Communication Langagière

Ingénierie des langues et de la parole

1. **Introduction générale**
2. **Ingénierie des langues**
   - 2.1. Représentation et codage des textes
   - 2.2. Applications du TALN :
     - 2.2.1. Dictionnaire et étiquetage de surface
     - 2.2.2. Traduction automatique statistique
   - 2.3. Introduction à l'apprentissage profond
   - 2.4. Encodeur/Décodeur
   - 2.5. BERT
3. **Ingénierie de la parole**
   - 3.1. Rappels de traitement numérique du signal
   - 3.2. Codage et représentation de la parole
   - 3.3. Approches auto-supervisées

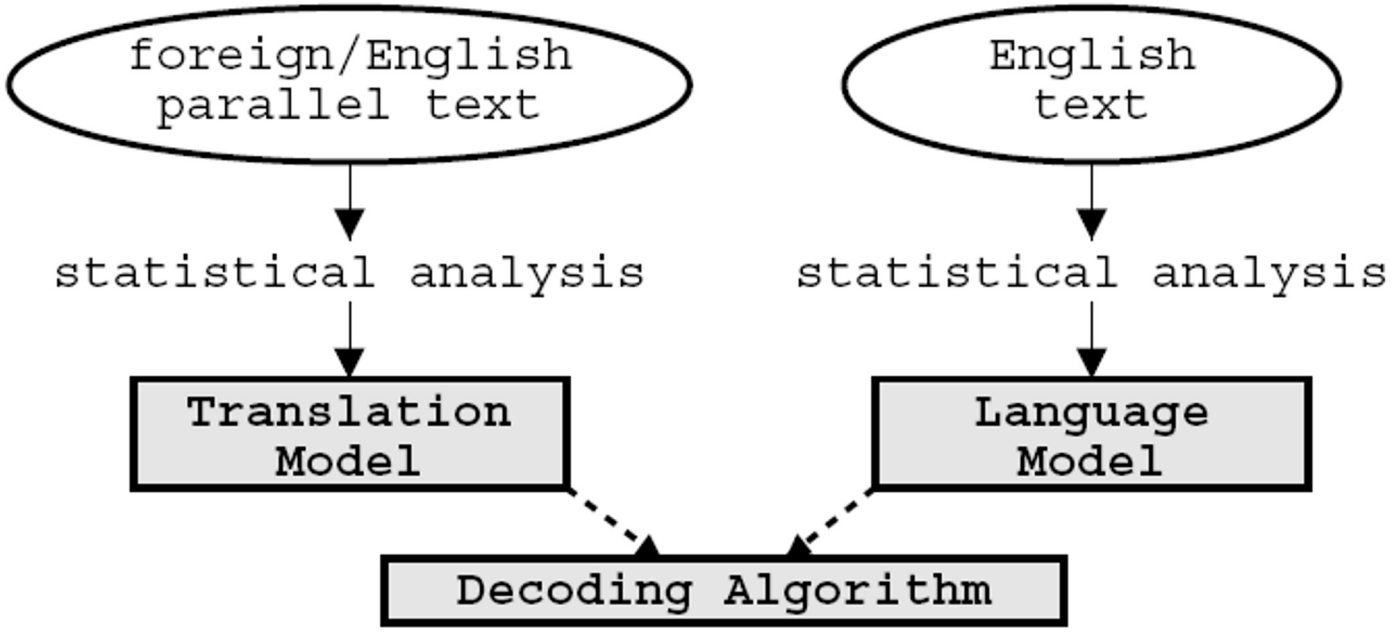
Quelques citations…

- *It must be recognized that the notion ”probability of a sentence” is an entirely useless one, under any known interpretation of this term. Noam Chomsky, 1969*

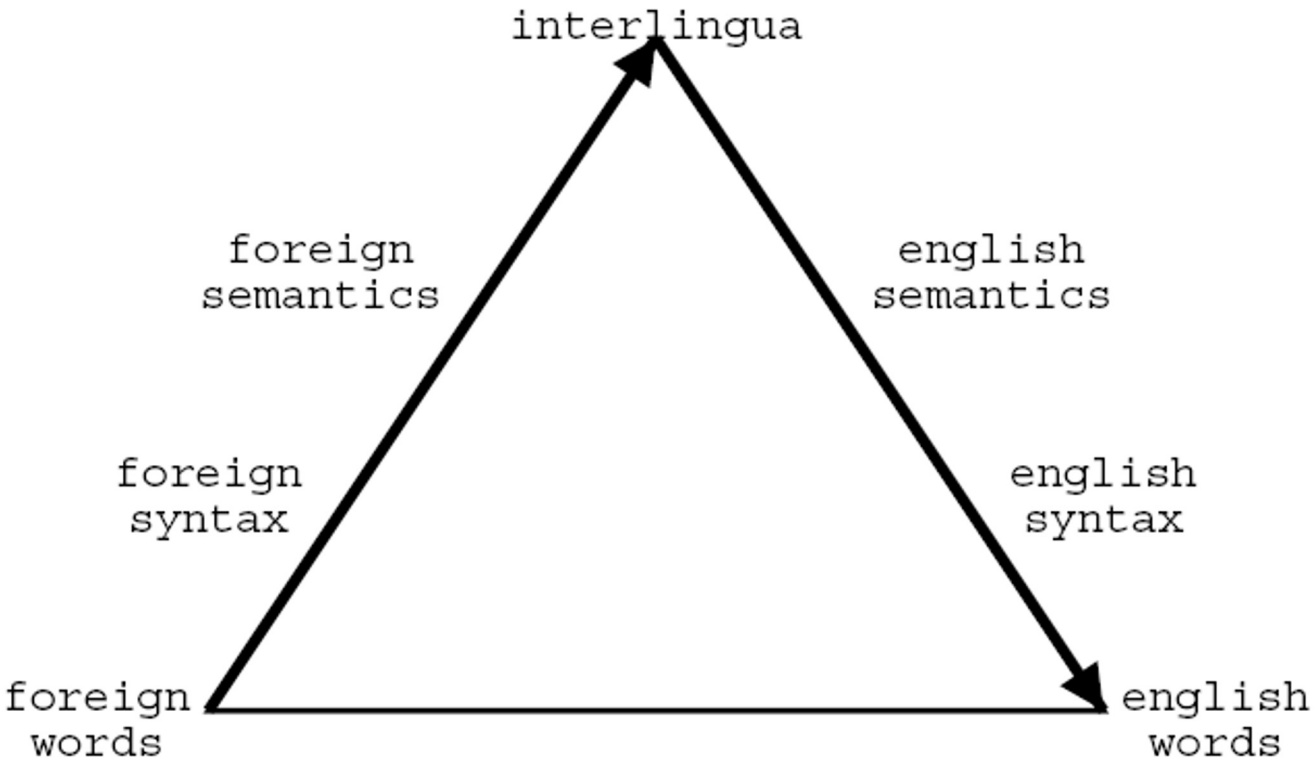
- *Whenever I fire a linguist our system performance improves. Frederick Jelinek, 1988*

Traduction automatique statistique

- Components: **Translation model**, **language model**, **decoder**



Le triangle de Vauquois

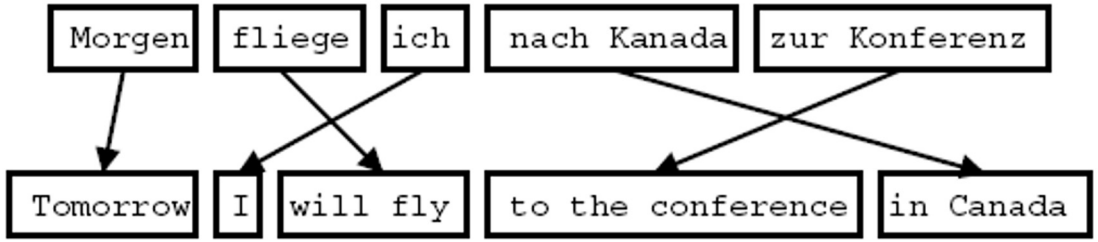


Méthodes à base de mots

[from Knight, 1997]

Premiers modèles pour la TA statistique [Brown et al., 1993]

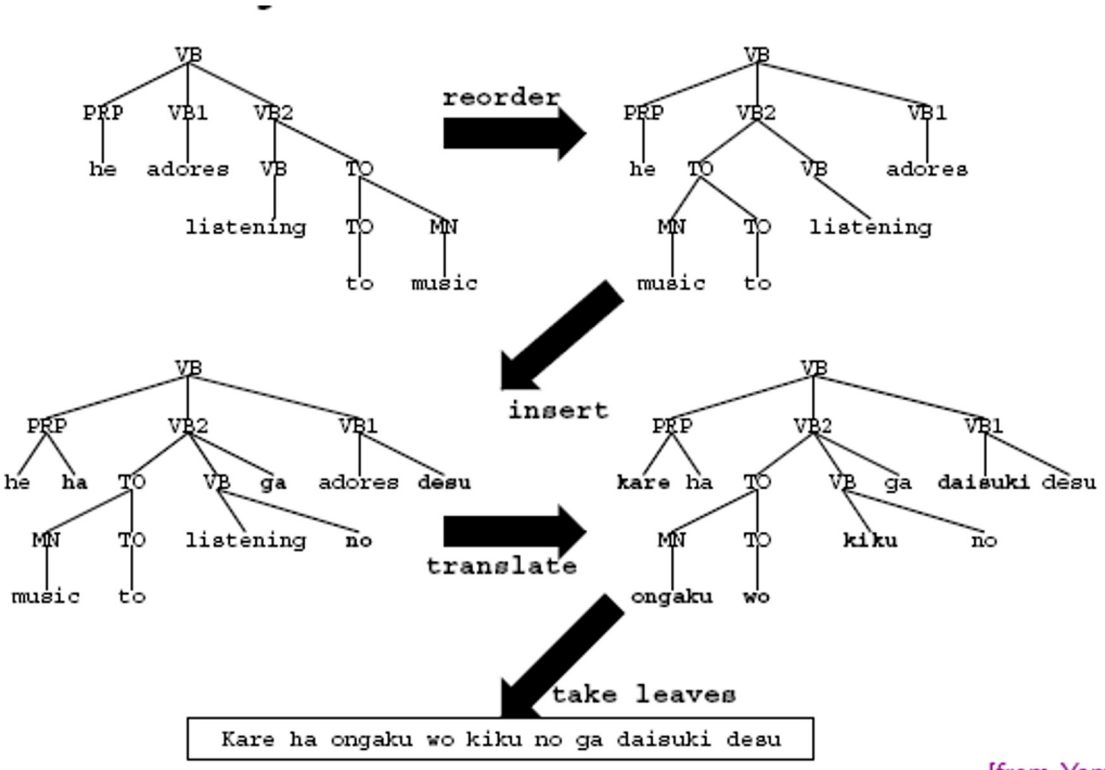
Méthodes à base de séquences



[from Koehn et al., 2003, NAACL]

- Une entrée est segmentée en séquences
- Chaque séquence est traduite
- Les séquences sont ré-ordonnées

Modèles fondés sur la syntaxe



[from Yamada and Knight, 2001]

Systèmes de traduction en ligne

- **Exemple 1 : google**
  - **https://translate.google.com/**
- **Exemple 2 : deepL**
  - **https://www.deepl.com/translator**

Evaluation automatique

- Pourquoi une évaluation automatique ?
  - **Evaluation manuelle trop lente**
  - **Doit être faite sur de larges ensembles de test**
  - **Permet un tuning automatique des systèmes de TA**
- Historique
  - **Word Error Rate**
  - **BLEU depuis 2002**
- BLEU : Recouvrement avec des traductions de référence

Evaluation automatique

- Reference Translation

–the gunman was shot to death by the police

- the gunman was police kill .
- wounded police jaya of
- the gunman was shot dead by the police .
- the gunman arrested by police kill .
- the gunmen were killed .
- the gunman was shot to death by the police .
- gunmen were killed by police ?SUB>0 ?SUB>0
- al by the police .
- the ringer is killed by the police .
- police killed the gunman .

- **Vert : 4-gram match => bien**
- **Rouge : no match => mauvais**

BLEU ! La formule

$$BLEU(t,r) = BP \times \exp\left(\sum_{n=1}^{N} \omega_n \times \log p_n\right)$$

où

$$p_n = \frac{count(n\_grammes_t \cap n\_grammes_r)}{count(n\_grammes_t)}$$

Problème de l'évaluation automatique

- Source :

–The rapporteurs have already stressed the quality of the debate and the need to progress further, and I can only agree with them.

- • Trad. Auto.

–Les rapporteurs ont déjà souligné la qualité du débat et la nécessité de progresser, et je ne peux qu'être d'accord avec eux.

- • Réf.

–Les rapporteurs ont souligné la qualité de la discussion et aussi le besoin d'aller plus loin. Bien sûr, je ne peux que les rejoindre.

Problème de l'évaluation automatique

- Source :

–**The rapporteurs have already stressed the quality of the debate and the need to progress further, and I can only agree with them.**

- • Trad. Auto.

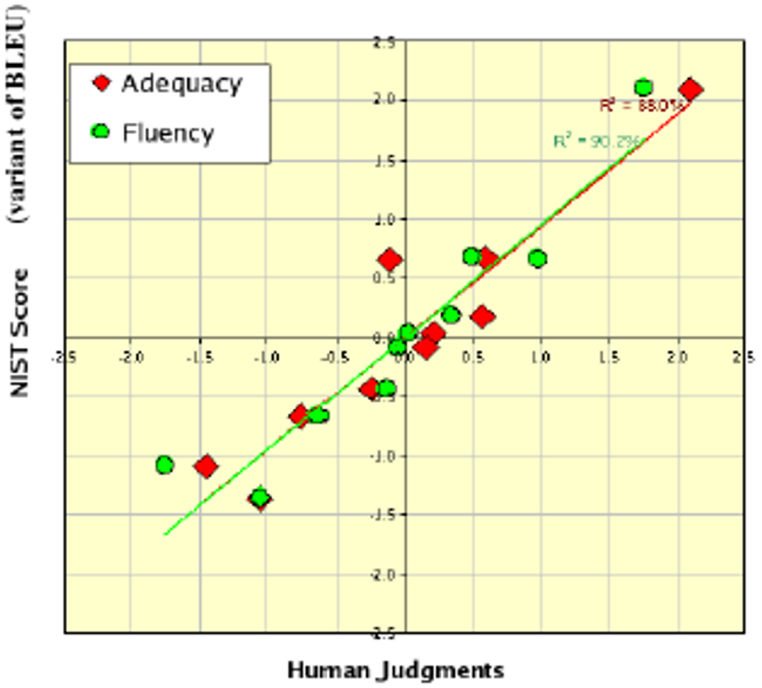
–**Les rapporteurs ont déjà souligné la qualité du débat et la nécessité de progresser, et je ne peux qu'être d'accord avec eux.**

- • Réf.

–**Les rapporteurs ont souligné la qualité de la discussion et aussi le besoin d'aller plus loin. Bien sûr, je ne peux que les rejoindre.**

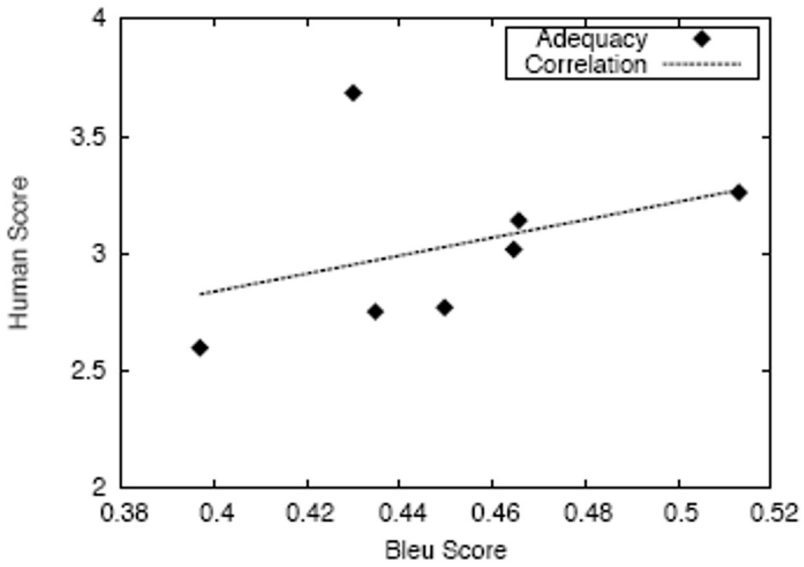
=> BLEU proche de 0 !!!

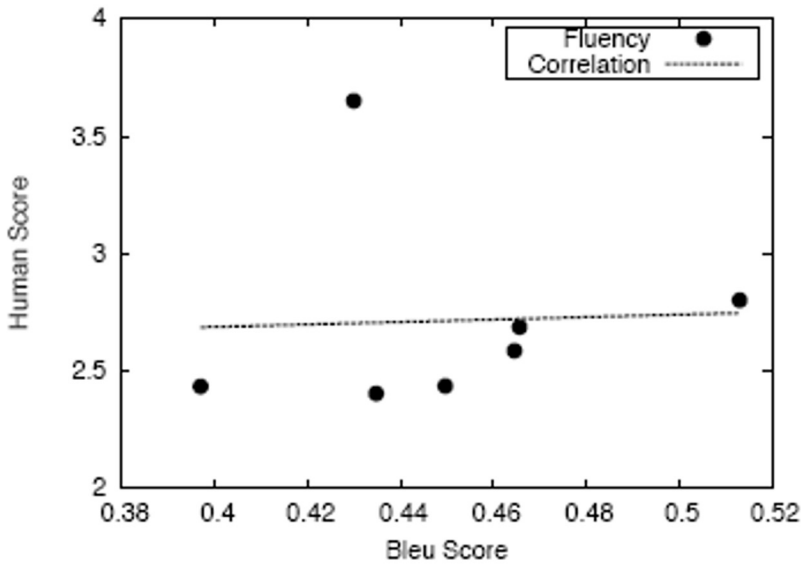
Correlation ou pas avec des jugements humains



[from George Doddington, NIST]

Correlation ou pas avec des jugements humains





[from Callison-Burch et al., 2006, EACL]

Campagnes d'évaluation

- NIST/DARPA:
  - **Campagnes annuelles : Arabic-English, Chinese-English, nouvelles, depuis 2001**
- IWSLT:
  - **Campagnes annuelles : Chinois, Japonais, Italien, Arabe => Anglais , domaine « tourisme » depuis 2003**
- WPT/WMT:
  - **Langues européennes , archives du parlement européen, depuis 2005**

Euromatrix

- **110 systèmes !!**

	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

[from Koehn, 2005: Europarl]

Traduire depuis/vers une langue

Language	From	Into	Diff
da	23.4	23.3	0.0
de	**22.2**	**17.7**	**-4.5**
el	23.8	22.9	-0.9
en	**23.8**	**27.4**	**+3.6**
es	26.7	29.6	+2.9
fr	26.1	31.1	+5.1
fi	19.1	12.4	-6.7
it	24.3	25.4	+1.1
nl	19.7	20.7	+1.1
pt	26.1	27.0	+0.9
sv	24.8	22.1	-2.6

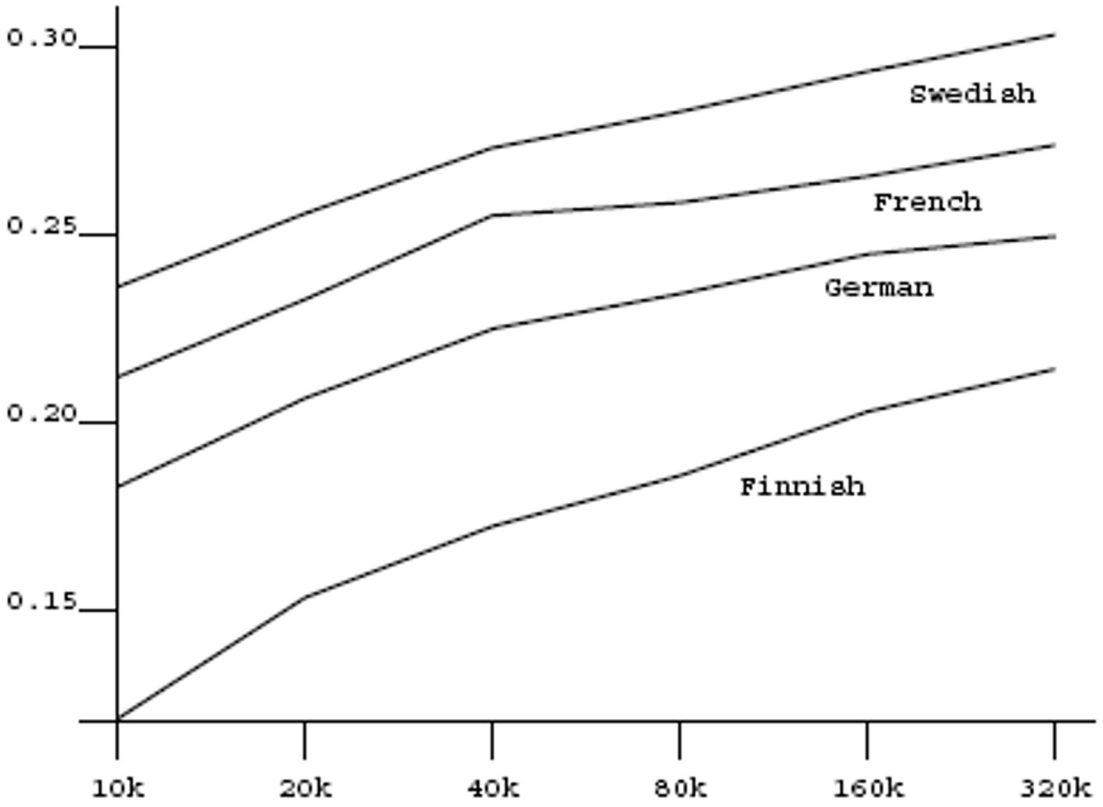
[from Koehn, 2005: Europarl]

Difficile de traduire vers les langues à riche morphologie (Allemand, Finnois)

Données disponibles

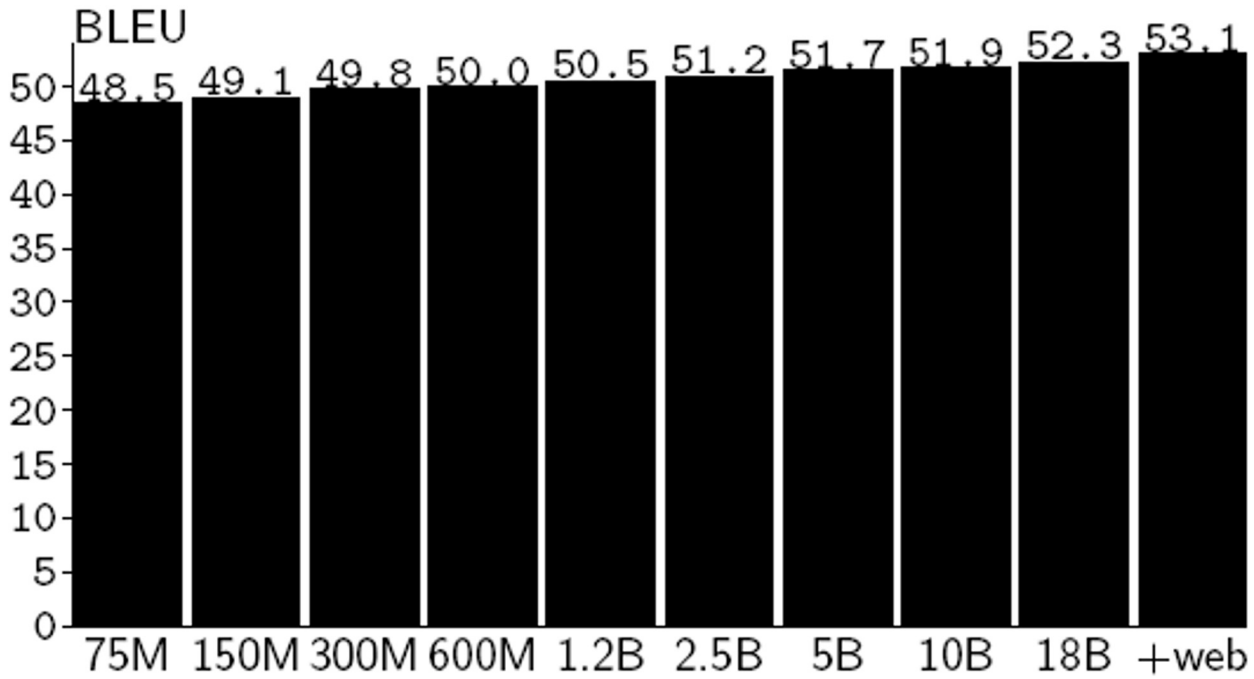
- Corpus parallèles
  - **Europarl: 30 millions de mots en 11 langues http://www.statmt.org/europarl/**
  - **Acquis Communotaire: 8-50 million mots (20 langues EU)**
  - **Canadian Hansards: 20 million mots**
  - **Chinese/Arabic to English: plus de 100 million mots (LDC)**
- Corpus monolingues
  - **2.8 milliards de mots (Anglais, LDC)**
  - **Web**

Plus de données parallèles…



[from Koehn, 2003: Europarl]

Plus de données en langue cible…



[from Och, 2005: MT Eval presentation]

Exemple de sortie d'un système chinois/anglais

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Systèmes de TA statistique à base de mots

- Traduire un mot
  - **Dictionnaires bilingues**
- Exemple
  - **Haus — house, building, home, household, shell.**
  - **Traduction multiples**

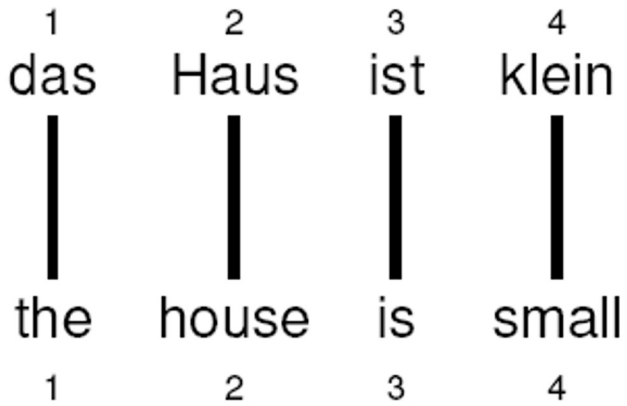
Collecter des statistiques

Translation of $Haus$	Count
house	8,000
building	1,600
home	200
household	150
shell	50

Maximum likelihood estimation
(Maximum de vraisemblance)

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \textit{house}, \\ 0.16 & \text{if } e = \textit{building}, \\ 0.02 & \text{if } e = \textit{home}, \\ 0.015 & \text{if } e = \textit{household}, \\ 0.005 & \text{if } e = \textit{shell}. \end{cases}$$

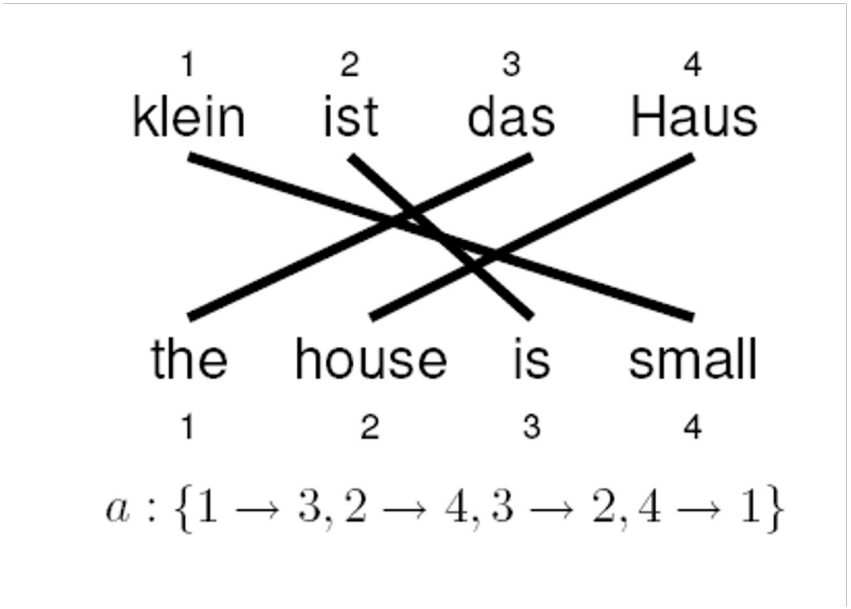
Alignement



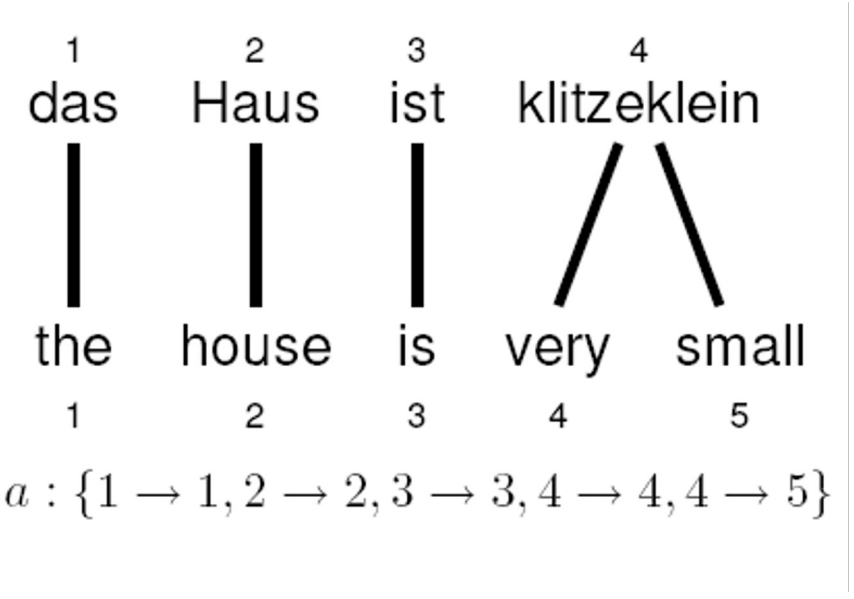
Fonction d'alignement : mot anglais cible en position i associé à mot étranger source en position j par la fonction a: i->j

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

Exemples d'alignement (1)

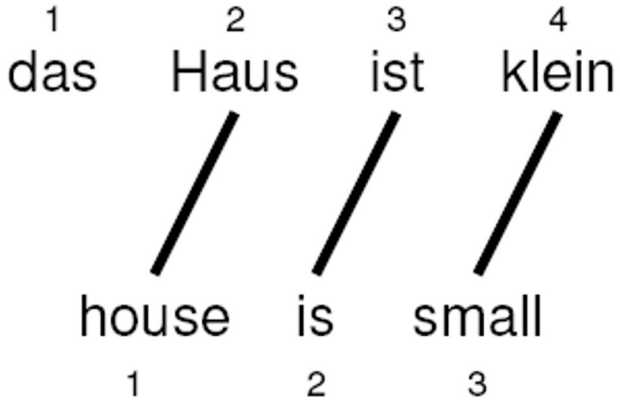


$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$

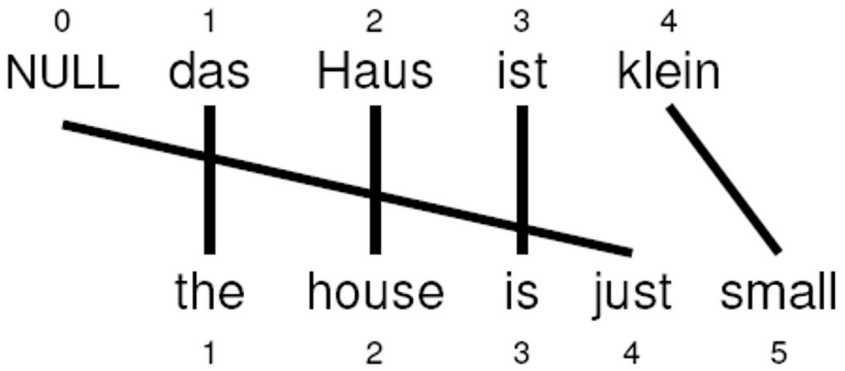


$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 4 \rightarrow 5\}$

Exemples d'alignement (2)



$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$



$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$

Modèle IBM-1

- *Generative model*: break up translation process into smaller steps
  - **IBM Model 1** only uses *lexical translation*
- Translation probability
  - for a foreign sentence $\mathbf{f} = (f_1, ..., f_{l_f})$ of length l_f
  - to an English sentence $\mathbf{e} = (e_1, ..., e_{l_e})$ of length l_e
  - with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \rightarrow i$

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

  - parameter ϵ is a *normalization constant*

Exemple

das	
e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus	
e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist	
e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein	
e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$p(e,a|f) = \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein})$$

$$= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4$$

$$= 0.0028\epsilon$$

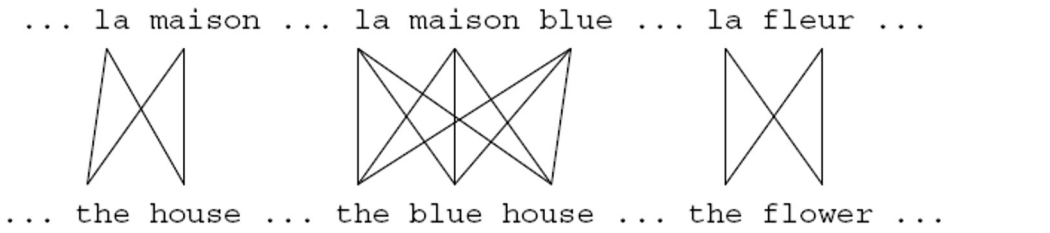
Apprendre un tel modèle (IBM-1)

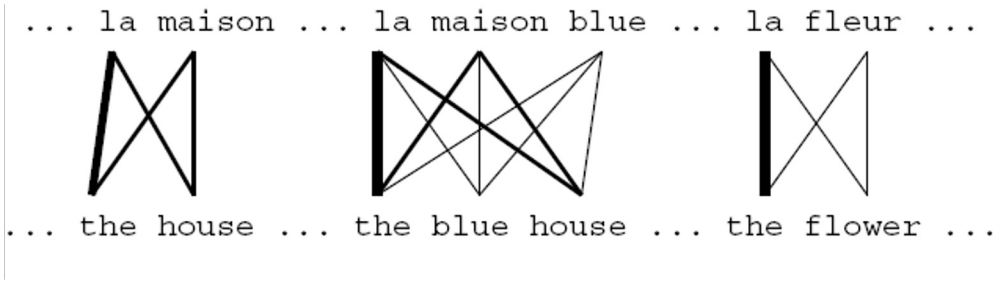
- **A partir d'un corpus parallèle**
- **Sans avoir les alignements**

⇒ **Algorithme EM**

Algorithme EM

- **Au départ, tous les alignements sont equi-probables**



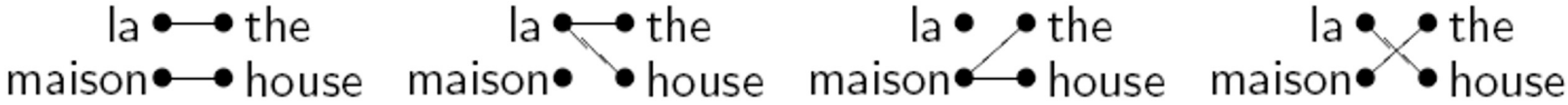


Algorithme EM

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \qquad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \qquad p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**



$p(a) = 0.56$ $p(a) = 0.035$ $p(a) = 0.08$ $p(a) = 0.005$

0.824 0.052 0.118 0.007

- **Counts**

$$c(\text{the}|\text{la}) = 0.824 + 0.052 \qquad c(\text{house}|\text{la}) = 0.052 + 0.007$$
$$c(\text{the}|\text{maison}) = 0.118 + 0.007 \qquad c(\text{house}|\text{maison}) = 0.824 + 0.118$$

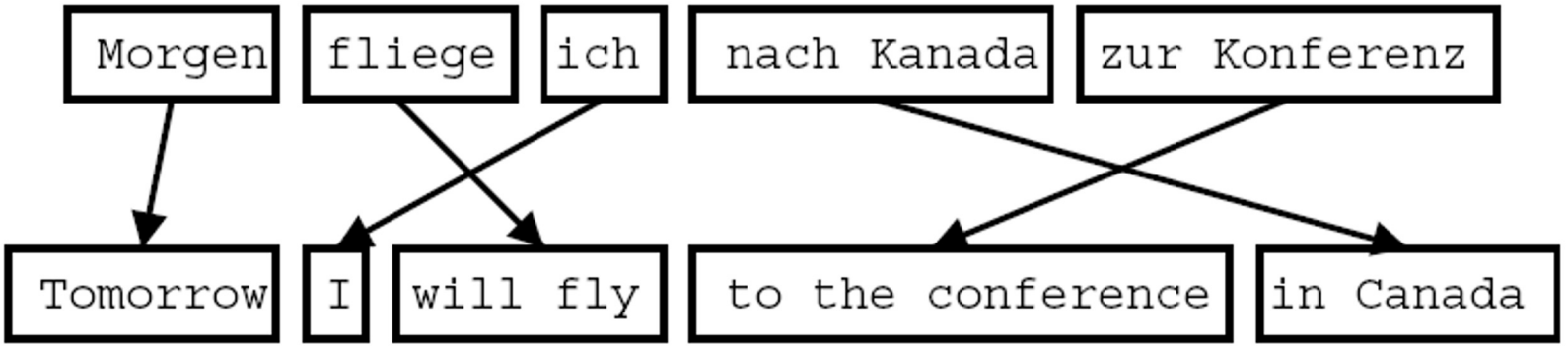
Pseudo-code

```
initialize t(e|f) uniformly
do
  set count(e|f) to 0 for all e,f
  set total(f) to 0 for all f
  for all sentence pairs (e_s,f_s)
    for all words e in e_s
      total_s = 0
      for all words f in f_s
        total_s += t(e|f)
    for all words e in e_s
      for all words f in f_s
        count(e|f) += t(e|f) / total_s
        total(f)   += t(e|f) / total_s
  for all f in domain( total(.) )
    for all e in domain( count(.|f) )
      t(e|f) = count(e|f) / total(f)
until convergence
```

Modèles suivants (IBM-2,3,4)

- **IBM-2 intègre une loi de distorsion de la forme *$p(a_j/j)$: proba. que le mot cible f_j soit aligné à e_i avec $i=a_j$***
- **IBM-3 intègre une loi de fertilité qui modélise le nombre de mots dans la phrase cible connectés à un mot dans la phrase source**
- **IBM-4 intègre un modèle de distorsion plus fin**
- **Il existe un outil (GIZA++) pour entraîner les paramètres de ces modèles**
  - **Les paramètres d'un modèle permettent d'initialiser l'apprentissage du modèle suivant**

Approche à base de séquences (*phrase-based approach*)



- Entrée en langue source ségmentée en séquences (*phrase* en anglais)
- Chaque séquence est traduite en anglais
- Les séquences sont réordonnées

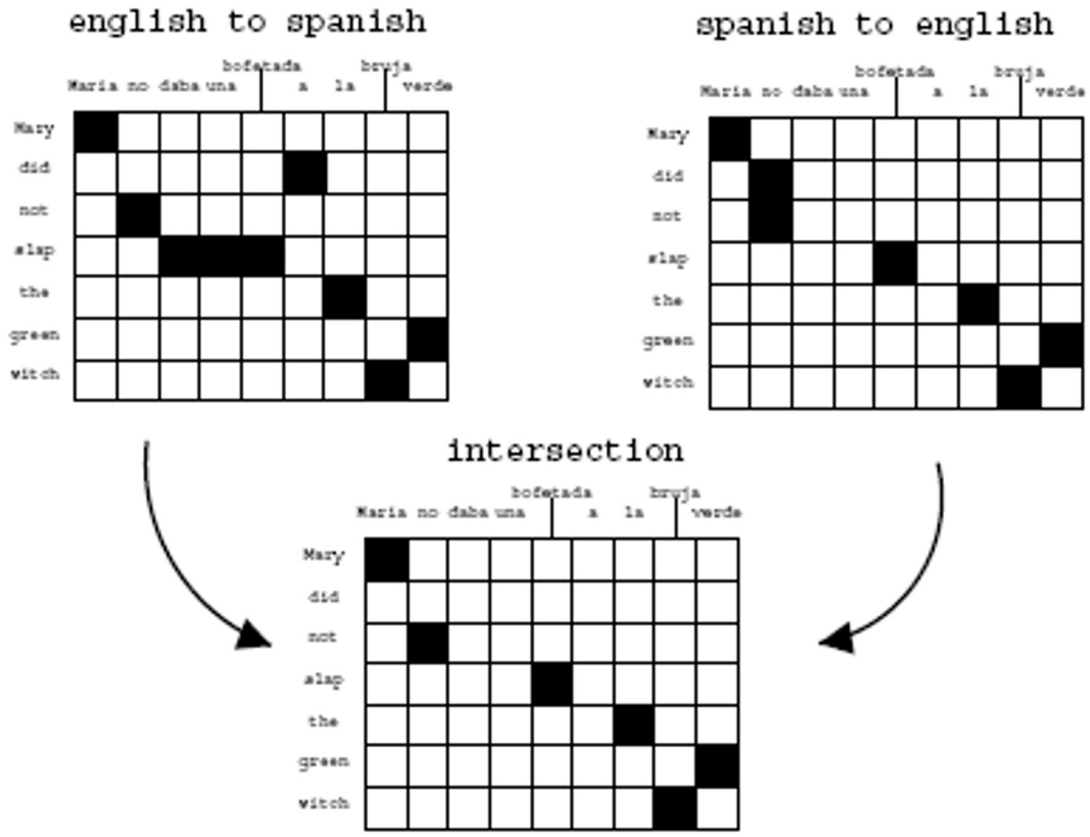
Table de traduction (*phrase table*)

- **Table pour la séquence allemande « den Vorschlag »**

English	ϕ(e|f)	English	ϕ(e|f)
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Obtention de la table

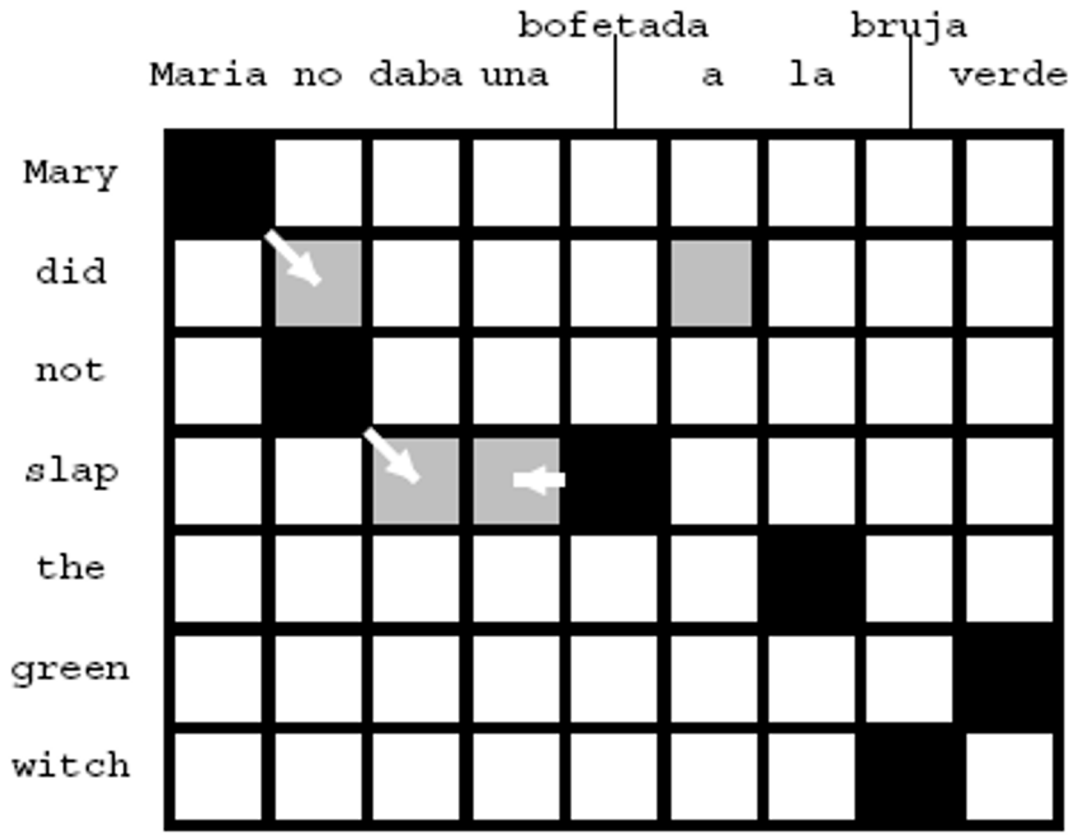
- **Obtenue à partir d'alignements bidirectionnels de mots (obtenus avec GIZA++)**



Obtention de la table

- **Expansion**

[Och and Ney, CompLing2003]

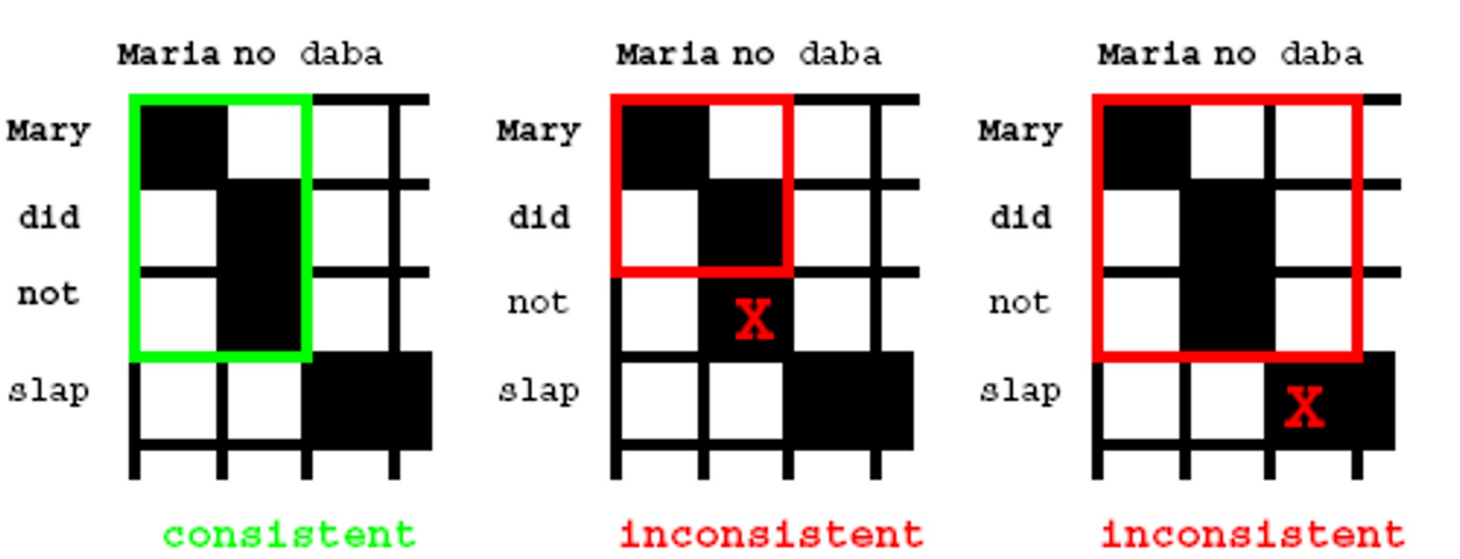


Expansion : algo

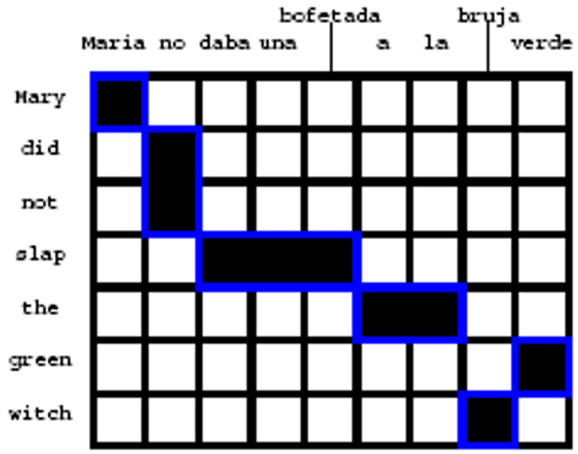
```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Obtention de la table

- **Collecter les paires de séquences consistantes avec l'alignement**

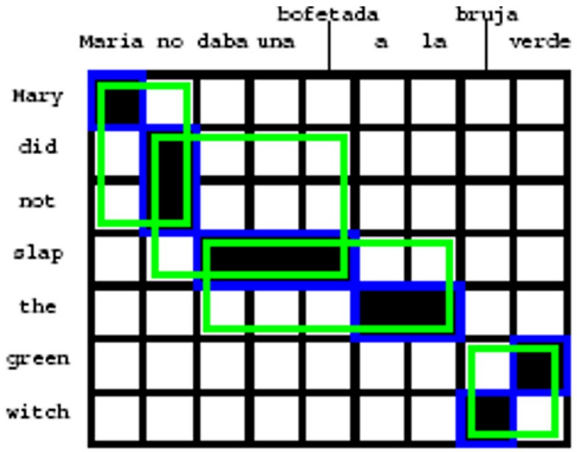


Obtention de la table



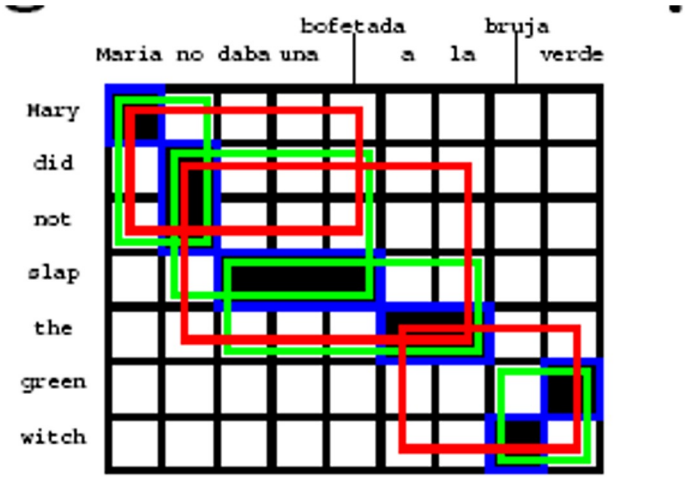
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Obtention de la table



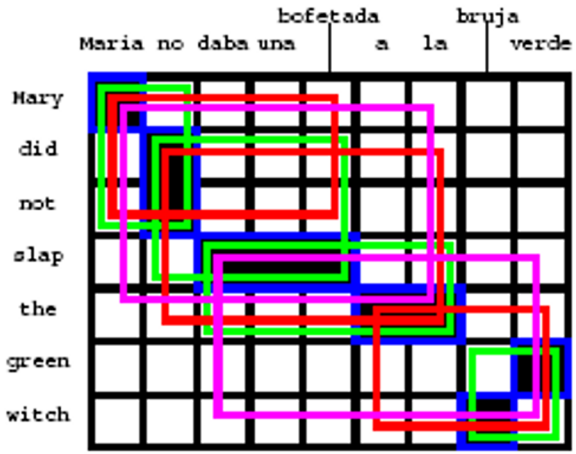
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch)

Obtention de la table



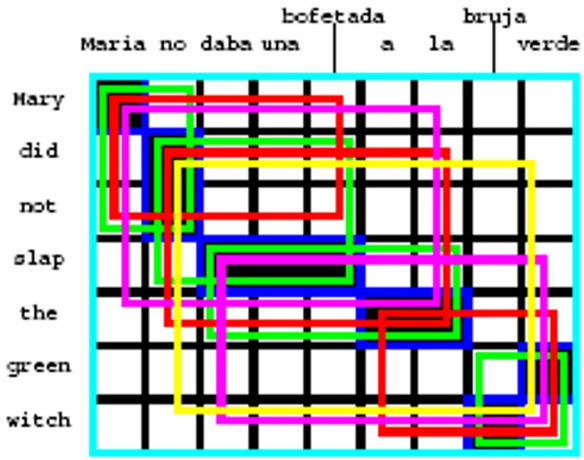
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada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

Obtention de la table



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

Obtention de la table



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

Obtention de la table

$$\phi(\overline{f}|\overline{e}) = \frac{\text{count}(\overline{f},\overline{e})}{\sum_{\overline{f}} \text{count}(\overline{f},\overline{e})}$$

Calculer aussi $\phi(\overline{e}|\overline{f})$

Garder aussi les probabilités lexicales (au niveau mot)

Modèle de traduction par séquences

- Major components of phrase-based model
  - **phrase translation model** $\phi(\mathbf{f}|\mathbf{e})$
  - **reordering model** $\omega^{\text{length}(\mathbf{e})}$
  - **language model** $p_{\text{LM}}(\mathbf{e})$
- Bayes rule

$$\begin{aligned} \text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) &= \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p(\mathbf{e}) \\ &= \text{argmax}_{\mathbf{e}} \phi(\mathbf{f}|\mathbf{e}) p_{\text{LM}}(\mathbf{e}) \omega^{\text{length}(\mathbf{e})} \end{aligned}$$

- Sentence $\mathbf{f}$ is decomposed into I phrases $\bar{f}_1^I = \bar{f}_1, ..., \bar{f}_I$
- Decomposition of $\phi(\mathbf{f}|\mathbf{e})$

$$\phi(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

Modèles log-linéaires

- IBM Models provided mathematical justification for factoring *components* together

$$p_{LM} \times p_{TM} \times p_D$$

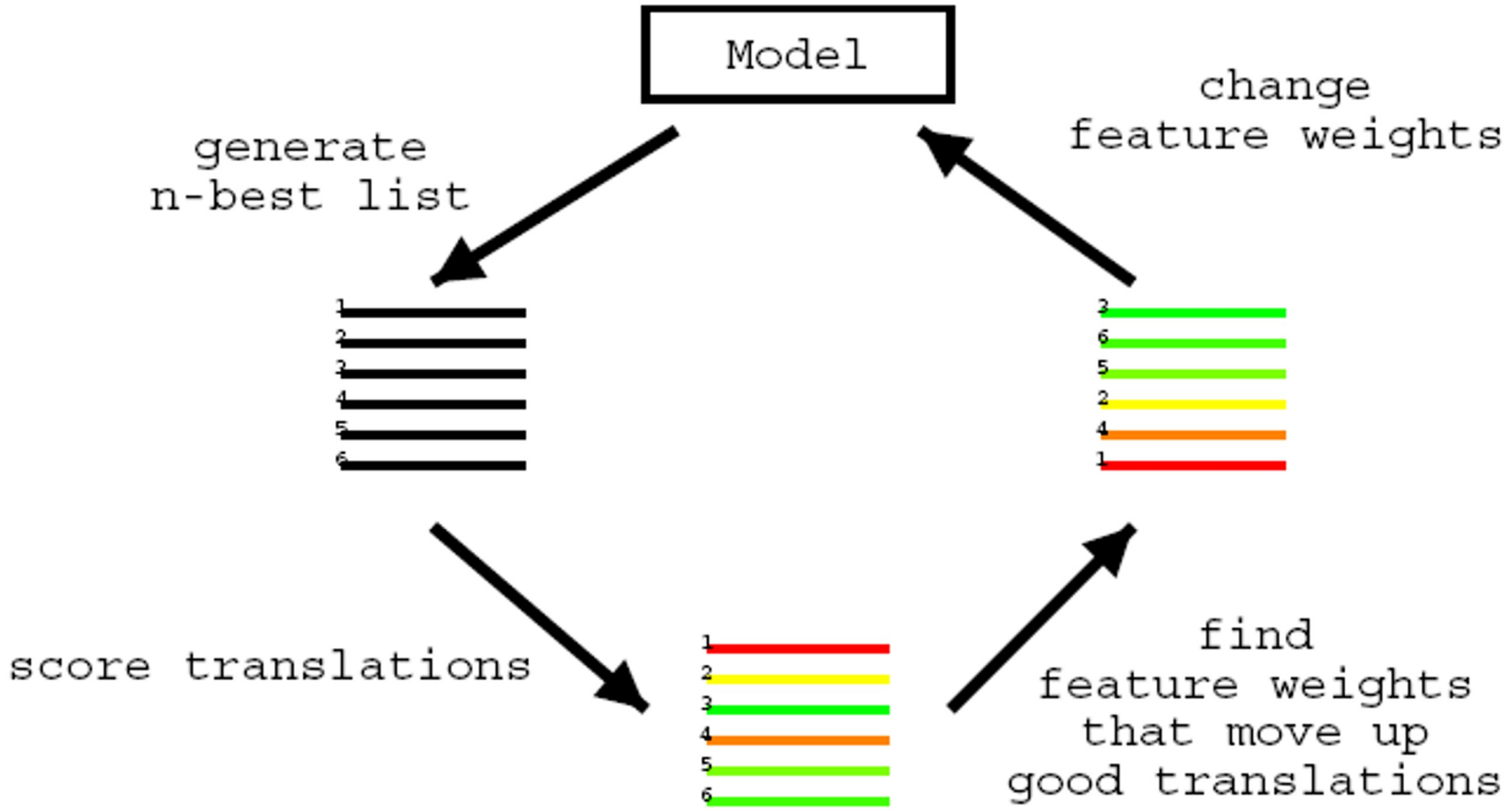
- These may be *weighted*

$$p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D}$$

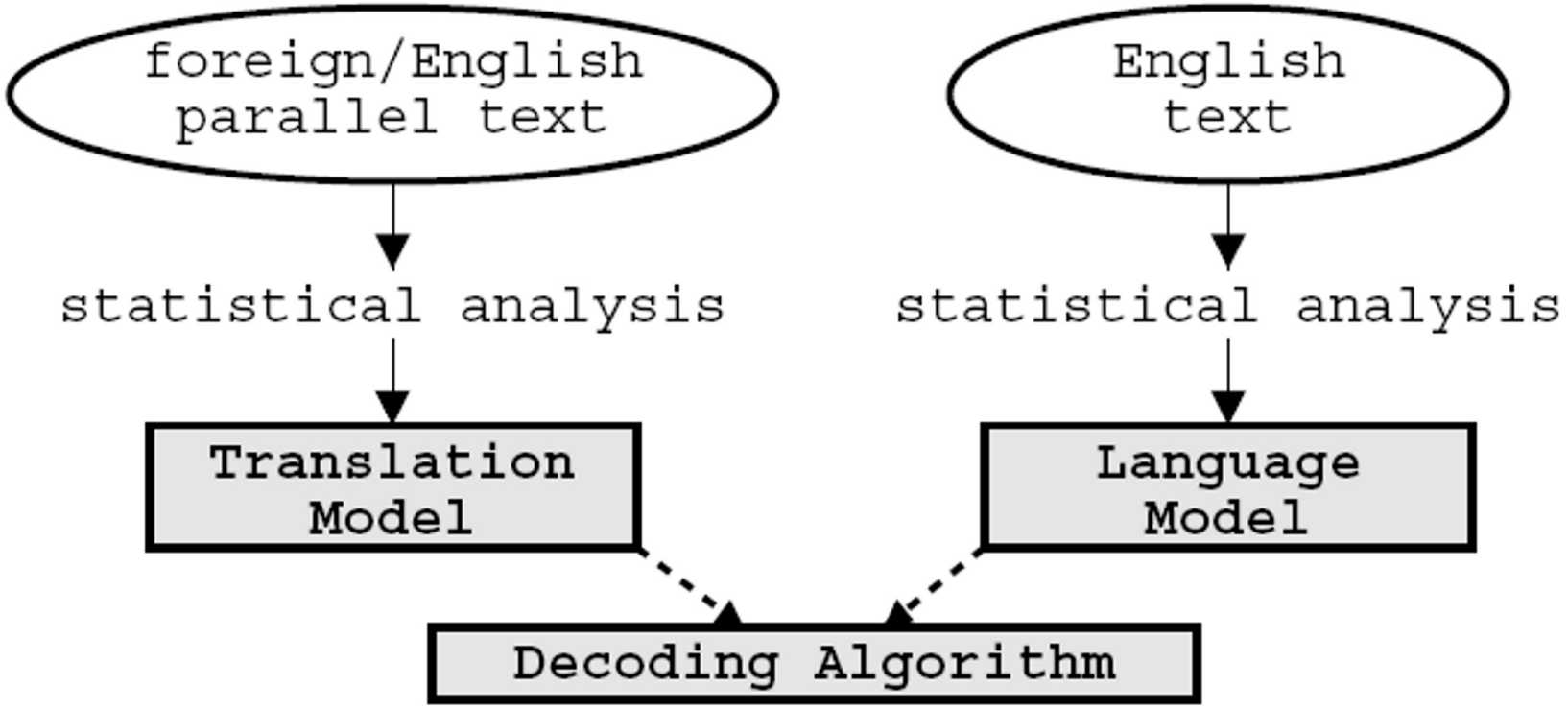
- *Many components* p_i with weights λ_i

$\Rightarrow \prod_i p_i^{\lambda_i} = exp(\sum_i \lambda_i log(p_i))$
$\Rightarrow log \prod_i p_i^{\lambda_i} = \sum_i \lambda_i log(p_i)$

Apprendre les poids

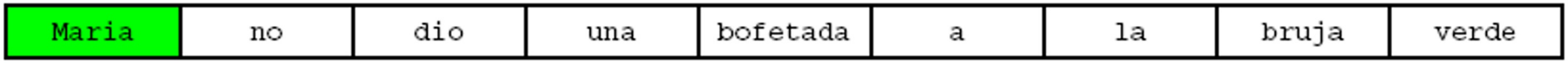


Décodage


foreign/English
parallel text
English
text
statistical analysis
statistical analysis
Translation
Model
Language
Model
Decoding Algorithm

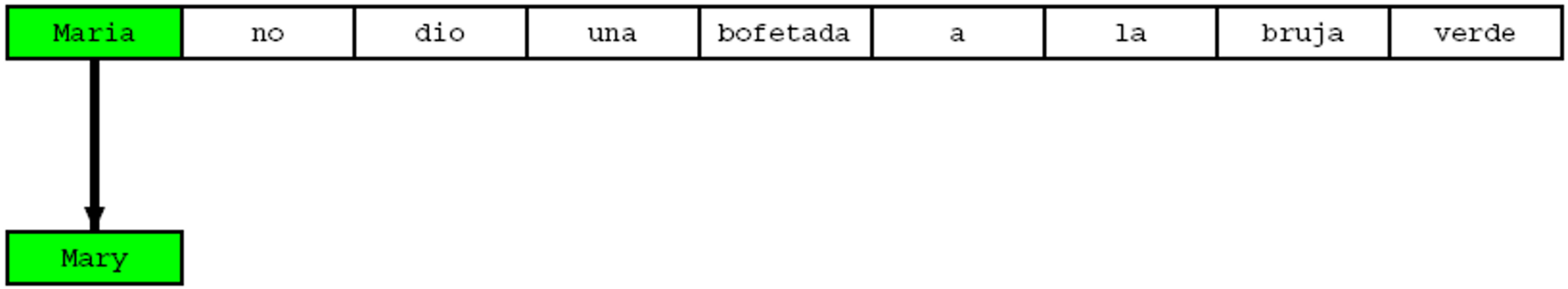
Processus de décodage

- **Construit la traduction de gauche à droite**
  - **Sélectionne les mots source à traduire**



Processus de décodage

- **Construit la traduction de gauche à droite**
  - **Sélectionne les mots source à traduire**
  - **Trouve la séquence anglais correspondante**
  - **Ajoute la séquence anglais à la fin de la traduction partielle courante**



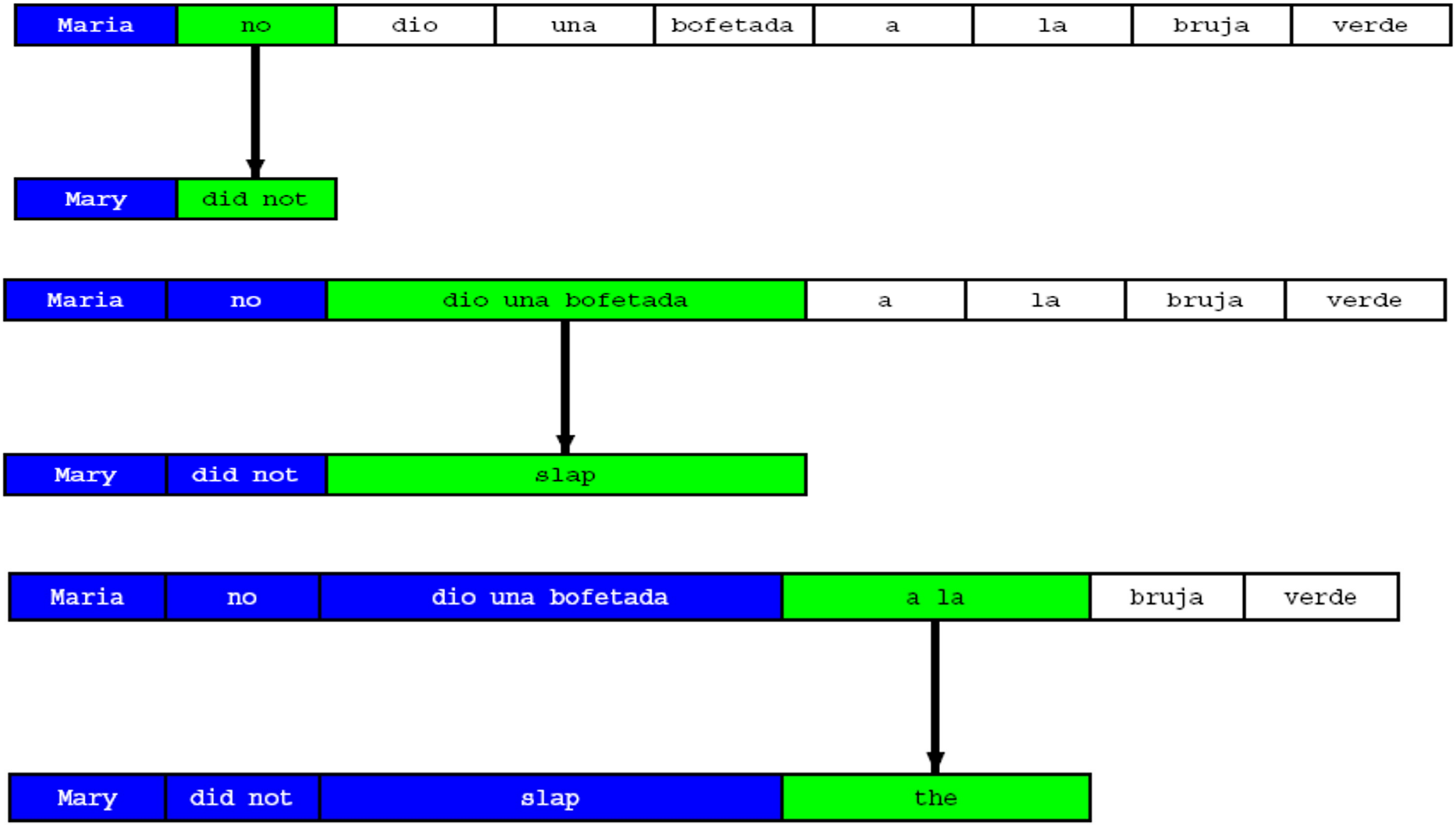
Processus de décodage

- **Construit la traduction de gauche à droite**
  - **Sélectionne les mots source à traduire**
  - **Trouve la séquence anglais correspondante**
  - **Ajoute la séquence anglais à la fin de la traduction partielle courante**
  - **Marque les mots sources « traduits »**

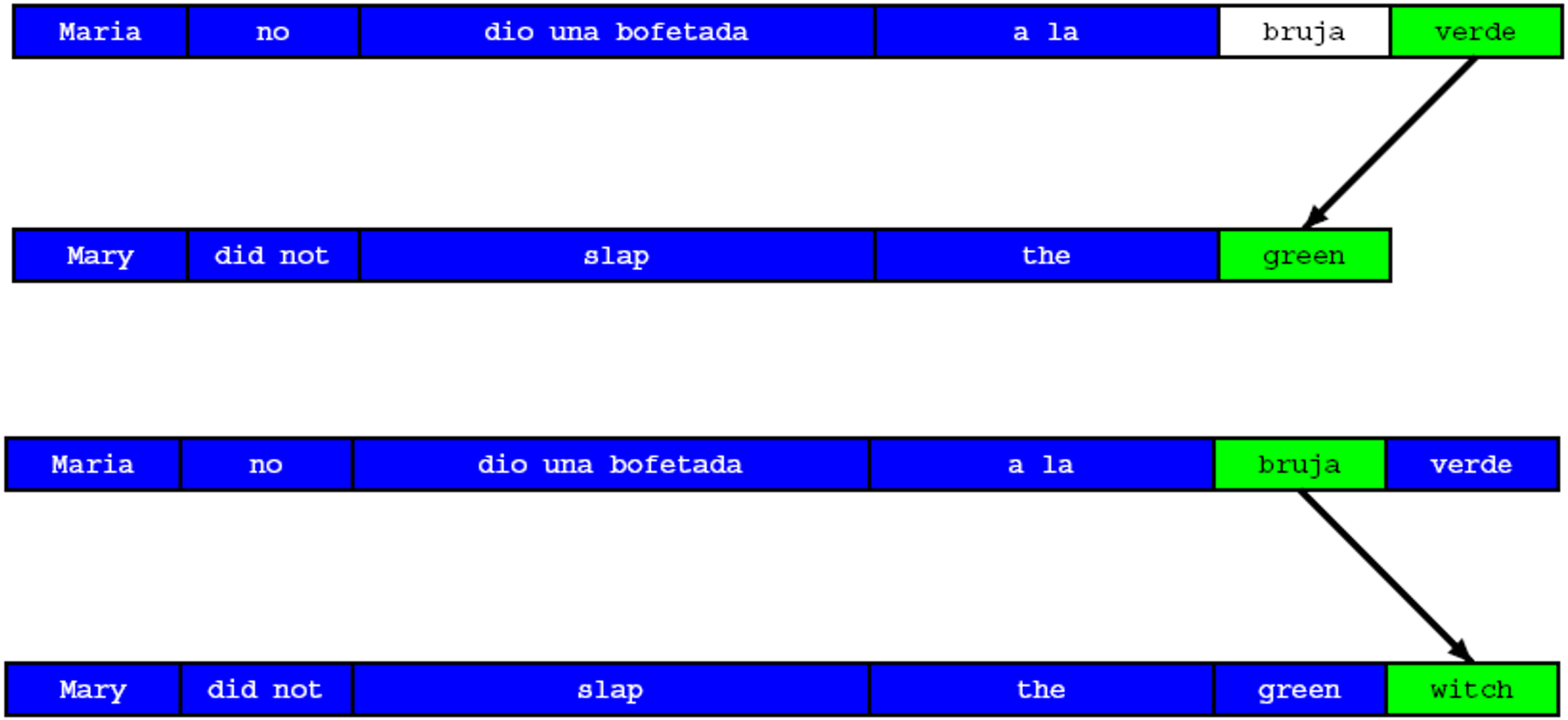
Maria	no	dio	una	bofetada	a	la	bruja	verde

Mary

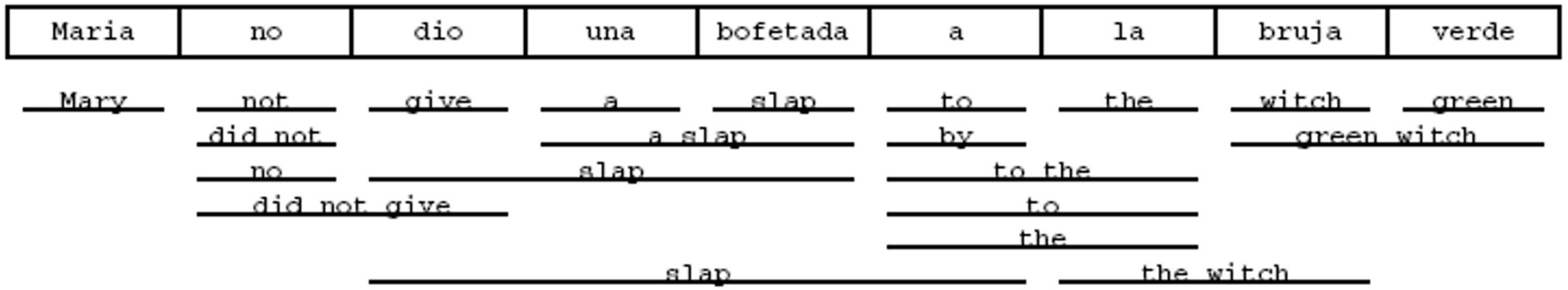
Processus de décodage



Processus de décodage



Options de traduction



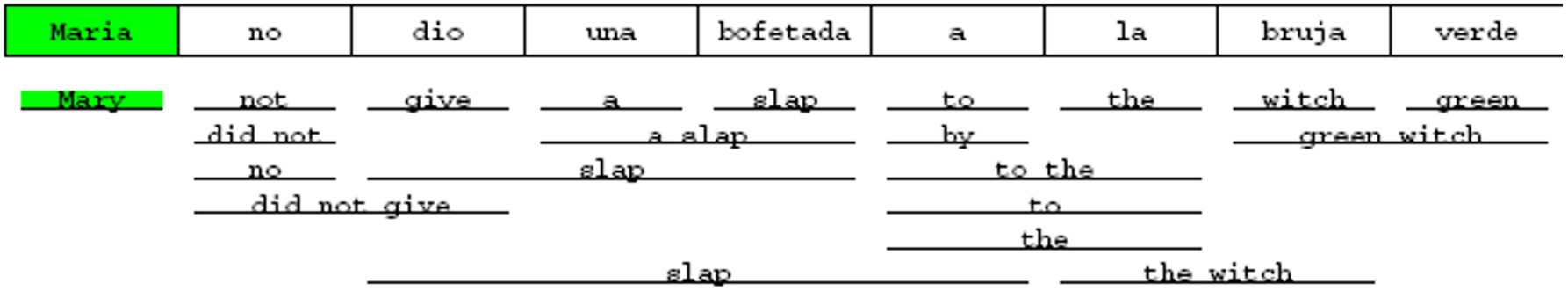
-différentes façons de segmenter une phrase source en séquences
-différentes façons de traduire chaque séquence

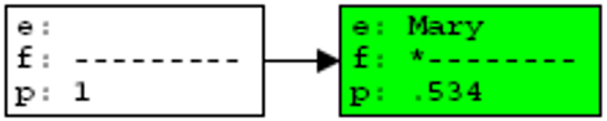
Expansion d’hypothèses

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		by		green witch	
	no	slap			to the			
	did not give				to			
					the			
		slap				the witch		

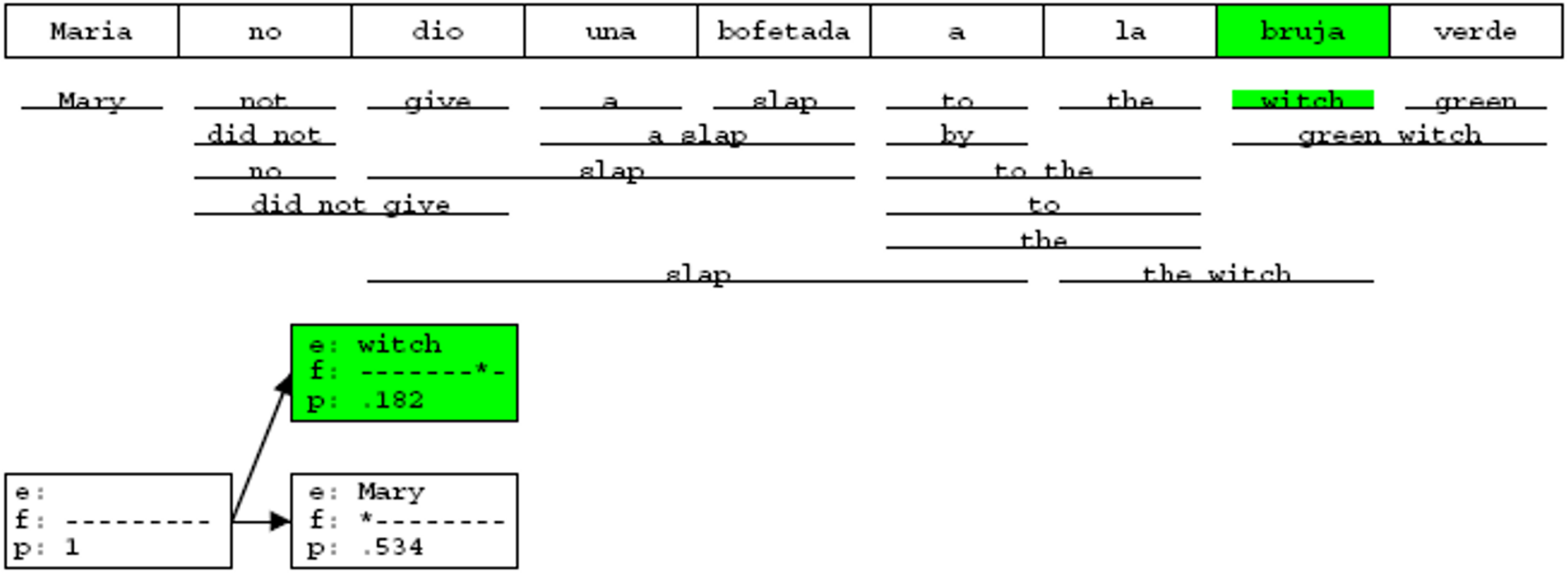
e:
f: ---------
p: 1

Expansion d’hypothèses

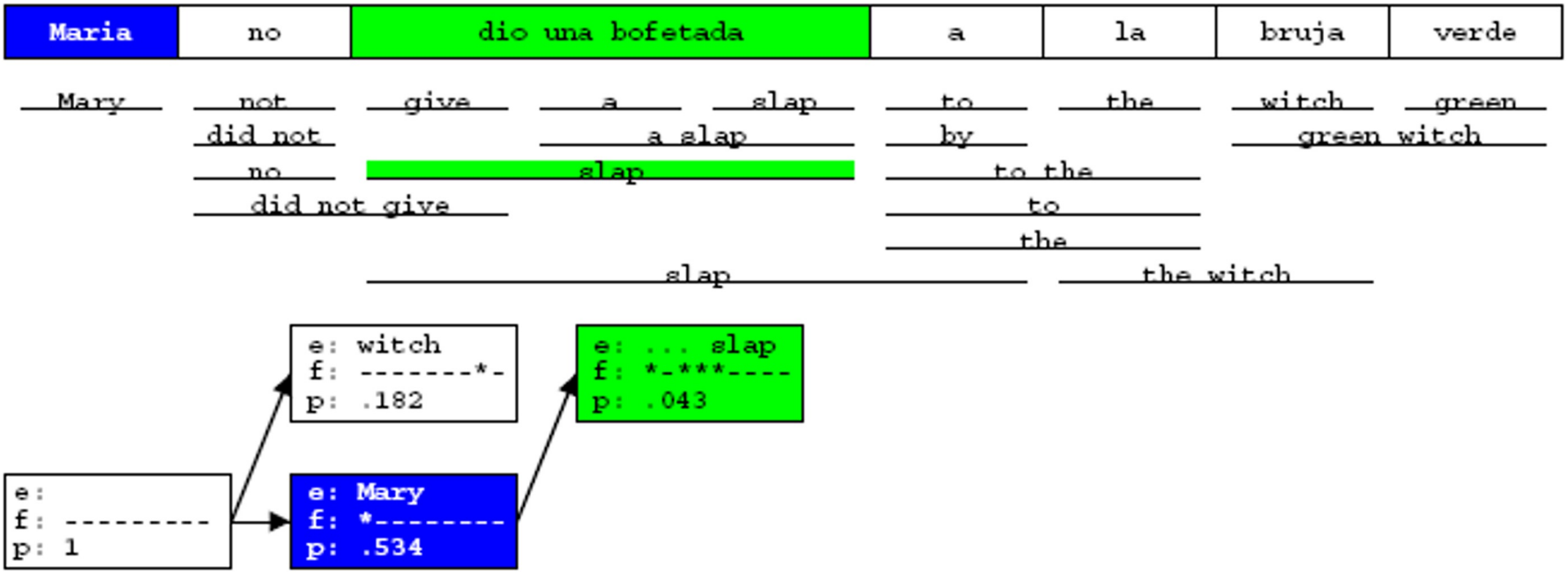




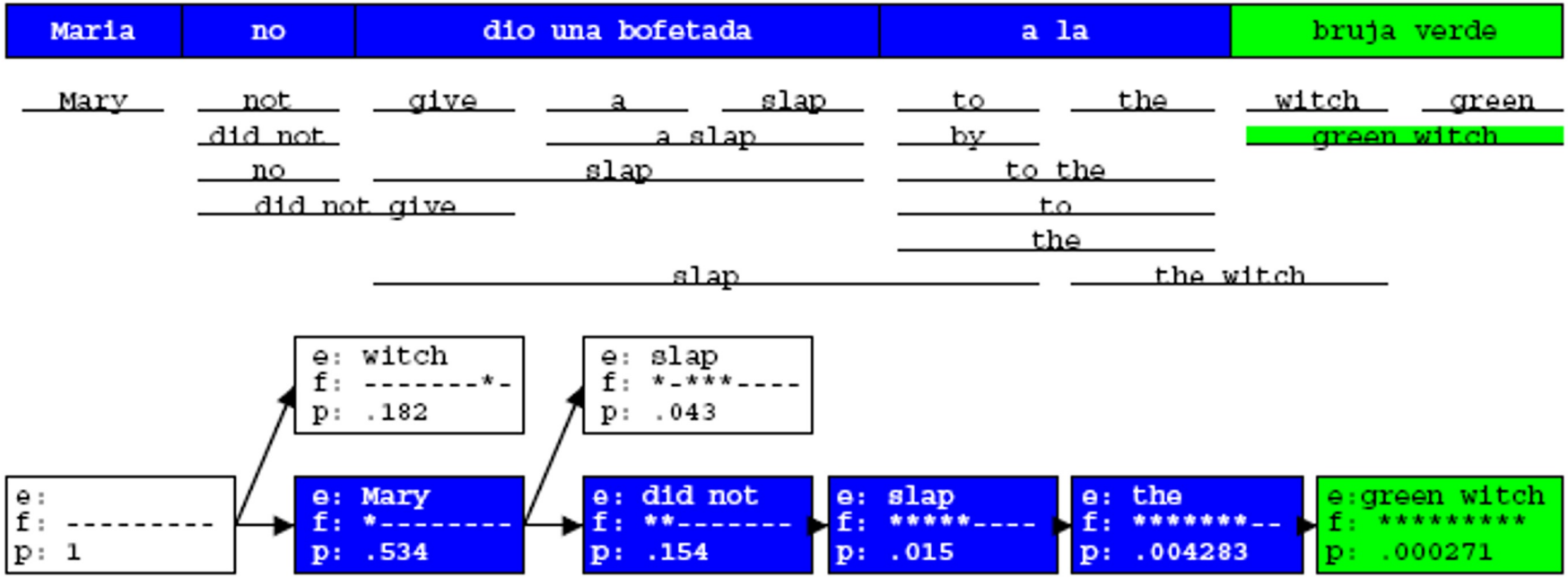
Expansion d’hypothèses



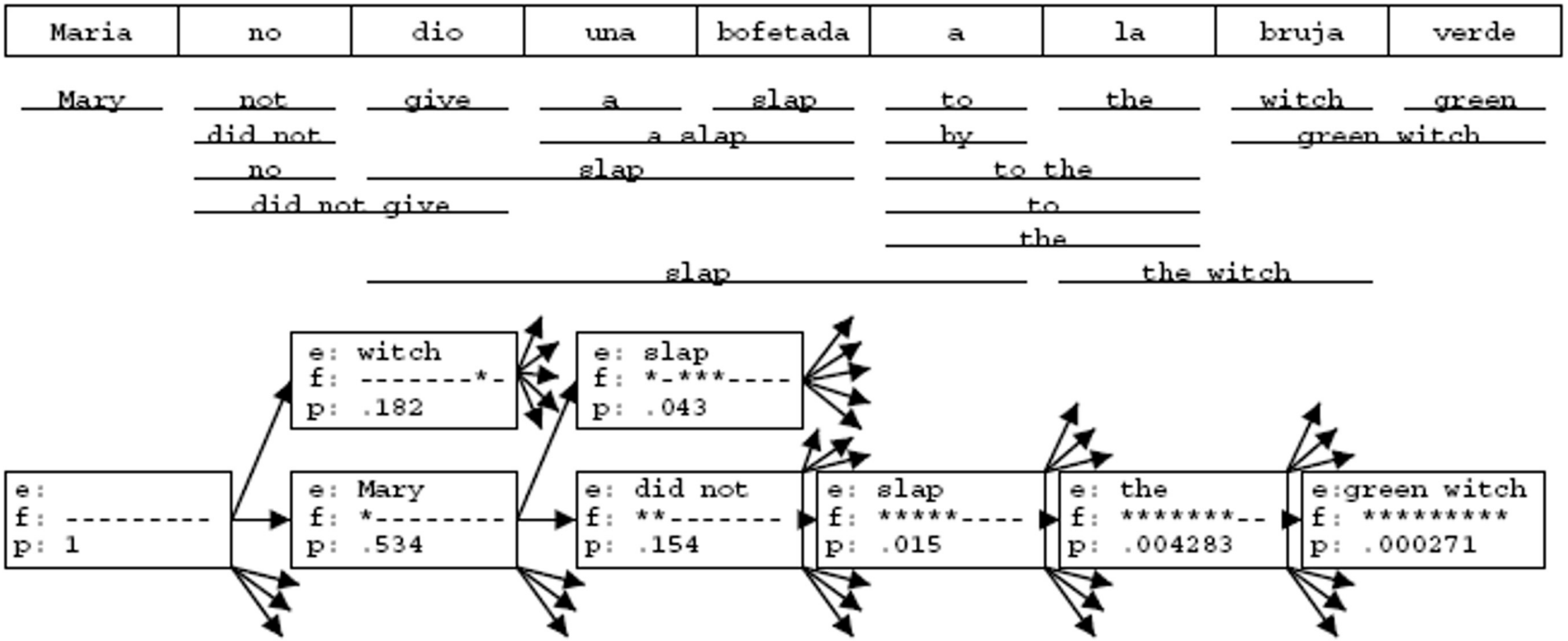
Expansion d’hypothèses



Expansion d'hypothèses



Expansion d'hypothèses

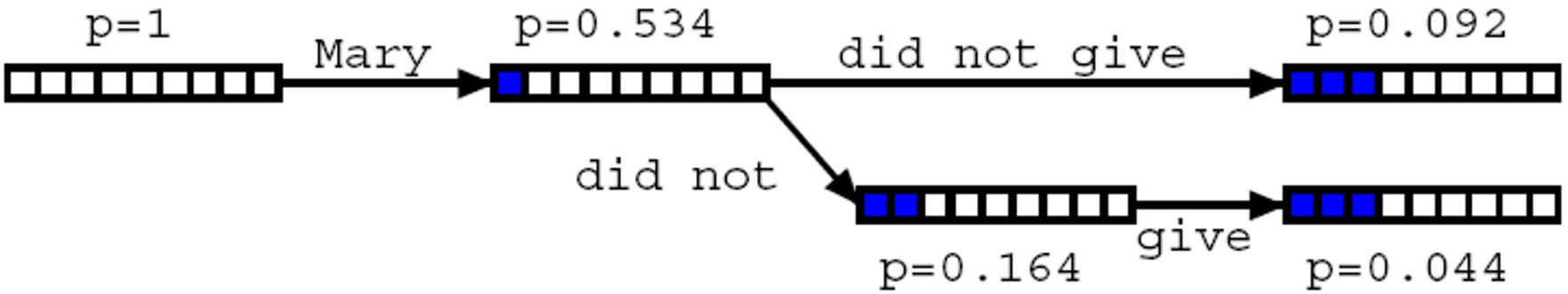


Explosion de l'espace de recherche

- **Le nombre d'hypothèses croit exponentiellement avec le nombre de mots dans la phrase source**
- **Le processus de décodage est un problème NP-complet [Knight, 1999]**
- **Besoin de réduire l'espace de recherche**
  - **Recombinaison d'hypothèses**
  - **Elagage (pruning)**

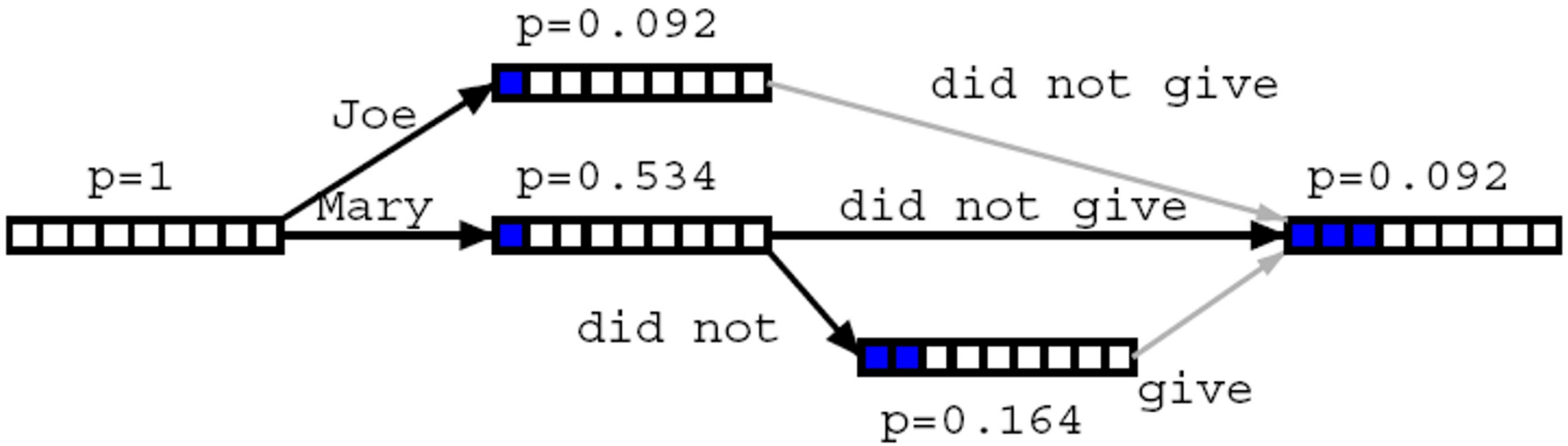
Recombinaison d'hypothèses

- **Différents chemins mènent à la même hypothèse partielle**
  - **Supprimer le chemin le moins probable**



Recombinaison d'hypothèses

- Les chemins n'ont pas besoin d'être strictement identiques
- On peut supprimer un chemin si
  - Les 2 derniers mots cible (anglais) correspondent (ML)
  - La couverture en mots source correspond (futurs chemins)

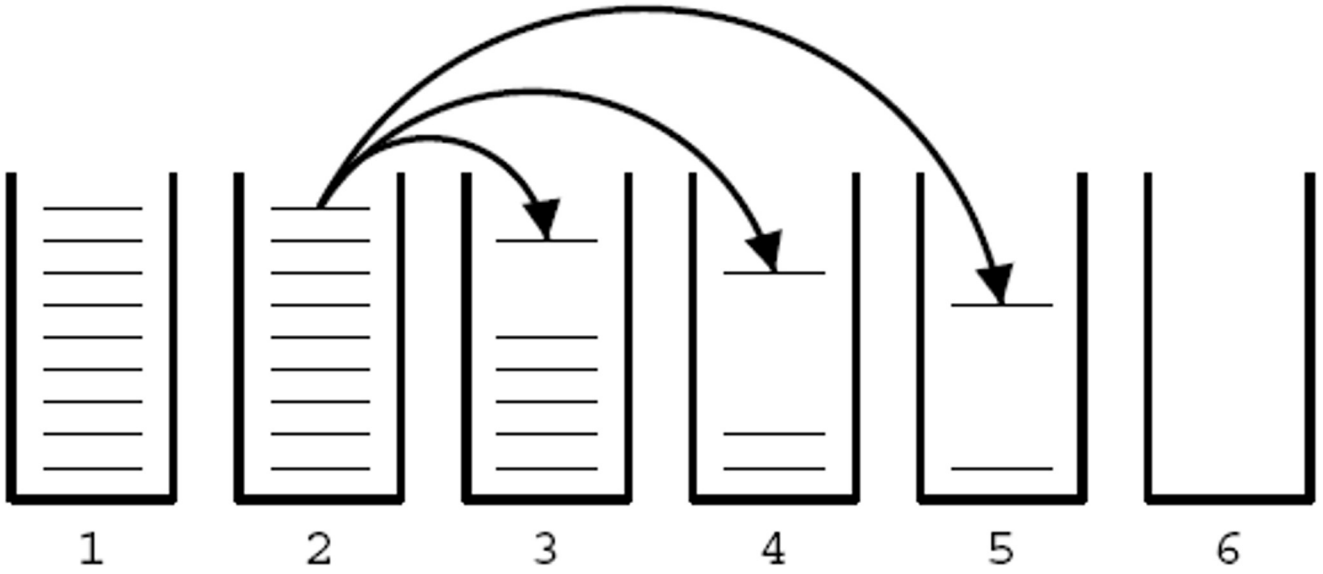


Elagage (pruning)

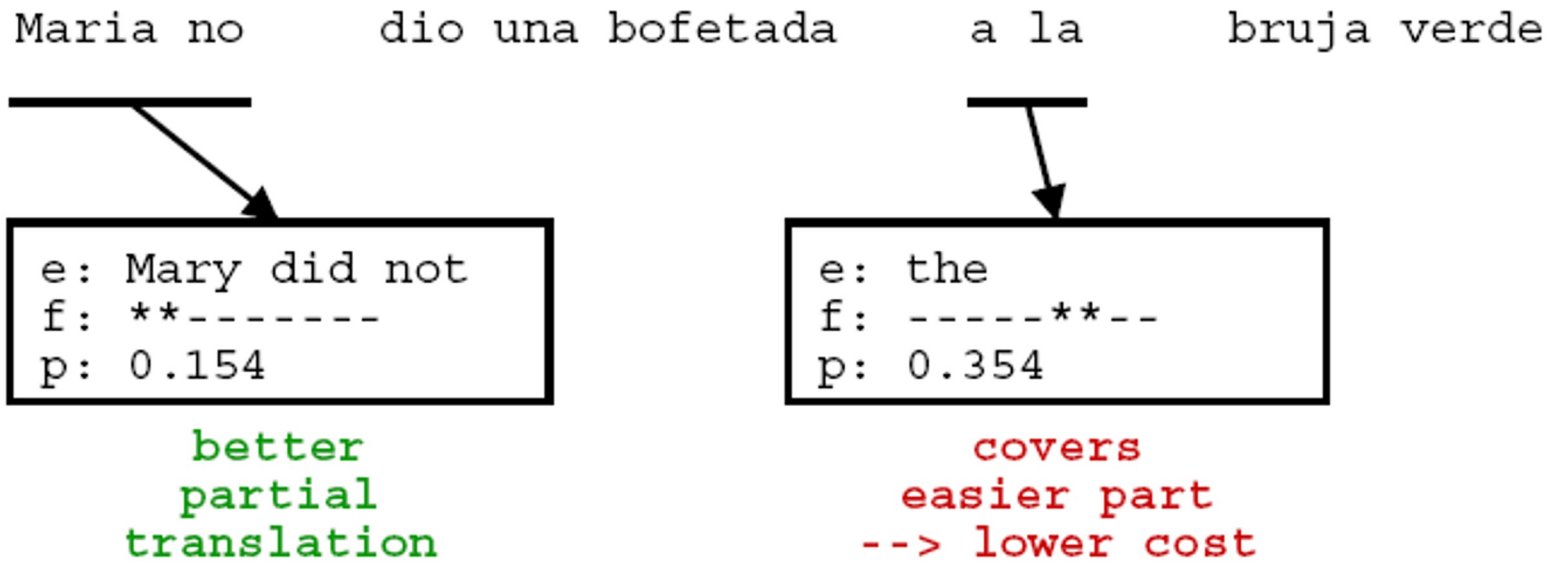
- Organiser les hypothèses en piles par
  - **mêmes mots sources couverts**
  - **même nombre de mots sources couverts**
  - **même nombre de mots cibles traduits**
- Comparer les hypothèses dans les piles, supprimer les mauvaises
  - ***histogram pruning*: garder les n meilleures hypothèses pour chaque pile (e.g., n=100)**
  - ***threshold pruning*: garder les hypothèses qui ont au plus un score égal à x fois le score de la meilleure hypothèse (e.g. x= 0.001)**

Exemple

- **Piles fondées sur le nombre de mots étrangers traduits**

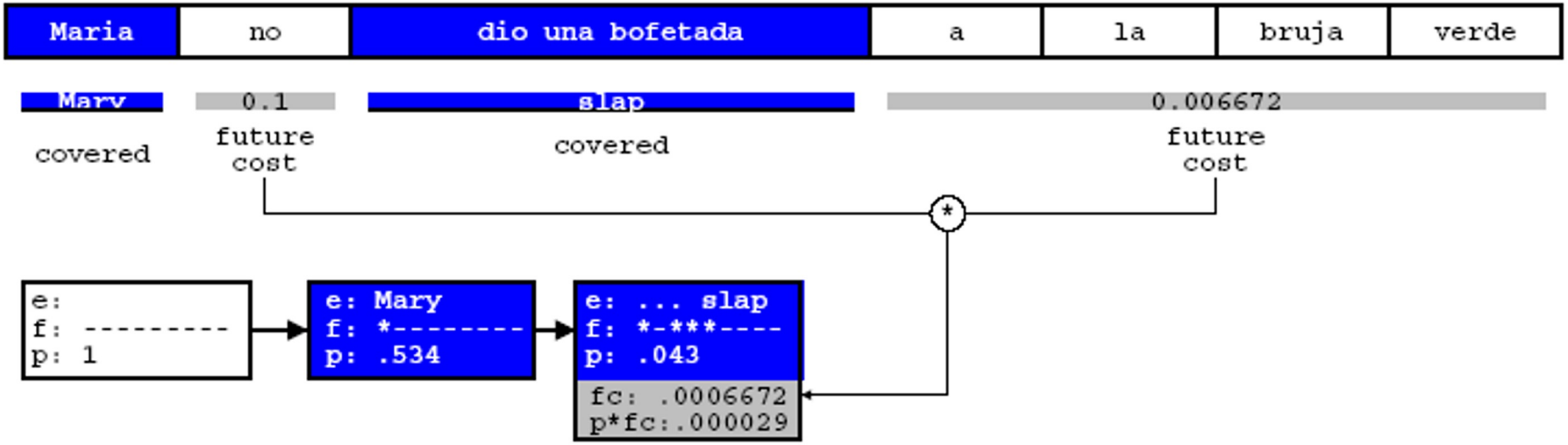


Comparer les hypothèses



Une hypothèse qui couvre une partie « facile » risque d'être préférée au détriment d'une bonne traduction partielle
=>Il faut considérer les coûts futurs des parties non traduites

Estimer les scores futurs



-Utiliser le score futur pour élaguer les hypothèses
-Ajouter score futur & score passé pour décider d'élaguer une hypothèse ou pas

Outils disponibles…

- **GIZA++**
- **Moses**
- **SRI-LM**

Outils commerciaux

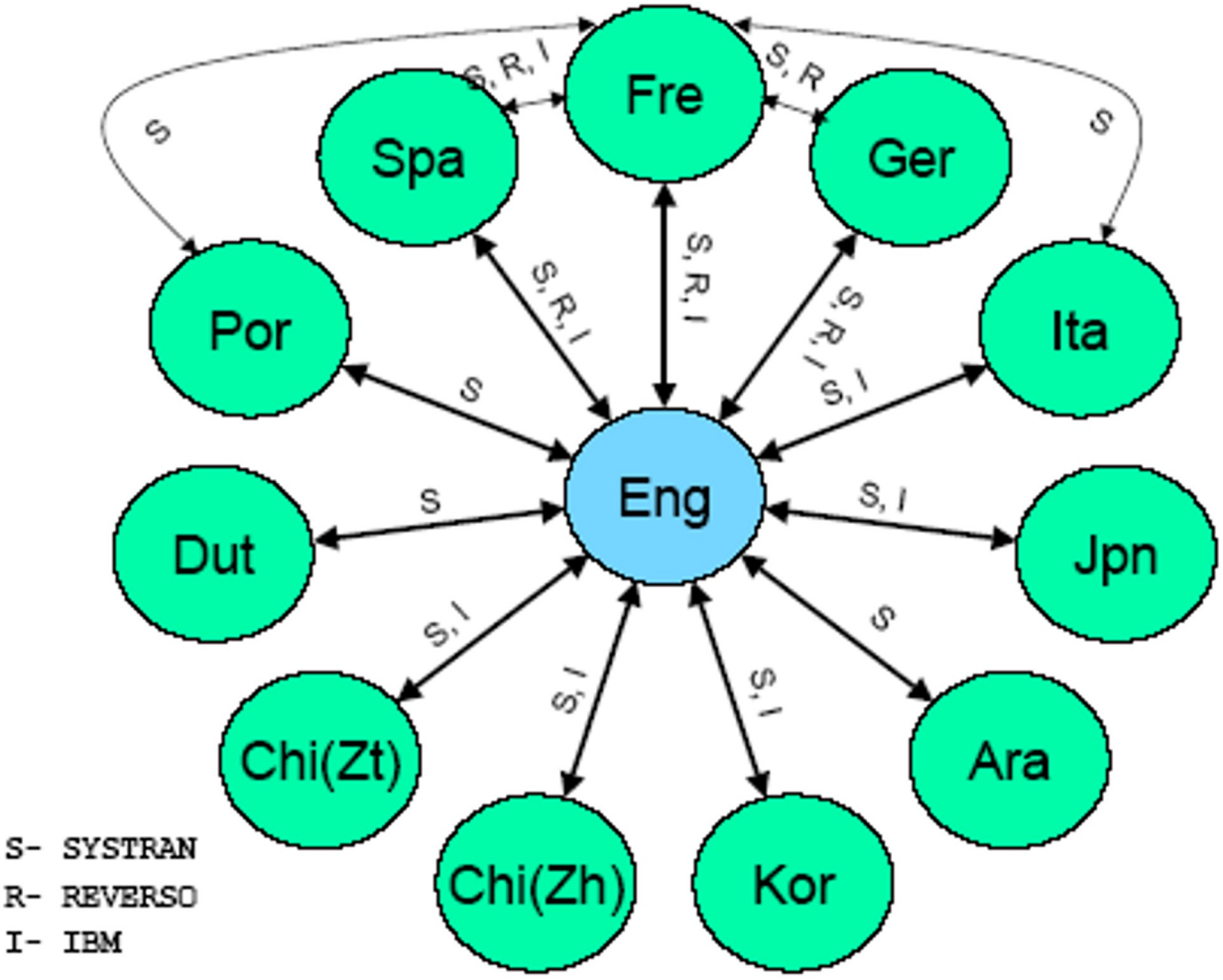


Fig. 1 : couples de langues de Systran, Reverso, et IBM disponibles sur le web