



Bevezetés a nyelvtechnológiába

9.

Gépi fordítás

Ezt olvassuk a gépi fordításról...

- ❑ A gépi fordítással foglalkozni kidobott idő, mert egy gép sosem fog Shakespeare-t fordítani.
- ❑ Állítólag volt egyszer egy gépi fordító rendszer, amelyik arra az angol mondatra, hogy *The spirit is willing, but the flesh is weak*, azt fordította oroszra, hogy: *A vodka jó, de a hús gyenge*. Ebből jól látszik, hogy a gépi fordítás haszontalan dolog.
- ❑ Általánosságban, a gépi fordítás minősége annyira gyenge, hogy semmire sem használható a gyakorlatban.
- ❑ A gépi fordítás elveszi a fordítók munkáját.
- ❑ A japánok kifejlesztettek egy rendszert, amihez csak bele kell beszélni a telefonba, és lefordítja, amit mondunk, és ezt a másik angolul hallja.
- ❑ Van egy dél-amerikai indián nyelv, amelyiknek a szerkezete logikailag annyira tökéletes, hogy nagyon könnyen lehet gépi fordító rendszert csinálni a segítségével.

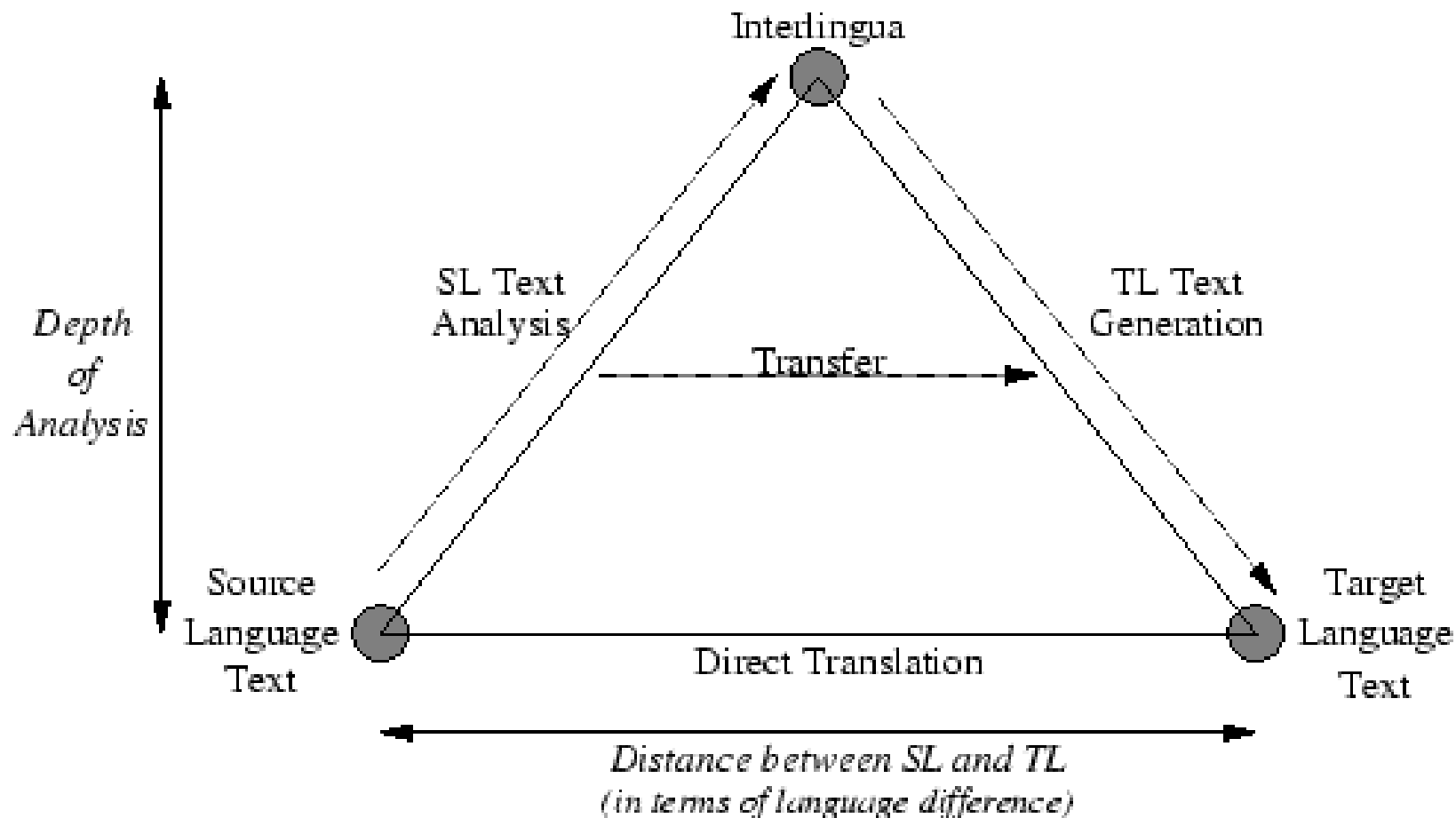
Ez igaz a gépi fordításról...

- ☐ Igaz, hogy a minőség még sokszor hagy kívánnivalót, de a sebesség igen értékes tulajdonsága a GF-rendszereknek.
- ☐ Bizonyos körülmények közt a GF minőségi fordítást is képes produkálni: a METEO 4%-nál kevesebb kimenete igényel emberi korrekciót.
- ☐ A GF hasznos dolog. A fenti METEO rendszer például napi használatban van 1977 óta. 1990 óta 45.000 szót fordít gyakorlatilag minden nap.
- ☐ A GF nem fenyegeti a fordítók munkáját.
- ☐ A beszéd-beszéd GF továbbra is kutatási téma. Általánosságban is igaz: sok kutatási téma van még a GF területén.
- ☐ A GF-rendszerek fejlesztése időigényes munka. A gyakorlatban egy GF-rendszer csak akkor tud megvalósulni, ha a felhasználó komoly munkát fektet bele az egyedi felépítésébe.
- ☐ A statisztikai és a neurális rendszerek készítéséhez nem kell nyelvtudás (csak a teszteléséhez)



Szabályalapú fordítás

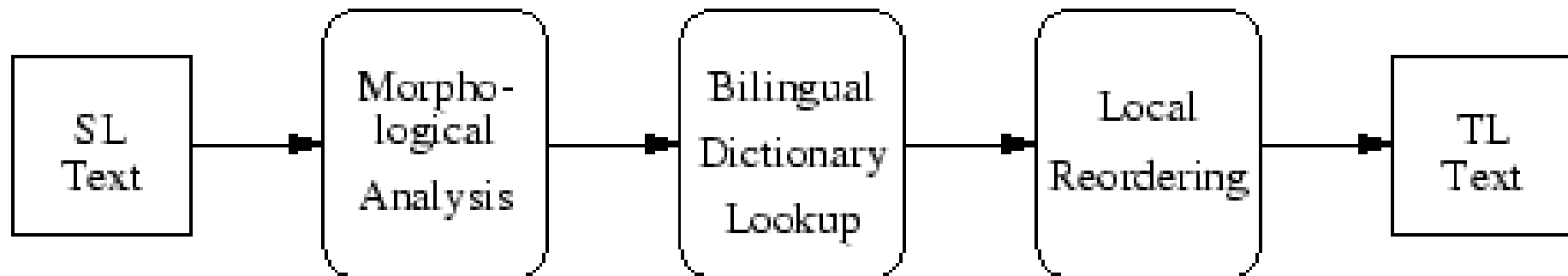
Szabályalapú fordítási stratégiák





Közvetlen fordítás

Közvetlen fordítás



A közvetlen fordítás lépései

Stage	Action
1.	morphological analysis
2.	lexical transfer of content words
3.	various work relating to prepositions
4.	SVO rearrangements
5.	miscellany
6.	morphological generation

Input:	watashihatsukuenouenopenwojonniageta.
After stage 1:	watashi ha tsukue no ue no pen wo jon ni ageru PAST.
After stage 2:	I ha desk no ue no pen wo John ni give PAST.
After stage 3:	I ha pen on desk wo John to give PAST.
After stage 4:	I give PAST pen on desk John to.
After stage 5:	I give PAST the pen on the desk to John.
After stage 6:	I gave the pen on the desk to John.



METEO

METRO TORONTO.

LE GRAND TORONTO.

TODAY:

MAINLY CLOUDY AND COLD WITH
OCCASIONAL FLURRIES.
BRISK WESTERLY WINDS TO 50 KM/H.
HIGH NEAR MINUS 7.

AUJOURD HUI:

GENERALEMENT NUAGEUX ET FROID
AVEC QUELQUES AVERSES DE NIEGE.
VENTS VIFS D'OUEST A 50 KM/H.
MAXIMUM D'ENVIRON MOINS 7.

TONIGHT:

VARIABLE CLOUDINESS.
ISOLATED FLURRIES.
DIMINISHING WINDS.
LOW NEAR MINUS 15.

CETTE NUIT:

CIEL VARIABLE.
AVERSES DE NIEGE EPARSEES.
AFFAIBLISSEMENT DES VENTS.
MINIMUM D'ENVIRON MOINS 15.

FRIDAY:

VARIABLE CLOUDINESS.
HIGH NEAR MINUS 6.

VENDREDI:

CIEL VARIABLE.
MAXIMUM D'ENVIRON MOINS 6.

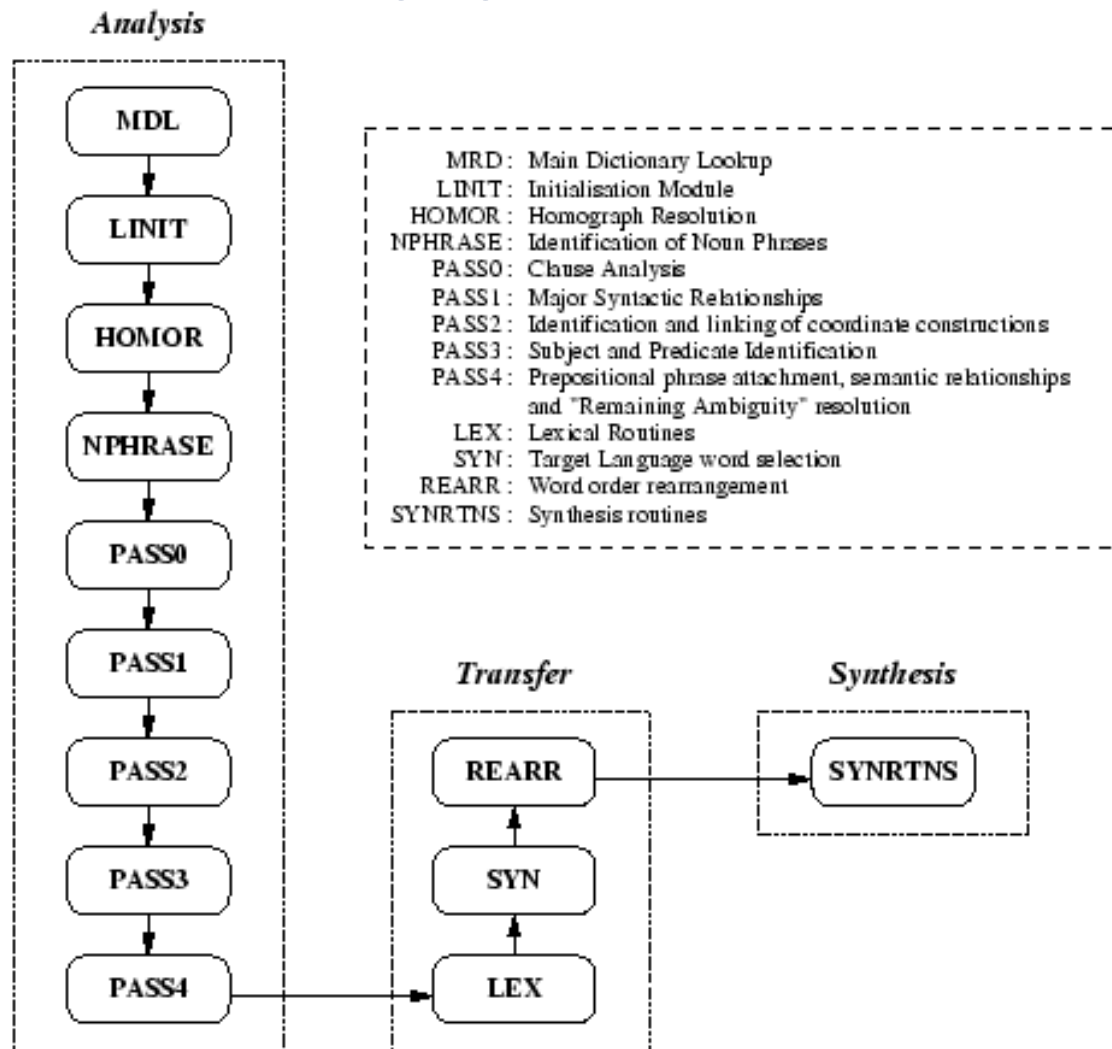
A közvetlen fordítás jellemzői

- ☐ Soha nincs kész semmilyen állapot sem, csak a végén
- ☐ Rengeteg lépésből áll
- ☐ Minden lépés egy adott nyelvi jelenséget dolgoz fel
- ☐ A legfontosabb eszköz a kétnyelvű szótár
- ☐ Komoly probléma a helyes kimeneti szórend és a megfelelő toldalékolás előállítása
- ☐ Ezt a feladatot átrendezésnek hívják

A közvetlen fordítás modulárisává válása

- ❑ Átrendezés: vonzatjellegű információk és általános szórendi kérdések „keveréke”
- ❑ Az átrendezésre jellemző a modularitás teljes hiánya
- ❑ Generálás: a célnyelv lexikális elemeinek morfológiailag megfelelő szósorozata mint kimenet
- ❑ A Systran és a Meteo ilyen átrendezéseit írták átkésőbb transzferrendszerre
- ❑ Az átrendezési műveletek tehát a később megjelenő transzfer és a generálás kombinációjának is tekinthetők

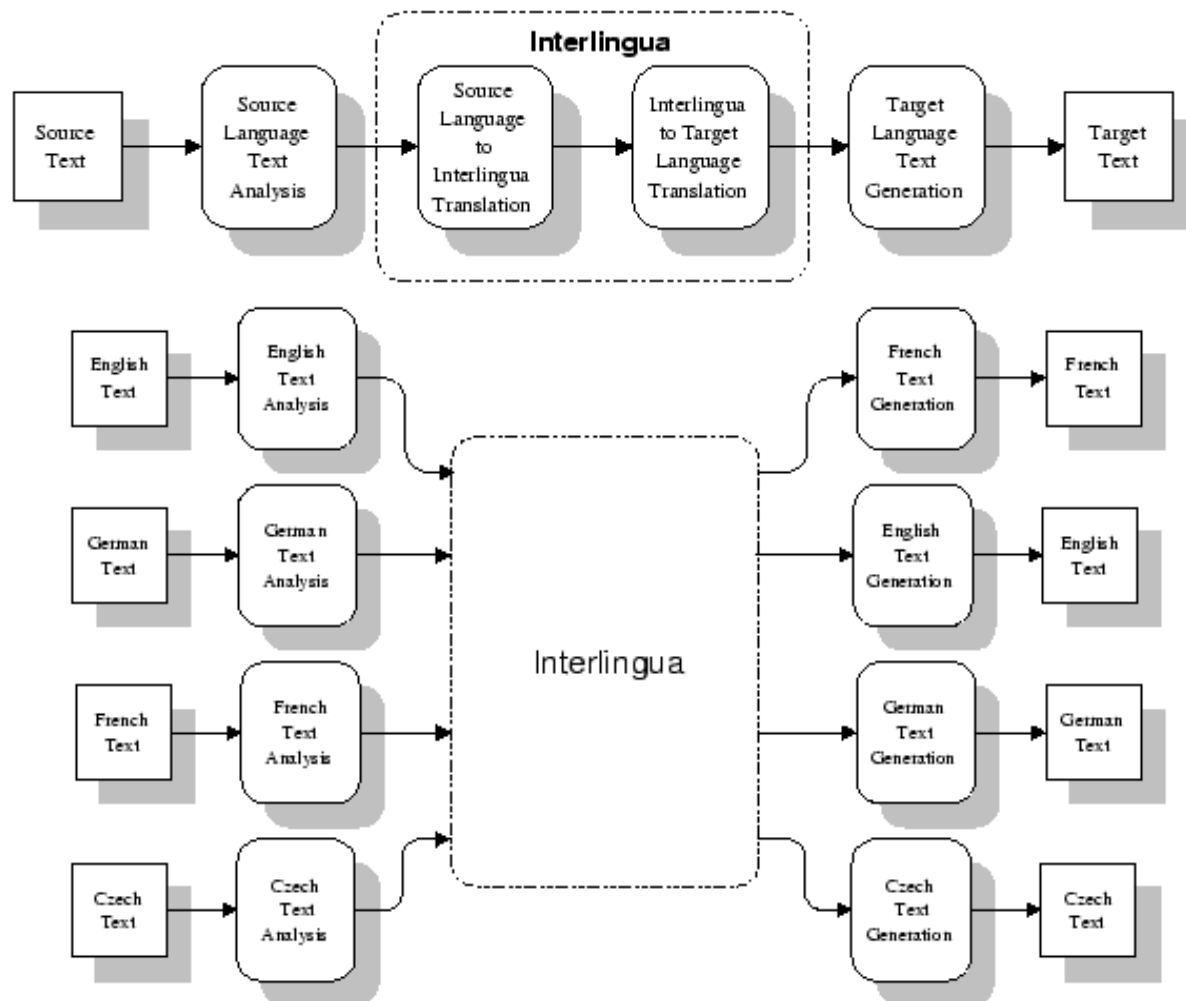
SYSTRAN





Közvetítőnyelves fordítás

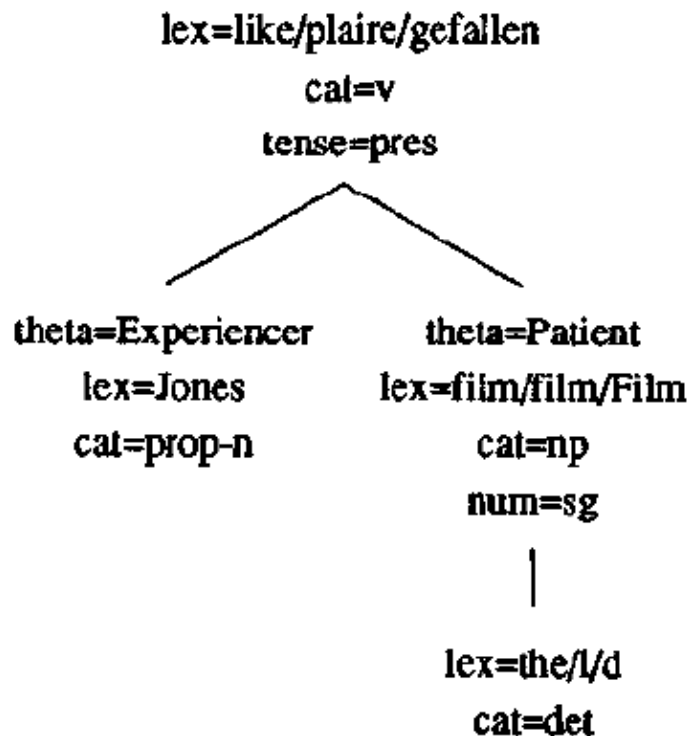
A közvetítőnyelves fordítás



A közvetítőnyelves fordítás jellemzői

- ☐ Nem létezik mindent kielégítő közvetítőnyelv
- ☐ Kísérletek: a logikai formalizmusoktól a formalizált eszperantóig
- ☐ A soknyelvű rendszerek esetén nagyon „vonzó”
- ☐ $n > 3$ esetén: $n \cdot (n-1) > 2 \cdot n$
- ☐ Végül is a két fordítási lépés nem más, mint egy-egy közvetlen fordítás

Példa a közvetítőnyelves fordításra

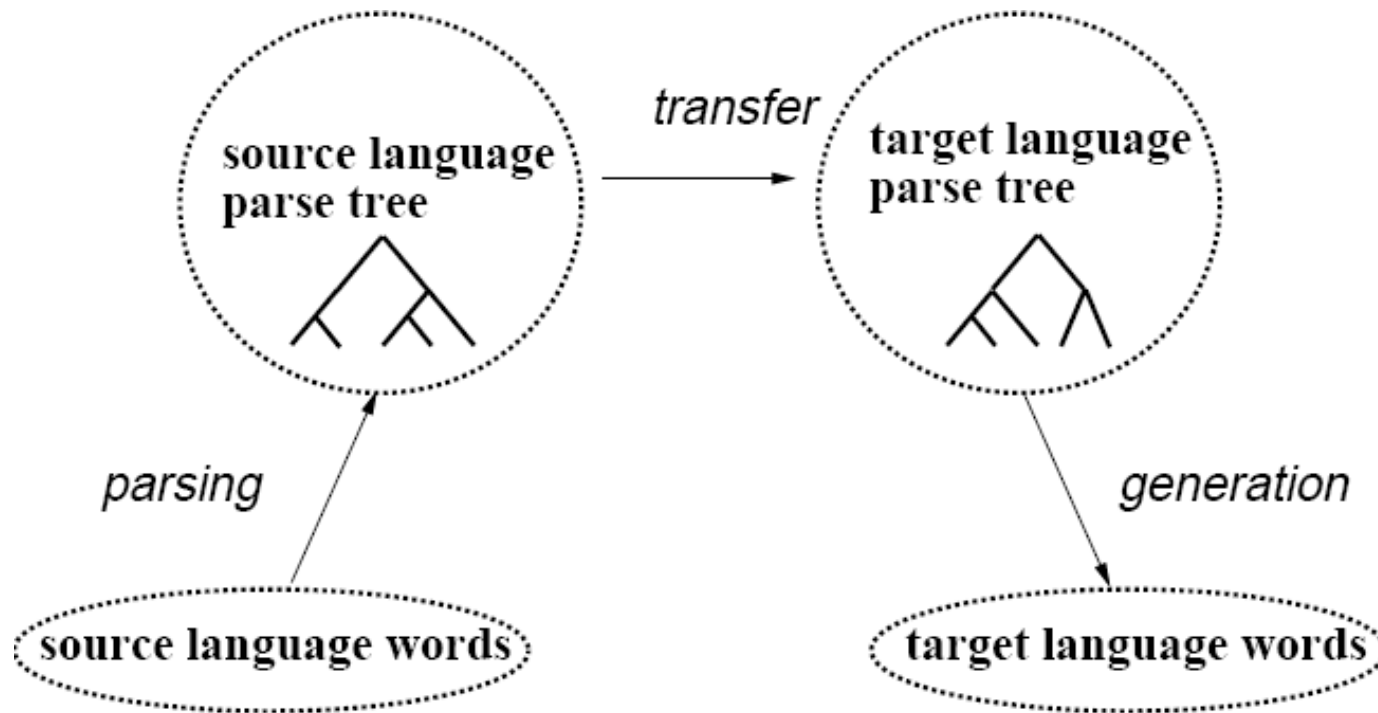


Jones likes the film. Le film plaît à Jones. Das Film gefällt Jones.



Transzfer fordítás

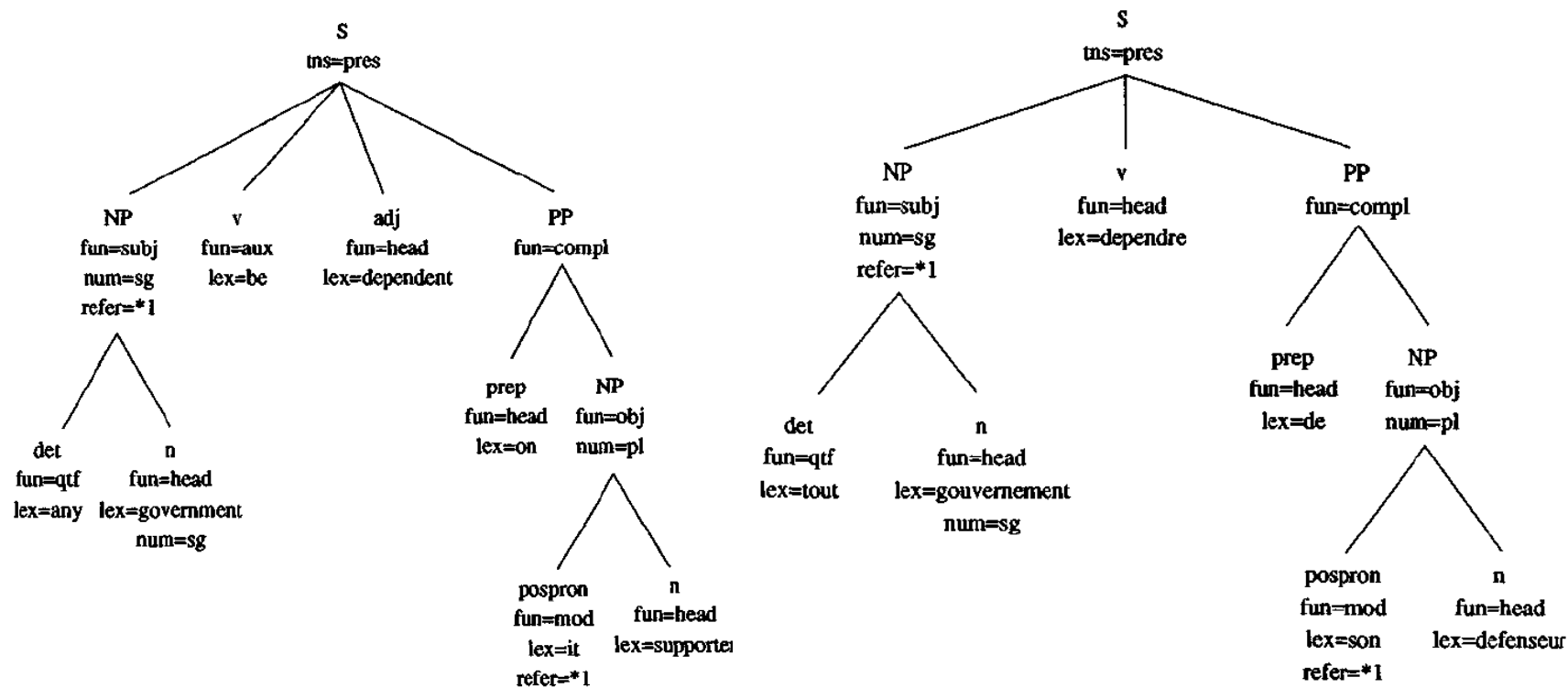
A transzfer fordítás menete



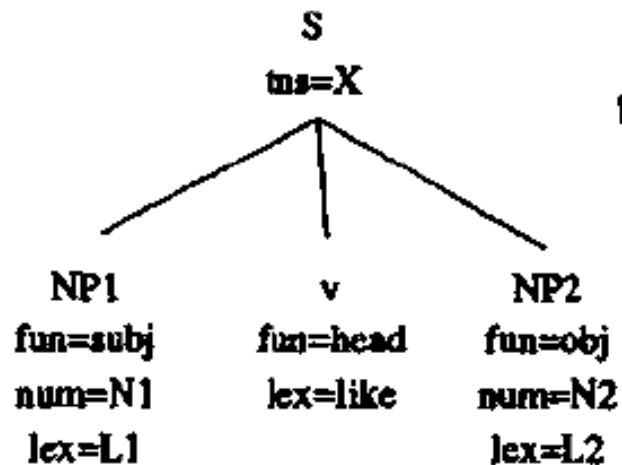
A transzfer fordítás menete és típusai

- ☐ Morfológiai elemzés
 - ☐ Egyértelműsítés (POS, WSD)
 - ☐ Lexikális transzfer: kétnyelvű szótárakkal
 - ☐ Szerkezeti transzfer: frázisok, csonkok
 - ☐ Morfológiai generálás
-
- ☐ Szintaktikai/felszíni transzfer: hasonló struktúrájú, különösen rokon nyelvek között
 - ☐ Szemantikus/mély transzfer: távolabbi rokonságban lévő nyelvek között

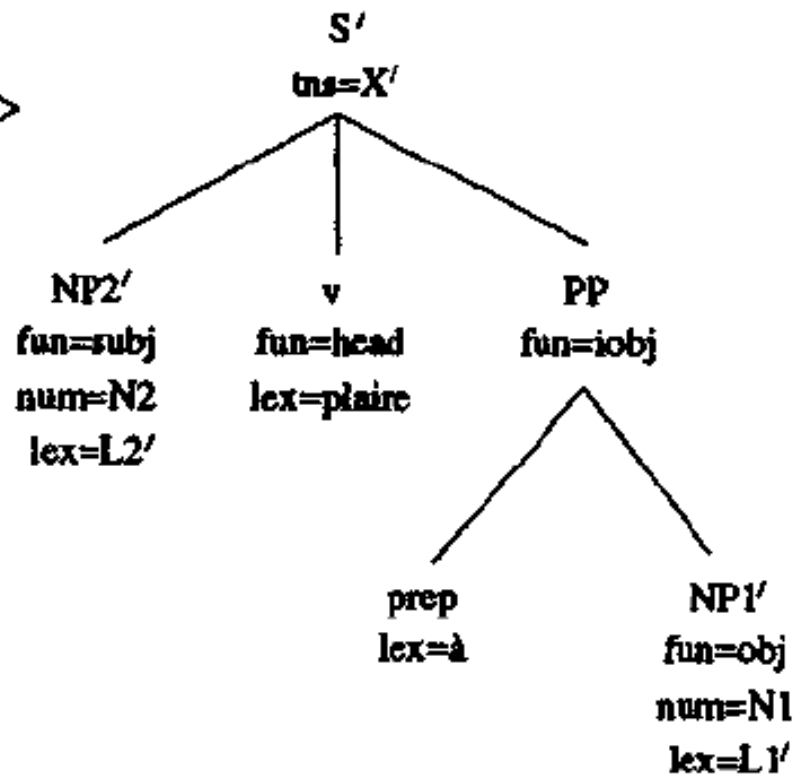
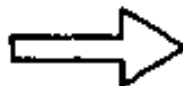
Példa a lexikális transzferre



Példa a szerkezeti transzferre



Jones likes the film.

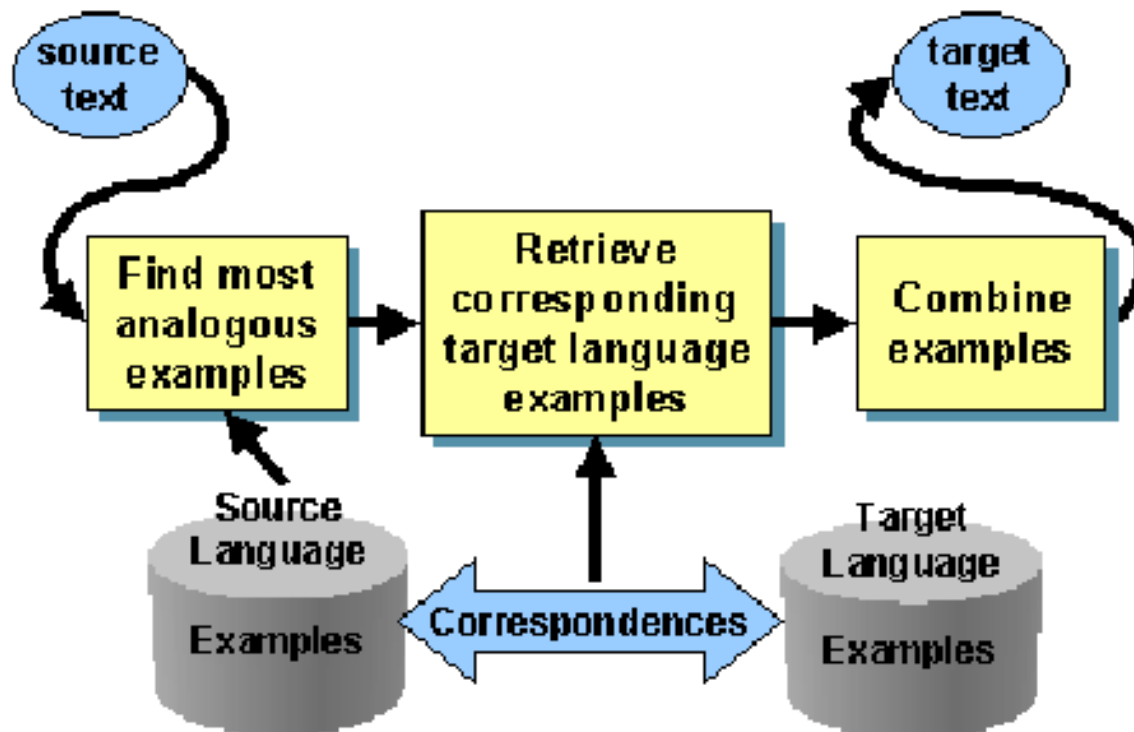


Le film plaît à Jones.



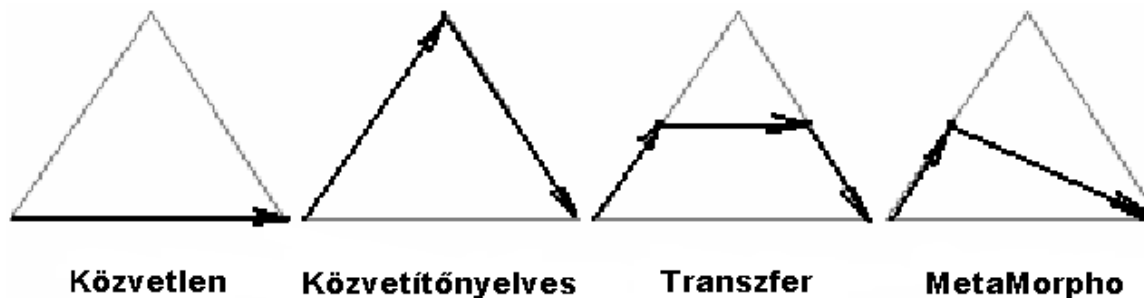
Más - de nem statisztikai - megoldások

Példa-alapú fordítás



MetaMorpho: egy hibrid szabály/példa-alapú fordító

- ❑ Nincs külön szótár és külön nyelvtan
- ❑ Csak minta-párok vannak: bemenet/interpretáció szerkezet-párok
- ❑ Egyetlen elemzési menet: nincs rákövetkező művelet (pl. transzfer)
- ❑ Célszerkezet-generálás: az elemzés „melléktermékeként”
- ❑ Új:





A gépi fordítás néhány tipikus nehézségéről



Lexikai többértelműségek kezelése

One minute has sixty seconds.

There is only minute difference between these pictures.

Part of the job of a secretary is to minute meetings.



PP-kapcsolások kezelése

The man saw the girl with the telescope.



Strukturális többértelműségek kezelése

Time flies like an arrow.



Többszavas kifejezések kezelése

Flying planes can be dangerous.



Idiómák kezelése

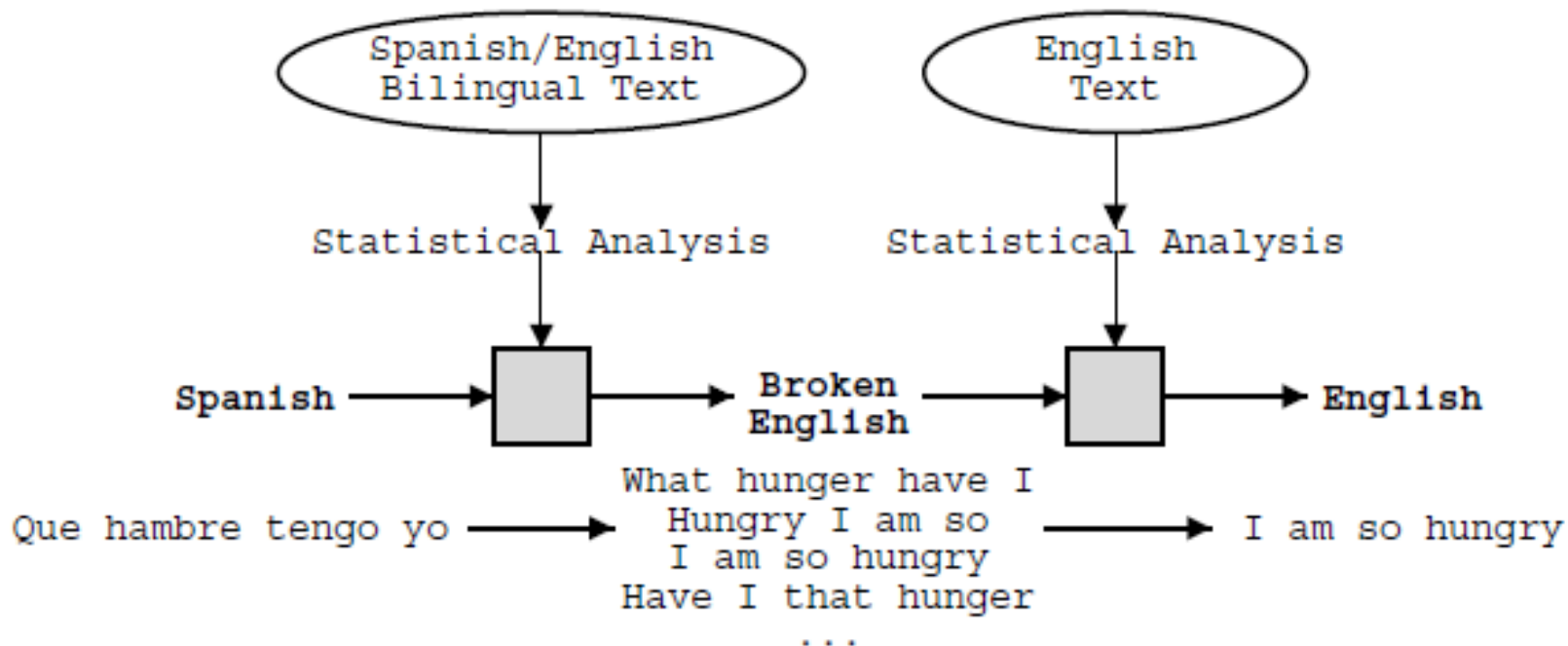
The country has gone to the dogs.



Statisztikai fordítás

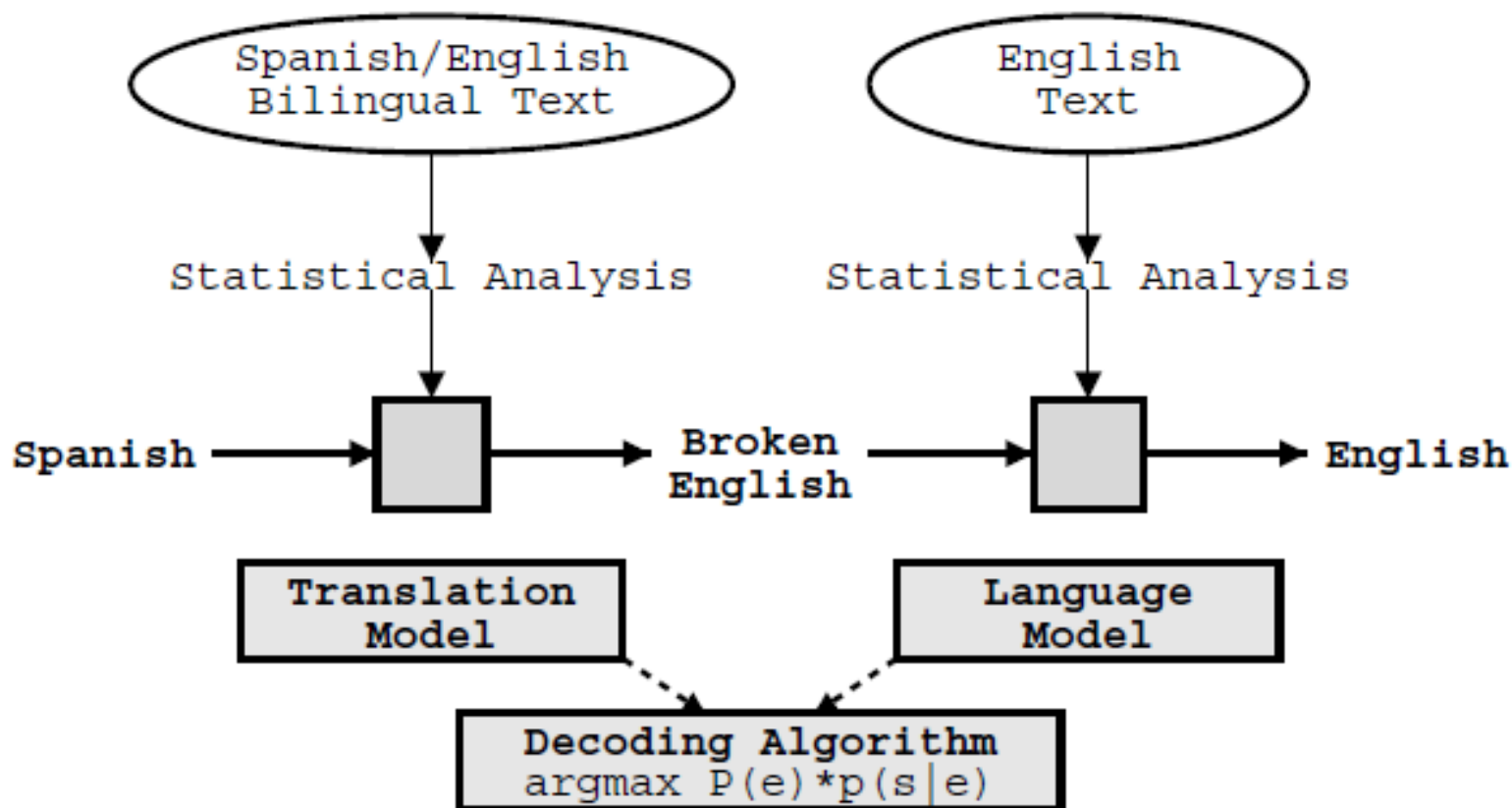
A statisztikai fordítás alapjai

(pl. spanyolról angolra fordítás)



A statisztikai fordítás alapjai (2)

(pl. spanyolról angolra fordítás)



A három fő modul

☐ Nyelvmodell:

adott egy e angol sztring, hozzá kell rendelni egy $P(e)$ -t

jó angol sztring: magas $P(e)$

rossz angol sztring: alacsony $P(e)$

☐ Fordítási modell:

adott egy $\langle f, e \rangle$ sztringpár, a hozzárendelt $P(f|e)$ valószínűséggel

ha $\langle f, e \rangle$ egymás fordítása: magas $P(f|e)$

ha $\langle f, e \rangle$ nem egymás fordítása: alacsony $P(f|e)$

☐ Dekódoló algoritmus:

nyelvmodell + fordítási modell + bemenő f mondat, amikhez meg kell találni az e fordítást, a $P(e) \cdot P(f|e)$ maximalizálásával

A statisztikai GF lényege

- ❑ A naiv Bayes-tétel:

$$\tilde{e} = \arg \max_{e \in E^*} p(e|f) = \arg \max_{e \in E^*} p(f|e)p(e)$$

- ❑ Az ötlet a beszédfelismerésből jön
- ❑ A legtöbb „minőségi” probléma a korpusz méretéből adódik
- ❑ Csak egy fordítási modell
(feltételes valószínűségi rész) és
egy nyelvmmodell
(teljes valószínűségi rész) kell,
meg a dekóder, ami számol

A fordítási modell

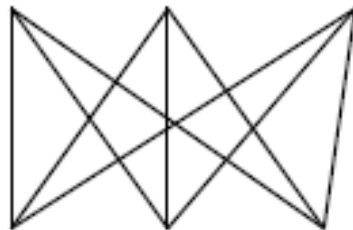
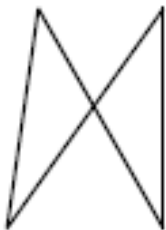
- ☐ A fordítási modell feladata megtalálni a legjobban illeszkedő bemenetet az e (angol) kimenethez
- ☐ $P(f|e)$ tanulását párhuzamos korpuszból lehet csak megoldani
- ☐ Tipikus helyzet: sokszor nincs elég adat a $P(f|e)$ közvetlen becslésére

Hogyan hozunk létre fordítási modellt?

- ☐ EM-algoritmus (expectation maximization)
- ☐ Inicializálás: minden kapcsolat azonos súlyú
- ☐ Valószínűség hozzárendelése a hiányzó adatokhoz
- ☐ Paraméterbecslés a teljes adatokból
- ☐ Iterálás

Inicializálás az EM algoritmusban

... la maison ... la maison blue ... la fleur ...

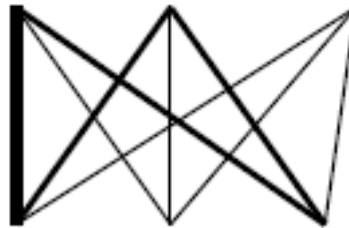


... the house ... the blue house ... the flower ...

Kezdőlépés: minden kapcsolat egyenlő valószínűségű

Iteráció az EM algoritmusban

... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

Egy iterációs lépés után a modell megtanulja, hogy bizonyos kapcsolatok gyakoribbak, tehát valószínűbbek (pl. a *la* és a *the* közötti)

További iterálás az EM algoritmusban

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

... la maison ... la maison bleu ... la fleur ...



... the house ... | the blue house ... the flower ...

Paraméterbecslés az EM algoritmusban

... la maison ... la maison bleu ... la fleur ...

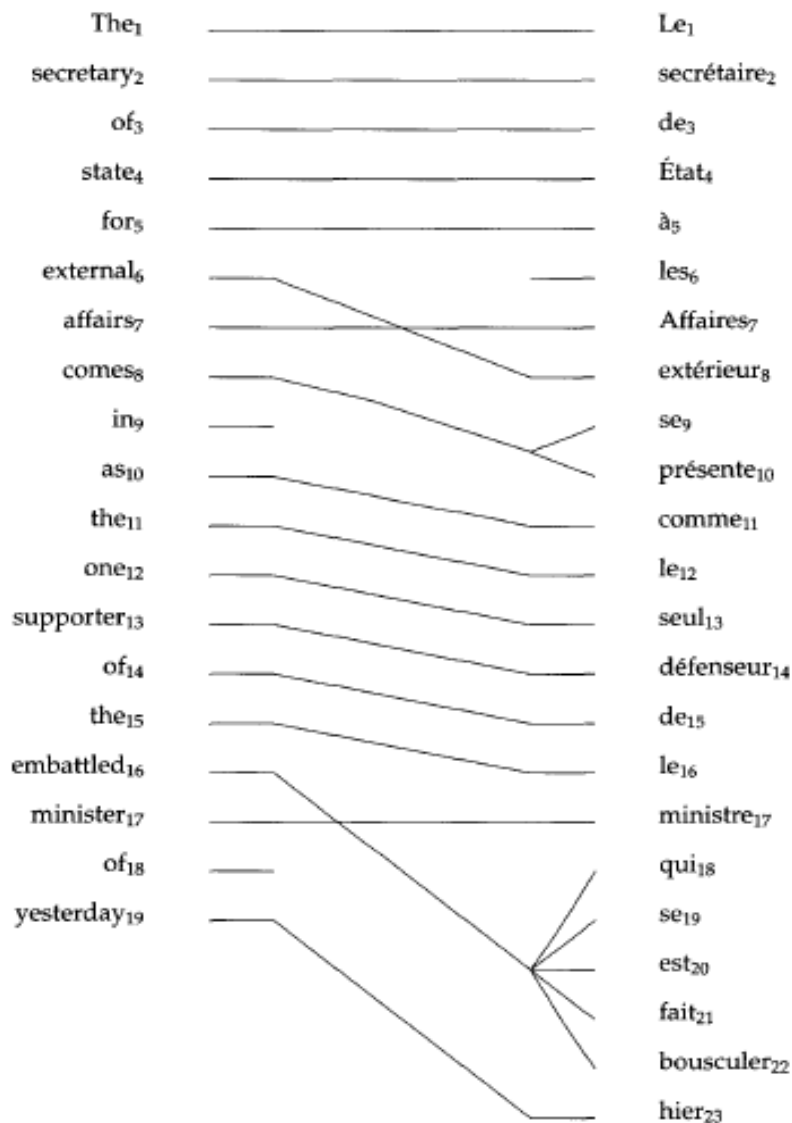


... the house ... the blue house ... the flower ...



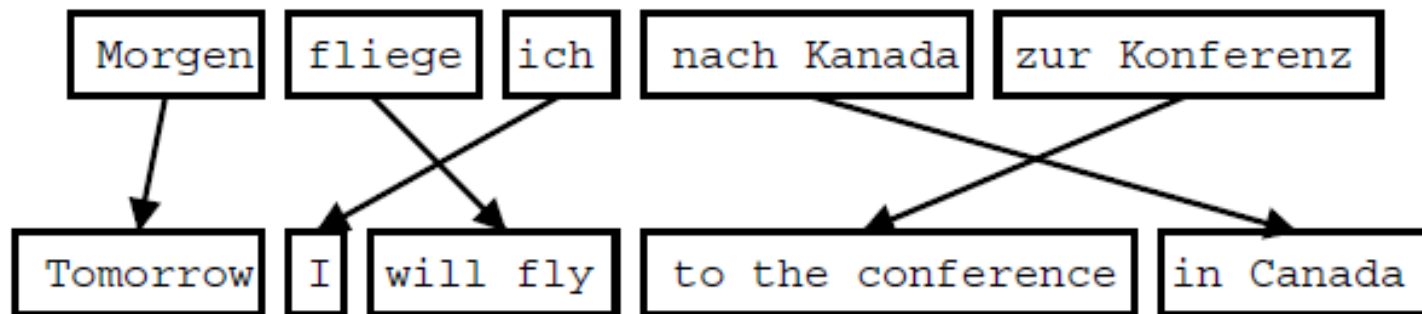
$$\begin{aligned}p(\text{la}|\text{the}) &= 0.453 \\p(\text{le}|\text{the}) &= 0.334 \\p(\text{maison}|\text{house}) &= 0.876 \\p(\text{bleu}|\text{blue}) &= 0.563 \\&\dots\end{aligned}$$

Működhet-e a szószintű szinkronizálás?



Megjelennek a frázis-alapú fordítási modellek

- ❑ Szószinten nehéz a törlés és a beszúrás, ezért megéri a bemenetet frázisokra szegmentálni
- ❑ Itt a frázis nem nyelvtani fogalom, hanem csak a gyakran előforduló összefüggő szószorozatok
- ❑ Először minden frázist lefordítunk a célnyelvre
- ❑ Ezután a frázisokat átrendezzük:



- ❑ Így nem kompozicionális frázisok is fordíthatók
- ❑ Minél több az adat, annál több hosszú frázis tanulható meg

A nyelvmodell kialakítása

- ☐ Ez azoknak a mondatoknak a halmazából indul ki, amit az adott nyelven mondani szoktak
- ☐ Milyen a „jó” nyelvmodell? Milyen a „jó” nyelv?
- ☐ Trigram-valószínűségekkel szokás közelíteni
- ☐ Pl. $p(\text{witch} | \text{the green}) > p(\text{green} | \text{the witch})$
- ☐ A nyelvmodell kiindulhat a web n-gramjaiból is!

Ha nem elég nagy a korpusz

- ☐ Sokszor szükséges ún. simítás a nyelvmodellhez
- ☐ Ha z sose követte xy -t a szövegeinkben, akkor még mindig megkérdezhetjük, hogy z legalább y -t követte-e?
- ☐ Ha igen, akkor xyz talán nem olyan rossz
- ☐ Ha nem követte, még mindig megkérdezhetjük, hogy z elfogadható hétköznapi szó-e?
Ha nem, akkor xyz tényleg nagyon kis valószínűségű kell legyen!
- ☐ $a = \text{előfordulás}("xyz") / \text{előfordulás}("xy")$ simítva:
 $b(z|x y) =$
 $0.95 * \text{előfordulás}("xyz") / \text{előfordulás}("xy") +$
 $0.04 * \text{előfordulás}("yz") / \text{előfordulás}("z") +$
 $0.008 * \text{előfordulás}("z") / \text{összes-látott-szó} +$
 0.002

A statisztikai GF problémái és a továbblépés

- ☐ Tulajdonnevek: másolás, transliterálás, szótárból újraalkotás
- ☐ Számmal alkotott kifejezések, dátumok, mennyiségek: saját fordítási táblák kellenek
- ☐ Főnévi csoportok általában: ezeket érdemes nyelvileg „elő-elemezni”
- ☐ A hibrid rendszerek előtérbe kerülnek (pl. EuroMatrix)
- ☐ Alternatív elképzelések: létező rendszerek/ szolgáltatások kombinációja
- ☐ webforditas.hu → iTranslate4.eu
- ☐ Előtérbe kerül a géppel támogatott utószerkesztés
- ☐ 2010-es évek: elindul a világ a neurális hálós fordítás felé...

Neurális gépi fordítás

