Représentation et analyse automatique des discontinuités syntaxiques dans les corpus arborés en constituants du français

Maximin Coavoux[1,2] – Benoît Crabbé[1,2,3]

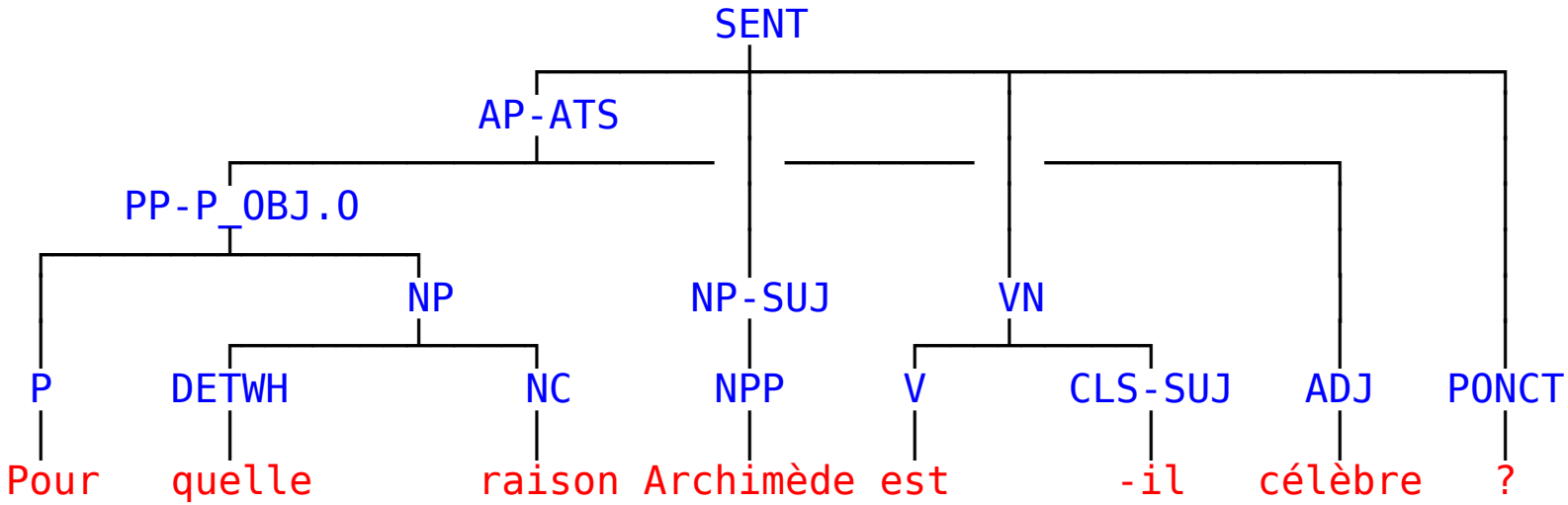
[1]Univ Paris Diderot – Sorbonne Paris Cité (SPC)
[2]Laboratoire de Linguistique Formelle (LLF, CNRS)
[3]Institut Universitaire de France (IUF)

TALN – Orléans – Juin 2017

Laboratoire de linguistique formelle

Introduction

Contexte: Analyse syntaxique en **constituants discontinus**



Contributions:

- Corpus en constituants discontinus pour le français
  - Obtenus par conversion de corpus existants (French Treebank, French Question Bank, Sequoia Treebank)
- Analyseur syntaxique en constituants discontinus
  - Analyse morphologique et fonctionnelle réalisée conjointement
  - Architecture multi-tâche

Outline

Arbres discontinus: motivations

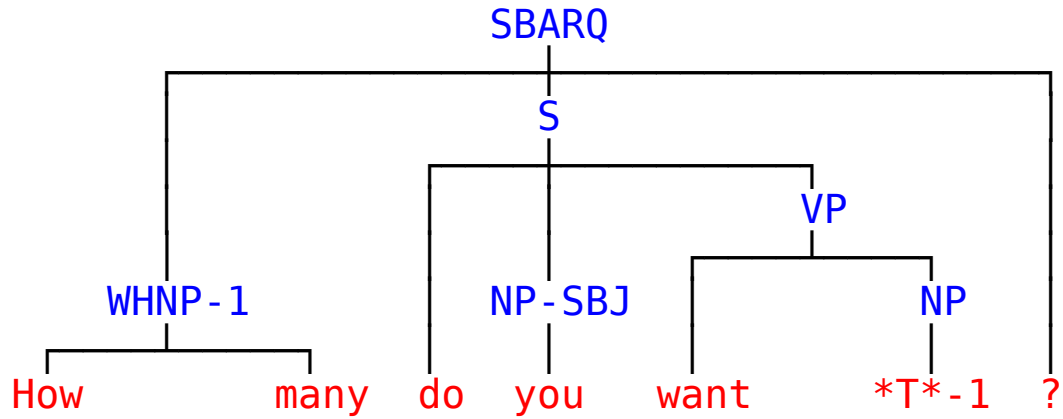
- Représentation unifiée pour les phénomènes de variation d'ordre des mots et d'extraction
- Dépendances à longue distance

Mais ...

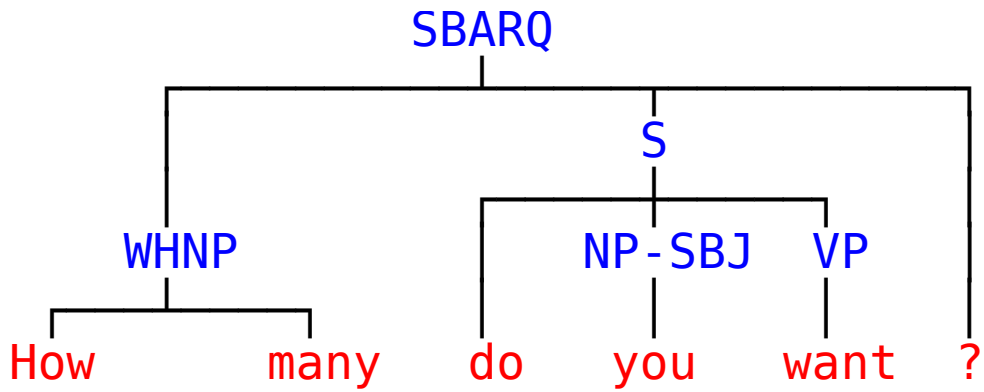
- Plus difficile à analyser en théorie
  - Grammaires légèrement sensibles au contexte ($\supset$ CFG)
  - LCFRS binaires: parsing tabulaire **exact** en $\mathcal{O}(n^{3f})$ ($f > 1$: fan-out, mesure le 'degré' de discontinuité)

Représentations alternatives

- Chemins fonctionnels (LFG)
- Traces indexées (Penn Treebank)



Habituellement, les parsers en constituants utilisent des versions prétraitées des corpus où ces informations sont retirées



Conversions

- Données:
  - French Treebank (FTB, Abeillé et al. 2003)
  - French Question Bank (FQB, Seddah et Candito, 2016)
  - Sequoia Treebank (SEQ, Candito et al. 2014)

- Schéma d'annotation du FTB

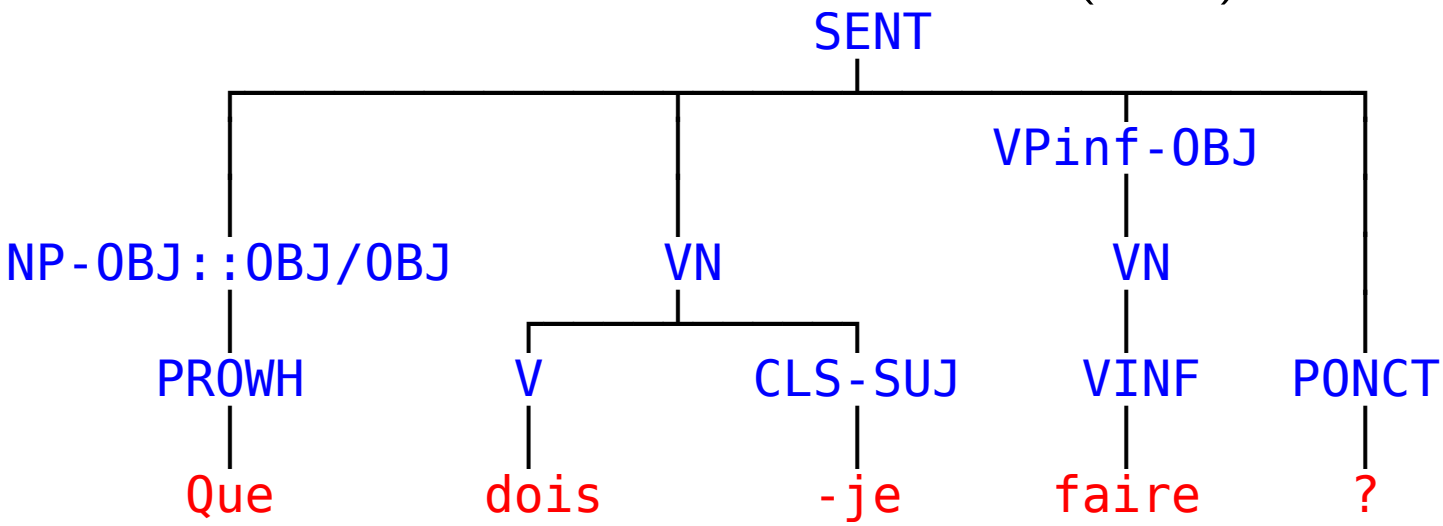
- Candito et Seddah (2012): ajouts de **chemins fonctionnels** sur les constituants pour certains types de dépendances à longue distance.
  - On utilise ces chemins fonctionnels pour convertir les corpus vers un format en constituants discontinus.

Phénomènes cibles

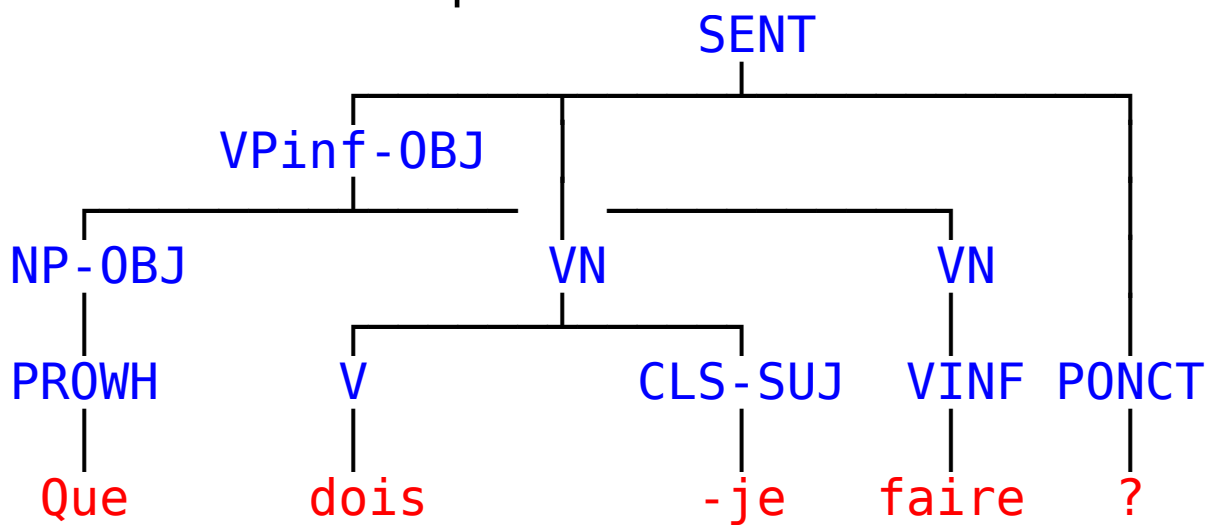
- Relatives
  - *Conseiller municipal socialiste, il était déjà cependant directeur général de la banque* **qu**'*il va* **présider**. FTB
- Questions
  - **Pour quel type de logement** *puis-je* **bénéficier** *d'une aide au logement ?* FQB
- Clivées
  - *C'est donc toute la vie industrielle du bassin de Saint-Dizier, sans oublier les papeteries de Jeand'Heurs, les carrières de Savonnières,* **que** *les visiteurs du lavoir pourront* **découvrir**. SEQ
- Dislocations
  - **À un "déterminisme technologique", développé notamment par Alain Touraine, où l'histoire des techniques s'impose à l'organisation du travail et à l'emploi,** *on doit* **opposer** *"une dialectique à trois termes, technologie, organisation et travail".* FTB
- Clitiques
  - *La crise, tout le monde la sentait, mais ce mois terrible* **en** *fait prendre la* **mesure**. FTB

Algorithme de conversion

Annotations de Candito et Seddah (2012):



Arbre après transformation:



Quelques statistiques

Quels phénomènes sont à l'origine des discontinuités ?

Phénomène	FTB-TRAIN		FQB		SEQUOIA	
Propositions relatives	183	72%	4	5%	36	77%
Questions	8	3%	83	95%	2	4%
Constructions clivées	5	2%	0	0%	4	9%
Dislocations	1	$< 1\%$	0	0%	1	2%
en	57	22%	0	0%	4	9%
Total	254	100%	87	100%	47	100%

Phénomènes les plus fréquents

- Relatives
- Questions (FQB)
- *en*

Quelques statistiques

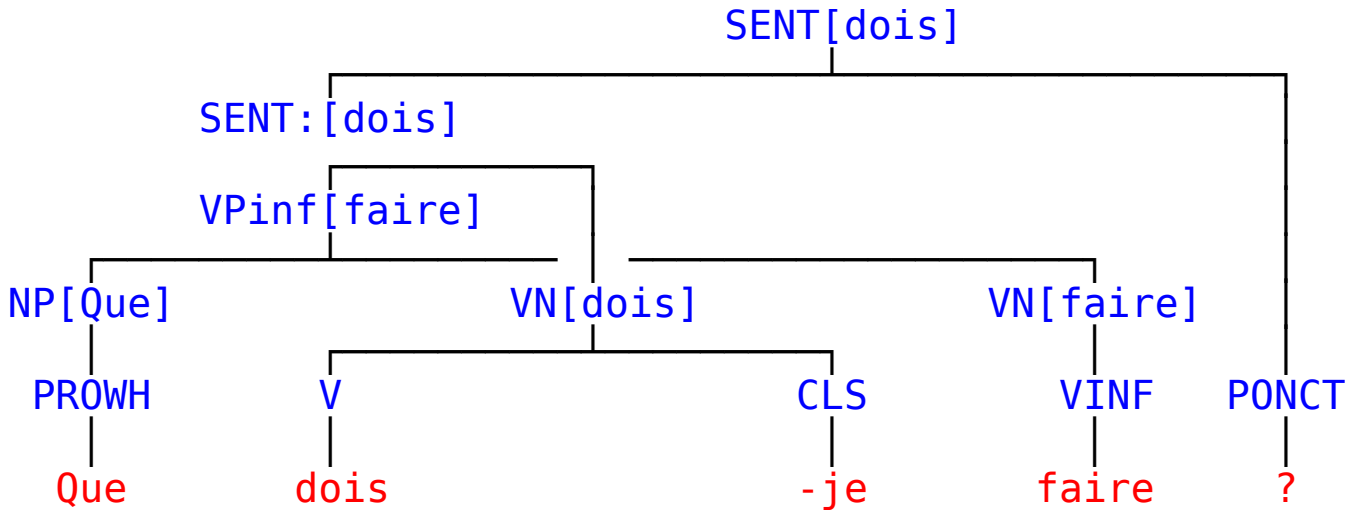
	FTB-TRAIN	FQB-ALL	SEQUOIA
Tokens	443,113	23,222	67,038
Phrases	1,4759	2,289	3,099
Phrases avec discontinuité	253 (**1.71%**)	88 (**3.84%**)	46 (**1.48%**)
Constituants	298,025	15,966	47,586
Constituants discontinus	374 (**0.13%**)	94 (**0.59%**)	70 (**0.15%**)

- Discontinuités très rares dans les corpus obtenus
  - 4 fois plus fréquentes dans le French Question Bank que dans les autres corpus
  - mais seuls certains phénomènes sont pris en compte

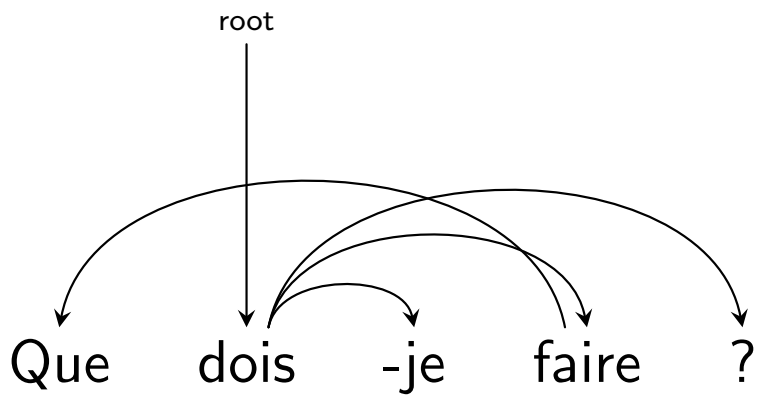
Plan

Hypothèses

- Arbres binaires et lexicalisés

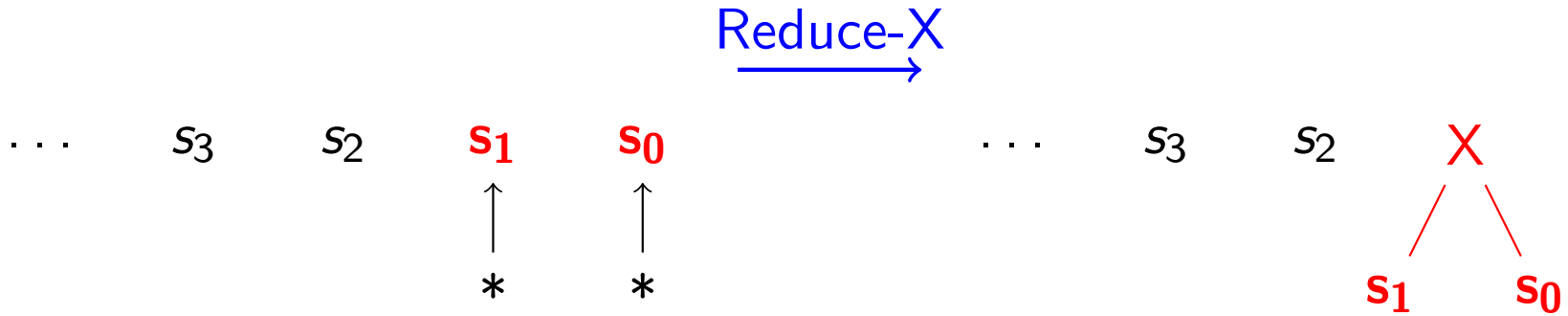


- `SENT:` symbole temporaire (introduit par binarisation)
- Encode implicitement un arbre en dépendances (non projectif)

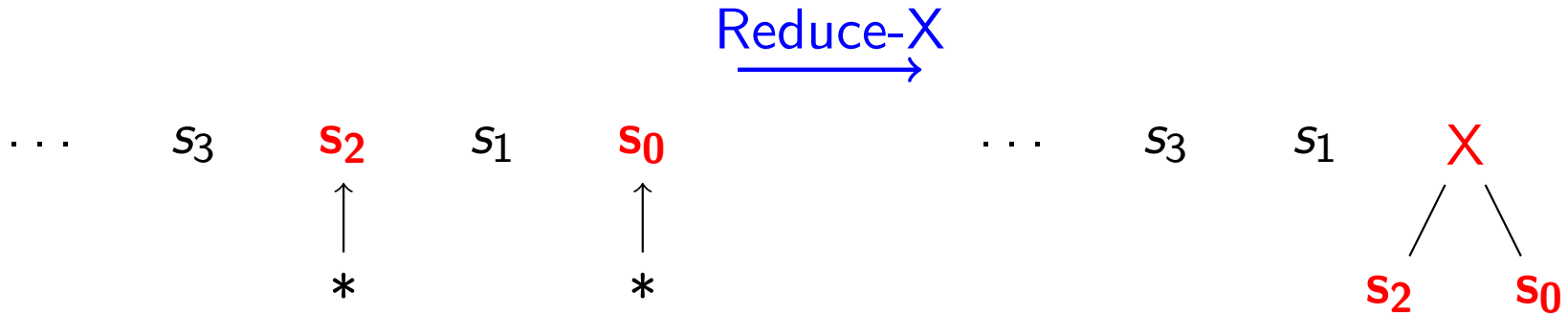


Construire des constituants discontinus

Shift-Reduce standard: les réductions s'appliquent aux 2 éléments en sommet de pile



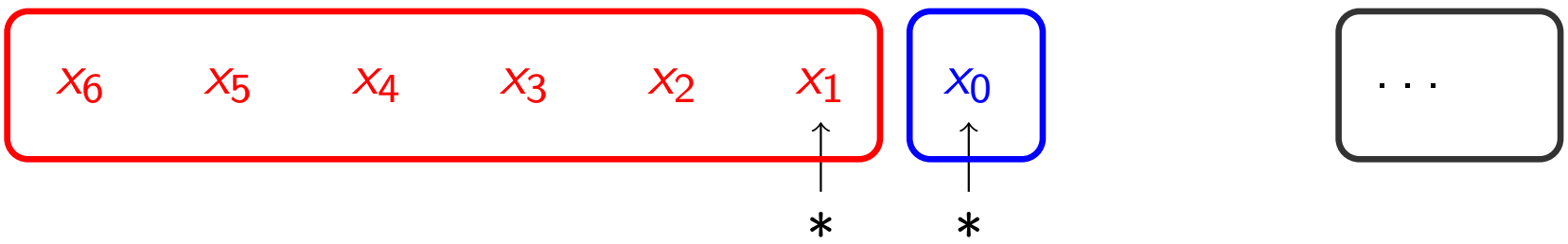
Pour les discontinuités: on voudrait faire des réductions avec n'importe quel constituant de la pile



Shift-Reduce+Gap (Coavoux et Crabbé, 2017)

Solution :

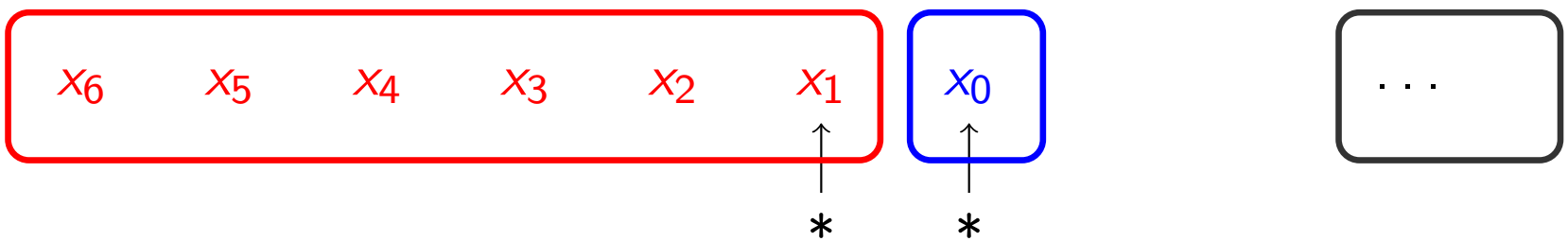
- Extension de Shift-Reduce: **Shift-Reduce+Gap**
- On divise la pile habituelle en 2 structures de données (Pile+File).



- Les réductions s'appliquent aux sommets respectifs des 2 structures
- Nouvelle action (GAP) pour contrôler les mouvements entre les 2 structures

Shift-Reduce+Gap: Pile File Buffer

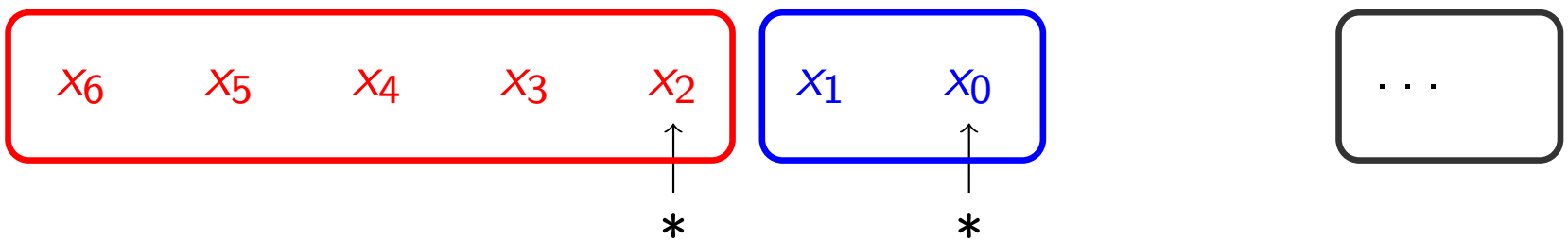
Exemple: créer un constituant X avec pour descendants x_0 et x_3



Shift-Reduce+Gap: Pile File Buffer

Exemple: créer un constituant X avec pour descendants x_0 et x_3

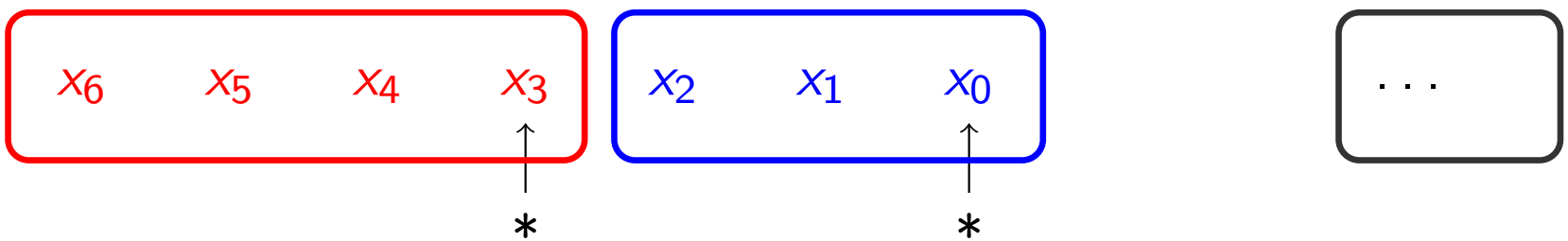
$\rightarrow$ GAP



Shift-Reduce+Gap: Pile File Buffer

Exemple: créer un constituant X avec pour descendants x_0 et x_3

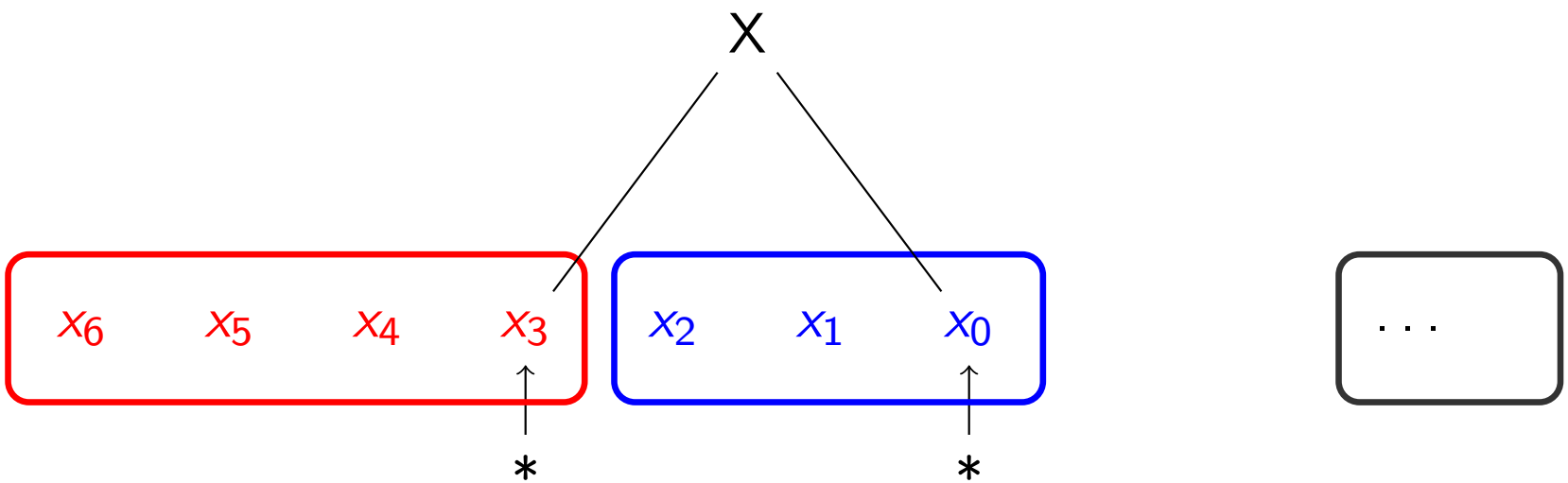
$\rightarrow$ GAP, GAP



Shift-Reduce+Gap: Pile File Buffer

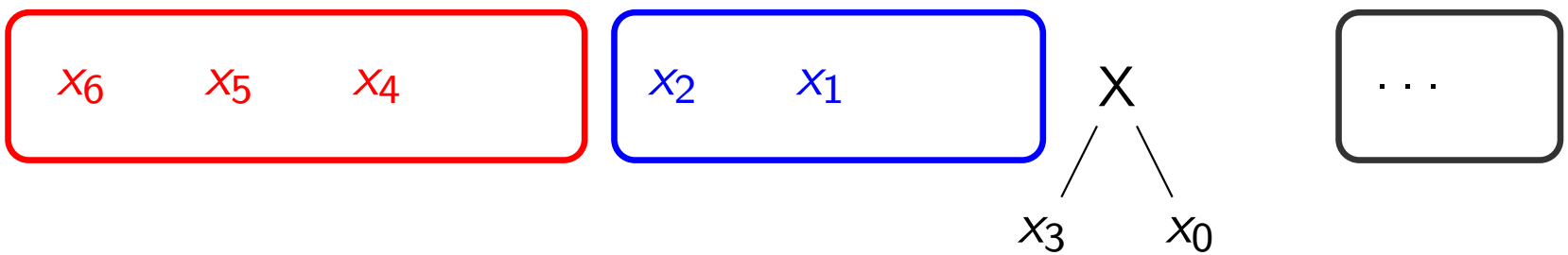
Exemple: créer un constituant X avec pour descendants x_0 et x_3

$\rightarrow$ GAP, GAP, REDUCE avec x_0 et x_3



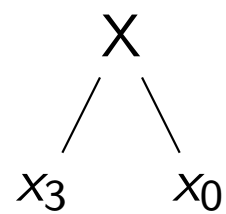
Shift-Reduce+Gap: Pile File Buffer

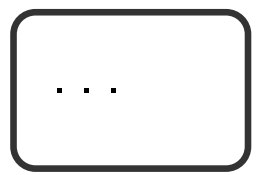
Exemple: créer un constituant X avec pour descendants x_0 et x_3
$\rightarrow$ GAP, GAP, REDUCE avec x_0 et x_3. Créer un nouveau noeud.



Shift-Reduce+Gap: Pile File Buffer

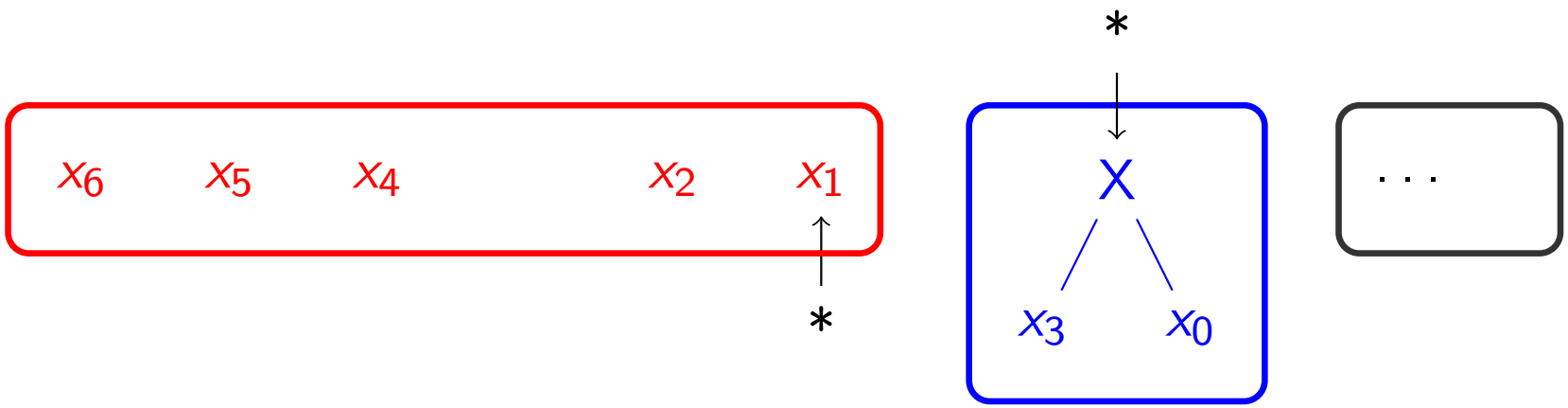
Exemple: créer un constituant X avec pour descendants x_0 et x_3 $\rightarrow$ GAP, GAP, REDUCE avec x_0 et x_3. Créer un nouveau noeud. Vider la File sur la Pile.





Shift-Reduce+Gap: Pile File Buffer

Exemple: créer un constituant X avec pour descendants x_0 et x_3 $\rightarrow$ GAP, GAP, REDUCE avec x_0 et x_3. Créer un nouveau noeud. Vider la File sur la Pile. Ajouter le nouveau noeud sur la File.



Système de transitions

- 3 structures de données: Pile et File (stockent des sous-abres), Buffer (stocke les tokens)

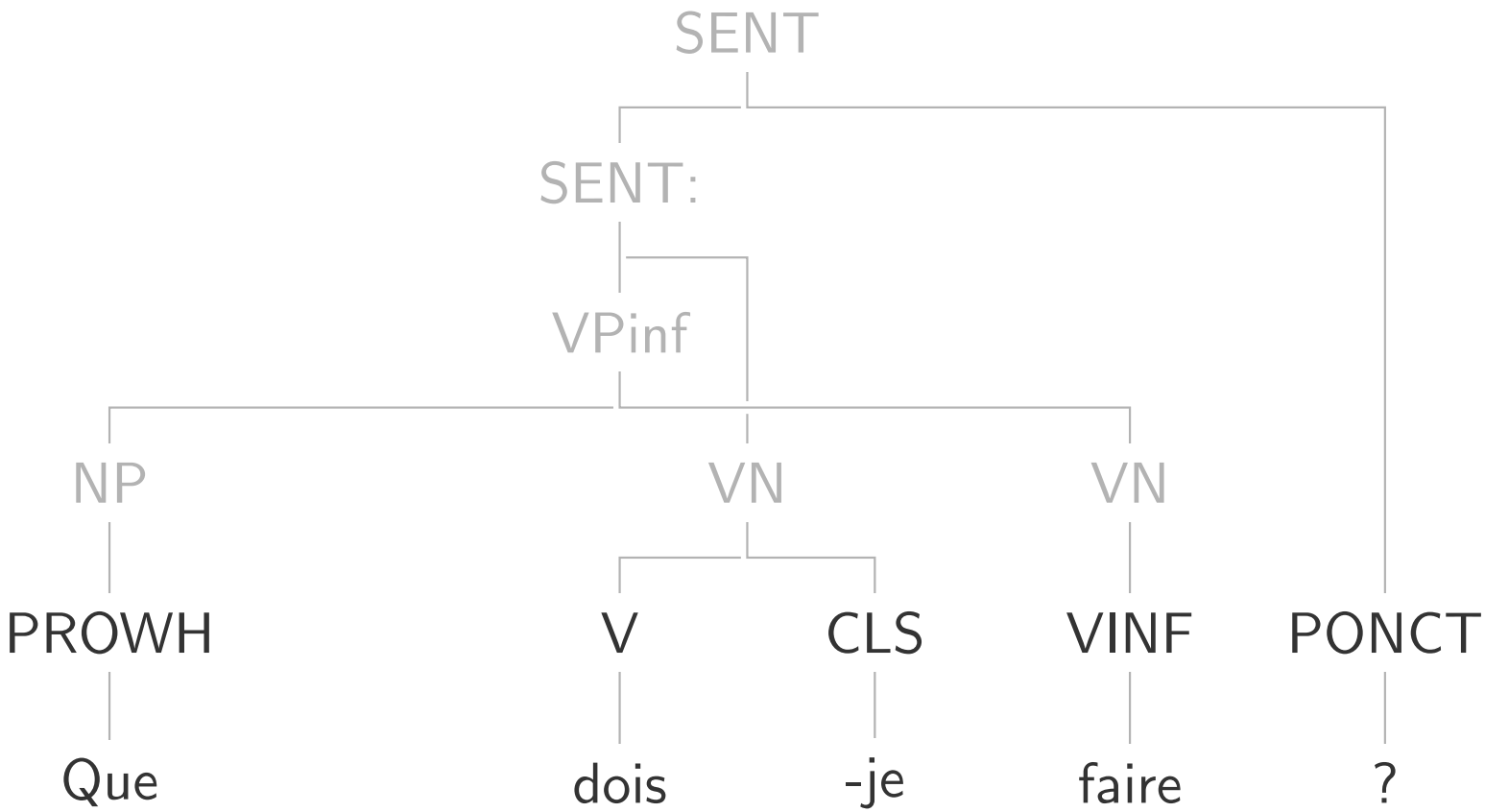
- Configuration = (Pile, File, Buffer)
  - Configuration initiale = ($\emptyset$, $\emptyset$, $[w_1, w_2 \dots w_n]$)
  - Configuration finale = ($\emptyset$, [A], $\emptyset$)
    - A = axiome

Transitions

	Input			Output		
Shift	(S,	D,	b_0|B)	(S|D,	$[b_0]$,	B)
Reduce-Left/Right(X)	(S|s_0,	D|d_0,	B)	(S|D,	[X],	B)
Gap	(S|s_0,	D,	B)	(S,	s_0|D,	B)

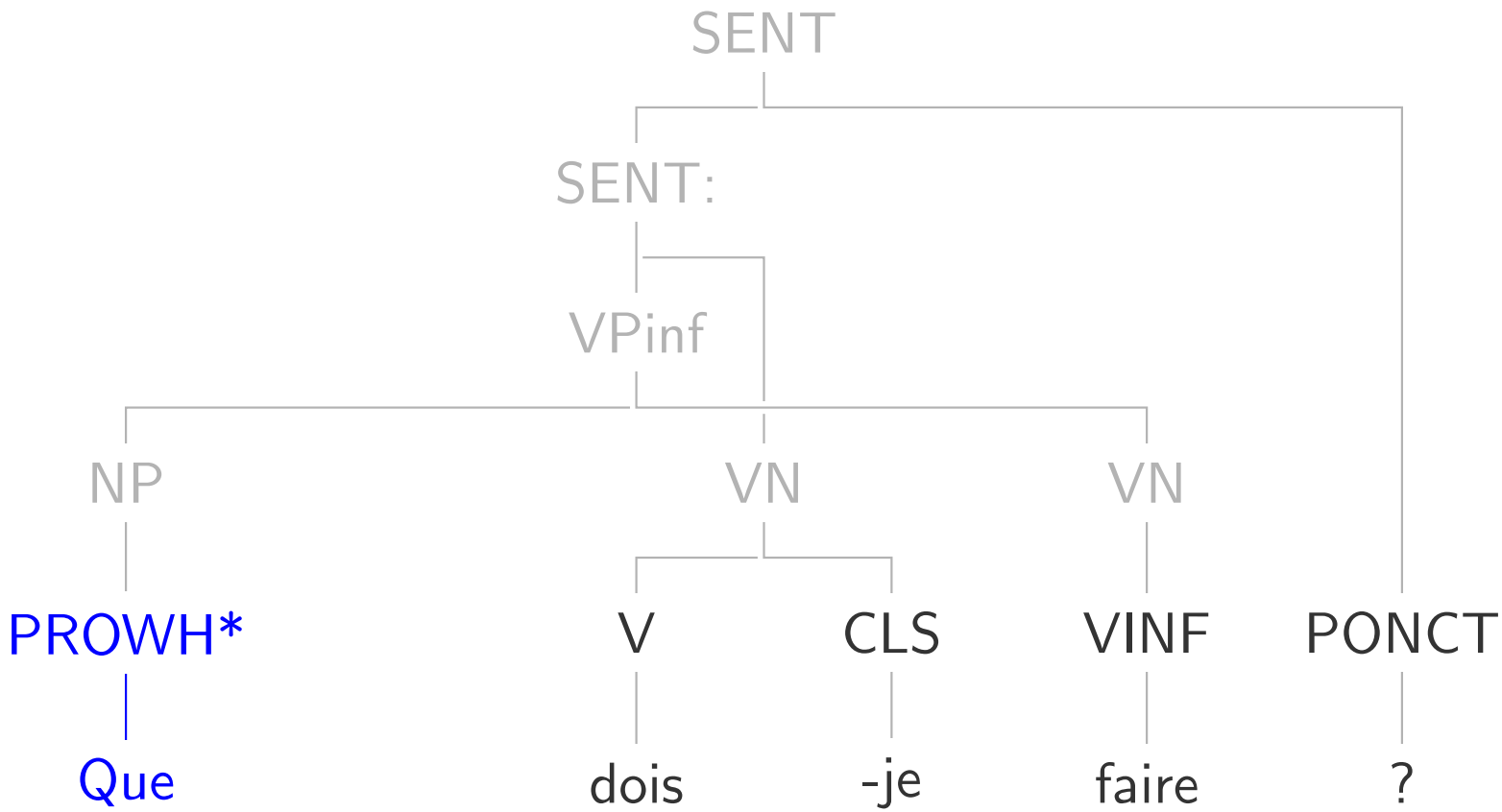
Shift-Reduce-Gap : Pile – File – Buffer

Initialisation



Shift-Reduce-Gap : Pile – File – Buffer

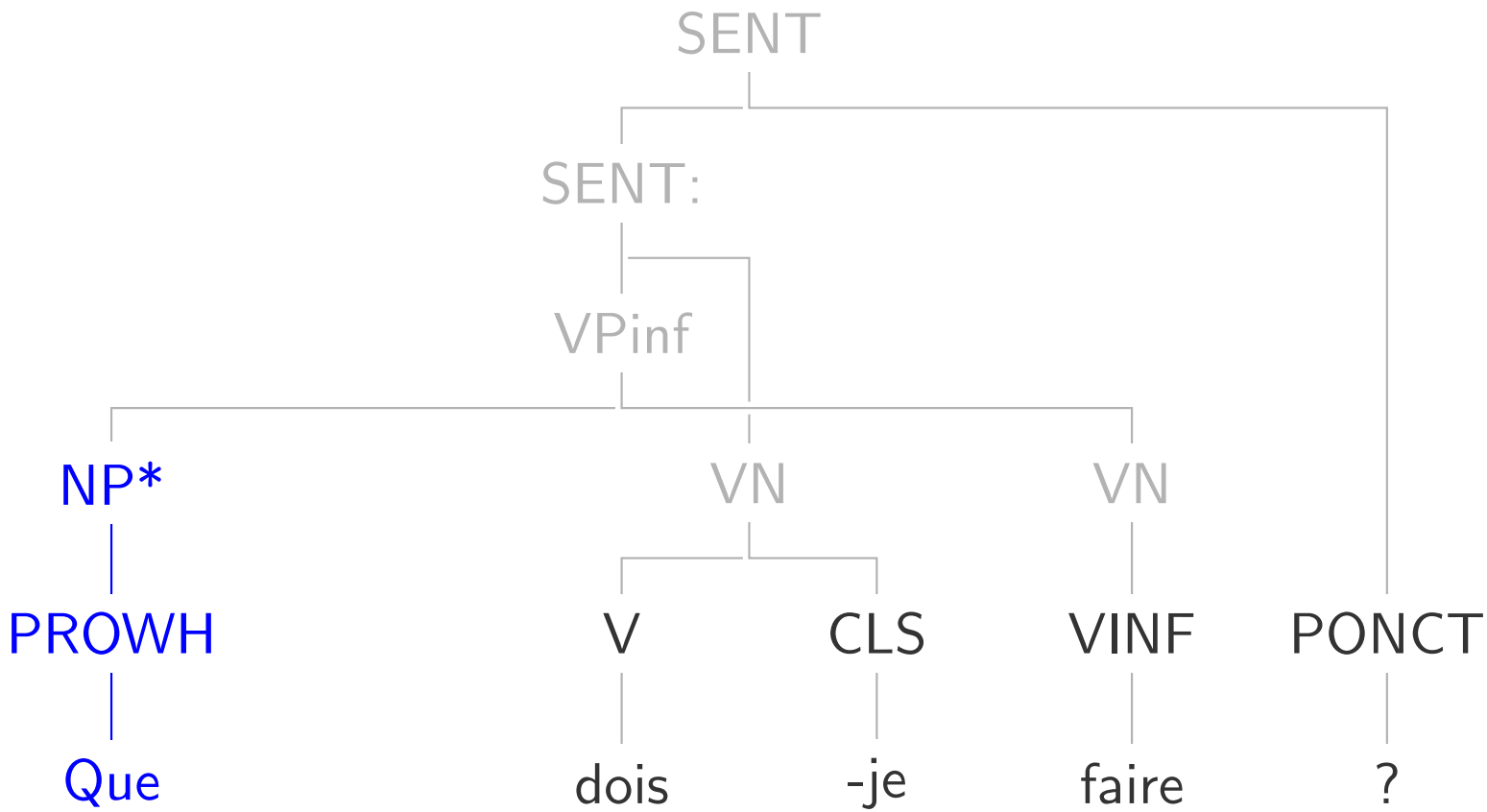
Shift



sh

Shift-Reduce-Gap : Pile – File – Buffer

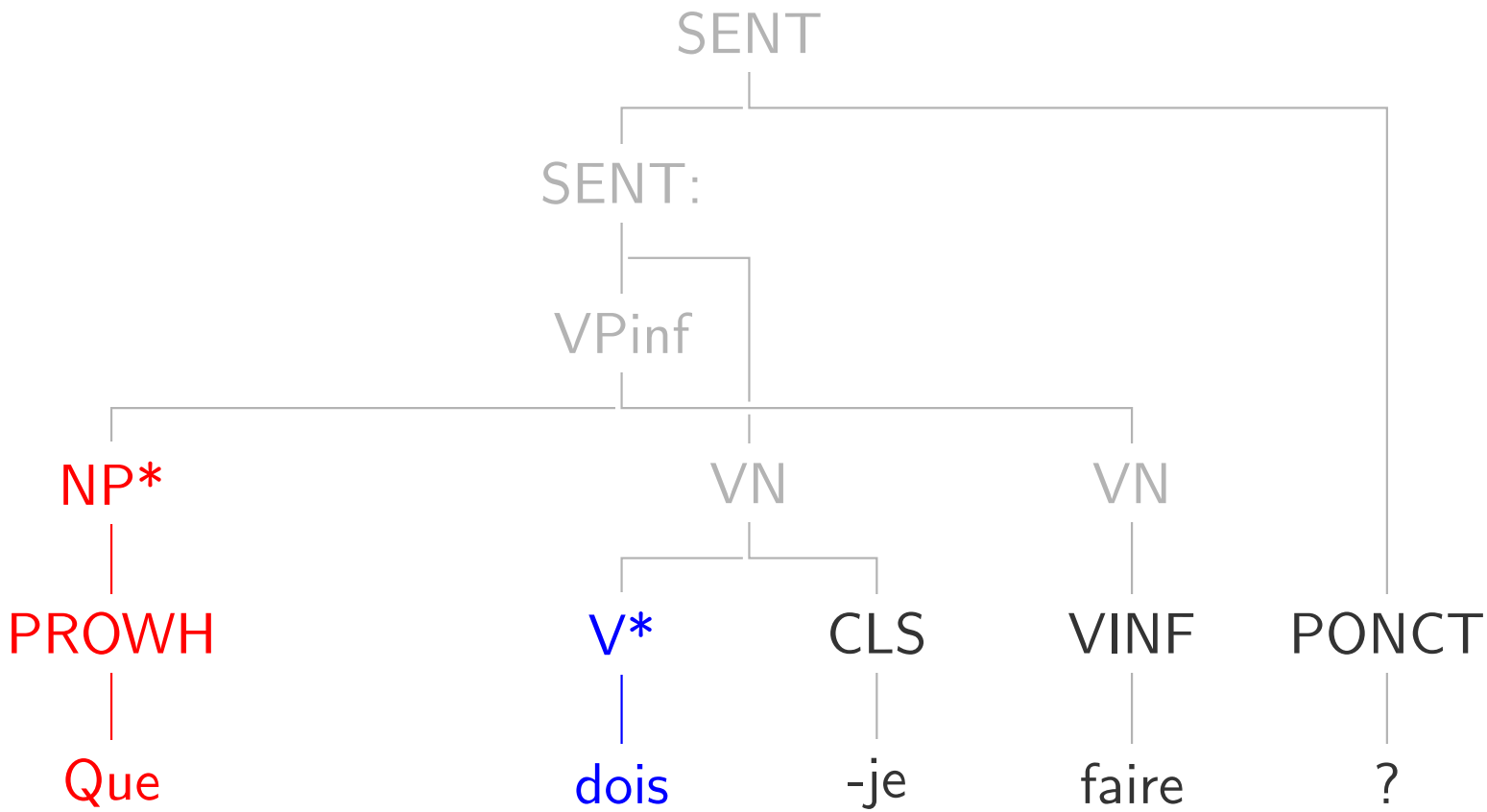
Reduce-Unary(NP)



sh, ru(NP)

Shift-Reduce-Gap : Pile – File – Buffer

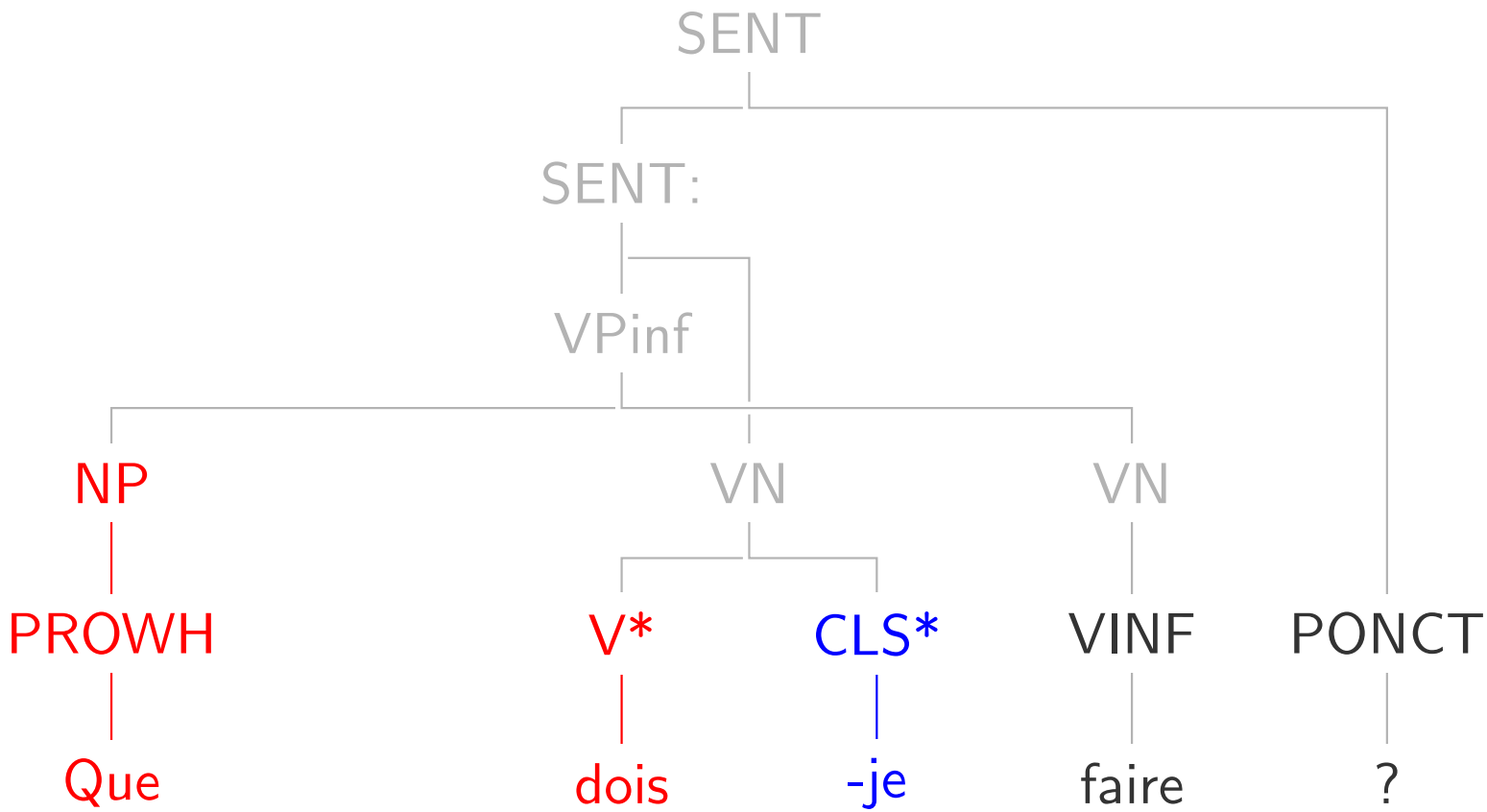
Shift



sh, ru(NP), sh

Shift-Reduce-Gap : Pile – File – Buffer

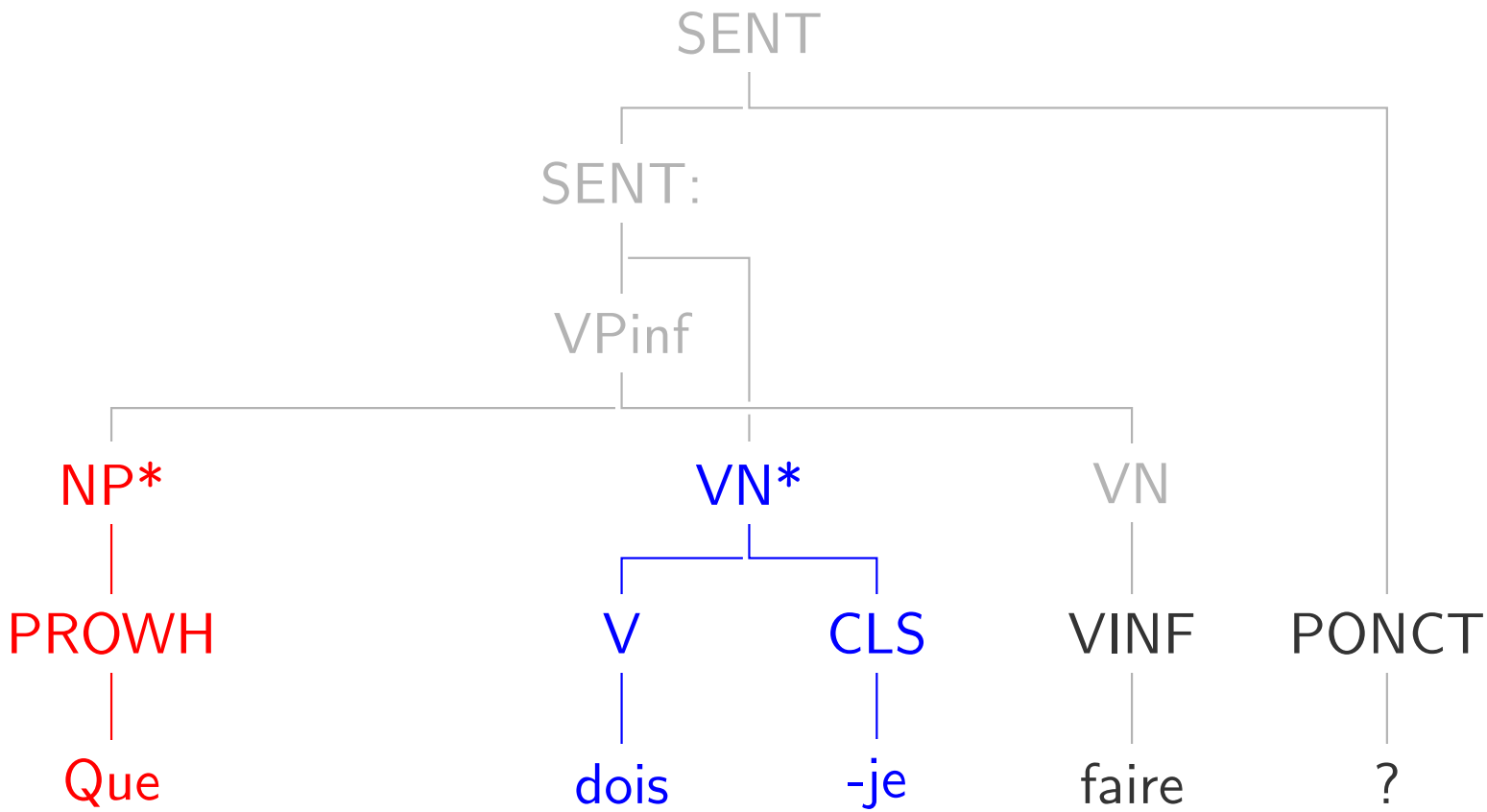
Shift



sh, ru(NP), sh, sh

Shift-Reduce-Gap : Pile – File – Buffer

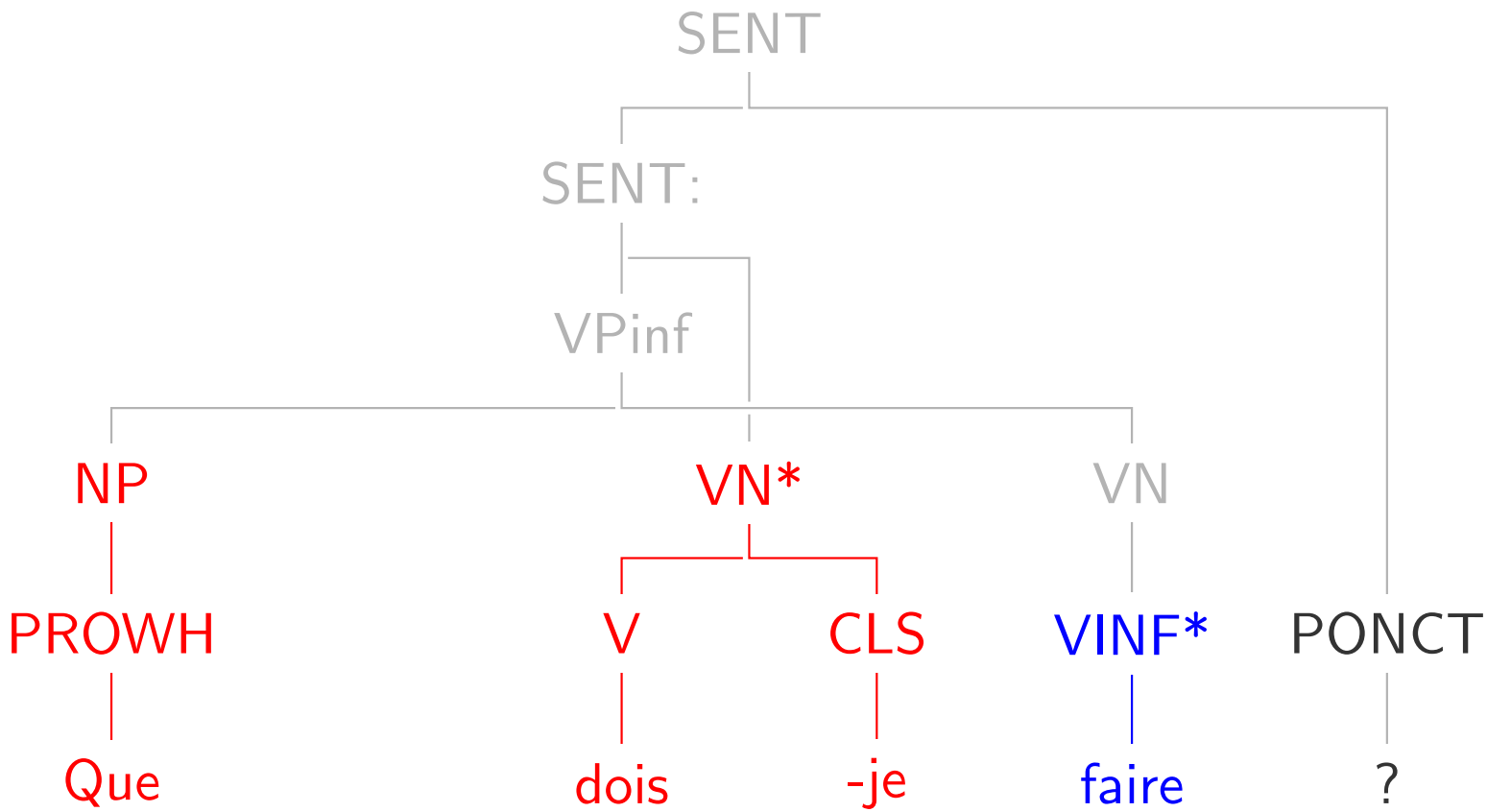
Reduce(VN)



sh, ru(NP), sh, sh, rl(VN)

Shift-Reduce-Gap : Pile – File – Buffer

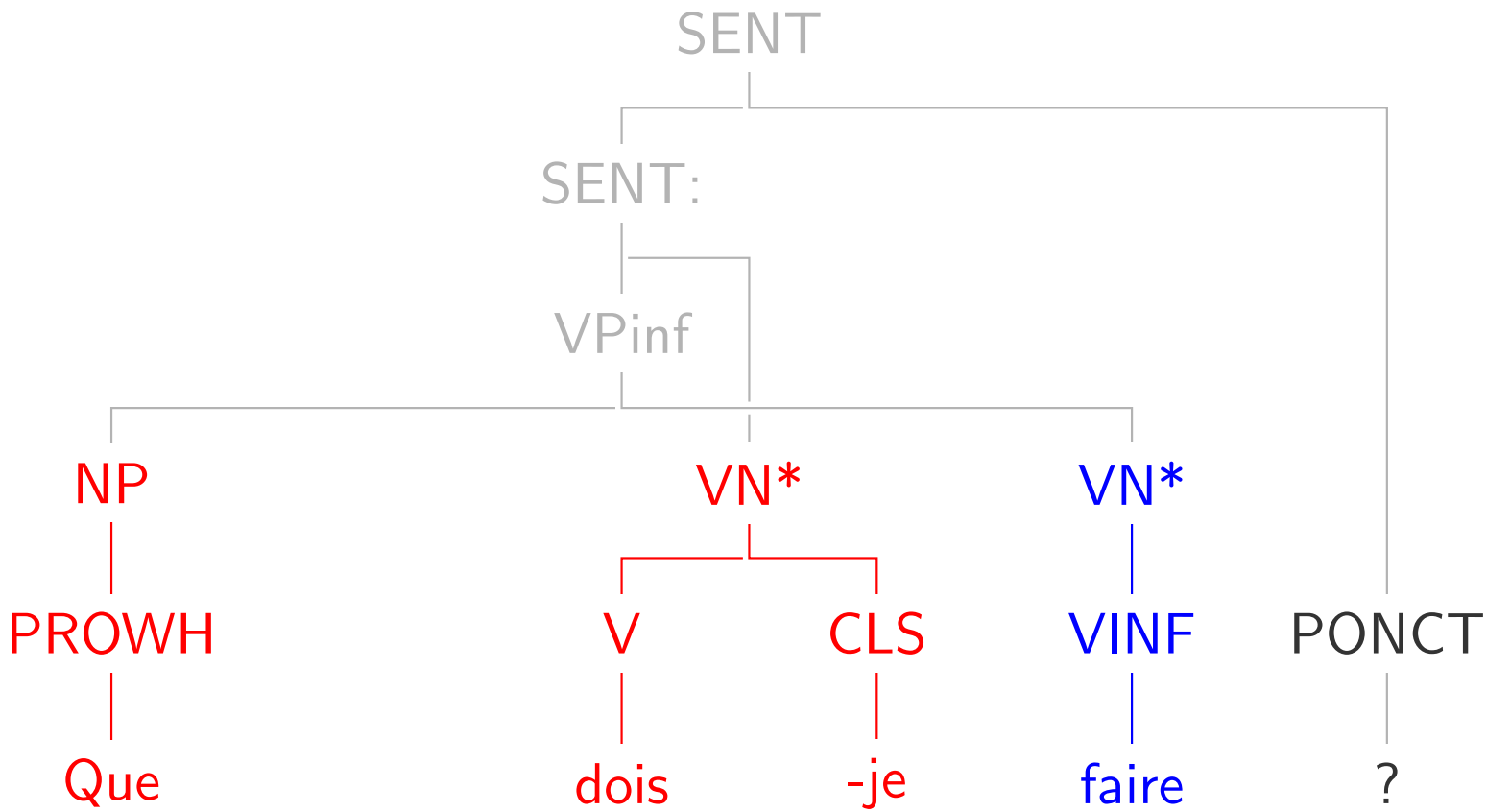
Shift



sh, ru(NP), sh, sh, rl(VN), sh

Shift-Reduce-Gap : Pile – File – Buffer

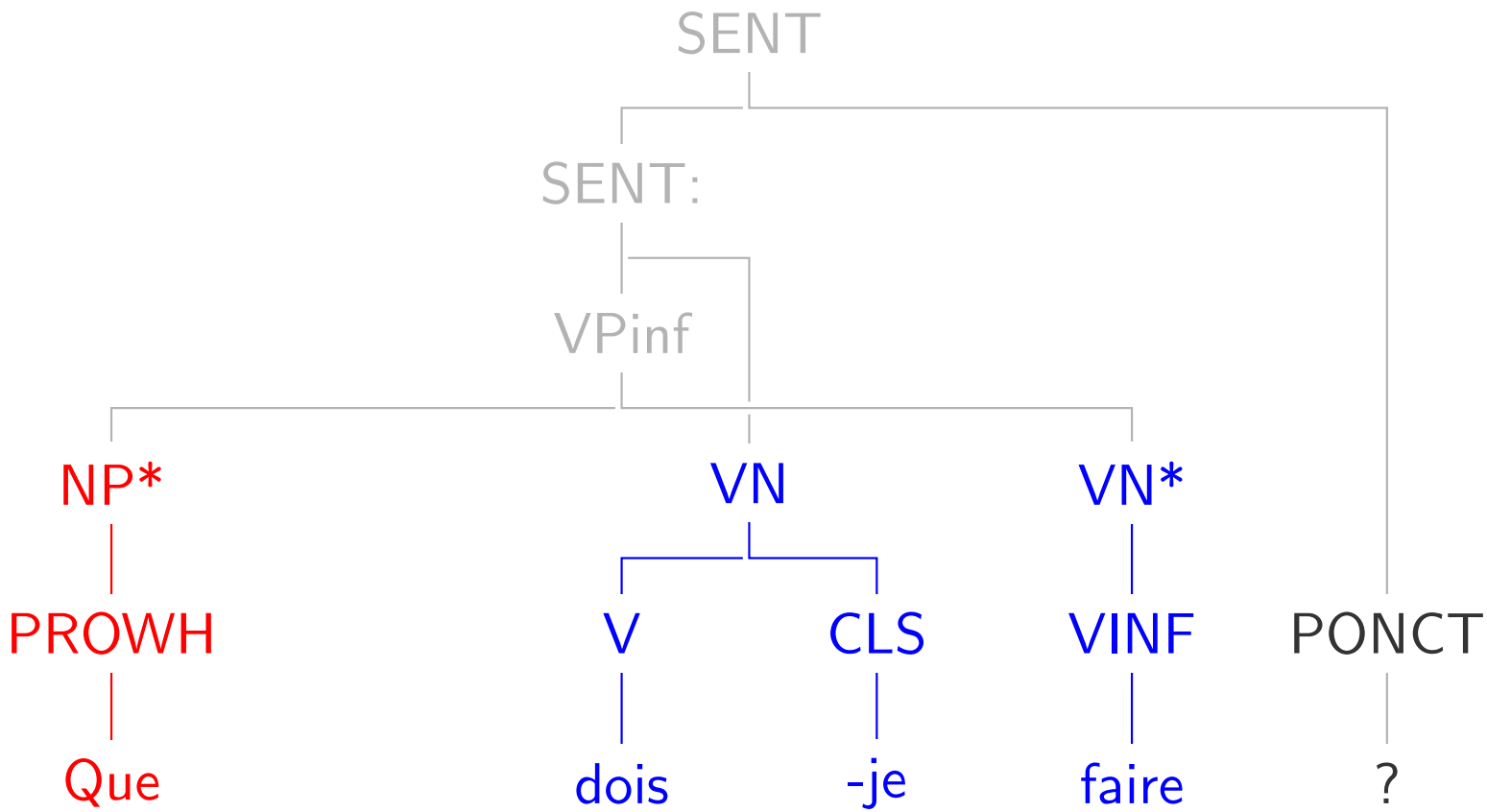
ReduceUnary(VN)



sh, ru(NP), sh, sh, rl(VN), sh, ru(VN)

Shift-Reduce-Gap : Pile – File – Buffer

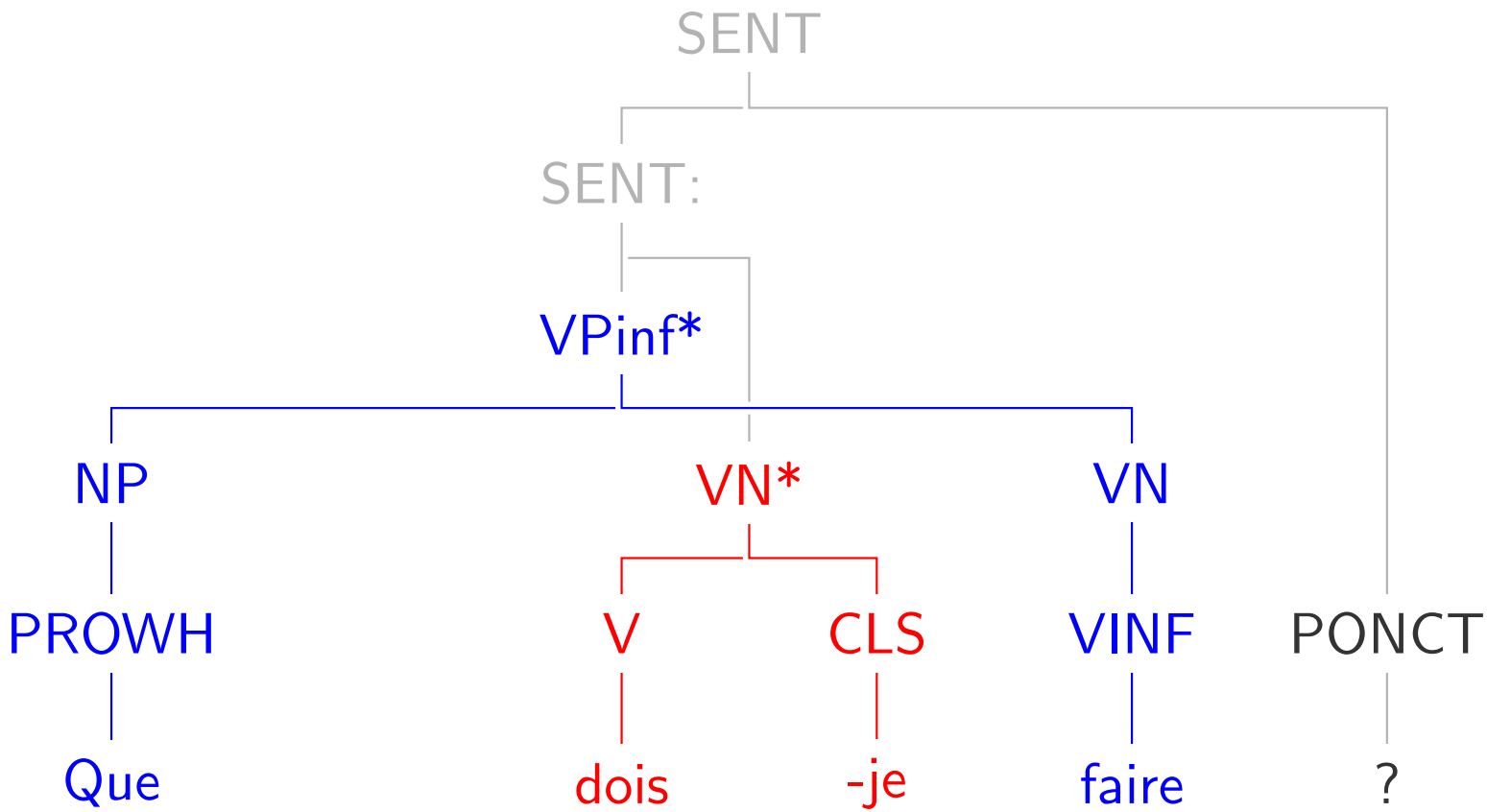
Gap



sh, ru(NP), sh, sh, rl(VN), sh, ru(VN), gap

Shift-Reduce-Gap : Pile – File – Buffer

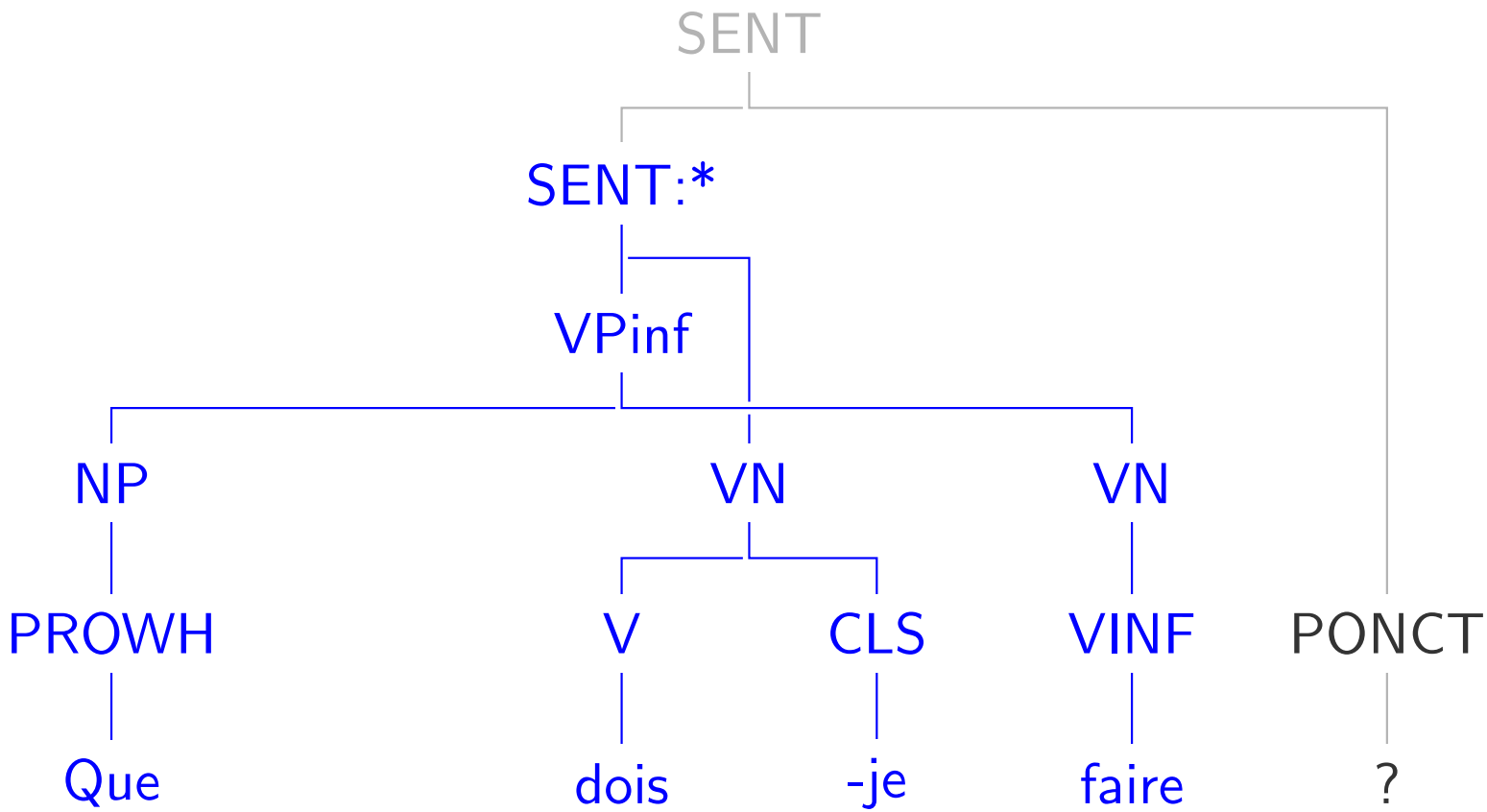
ReduceLeft(VPinf)



sh, ru(NP), sh, sh, rl(VN), sh, ru(VN), gap, rl(VPinf)

Shift-Reduce-Gap : Pile – File – Buffer

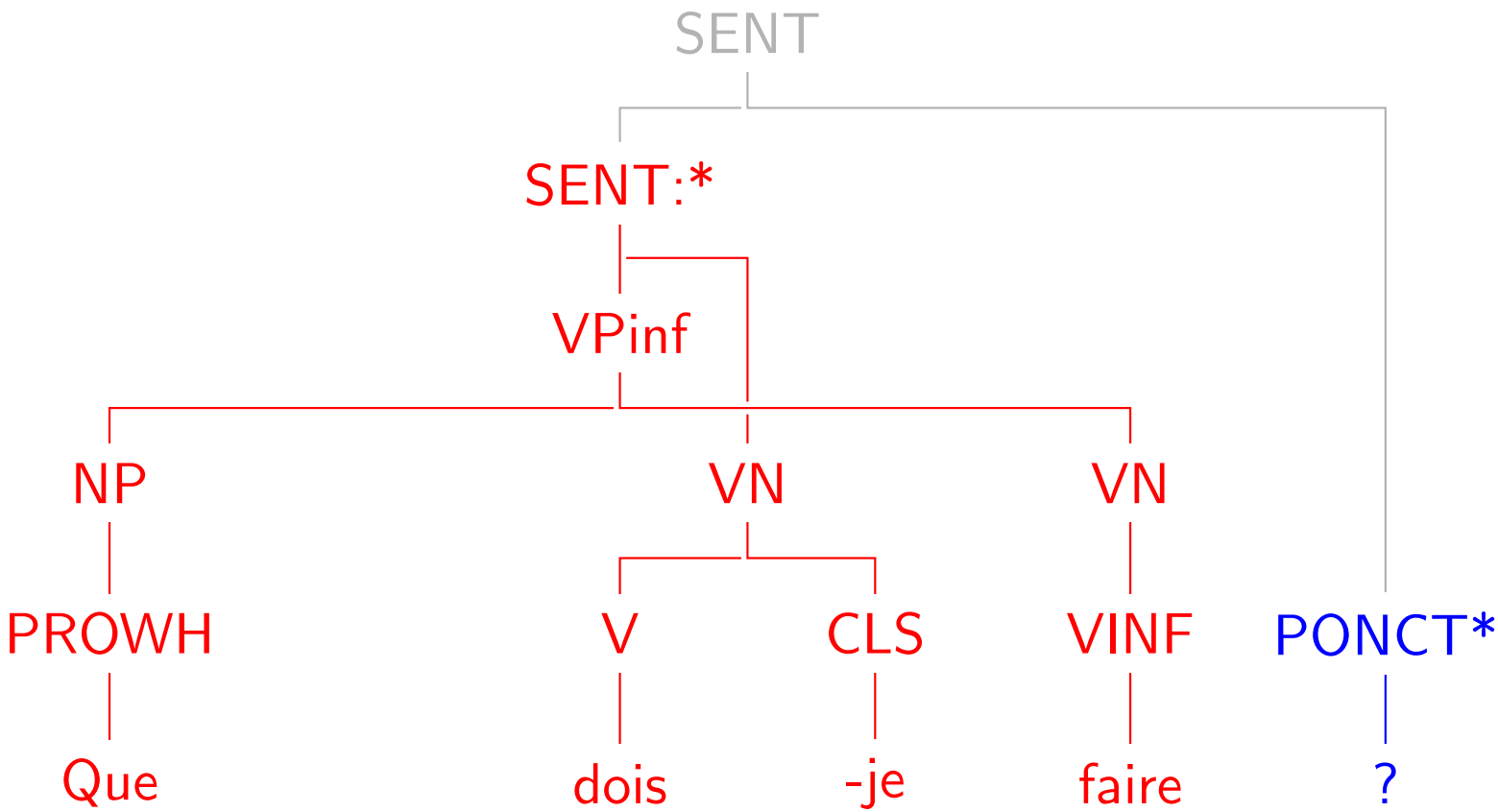
ReduceRight(SENT:)



sh, ru(NP), sh, sh, rl(VN), sh, ru(VN), gap, rl(VPinf), rr(SENT:)

Shift-Reduce-Gap : Pile – File – Buffer

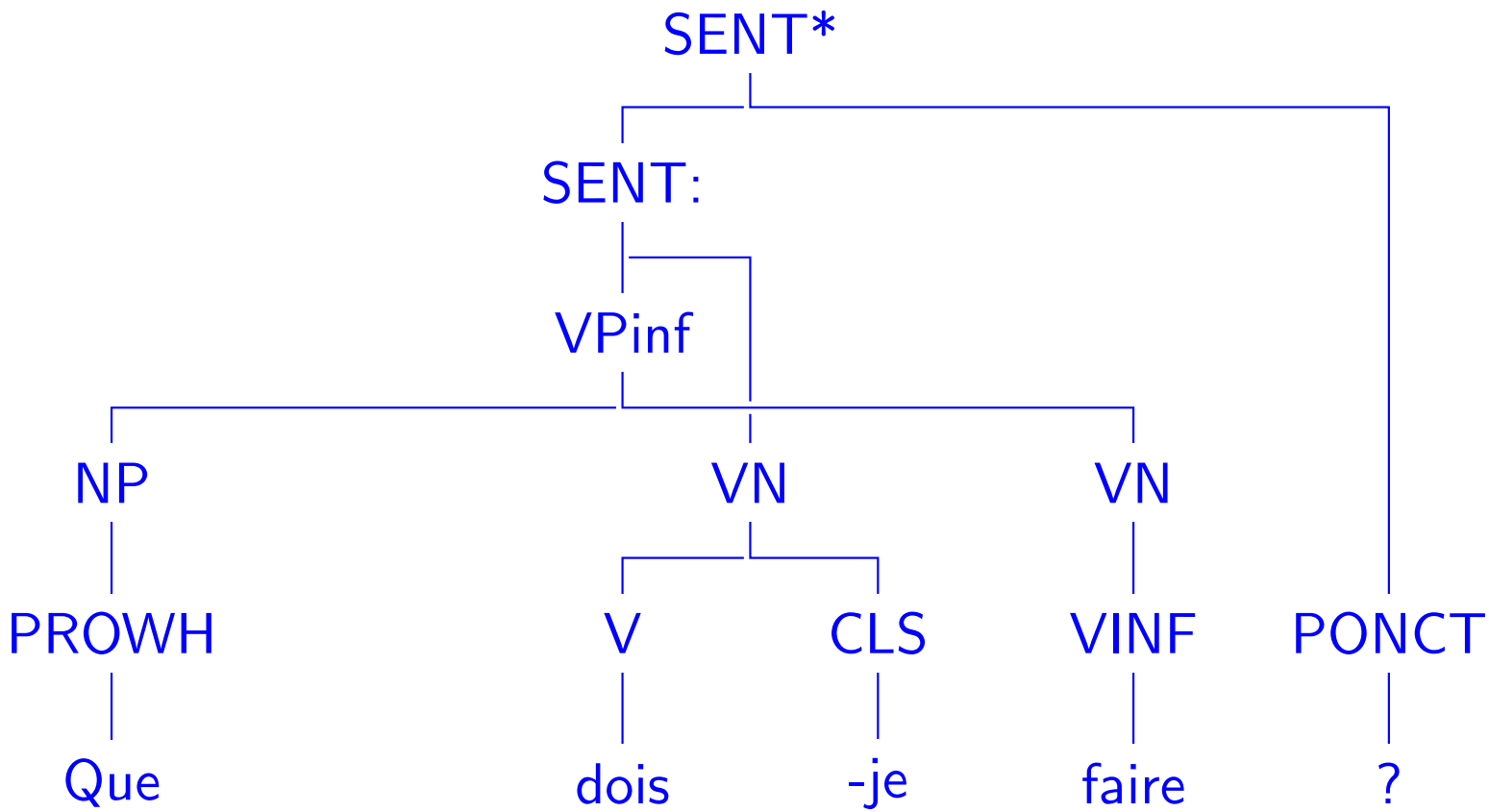
Shift



sh, ru(NP), sh, sh, rl(VN), sh, ru(VN), gap, rl(VPinf), rr(SENT:), sh

Shift-Reduce-Gap : Pile – File – Buffer

ReduceLeft(SENT)



sh, ru(NP), sh, sh, rl(VN), sh, ru(VN), gap, rl(VPinf), rr(SENT:), sh, rl(SENT)

Modèle statistique

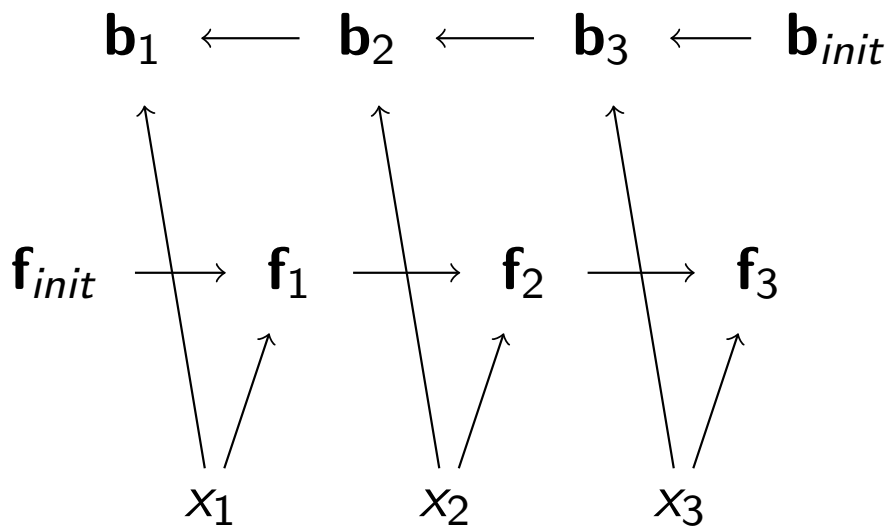
- Architecture neuronale **multitâches** (Caruana 1997) qui modélise conjointement:
  - $P(t|w_1^n)$: la probabilité d'un arbre t pour la phrase w_1^n
  - $P(M_1^n|w_1^n)$: la probabilité de la matrice de tags M_1^n

	POS	nombre	genre	temps	mode	fonction
Le	D	sg	m	NA	NA	det
chat	N	sg	m	NA	NA	suj
dort	V	sg	NA	pres	ind	root

- Multitâches: partage de représentations entre le tagger et le parser
  - intuition: les tâches se profitent mutuellement
  - biais inductif: on contraint les représentations apprises pour le parser à être de bons prédicteurs de la morphologie

Codeur LSTM bi-directionnel

- Réseau récurrent permettant de représenter une séquence
- LSTM avant: calcule des représentations pour les préfixes $x_1^k = (x_1, x_2, \ldots, x_k)$ $(k \in \{1, \ldots, n\})$
- LSTM arrière: calcule des représentations pour les suffixes $x_k^n = (x_k, x_{k+1}, \ldots, x_n)$ $(k \in \{1, \ldots, n\})$



- $[\mathbf{f}_n; \mathbf{b}_1]$ représente toute la séquence x_1^n
- $[\mathbf{f}_i; \mathbf{b}_i]$ représente le token x_i en contexte (embedding contextuel)

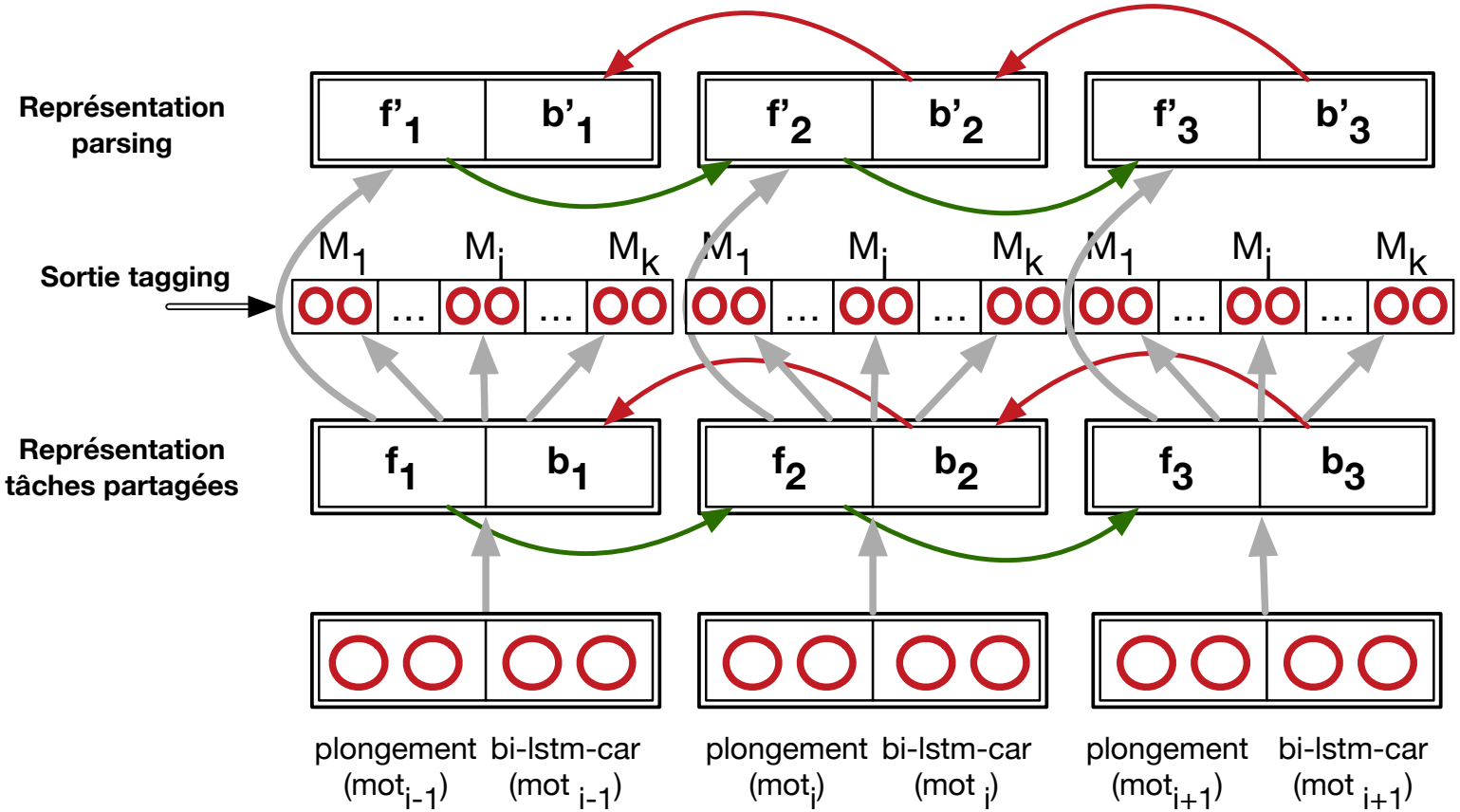
Représentations partagées entre tagger et parser

Réseau bi-LSTM hiérarchique (Plank et al. 2016):

- Une entrée lexicale x est représentée par la concaténation $[\mathbf{w}; \mathbf{c}]$
  - $\mathbf{w}$: embedding de mot
  - $\mathbf{c}$ = bi-LSTM(caractères)

- Un token x_i en contexte dans une phrase x_1^n est représenté par un second bi-LSTM
  - $\mathbf{h}_i^{(1)} = [\mathbf{f}_i; \mathbf{b}_i]$

Architecture



En pratique: bi-LSTM profond (2 étages), supervision du tagging sur la 1ère couche (Søgaard et al. 2016)

Prédire les tags

- Hypothèses d'indépendance entre les différents types de tags (POS, nombre, fonction ...) et entre les tags de chaque token:

$$P(M_1^n|w_1^n) = \prod_{i=1}^{n}\prod_{j=1}^{k} P(M_{i,j}|w_1^n)$$

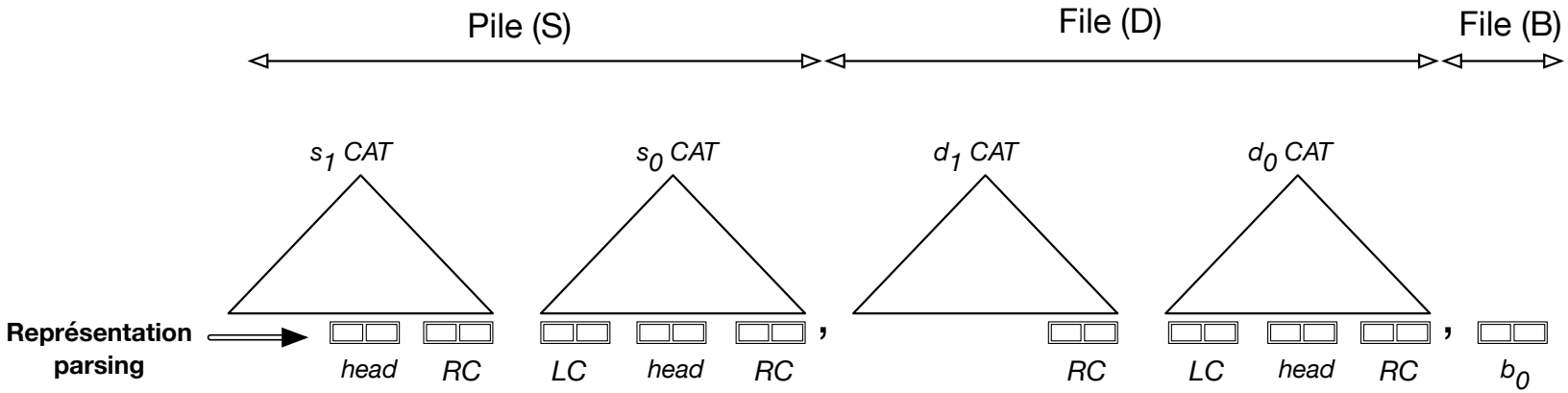
- Pour le type j et le token en position i:

$$P(m_{i,j} = \cdot|w_1^n) = \text{Softmax}(\mathbf{W}^{(j)} \cdot \mathbf{h}_i)$$

$$j \in \{1, \ldots k\}$$

Prédire les actions de parsing

- Réseau feed-forward (Chen et Manning 2014)



- Pour prédire une action à partir d'une configuration, l'input est la concaténation $\mathbf{v}$:
  - Des embeddings simples pour les non-terminaux
  - Les embeddings contextuels du bi-LSTM $\mathbf{h}_i$ pour les éléments lexicaux
- $P(a_i|a_1^{i-1}, w_1^n) = \text{Softmax}(\mathbf{W}^{(p)} \cdot \mathbf{v})$

Plan

Expériences

Données:

- 3 corpus : FTB, FQB, Sequoia
- Entraînement: soit FTB (split standard), soit FQB (80%)
- Prétraitements standards (binarisation)

Optimisation:

- SGD moyennée (vraisemblance des tags et des arbres golds)

$$-\log \prod_{i=1}^{n} \prod_{j=1}^{k} P(M_{i,j}|w_1^n) - \log \prod_{a=1}^{K} P(a_i|a_1^{i-1}|w_1^n)$$

- *Grid search* pour les hyperparamètres (tailles des couches cachées, des embeddings, learning rate), sélection du meilleur modèle sur le corpus de développement

Décodage:

- Recherche gloutonne ($\approx$ 480 tokens / seconde)

Évaluation:

- F1 standard / F1 calculé uniquement sur les constituants discontinus

Résultats: Expérience 1

Les discontinuités sont plus faciles à prédire sur le French Question Bank.

Corpus (dev)	All F1	Constituants discontinus F1	P	R
French Treebank	82.3	**17.4**	36.4	11.4
French Question bank	95.2	**62.5**	55.6	71.4

Hypothèses:

- Discontinuités beaucoup plus fréquentes
- Types de discontinuités homogènes (questions à 95%)
- Phrases plus courtes en moyenne

Résultats: Expérience 2

Variabilité des résultats (sur 64 modèles avec différents hyper-paramétrages, Dev)

	Constituants					
		All			Disc.	
	F1	P	R	F1	P	R
FTB-DEV – Entraînement sur FTB-TRAIN						
Maximum	82.33	82.3	82.39	32.0	60.0	22.86
Minimum	80.2	80.11	80.3	3.85	5.88	2.86
Écart-type	**0.428**	0.431	0.433	**6.931**	12.8	4.853
FQB-DEV – Entraînement sur FQB-TRAIN						
Maximum	95.18	95.23	95.26	75.0	71.43	85.71
Minimum	93.75	93.73	93.76	40.0	30.77	57.14
Écart-type	**0.317**	0.324	0.33	**8.838**	11.42	7.785

- Résultats très stables pour F-score sur tous les constituants
- Résultats très instables sur les constituants discontinus
- Trop peu de données pour évaluation fiable

Résultats: Expérience 3 – Évaluation finale (corpus de test)

- Comparaison à un perceptron structuré (même système de transitions)
- Comparaison à des analyseurs en dépendances non projectifs

	Constituants		Dépendances		Tagging
	All	Disc.	UAS	LAS	
	F1	F1			
Entraînement sur FTB-TRAIN					
Analyseur bi-LSTM, faisceau=1 (glouton)					
FTB-TEST[a]	82.04	14.46	88.24	**83.40**	97.66
FQB-ALL	85.14	11.43	89.08	**79.39**	93.89
Perceptron structuré, faisceau=16					
FTB-TEST[a]	79.42	**19.05**	-	-	97.35
FTB-TEST: Michalon et al. (2016)	-	-	86.6	83.3	
FQB-ALL: Seddah et Candito (2016)	-	-	87.70	76.48	

Plan

Conclusion

Contributions

- Corpus discontinus pour le français obtenus par conversion de corpus existants
  - `https://github.com/mcoavoux/french_disco_data`
- Expériences d'analyse en constituants discontinus
  - `https://github.com/mcoavoux/mtg`

Perspectives

- Représentations discontinues pour d'autres phénomènes (incises, relatives extraposées)
- Évaluation multilingue (Anglais, Allemand, Français)

https://github.com/mcoavoux/mtg/
mcoavoux@linguist.univ-paris-diderot.fr

Merci !

Questions ? Commentaires ?

Merci à Marie Candito et Djamé Seddah.

Résultats multi-lingues (Dev)

Transition System	English (PTB)		Allemand (Tiger)		Français (FTB)	
	F	Disc. F	F	Disc. F	F	Disc. F
sr-gap	90.71	71.93	86.7	58.89	81.7	13.64
unlex-cl-gap	91.13	72.71	87.39	62.64	82.3	18.18