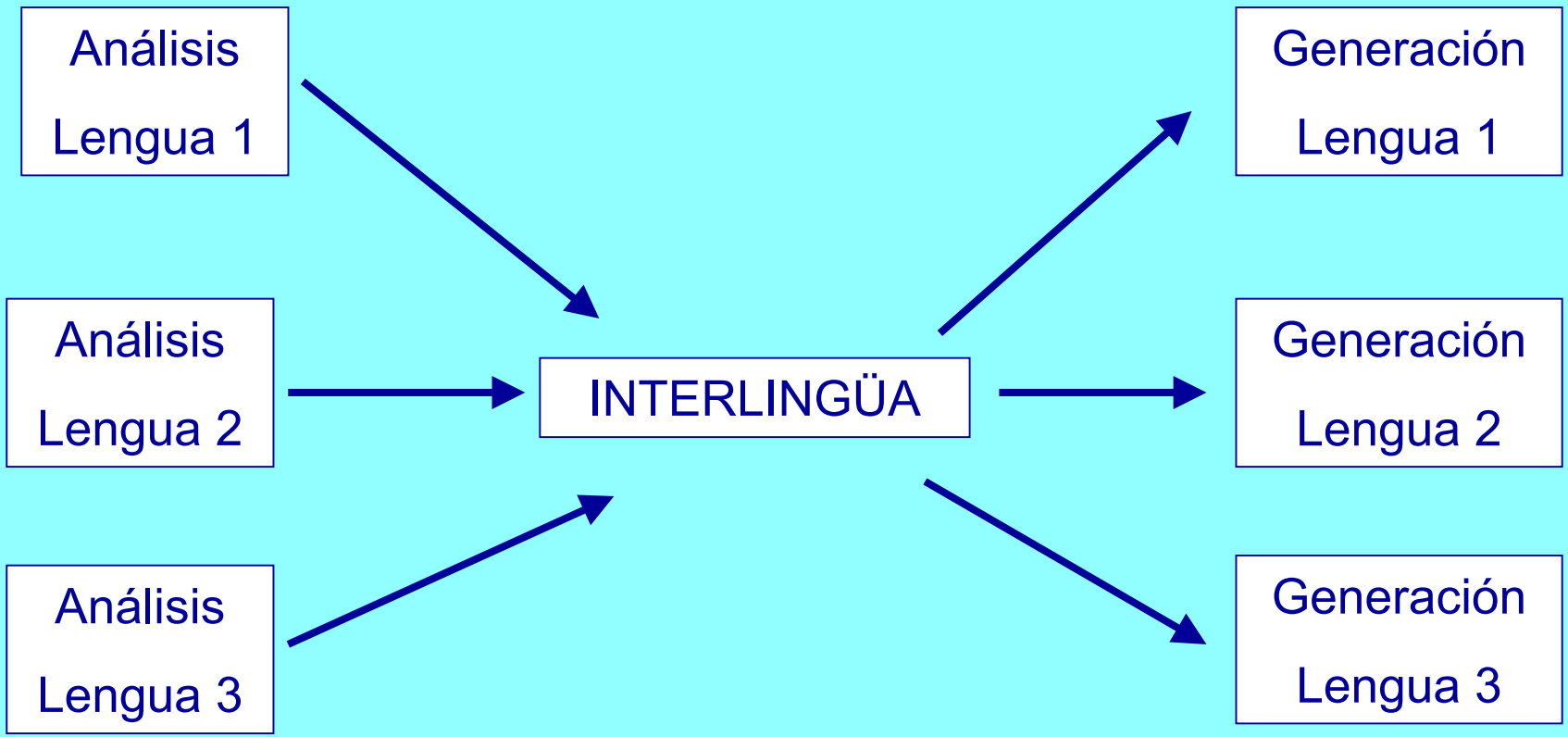
Ventajas:

- Facilita el desarrollo de sistemas multilingües ya que el módulo de análisis es independiente del de generación

- Incorpora los niveles del análisis lingüístico

1.4 Métodos básicos en la TA

Método indirecto: interlingüa



1.4 Métodos básicos en la TA

Método indirecto: transferencia

Características:

- utiliza una representación intermedia dependiente del par de lenguas

1.4 Métodos básicos en la TA

Método indirecto: transferencia

Dificultad:

- El desarrollo de sistemas multilingües es más dificultoso que en el modelo de interlingüa ya que hay que diseñar módulos de transferencia en cada par de lenguas y sentido de la traducción

Ventajas:

- El diseño del módulo de transferencia es menos complejo que la representación interlingüe ya que la representación intermedia es una **abstracción dependiente de la lengua**

1.4 Métodos básicos en la TA

Sistemas basados en conocimiento

- Parten del modelo interlingüa
- Fundamento:
  - La traducción se basa en volcar “el significado” de un texto en un nuevo texto en la otra lengua
  - Un sistema de TA debe “entender” los significados de los textos
  - Sin comprensión un programa no podrá decidir cuál de las expresiones de la lengua meta corresponde al significado del texto original

1.4 Métodos básicos en la TA

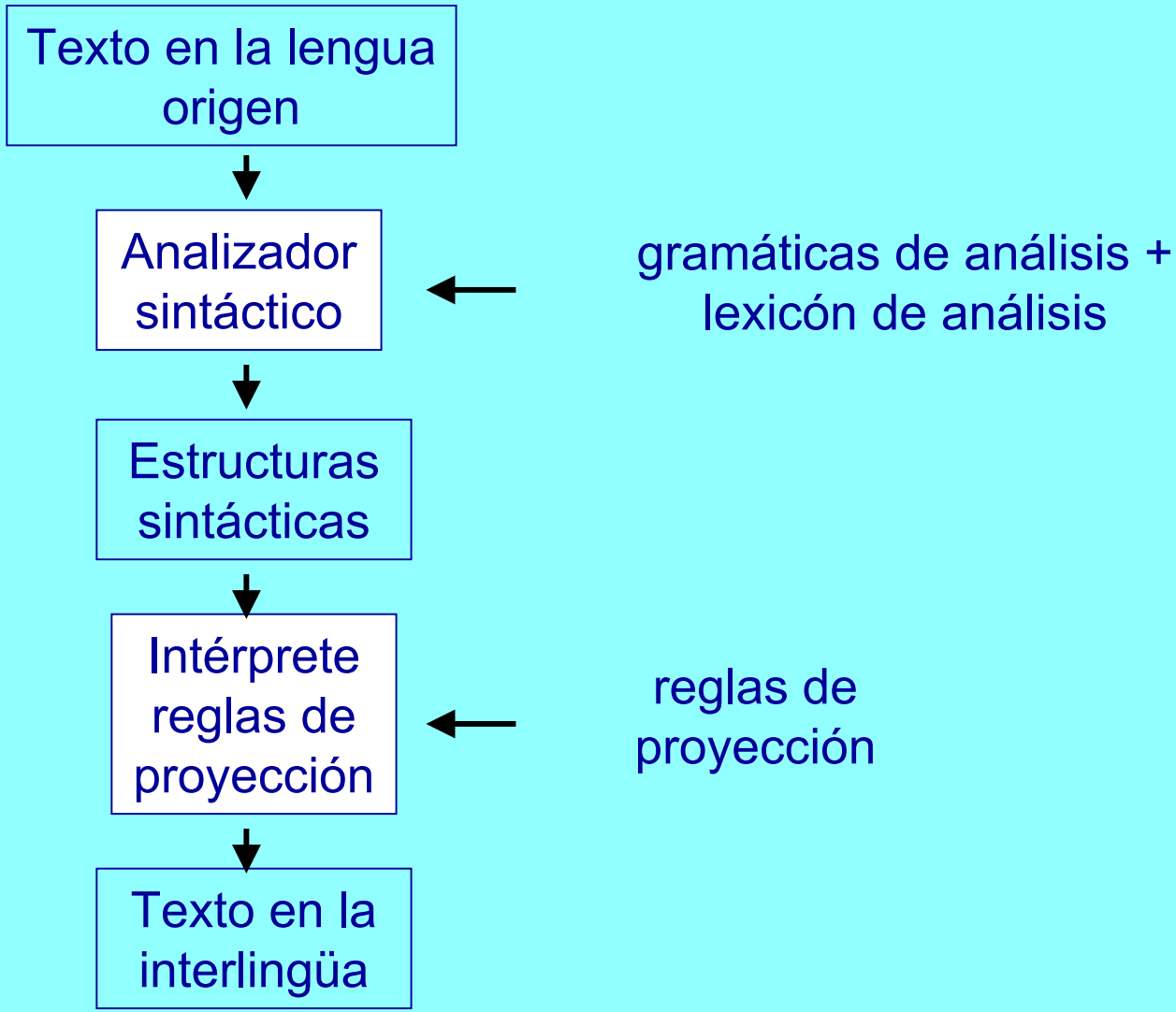
Sistemas basados en conocimiento

Características:

- análisis orientado a la semántica
- interpretación de los textos utilizando bases de conocimiento
- uso de mecanismos de inferencia y representaciones del significado independientes de toda lengua

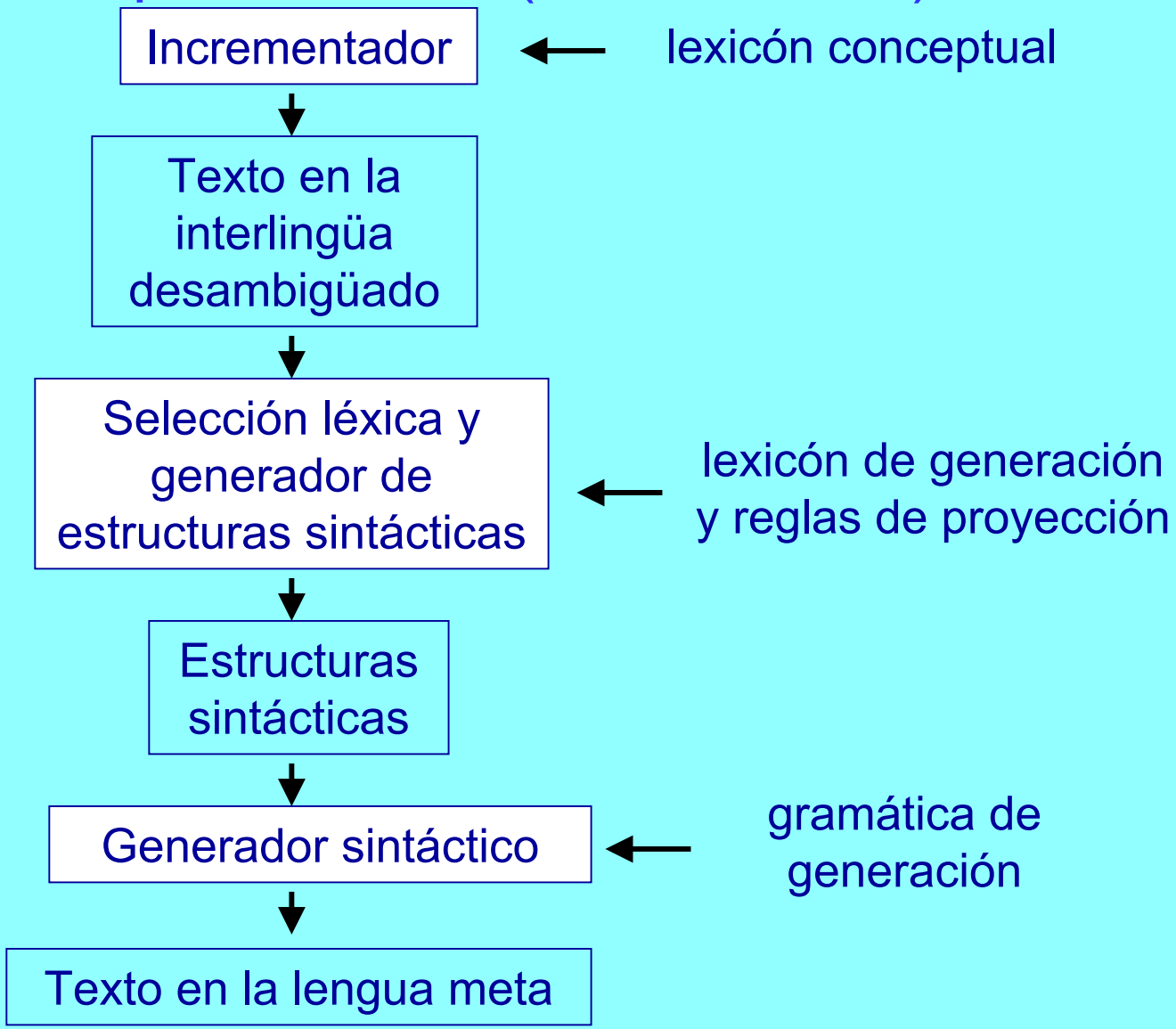
1.4 Métodos básicos en la TA

Esquema de un típico STABC:



1.4 Métodos básicos en la TA

Esquema de un típico STABC (continuación):



1.4 Métodos básicos en la TA

Recursos necesarios:

- lexicones de análisis y generación (dependientes de las lenguas y del dominio)
- lexicón conceptual (específico del dominio)
- reglas de proyección (dependientes de las lenguas y del dominio)
- gramáticas de análisis y generación (dependientes de las lenguas y del dominio)

1.4 Métodos básicos en la TA

Lexicón conceptual: base de datos de conocimiento sobre los eventos y entidades comprendidos en el dominio

Por ejemplo:

tulipán concepto: flor

color: (blanco, negro, rojo, amarillo, azul, ...)

Representación interlingüe común**: redes de proposiciones**: eventos o estados con sus correspondientes argumentos y con conexiones causales, temporales, espaciales, etc., a otros eventos o estados

1.4 Métodos básicos en la TA

Métodos basados en conocimiento

Dificultades:

- Viabilidad de la elaboración de bases de conocimiento independientes de las lenguas para dominios que no muestren un alto grado de especificidad

- Alto coste computacional

Principal aplicación:

- TA en dominio restringidos

1.4 Métodos básicos en la TA

RBMT (Rules-Based MT)

- M. Directo
- M. Indirecto
- M. Basados conocimiento

Limitaciones (según algunos autores):

- Requieren de la formalización de los fenómenos lingüísticos mediante reglas
- Es difícil hacer uso de información situacional o de dominio
- Se basan en gramáticas que no siempre contemplan los usos reales del lenguaje

1.4 Métodos básicos en la TA

Métodos basados en el estudio y procesamiento de corpus:

Características:

- Corpus o BD de textos en diversas lenguas
- Uso de modelos estadísticos
- Resolución de problemas de optimización

Modelos:

- TA basada en ejemplos (*Example-Based MT*)
- TA estadística

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Características:

- El proceso de traducción se ve como un proceso de "encontrar" ejemplos análogos traducidos anteriormente
- Se fundamenta en la reutilización de traducciones humanas validadas una vez han sido analizadas
- Suponen una alternativa a los enfoques basados en conocimiento
- Suponen un apoyo a los métodos tradicionales de transferencia e interlingüa

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Ejemplo. Supongamos el siguiente conjunto de sintagmas bilingües que impliquen a la palabra inglesa *field*:

the main fields	*los campos principales*
the coal fields	*los yacimientos de carbón*
the corn fields	*los campos de maíz*

(BD de sintagmas alineados)

La traducción de *field* (*campo, yacimiento, área)* vendrá determinada por la frecuencia de aparición de los sintagmas cuyos contextos son más parecidos al del ejemplo

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Ejemplo. Si quisiéramos traducir *the gold field* (yacimiento aurífero), no encontraríamos un **emparejamiento exacto** en la BD de ejemplos ⇒ surge el concepto de **similitud**

La similitud entre una entrada y un ejemplo de la BD vendrá dada por una **medida** de la **distancia del significado**

Si los términos léxicos se clasifican por jerarquías semánticas, la jerarquía indicaría una distancia menor entre *gold* y *coal* que entre *gold* y *corn* ⇒

Mayor probabilidad de que *field* se traduzca por *yacimiento* que por *campo*

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

- Se puede utilizar para determinar qué **estructura oracional** meta es la que le corresponde a una oración origen

- En este caso la similitud se podría referir a:

    - la distribución de los elementos gramaticales

    - secuencias de ciertas categorías gramaticales

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Si tenemos la oración:

> *Remove the bulb and replace it with a new one*

la podríamos contrastar con un ejemplo en nuestra BD:

> *Remove the X and replace it with Y*
>
> *Quite X y sustitúyalo por Y*

donde:

X: podrá ser un sustantivo o adjetivo+sustantivo cualquiera

Y: podrá ser cualquier sintagma nominal

Los elementos no comparables se han traducido aparte

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Requiere: Banco de conocimiento bilingüe o multilingüe (BKB)

Un BKB parte de: - un corpus de textos equivalentes en dos o más lenguas analizados estructuralmente (anál.morfosint.)

puede contener además: - diccionarios

- otras bases de conocimiento

Los textos equivalentes se estructuran es:

- ***unidades de traducción*** (UT) alineadas

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Unidad de traducción (UT): fragmento de texto

- Granularidad de las UT: oración, SN, SV, ...

- **UT alineada**: un fragmento de texto relacionado con una traducción de dicho fragmento en al menos otra lengua

Memorias de traducción: recopilaciones de UT alineadas

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

- Puede integrarse en cualquiera de los modelos básicos (directo, transferencial, interlingüa)
- En TA basada en conocimiento no puede integrarse: depende del análisis semántico hasta un grado muy alto de abstracción
- TA basada en ejemplos puede aplicarse a cualquier nivel de transferencia (morfológico, sintáctico, semántico)

1.4 Métodos básicos en la TA

TA basada en ejemplos (Example-Based MT) :

Ventajas:

- La tarea de construir BKB es una tarea con un grado de viabilidad aceptable

- Los textos (la cobertura léxica, sintáctica, etc) pueden seleccionarse para cubrir diversos dominios o necesidades específicas del usuario

- Las BD pueden actualizarse con facilidad para tratar neologismos añadiendo textos y mediante el “aprendizaje”

- La información contextual que proporcionan los ejemplos es difícil encontrarla en los diccionarios

1.4 Métodos básicos en la TA

TA estadística :

Objetivo: desarrollar sistemas de TA basados casi exclusivamente en técnicas estadísticas

Parten de:

- Un corpus paralelo
- Alineación de oraciones, locuciones, palabras basada en métodos estadísticos

1.4 Métodos básicos en la TA

TA estadística :

- El análisis estadístico se utilizó en los primeros años de la investigación en TA:

  - clasificación automática de los datos lingüísticos

- Ha continuado vigente hasta nuestros días para:

  - dirigir la selección de reglas de transferencia

  - selección y resolución de ambigüedades léxicas

- IBM impulsó el uso de técnicas estadísticas como única herramienta de análisis y generación

1.4 Métodos básicos en la TA

TA estadística :

- **Impulso**: el desarrollo y éxito de métodos estadísticos en procesamiento y reconocimiento del habla

Experimento inicial:

- Corpus paralelo bilingüe inglés/francés (debates del parlamento canadiense) llamado *Hansard* canadiense (40.000 pares de oraciones)
- Alineado a nivel de oraciones
- Cálculo de las probabilidades de que una palabra cualquiera situada en una oración de una lengua corresponda a 2, 1 o 0 palabras en la oración de la otra lengua

1.4 Métodos básicos en la TA

TA estadística :

- Las probabilidades se calculan mediante el cotejo de "bigramas" (dos palabras consecutivas) en cada oración en inglés frente a "bigramas" en las oraciones francesas equivalentes
- Se obtuvieron 2 conjuntos de probabilidades:

  - cada palabra inglesa junto con las probabilidades de sus correspondencias con respecto a un conjunto de palabras francesas

  - de que 2, 1 o 0 palabras francesas correspondan a una única palabra inglesa

1.4 Métodos básicos en la TA

TA estadística :

- Con un vocabulario de las 1000 palabras más frecuentes en inglés y sus correspondientes en francés (1700)
- Se tradujeron 73 oraciones de otras secciones del *Hansard* con un 48% de éxito

  Éxito: (traducciones idénticas a las del Hansard+el mismo significado pero ligera variación en las palabras+ traducción legítima pero no expresa el mismo significado)

1.4 Métodos básicos en la TA

TA estadística :

Posibles mejoras:

- Utilizar un corpus más extenso
- Segmentación probabilística de las oraciones en sintagmas
- Uso de trigramas
- Incluir información sobre morfología flexiva

Tendencia actual: sistemas híbridos que incorporan métodos estadísticos y conocimiento lingüístico

1.5 Principales problemas de la TA

Características de la mayoría de los sistemas de TA:

- No se basan en un único modelo lingüístico (los que lo usan)

- Se basan parcialmente en una teoría general modificada por préstamos de otras teorías y por las exigencias computacionales

Los principales problemas son:

- morfológicos
- ambigüedad léxica
- ambigüedad estructural
- resolución de anáforas
- ambigüedad en el alcance de los cuantificadores

1.5 Principales problemas de la TA

Morfológicos:

• El análisis morfológico es un instrumento para resolver problemas de:

- análisis y generación sintáctica
- " " léxica
- " " semántica

(Revisar aspectos y problemas del análisis morfológico la primera parte del curso)

1.5 Principales problemas de la TA

Ambigüedad léxica:

• Se presenta cuando una palabra puede tener más de una interpretación

• Pueden ser de tres tipos:

- categoriales

- homógrafos y polisemias

- de transferencia o de traducción

1.5 Principales problemas de la TA

Ambigüedad léxica categorial:

• Posibilidad de asignar a una palabra más de una categoría gramatical o sintáctica (p.ej. sustantivo, verbo o adjetivo) dependiendo del contexto

Ejemplo:	*vino* (sustantivo o verbo)
	como (verbo o adverbio conjuntivo)

• A menudo puede resolverse:

- atendiendo a la flexión morfológica
- mediante el análisis sintáctico

1.5 Principales problemas de la TA

Ambigüedad léxica categorial:

Ejemplo: *Gas pump prices rose last time oil stocks fell*

El precio del gas subió la ultima vez que bajaron las reservas del petroleo

- Cada palabra de la oración en inglés tiene al menos una ambigüedad categorial (sustantivo o verbo)

- l*ast* puede ser: sustantivo, verbo, adjetivo y adverbio

- Solo existe un modo de analizar correctamente esta oración y requiere de un análisis sintáctico “intenso”

1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

• Se da cuando una palabra tiene dos o más significados diferentes posibles

Conceptos relacionados:

• Dos o más palabras son **homógrafas** si tienen la misma forma escrita y significados diferentes (*banco, gato*)

• Dos o más palabras son **homófonas** si se pronuncian igual tienen significados diferentes y se escriben diferente (*vaca,baca*)

- irrelevante para TA de textos escritos, relevante para la TA de habla

1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

• Dos o más palabras son **polisémicas** si muestran una variedad de significados relacionados de algún modo entre sí

• En la práctica las diferencias entre palabras homógrafas y polisémicas son difíciles de tratar

Ejemplo: *ear* (*oreja*), *ear of corn* (*espiga de cereal*)

puede considerarse un caso de polisemia (similitud física)

son homógrafas tienen diferentes derivaciones (auris, acus)

1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

• La homografía y la polisemia pueden recibir el mismo tratamiento en un análisis de TA

Métodos para eliminar este tipo de ambigüedad léxica:

• Los homógrafos de diferentes categorías sintácticas pueden resolverse como un ambigüedad léxica categorial

• Con los homógrafos de la misma categoría hay que recurrir a información semántica

1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

Homografos de la misma categoría sintáctica:

• Se identifica el tipo o dominio del texto y se selecciona la acepción del diccionario que más se adecue a dicho tipo (información contextual)

• Uso de un diccionario en el que se asignan **rasgos semánticos** (“humano”, “femenino”, “líquido”, etc.) y se especifican qué rasgos son compatibles en determinadas construcciones sintácticas mediante **restricciones de selección**

Ejemplo: *beber* requiere un sujeto animado

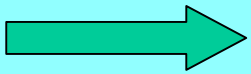
1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

Existen dificultades para:

• Determinar un grupo de rasgos semánticos que puedan aplicarse siempre

• Especificar las restricciones de selección de sustantivos y verbos en función de dichos rasgos



los rasgos semánticos no pueden resolver todos estos problemas de ambigüedad

1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

Ejemplo: homógrafo *ball* (*objeto esférico* | *fiesta de baile*)

Ambas acepciones se podrían distinguir en las siguientes oraciones con restricciones de selección apropiadas:

(1) *The ball rolled down the hill* (*La pelota rodó colina abajo*)

(2) *The ball lasted until midnight* (*El baile duró hasta la media noche*)

en (1) *rolled* requiere un objeto redondo como sujeto

en (2) *last* exige un sujeto con duración temporal

1.5 Principales problemas de la TA

Ambigüedad léxica: homografía y polisemia

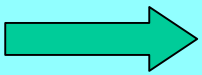
Ejemplo: homógrafo *ball* (*objeto esférico* | *fiesta de baile*)

Si una oración empieza con:

(3) *When you hold a ball ...*

hold es una palabra ambigua (*agarrar*|*organizar*)

puede referirse a cualquiera de las acepciones de *ball*



no se podrá desambigüar hasta que el resto de la oración proporcione más información lingüística o contextual

1.5 Principales problemas de la TA

Ambigüedad léxica de transferencia:

• Las ambigüedades categoriales, de homografía y de polisemia se producen en el análisis del texto en la lengua origen (ambigüedades monolingües)

• La ambigüedad de transferencia (de traducción) se produce cuando una palabra de la lengua origen puede traducirse a diversas palabras o expresiones en la lengua meta

• La ambigüedad no se produce con respecto a la lengua origen sino con respecto a la traducción (lo veremos más adelante)

1.5 Principales problemas de la TA

Ambigüedad estructural:

• Los problemas relativos a las estructuras y representaciones sintácticas de las oraciones

• Surge cuando la estructura profunda de una oración puede analizarse de más de un modo según esté definida la gramática empleada por el sistema

Tipos:

- Real: oraciones para las que una persona encontraría varias interpretaciones

- Accidental: si es la gramática del sistema de TA la que encuentra varias interpretaciones

1.5 Principales problemas de la TA

Ambigüedad estructural real:

Ejemplo: *The man saw the girl with the telescope*

• Un humano puede encontrar 2 interpretaciones:

- *The man saw the girl who posessed the telescope*
  El hombre vio a la niña que tenía el telescopio
- *The man saw the girl with the aid of the telescope*
  El hombre vio a la niña con la ayuda del telescopio

• Si la oración aparece en un relato, se podría deducir la interpretación de la línea argumental (un humano)

• Los sistemas de TA evalúan el contexto de una manera muy limitada

1.5 Principales problemas de la TA

Ambigüedad estructural accidental:

• Un sistema de TA no distingue entre ambigüedades reales o accidentales

• Se deben a :

- combinación accidental de palabras que tienen ambigüedades categoriales

- usos gramaticales alternativos de los constituyentes sintácticos (sintagma que puede modificar a varios elementos)

- distintas combinaciones posibles de los constituyentes

1.5 Principales problemas de la TA

Ambigüedad estructural:

- Difiere de una lengua a otra
- La accidental difiere dentro de una lengua de una gramática a otra

Resolución de la ambigüedad estructural:

- Escoger una de las posibles interpretaciones de una oración
- La traducción a la lengua meta puede variar según la interpretación escogida en la lengua origen

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural:

- uso de conocimiento lingüístico
- uso de conocimiento contextual
- uso del conocimiento del “mundo real”
- otras estrategias

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante el uso de conocimiento lingüístico:

- Información relativa a las palabras y al modo en que éstas se combinan
- Se proporciona a los analizadores información sobre las **restricciones de coaparición** (indicadores sobre cómo la presencia de ciertos elementos en una estructura influye en la probabilidad de que aparezcan otros)

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante el uso de conocimiento lingüístico:

Ejemplos:

- **marcos de subcategorización** para verbos

indican qué tipo de complementos corresponden a un verbo determinado

(verbo *dar*, sujeto sustantivo “donante”, OD sustantivo cosa “dada”, OI sustantivo “receptor”)

- **rasgos semánticos** para los sustantivos:

(“donante”, “animado”, etc.)

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante el uso de conocimiento lingüístico:

Ejemplos:

- En la **gramática de casos** los complementos dependientes reciben nombres como “agente”, “paciente”, “instrumento”, “locativo”, ... etc.

> *Leí lo del accidente de aviación en Francia*
>
> *Leí lo del accidente de aviación en el periódico*

leer puede ser modificado por un sintagma proposicional que comienza por *en* si el sustantivo que sigue tiene el rasgo “legible”

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante el uso de conocimiento contextual:

- La mayoría de las ambigüedades estructurales se podrían resolver con información contextual
- Son escasos los sistemas de TA que lo utilizan por:

  - no hay reglas para definir dónde buscar la porción de conocimiento necesario

  - cuánto tiempo sería necesario almacenarla (vigencia del conocimiento extraído)

  - coste computacional asociado

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante el uso de conocimiento del "mundo real":

- Información relativa a los acontecimientos de la vida real descritos en las oraciones

> *El hombre vio el caballo con el telescopio*

con el telescopio modifica obligatoriamente a *vio* porque nuestro conocimiento del mundo nos dice que no puede ser de otra manera

- A veces resulta difícil distinguir entre conocimiento del "mundo real" y conocimiento lingüístico

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante el uso de conocimiento del "mundo real":

- Hoy resulta imposible en la práctica codificar e incorporar todo el conocimiento del mundo real necesario para resolver todas las posibles ambigüedades de un sistema concreto
- La complejidad del conocimiento del mundo real y de su manejo no lo permiten

1.5 Principales problemas de la TA

Resolución de la ambigüedad estructural mediante otras estrategias:

- Seleccionar la estructura más probable o usual (puede producir errores)
- Si es sistema de TA es interactivo puede pedir al usuario que seleccione él la interpretación
- Si las lenguas tienen una estructura y vocabulario semejantes se puede recurrir al ***free ride*** (pase gratuito), no resolver la ambigüedad porque puede mantenerse como tal en la lengua meta

1.5 Principales problemas de la TA

La resolución de anáforas:

Anáfora: una referencia indirecta a una entidad mencionada de forma explícita en otro lugar del texto

- Recursos lingüísticos para realizar una referencia indirecta:
  - pronombres (*él, ellos, lo*, etc.)
  - demostrativos (*esto, aquello*, etc.)
  - expresiones (*el último, el anterior*, etc.)
- El objeto al que se refiere la referencia indirecta se denomina **antecedente**
- En muchos casos es importante identificar el antecedente de la anáfora para traducir correctamente

1.5 Principales problemas de la TA

La resolución de anáforas:

Ejemplo: la lengua meta distingue el género de las referencia indirectas:

(1)- *The monkey ate the banana because it was hungry*
El mono comió el plátano porque estaba hambriento
(2)- *The monkey ate the banana because it was ripe*
El mono comió el plátano porque estaba maduro
(3)- *The monkey ate the banana because it was tea-time*
El mono comió el plátano porque era la hora de la merienda

Si la lengua meta es el alemán: los pronombres adoptan el mismo género que sus antecedentes, habrá que identificar:

(1) *it* (mono); (2) *it* (banana); (3) *it* (frase temporal sin antecedente)

1.5 Principales problemas de la TA

La resolución de anáforas:

- Se requiere el mismo tipo de conocimientos (lingüístico, contextual, mundo real) que para despejar otros tipos de ambigüedad
- Una anáfora puede considerarse un tipo de ambigüedad en el que el antecedente no se conoce con certeza
- El conocimiento lingüístico no siempre será suficiente

(1) *The soldiers shot at the women and some of them fell*

(2) *The soldiers shot at the women and some of them missed*

some of them : *algunos de ellos* o *algunas de ellas*

1.5 Principales problemas de la TA

Ambigüedad en el alcance de los cuantificadores:

- Se produce cuando el alcance de cuantificadores como (*algunos/as, todos/as, ninguno/a*) es impreciso
- En algunas lenguas (inglés, castellano) se recurre al fenómeno sintáctico "**elevación del cuantificador**" que intenta expresar el verdadero significado

No smoking seats are available on domestic flights

puede interpretarse:

(1) *There are no seats where you may smoke on domestic flights*

(2) *There are "no smoking" sections on domestic flights*

se requiere conocimiento contextual y del "mundo real"

1.5 Principales problemas de la TA

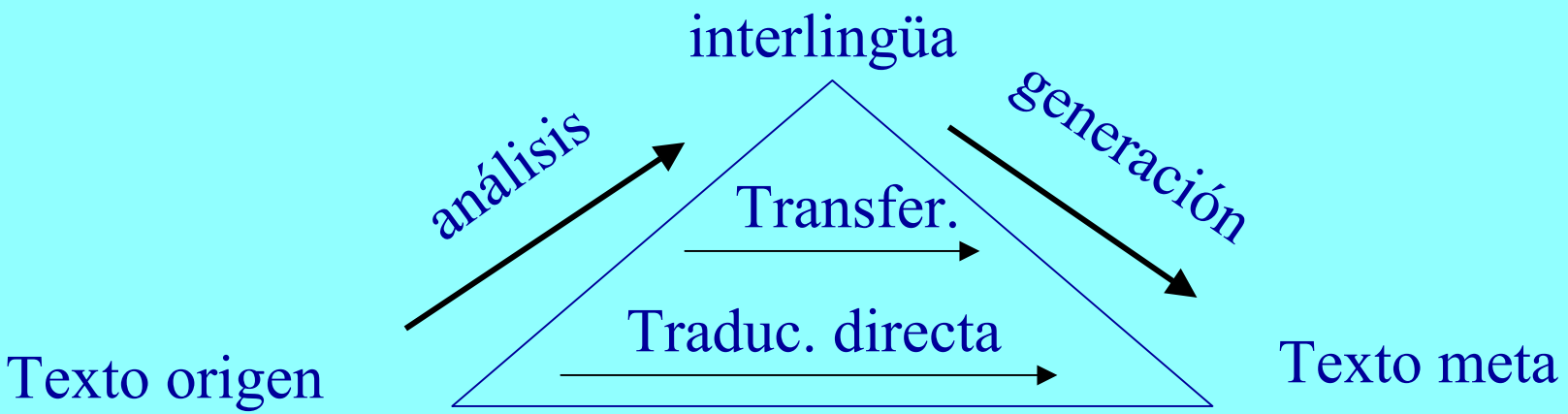
Ambigüedad en el alcance de los cuantificadores:

- Si existe el mismo tipo de ambigüedad en ambas lenguas, se podrá mantener en la lengua meta
- En caso contrario habrá que resolverla

1.6 Problemas de transferencia e interlingüa

- Los problemas vistos hasta ahora se referían fundamentalmente al análisis de la lengua origen (dificultades monolingües)
- Existen problemas con respecto al componente intermedio entre la lengua origen y meta

Diagrama de transferencia e interlingüa:



1.6 Problemas de transferencia e interlingüa

Sistemas de transferencia:

• Vamos a utilizar el siguiente ejemplo:

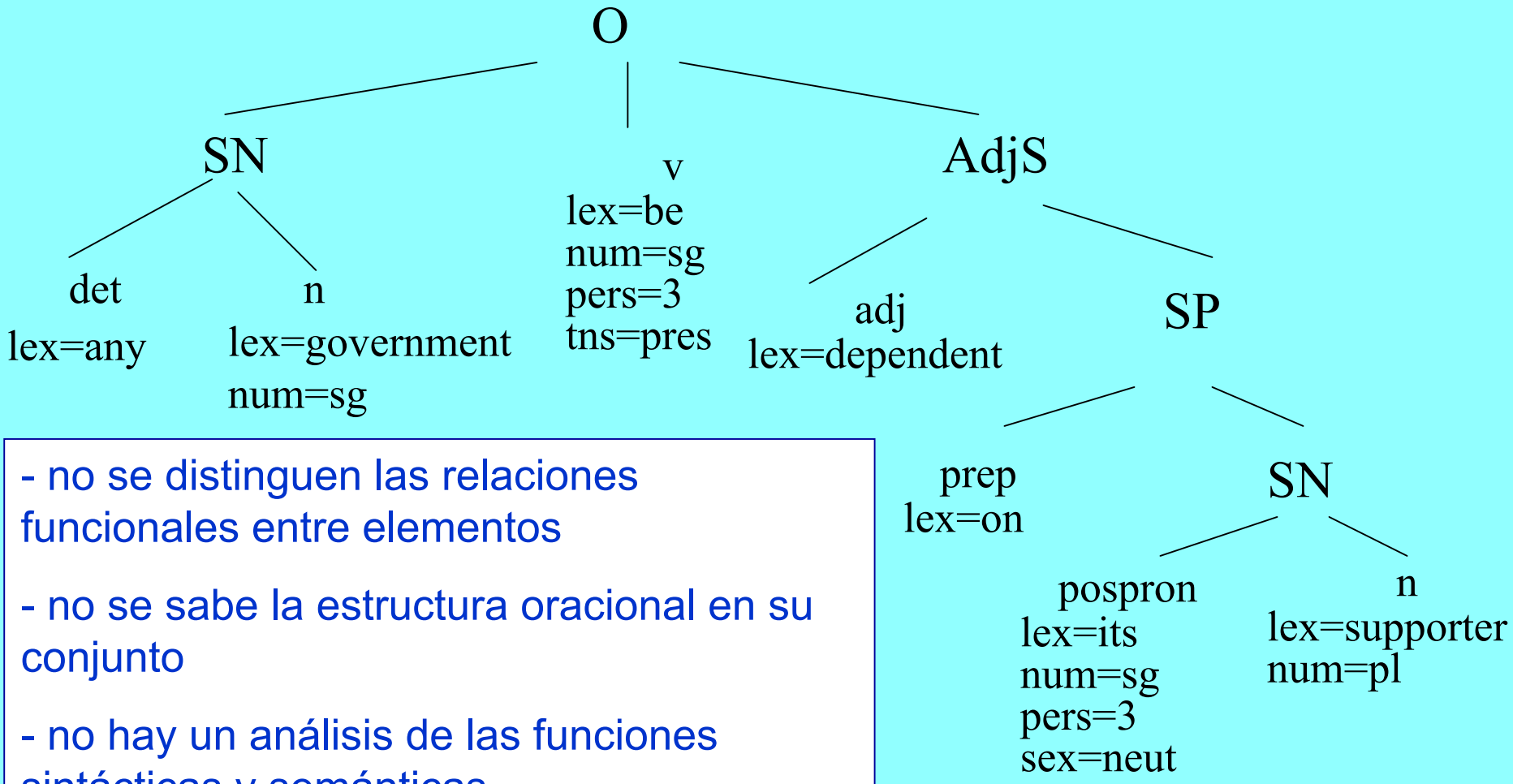
> *Any government is dependent upon its supporters*
>
> *(lit. Todo gobierno es dependiente de sus partidarios)*

y el francés como lengua meta

1.6 Problemas de transferencia e interlingüa

Sistemas de transferencia:

- Veamos una representación de la estructura superficial:



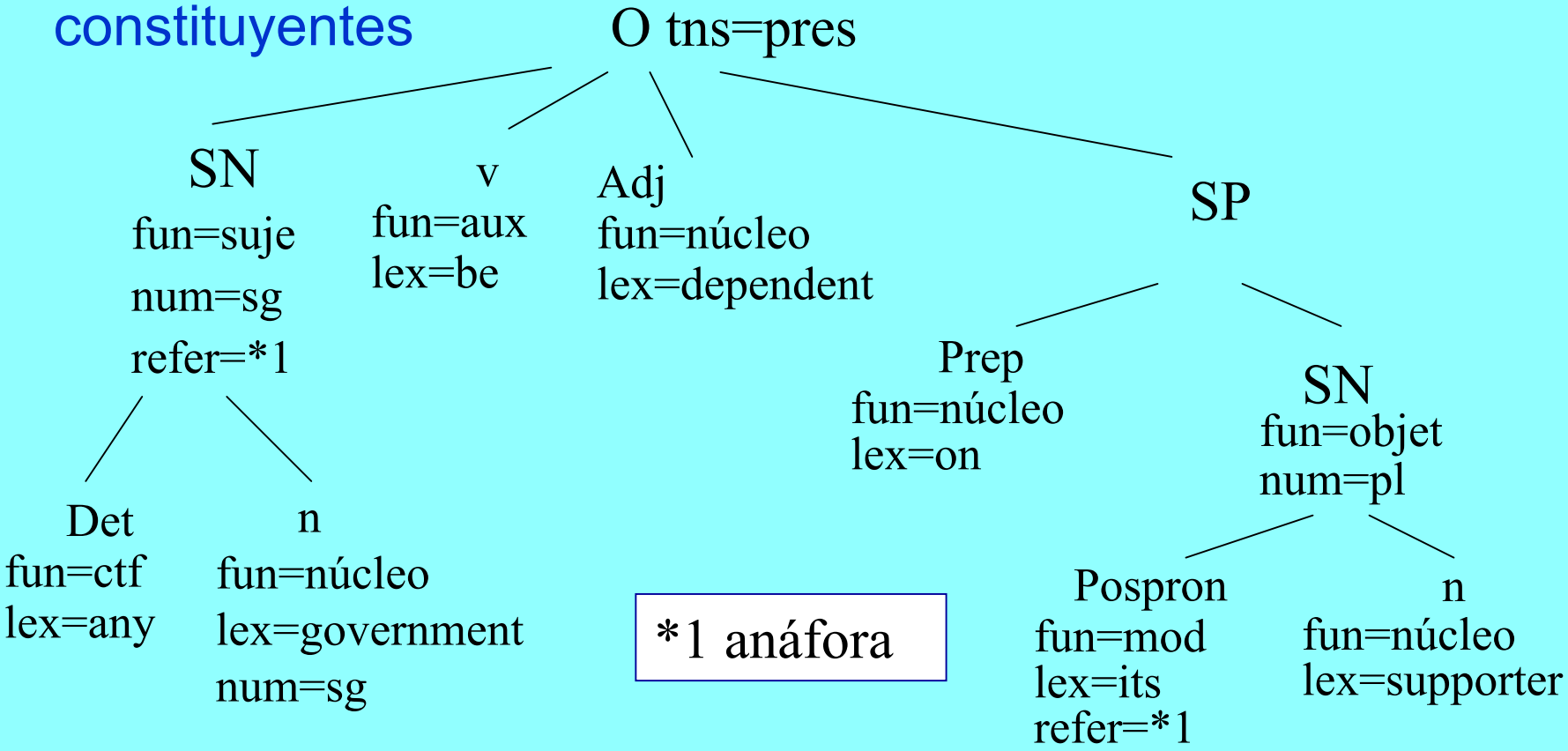
- no se distinguen las relaciones funcionales entre elementos

- no se sabe la estructura oracional en su conjunto

- no hay un análisis de las funciones sintácticas y semánticas

1.6 Problemas de transferencia e interlingüa

- Salvo para oraciones muy simples, la transferencia basada en un análisis superficial no será suficiente
- Será necesario identificar las funciones sintácticas de los constituyentes



1.6 Problemas de transferencia e interlingüa

- A continuación se llevarían a cabo:
  - la transferencia léxica
  - la transferencia estructural

Transferencia léxica: sustituir un componente léxico de la lengua origen por otro de la lengua meta

(ya se han comentado los problemas asociados)

1.6 Problemas de transferencia e interlingüa

Transferencia estructural: necesaria cuando la estructura "heredada" de la lengua origen es inapropiada para la lengua meta

- El objeto de profundizar en el análisis es neutralizar las diferencias entre las lenguas
- Las **reglas de transferencia** se encargan de construir la estructura en la lengua meta correspondiente a la entrada en la lengua origen

Jones likes the film

A Jones le gusta la película

Le film plaît a Jones

- deben tener en cuenta aspectos estilísticos
- problemas de cobertura

1.6 Problemas de transferencia e interlingüa

Transferencia con una interlingüa estructural

- En un tiempo se pensó que las estructuras “profundas” de la gramática generativo-transformacional podrían servir de representaciones interlingües
- Esas representaciones no neutralizan las idiosincrasias de las lenguas respectivas

Gramática de caso ≈ representación estructural interlingüe

- Los roles de caso (“roles semánticos”, “casos profundos”, “roles theta”) son relaciones semánticas tales como “agente”, “paciente”, “experimentador”, “instrumento”, ...

1.6 Problemas de transferencia e interlingüa

Transferencia con una interlingüa estructural

Gramática de caso ≈ representación estructural interlingüe

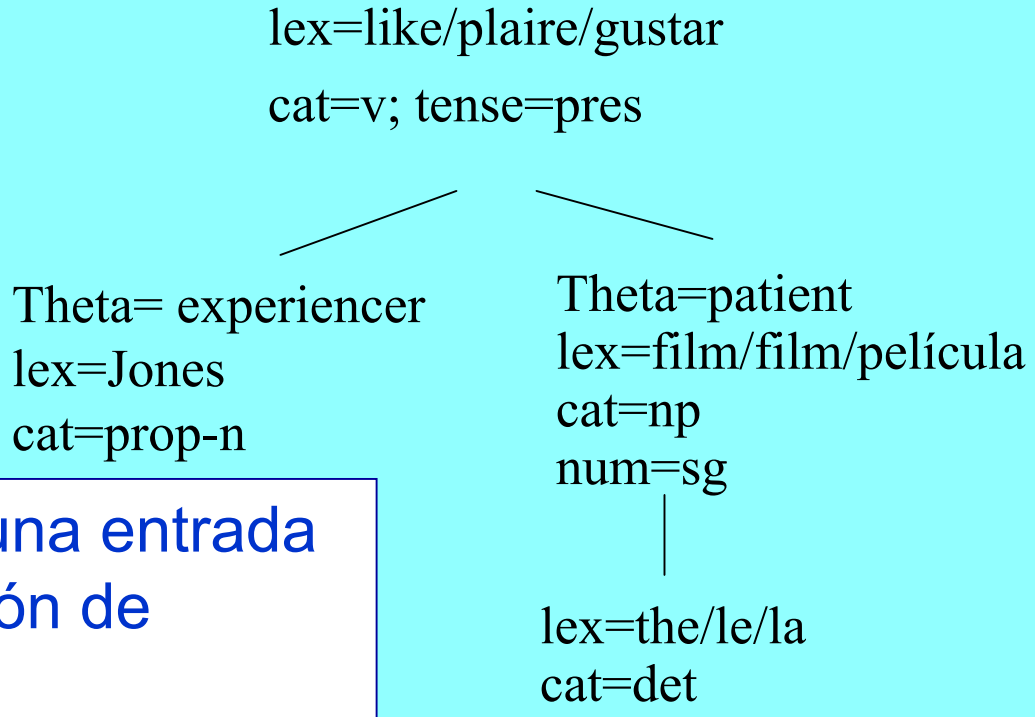
- Algunos problemas pocos complejos de la transferencia estructural desaparecen con este tipo de representación

Ejemplo:

Jones likes the film

A Jones le gusta la película

Le film plaît a Jones



Esta representación sería una entrada apropiada para la generación de francés o castellano

1.6 Problemas de transferencia e interlingüa

Transferencia con una interlingüa estructural

- Se suele suponer que las **representaciones de estructura de caso** reflejan los **universales de la sintaxis** que podrían considerarse interlingües
- Los sistemas de TA basados en transferencia emplean en muchos casos los roles de caso (sobre todo si una de las lenguas es el japonés)
- No se ha logrado un acuerdo unánime sobre un posible conjunto de relaciones de caso
- Los investigadores en TA que usan este enfoque deben elaborar su propio conjunto de roles de caso

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa

- El resultado del análisis en un sistema de TA interlingüe es una representación:
  - independiente de la lengua del texto origen
  - sirve como base para la generación del texto en una o varias lenguas meta
- Es necesaria la separación del análisis y la generación
  - no se puede orientar el análisis a una lengua meta en particular
  - no es posible volver al texto origen durante la generación
- La interlingüa debe representar “el significado” del texto

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa

- El problema que plantea el enfoque interlingüe abarca dos aspectos:
  - decidir cuál es la representación neutra más apropiada
  - establecer los procedimientos para extraer de los textos la información necesaria

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa: representación estructural

Expresión de las representaciones interlingües

- **Lógica proposicional**

Ejemplo: *Any government is dependent upon its supporters*

podría quedar:

all(X), government(X), indefinite(Y), plural(Y), support(Y,X,T), depend-on (X,Y,T), timeless(T)

- el problema no es llegar a este tipo de representación para las oraciones (aunque en algunas si lo podría ser)

- la complejidad es generar texto a partir de la representación (múltiples oraciones meta posibles)

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa: representación estructural

- Casi todos los sistemas basados en interlingüa utilizan representaciones cuya estructura es, en esencia, semejante a la de los sistemas basados en transferencia
- La diferencia radica en el carácter abstracto de la representación de la estructura y en el tratamiento de los términos léxicos
- La interlingüa debe representar:

  - las relaciones sintácticas

  - la función textual

  - el rol de caso o cualquier otra interpretación dictada por el orden de las palabras

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa: representación estructural

- **Representación interlingüe neutra**

He walked across the road

El atravesó la calle a pie

$$
\begin{bmatrix}
\text{Pred} & = \langle \text{MOTION} \rangle \\
\text{Tense} & = \text{past} \\
\text{Agent} & = \begin{bmatrix} \text{Pred} = \text{Pron} \\ \text{Num} = \text{sing} \\ \text{Pers} = 3 \\ \text{Sex} = \text{male} \end{bmatrix} \\
\text{Instr} & = \{\text{Pred} = \langle \text{FOOT} \rangle\} \\
\text{Loc} & = \begin{bmatrix} \text{Pred} = \langle \text{CROSS} \rangle \\ \text{Obj} = \{\text{Pred} = \langle \text{ROAD} \rangle\} \end{bmatrix}
\end{bmatrix}
$$

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa: representación léxica

- Definir las representaciones neutras para las unidades léxicas es aún más complejo que para las estructuras sintácticas
- Las representaciones de las unidades léxicas han de representar conceptos
- Cualquier distinción que se exprese o pueda expresarse léxicamente en las lenguas del sistema debe aparecer de modo explícito en la representación interlingüe

Ejemplo: si el sistema incluye el japonés, la interlingüa debería distinguir 8 diferentes conceptos de <llevar-puesto> aunque las otras lenguas del sistema no los usen

1.6 Problemas de transferencia e interlingüa

Sistemas basados en interlingüa: representación léxica

- Hay muy pocos sistemas que en la práctica distingan todos los posibles conceptos
- Recurren a información contextual o conocimiento del mundo real para elegir las alternativas de la traducción

  transferencia léxica con más o menos representaciones estructurales de interlingüa
- Algunas propuestas apuntan a utilizar el esperanto como interlingüa, pero

  ¿sería entonces un sistema interlingüa?

1.7 La generación

Generación o síntesis:

obtención de textos meta a partir de representaciones intermedias

Se distingue:

- generación en los programas directos
- generación en los programas indirectos
  - transferencia
  - interlingüa

1.7 La generación

Generación en los programas directos:

- En estos sistemas no existe una generación tal y como la hemos definido
- La fase de reordenamiento local puede considerarse como una mezcla de transferencia y generación
- La generación depende sobre todo de las estructuras en la lengua origen
- Solo se realizan los cambios necesarios para producir una ordenación aceptable de las palabras en la lengua meta

1.7 La generación

Generación en los programas de transferencia:

- La generación suele estar dividida en dos módulos:
  - generación sintáctica
  - generación morfológica
- En la **generación sintáctica** la representación intermedia (rdo. de análisis y transferencia) se asemeja a un árbol de estructura profunda (ver transparencia 89)
- Ese árbol se convierte mediante reglas transformacionales en un árbol ordenado de estructura superficial en la lengua meta
- En las hojas del nuevo árbol se añaden etiquetas adecuadas para los rasgos y las funciones gramaticales en la lengua meta

1.7 La generación

Generación en los programas de transferencia:

- La tarea principal de la **generación sintáctica** consiste en ordenar los constituyentes en una secuencia correcta en la lengua meta

Ejemplo: si a una oración se le coloca la etiqueta de pasiva en la estructura profunda, la generación sintáctica generará un nodo para el verbo auxiliar con una etiqueta con:

- información de tiempo apropiada

un nodo para el verbo principal con una etiqueta con:

- “participio”

1.7 La generación

Generación en los programas de transferencia:

- La estructura superficial resultante es el punto de partida para la generación morfológica
- La **generación morfológica** interpreta las cadenas de los elementos léxicos etiquetados para obtener como resultado las oraciones meta

perro "plural"	→	*perros*
go "past"	→	*went*

- Deberá tener en cuenta tanto los casos regulares como los casos particulares o irregularidades

1.7 La generación

Generación en los programas de interlingüa:

- Al igual que en el sistema de transferencia, la generación suele estar dividida en dos módulos:
  - generación sintáctica
  - generación morfológica
- La principal diferencia es que el punto de partida no es una representación sintáctica de la estructura profunda
- El punto de partida es una representación interlingüe, por ejemplo, estructuras de predicado-argumento (transparencia 99)

1.7 La generación

Generación en los programas de interlingüa:

- A partir de la representación interlingüe se genera la estructura sintáctica profunda en una fase denominada **generación semántica**
- A continuación se suceden las fases de generación sintáctica y generación morfológica al igual que en el sistema de trasferencia

1.8 Principales programas comerciales de TA

- **TRADOS**

  Gama de productos:

  - gestión terminológica, **MultiTerm**
  - memorias de traducción, **Translation Workbench**

  Tiene contratos con:

  - Microsoft (una integración de estas herramientas en los entornos futuros de su caja de herramientas ofimáticas *Office*)
  - La Comisión Europea

  Demo:

`http://www.trados.com/products/download.asp`

1.8 Principales programas comerciales de TA

- **SYSTRAN**

  - Sistema de TA directo evolucionado hacia la transferencia

  - Primera versión: 1960 traducción del ruso al inglés para las fuerzas aéreas americanas

  - La Comisión lo compró en 1976 y en la actualidad se están desarrollando 16 pares de lenguas nuevos

  - Los diccionarios del sistema se han ido llenando con terminología propia de la Comunidad.

  - La versión que con el paso de los años se ha desarrollado en la Comisión es muy distinta de la que se comercializa en California por la casa matriz.

    `http://www.systranet.com/systran/net`

1.8 Principales programas comerciales de TA

- **METAL**

  - Sistema de TA basado en transferencia

  - Comenzó en Texas y fue adquirido por la empresa Siemens

  - Traducciones bidireccionales inglés-francés-castellano-alemán (futura inclusión lenguas asiáticas)

http://www.sail-labs.de/engl/index.htm

1.8 Principales programas comerciales de TA

- **GLOBALINK**

  - Esta empresa ha emprendido una ambiciosa campaña de adquisición de productos, entre los que destaca la colección **Language Assistant**.

  - Traducción por transferencia entre inglés y francés, alemán, italiano, castellano y portugués

  - No destaca precisamente por su calidad, pero ha alcanzado un considerable éxito entre los usuarios de WINDOWS

http://buymebuyme.com/product/translator.shtml#ProductInfoSection

1.8 Principales programas comerciales de TA

- **LOGOS**

  - Empresa americana que comenzó ofreciendo traducción del vietnamita al inglés

  - Los pares ahora incluyen alemán al inglés y francés, e inglés al francés, alemán y castellano

  - En la página que se indica se puede acceder a numerosos recursos bilingües

http://www.logos.it/owa-wt/html_logos.home?lang=en